The Design and Evaluation of Hybrid Controlled Trials that Leverage External Data and Randomization

Steffen Ventz (ventzer@yahoo.de)
Sean Khozin
Bill Louv
Jacob Sands
Patrick Y. Wen
Leah Comment
Brian M. Alexander
Lorenzo Trippa

Research Article

Keywords:

Posted Date: March 14th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1424988/v1

Abstract

Patient-level data from completed clinical studies or electronic health records can be used in the design and analysis of clinical trials. However, these external data can bias the evaluation of the experimental treatment when the statistical design does not appropriately account for potential confounders. In this work, we introduce a hybrid clinical trial design that combines the use of external control datasets and randomization to experimental and control arms, with the aim of producing efficient inference on the experimental treatment effects. Our analysis of the hybrid trial design includes scenarios where the distributions of measured and unmeasured prognostic patient characteristics differ across studies. Using simulations and datasets from clinical studies in extensive-stage small cell lung cancer and glioblastoma, we illustrate the potential advantages of hybrid trial designs compared to externally controlled trials and randomized trial designs.

Introduction

Randomized controlled trials (RCTs) are essential to demonstrate causal effects of an intervention on clinical outcomes. Randomization reduces the risk of bias by balancing potential confounders across treatment arms[1]. Though valuable, RCTs often require large samples sizes, resulting in long durations of accrual and high costs[2]. Non-randomized single-arm trials compare experimental treatments to historic benchmarks, and typically require smaller sample sizes than RCTs; however, they carry a risk of over- or underestimating treatment effects because of potential variations in patient populations across clinical trials[3–5]. The use of patient-level external control (EC) data from prior clinical studies has been proposed to reduce these risks and improve the evaluation of experimental treatments[6].

The integration of EC data in the design and analysis of clinical trials can take several forms, including testing/estimating treatment effects upon study completion[5], sample size re-estimation at interim analyses (IAs), and early decisions to terminate the study for futility or efficacy[7,8]. With the increasing availability of data from past trials, the prospective use of EC data in the design, conduct, and analysis of clinical trials has the potential to reduce the cost and time of evaluating new treatments [6,9,10].

In this work, we introduce and examine a hybrid trial (HT) design that combines the use of EC data and randomization (Fig. 1) to test experimental therapeutics. We evaluated pivotal operating characteristics of the HT design such as power, the control of the false positive rates, and the average sample size and study duration. To evaluate these operating characteristics, we used simulations and two collections of datasets from clinical trials in newly diagnosed glioblastoma (GBM) and extensive-stage small cell lung cancer (ES-SCLC). We compared the HT design to single-arm externally controlled trials[5] (ECTs), which leverage EC data, and RCTs. These comparisons illustrate the benefits, limitations, and risks of leveraging EC data using established metrics, such as the bias of treatment effects estimates and the average sample size.

Results

We examined the operating characteristics of the HT design described in Methods. As summarized in Figure 1, the first stage of the design randomizes patients to the experimental and internal control (IC) arms. The IA then determines if the study is closed for futility or not, and potentially updates the randomization ratio from 1:1 during the first stage of the study to for the second stage of the trial. These decisions are supported by an index of dissimilarity (see Methods) between the EC data and the early data from the IC arm. The same index of dissimilarity is recomputed at completion of the study and supports the decision to leverage the EC data for estimating the treatment effects of the experimental therapeutic or not.

We compared HT, ECT[5], and RCT designs using model-based simulations and *in silico* clinical trials generated with a resampling algorithm (see Methods) applied to ES-SCLC and GBM datasets.

Model-based simulations

We considered a study with a maximum sample size of patients, an IA after enrollments, and a targeted type I error rate of . ECTs and HTs utilized an EC dataset with patients. The size of the EC dataset is similar to the sample sizes of the ES-SCLC and GBM datasets. The simulated RCTs randomized all 120 patients to the IC and experimental treatment in a ratio, while all 120 patients in the ECT received the experimental treatment.

Table 1 summarizes the simulation scenarios that we used to compare the study designs. To examine the robustness and illustrate potential pitfalls of the trial designs, we included scenarios (2-5) where relevant pre-treatment variables were not available for interim and final analyses. Moreover, in scenarios 4 and 5, the conditional outcome distributions of the IC and EC populations were different. Table 2 reports the results for each scenario, the average study duration, the average sample size, the proportion of trials that were terminated early for futility, and the type I error rate and power across 2,000 (RCTs, ECTs, and HTs) simulations.

Scenario 1 (Tables 1), where all relevant pre-treatment patient characteristics are available for analysis, represents an ideal condition for leveraging EC data. Here, all designs have type I error rates close to the targeted level. As expected, the ECT has superior performance compared to HTs and RCTs. For instance, without a positive treatment effect, 44% of ECTs were terminated early for futility, compared to 7% and 15%-20% for RCTs and HTs, respectively. In scenario 1, the RCT had approximately 67% power, compared to 93% and 70-73% for the ECT and HT designs.

In scenarios 2-5, the set of available prognostic pre-treatment variables for the interim and final analyses is incomplete and statistical assumptions for inference in ECTs are therefore violated. In these scenarios, the ECT design performed worse than the HT and RCT designs. For instance, in scenarios 2 and 4, without a positive treatment effect, 71% (>99%) of the generated ECTs reported a false positive result

(type I error), compared to 5-8% for the HT design and 5% for the RCT design. Moreover, in scenario 3, the power of the ECT design declined to 10% compared to 53% for the RCT and HT designs.

In silico trials in ES-SCLC

We performed a literature review and identified 11 prognostic characteristics associated with overall survival (OS) in ES-SCLC (column 1 of Table S1). Only three of these variables (sex, age, and ECOG performance status) were available in the datasets (CALGB-9732[11], GALES[12] and Pirker et al.[13]) and were included in our analyses (Table S1).

The effects of pre-treatment variables on OS were estimated for patients treated with the standard of care (SOC) using a Cox model[14], with baseline survival stratified by studies (Table S3). Sex (male vs female, HR 1.45, p<0.001), age (<65 years vs ≥65, HR 0.7, p<0.001), and performance status (1 vs 0 HR 1.28, p=0.024, 2 vs 0 HR 2.54, p<0.001) had a significant association with OS. To investigate heterogeneity across studies, we estimated study-specific random effects in a Cox model for OS (column 3 of Table S3). These random effects represent differences of the outcome distributions across trial populations that are not attributed to the available patient pre-treatment characteristics. The estimates suggest differences in the conditional outcome distributions (i.e., given the available pre-treatment variables) between studies. The limited availability of pre-treatment patient characteristics, as well as the random-effects analyses, indicate limitations of the ES-SCLC datasets as EC for future ES-SCLC trials.

We considered a study with a size of 75 patients and OS at 9 months (OS-9) as primary endpoint. For the HT design, 50 and 25 patients were enrolled during the first stage (1:1 randomization) and second stage (equal to 0:1), respectively. We report results for additional values of the design parameters in the Supplementary Material. We used block randomization; for example, for RCTs, 12 patients per arm (experimental and control) were assigned during the second stage (25 patients) and the last patient was randomly assigned.

Figure 2 shows selected characteristics of the ECT, HT, and RCT designs based on 2,000 resampled trials. The resampling algorithm to generate these *in silico* trials is described in Methods. The bottom row of Figure 2 illustrates the operating characteristics when we apply the resampling algorithm. Each panel includes three columns that indicate the study (CALGB-9732[11], GALES[12] and Pirker et al.[13]) that was resampled to generate *in silico* trials. The results reflect the underlying study-to-study heterogeneity and the described limitations of the ES-SCLC datasets.

The top row of Figure 2 illustrates the same operating characteristics of the three trial designs under an ideal setting, without unmeasured confounders and differences of the conditional outcome distributions under the control treatment across studies. This was achieved by first randomly permuting the study membership labels of patients in the ES-SCLC datasets and then applying the resampling algorithm. These results serve as a reference to illustrate differences between the operating characteristics of ECTs and HTs under ideal settings for leveraging EC data (top row) and for the actual study-to-study differences (bottom row) in the ES-SCLC datasets.

Panels A and C of Figure 2 show the estimated type I error rates (solid vertical lines; target value 5%) and power (dotted vertical lines) of the ECT, HT, and RCT designs when we resampled the CALGB-9732[11], GALES[12] and Pirker et al.[13] studies. As expected, without cofounding (Figure 2a), the ECT was the most powerful design, with 94%, 97%, and 93% power for the CALGB-9732, Pirker et al., and GALES studies, respectively, compared to 76%, 80%, and 65%, and 54%, 62%, and 43% for the HT and RCT designs in each of the three studies, respectively. In contrast, because of study-to-study heterogeneity, the resampling algorithm (Figure 2C) showed that the ECT design inflates the type I error rates, which reaches 59% for the GALES study. The type I error rates were considerably lower for the HT design (5%, 8%, and 5% for CALGB-9732[11], GALES[12] and Pirker et al.[13]), as the dissimilarity analyses (see Methods) recognize the limitations of the EC data.

In silico trials in GBM

We used five GBM [5,15] datasets (see Methods and Table S4) to compare HT, ECT, and RCT designs. We considered a study with a sample size of 100 patients, OS-12 as the primary endpoint, and an IA after 50 enrollments. The initial randomization ratio was 1:1 for both the HT and RCT designs. For the HT design, the randomization ratio during the second stage remained 1:1 or was updated to (we considered 1:2 and 0:1). We generated *in silico* trials resampling from GBM datasets (Chinot et al.[16], Dana-Farber Cancer Institute[17] [DFCI], and University of California, Los Angeles[5] [UCLA]) with more than 100 patients treated with the current SOC[18], temozolomide in combination with radiation therapy.

In contrast to ES-SCLC, the GBM datasets included all major prognostic patient pre-treatment characteristics identified through a literature review[17]. This difference between the ES-SCLC and GBM datasets is consistent with results obtained from Cox regression models with study-specific random effects (Tables S3 and S5). The estimated model indicates lower study-to-study variability in the GBM datasets compared to the ES-SCLC datasets.

Table 3 shows selected operating characteristics of the ECT, HT, and RCT designs based on 2,000 *in silico* trials generated by resampling the SOC arms of the Chinot et al. (rows 4-8), DFCI (rows 4-8), and UCLA (rows 4-8) datasets. Columns 7-11 (2-6) correspond to *in silico* RCTs, ECTs, and HTs that evaluated an experimental treatment with (or without) a positive treatment effect.

All three study designs showed type I error rates across *in silico* trials close to the targeted 5% level. Both the ECT and HT designs had a higher probability (42%-50% for ECT and 24%-27% for the HTs) of stopping the study early when the treatment effect is null compared to the RCT design (6%-7%). This translates into reductions of the average sample size of the *in silico* ECTs and HTs compared to the RCTs, from 96 patients for the RCT design to 75-79 patients and 86-88 patients for the ECT and HT designs. Moreover, for the *in silico* GBM trials that evaluated an effective experimental treatment (columns 7-11), we observed gains in power for ECT (85%-92%) and HT (73%-77%, 78%-82%, and 74%-78% with 1:1, 1:2, and 0:1, respectively) designs compared to conventional RCTs (58%-63%).

Discussion

The increasing availability of patient-level data from completed clinical studies and electronic health records constitutes an opportunity for the development of novel trial designs that leverage EC data[7,9,15,17,19]. Recent contributions[8,17,19] have proposed methodologies to integrate EC data into the analysis of single-arm trials (ECTs). These methods replace published estimates of the SOC's efficacy used as a benchmark with patient-level EC data. The EC data in ECTs allow the analyst to account for variations in the distribution of prognostic pre-treatment characteristics across clinical studies. This approach has the potential to reduce bias, false positive/negative rates, and ultimately improve the evaluation of experimental treatments[4,17].

As illustrated in recent retrospective studies[17,20,21] and in Table 2, under ideal conditions—without unmeasured confounding and with moderate variations of the patient pre-treatment profiles across study populations—the ECT design is an attractive alternative to the RCT design. However, it is challenging to anticipate mechanisms, such as unmeasured confounding and variations of the trial population during the enrollment period, which can bias the primary findings of the study.

Statistical methods applicable to ECTs, such as MSMs[22], matching[23], and inverse-probability weighting[24] (IPW), rely on key assumptions that are difficult to validate. They assume that (a) all confounding pre-treatment variables are available and included in the analyses; (b) consistent definitions and standards are used to measure patient profiles and outcomes during the trial and in the EC; and (c) identical conditional outcome distributions, given the patient pre-treatment characteristics, under the control therapy for the EC and trial populations. If these assumptions are violated, then the treatment effects estimate can be biased, and the control of false positive rates can be compromised (see Table 2 and Fig. 2).

During the design phase of an ECT, it is challenging to quantify the risks associated with leveraging an EC dataset. For example, unexpected confounding variables may not be included in the EC data, or subtle differences in the definition or measurement standards of the patient characteristics and treatment outcomes may remain unnoticed. Importantly, in most cases, the data generated during the trial do not provide evidence in favor or against the ECT assumptions, as the study does not have a control arm.

In consideration of these challenges, we introduced a hybrid design that combines randomization and the use of EC data. We developed the design to achieve and balance two goals. First, we aimed for reliable inference of the treatment effects even in settings where the EC data have limitations. This included unmeasured confounding and other mechanisms that translate into poor operating characteristics of ECTs (see Table 2 and Fig. 2). Second, we sought to achieve efficiency levels comparable to ECTs in the ideal setting, when the EC data have no limitations and the ECT assumptions hold. In these scenarios, it is convenient to leverage the EC data to improve the trade-off between power and the resources for conducting the trial (Table 3). HTs prospectively compare the conditional outcome distributions of the IC

and EC groups. The EC data are used for inference on the treatment effects only when the resulting index of dissimilarity does not suggest different conditional distributions for these two groups.

We used datasets from completed clinical studies and electronic health records to create realistic scenarios that highlight potential risks and benefits of the ECT and HT designs. ES-SCLC and GBM datasets were used to compare ECT, HT, and RCT designs. The scenarios defined by resampling the control arms of the ES-SCLC datasets are representative of settings where ECTs have poor operating characteristics due to confounding. Scenarios defined through GBM datasets were markedly different. In the resulting *in silico* GBM trials, leveraging EC data translated into efficiency gains compared to RCTs while maintaining control of false positive rates.

The analyses based on model-based simulations (Table 2) and *in silico* trials obtained by resampling the GBM datasets[8] (Table 3) indicated potential efficiency gains of HTs compared to RCTs when EC data without substantial limitations are available. We showed improvements of power, average study duration, and sample size.

A limitation of our analyses is the relatively small number of GBM and ES-SCLC datasets used to evaluate the HT and ECT designs. A larger number of datasets could provide a more representative sample of outcome distributions and other important differences across SOC arms of recent RCTs in GBM and ES-SCLC. Moreover, only a small subset of known prognostic pre-treatment variables (Table S1) was available in the ES-SCLC datasets for statistical adjustments in ECTs and HTs. One study was open label (GALES[25]) and another one was only partially randomized (CALGB-30504[26]). Additionally, there were variations of the eligibility criteria across the ES-SLCLC studies, and etoposide with either platinum-based cisplatin or carboplatin chemotherapy were two SOC regimens in these trials. With these data limitations, the type I error rate of the ECT design in ES-SCLC, accounting for a limited set of available prognostic variables (Table S1), was as high as 59% in our analyses.

When there is uncertainty regarding the risks associated with available EC data, the proposed HT design can be an attractive alternative to the ECT and RCT designs. Limitations of the EC data can impact the operating characteristics of ECTs, while at the opposite end of the spectrum RCTs do not utilize EC data. HTs can be viewed as a compromise between ECTs and RCTs, as HTs can prospectively evaluate potential limitations of the EC data compared the IC arm.

The described limitations of the datasets (i.e. Table S1, different eligibility criteria, etc.), the random effects analysis (Table S3), and the *in silico* ECTs (Fig. 2) consistently associated the use of the ES-SCLC data as EC group with risks of bias and inadequate control of false positive/negative rates. We used the ES-SCLC datasets primarily to illustrate that HTs could substantially reduce these risks compared to ECTs.

ECTs have been considered previously in settings beyond ES-SCLC and GBM. Carrigan et al.[20] demonstrated the feasibility of generating external controls in non-small cell lung cancer (NSCLC) using real-world data from the Flatiron Health database. Similarly, in Project Switch[21], FDA investigators

showed that ECTs can estimate OS hazard ratios by exchanging the control arms between trials in second-line NSCLC with docetaxel controls.

The integration of EC data into clinical trials requires high-quality and up-to-date patient-level datasets representative of the current SOC. Factors such as changes in the SOC and the discovery of new prognostic biomarkers pose challenges in maintaining contemporaneous EC datasets. On the other hand, EC data with biomarker information can be useful for HT testing novel treatments in subpopulations with low enrollment rates. Moreover, HT designs can be extended to alternative study aims, such as testing non-interiority. Recent data sharing efforts[27], such as the National Cancer Institute (NCI) NCTN/NCORP Data Archive, Project Data Sphere[28], YODA[29], Vilvi[30], and CancerLinQ[31], provide valuable data sources for this endeavor.

Methods

We use Y to indicate the binary primary outcome. We also report results for time-to-event primary endpoints Y (e.g., OS) in the Supplementary Material. The binary variable A indicates whether the patient received the experimental (A = 1) or control (A = 0) therapy, and the vector X includes a fixed set of pre-treatment patient characteristics (e.g., age, sex, etc.). The indicator S distinguishes patients enrolled during the trial (S = 0) from patients in the external control (EC) dataset (S = 1). Patients in the EC group were treated with the control therapy (A = 0). We use $\Pr(\mathrm{Y} \mid \mathrm{X}, \mathrm{A}, \mathrm{S})$ to indicate the conditional outcome distribution of patients with pre-treatment characteristics X and treatment A in the trial population (S = 0) or in the EC group ($\mathrm{S} = 1$).

Hybrid design

Figure 1A describes a HT design that uses EC data and randomization to the experimental and control (internal control, IC) arms to estimate and test treatment effects. For simplicity, we focused on a two-stage design with sample size $n = n_1 + n_2$. During the first stage n_1 patients are randomized to the IC and experimental arms in the ratio $r_{1,C}$: $r_{1,E}$ (1:1 in our analyses). At completion of the first stage, after enrollment of the first n_1 patients, an IA is used to decide (a) if the clinical study continues to the second stage or is stopped for futility; and, if the study is not stopped for futility, (b) whether or not to update the randomization ratio to $r_{2,C}$: $r_{2,E}$ for the remaining n_2 patients during the second stage. These two decisions are supported by an index of dissimilarity (W_1, Supplementary Material), computed using early data from the trial and the EC dataset. The summary W_1 quantifies the evidence of differences between the conditional outcome distributions $Pr(Y|X, A = 0, S)$ of the IC ($S = 0$) and EC ($S = 1$) populations. Large values of W_1 indicate dissimilarity between the two conditional distributions. In particular,

1. *if W_1 exceeds a predefined threshold w_1 ($W_1 > w_1$)*, then the EC data are excluded from the futility analysis and, if the trial is not stopped for futility, the assignment ratio during the second stage remains 1:1, as in the first stage.

2. *If* $W_1 \leq w_1$, then the futility IA utilizes both IC and EC data. If the trial is not stopped for futility, the proportion of patients assigned to the IC during the second stage is decreased by updating the assignment ratio to the prespecified value $r_{2,C}: r_{2,E}$. We considered ratios of 1:1, 1:2, and 0:1. When $r_{2,C}: r_{2,E} = 0:1$, patients are not randomized during the second stage.

At completion of the trial, after the primary outcomes of all n patients become available, we re-compute the index of dissimilarity (W_2) using all the available data. If W_2 is larger than a pre-defined threshold w_2, then the EC data are excluded from the final analyses. If $W_2 \leq w_2$, the final trial analyses leverage the EC data.

Externally controlled trial (ECT) designs

ECTs[5] (Fig. 1B) are a particular case of the class of designs in Fig. 1A, without randomization. The design assumes identical SOC conditional outcome distributions $Pr(Y|X, A = 0, S)$ for the trial and EC populations, which makes the indicator S unnecessary. Patient-level data of the experimental arm and EC data are used to estimate the treatment effect (TE),

$$TE = \sum_x \{E[Y|X = x, A = 1] - E[Y|X = x, A = 0]\}\Pr(X = x). \quad (1)$$

Here, the expected outcome $E[Y|X = x, A]$ of patients receiving experimental (A = 1) and control (A = 0) treatments with pre-treatment characteristics *x* are weighted by a distribution $\Pr(X = x)$, for example, the distribution of pre-treatment variables X in the experimental arm.

We considered different procedures to estimate the TE in (1), including matching[23], IPW[24], and marginal structural models[22] (MSMs) (see Figure S1). We did not observe substantial differences between these methods, and used MSMs in our subsequent analyses.

Testing the null hypothesis of no treatment effects at completion of the study

For ECTs, as well as HTs when $W_2 \leq w_2$, we utilized MSMs[22] to estimate treatment effects and test the null hypothesis $H_0: TE \leq 0$, using the data available at completion of the trial and the EC data. Whereas for RCTs and for HTs with $W_2 > w_2$ we utilized only the trial data to estimate treatment effects (estimator: difference of the empirical response rates between the experimental and IC) and test H_0 (test: 2-sample z-test for proportions[32]).

Permutation test. We also considered an alternative permutation test (see Figure S2) for HT designs that utilize trial data and EC data (i.e., HTs with $W_2 \leq w_2$). The procedure controls the type I error rate at a predefined α-level, both when the standard assumptions of adjustment methods, such as MSM, holds or are violated, for example in settings with unmeasured confounders, or when the conditional outcome distributions $Pr(Y|X, A = 0, S)$ of the IC ($S = 1$) and EC ($S = 0$) groups differ. The procedure has three components:

1. First, a treatment effects estimate $\hat{TE}\left(D_{HT}, D_{EC}\right)$ is calculated using the HT data and the EC data. Here $D_{HT} = \left\{\left(Y_i, X_i, A_i, S_i = 1\right)\right\}_{i \le n}$ indicates the HT data, whereas $D_{EC} = \left\{\left(Y_i, X_i, A_i = 0, S_i = 0\right)\right\}_{n < i \le n + n_{EC}}$ includes information for n_{EC} EC patients. The index *i* identifies the patients.
2. Next, we randomly permute l=1,..., 1,000 times the treatment assignment variables $\left\{A_i\right\}_{i \le n}$ in the HT ($A_{\rho_{l,1}}, A_{\rho_{l,2}} \ldots A_{\rho_{l,n}}$), while the assignment variables $\left\{A_i = 0\right\}_{i > n}$ in the EC remain identical. For each $1 \le l \le 1{,}000$, we obtain a permuted dataset $D_{HT,\rho_l} = \left\{\left(Y_i, X_i, A_{\rho_{l,i}}, S_i = 0\right)\right\}_{i \le n}$ *and* compute the estimate $\hat{TE}_l = \hat{TE}\left(D_{HT,\rho_l}, D_{EC}\right)$.
3. We then estimate the p-value ($H_0 : TE \le 0$) as the proportion of permutations l with statistics $\hat{TE}_l$ larger than the actual estimate $\hat{TE}$.

Evaluation of the trial designs

We evaluated the operating characteristics of the HT, ECT, and RCT designs using model-based simulations and a leave-one-study-out resampling algorithm.

Model-based simulations

We generated clinical studies using a parametric model (Table 1) for

1. $Pr(X|S)$, the distributions of pre-treatment variables in the trial ($S = 0$) and EC ($S = 1$) populations, and
2. $Pr(Y|X, A, S)$, the conditional outcome distributions in the trial ($S = 0$) and EC ($S = 1$) populations.

We considered scenarios where the distributions of pre-treatment variables (a) and the conditional outcome distributions (b) differ between the two populations (S = 0,1), as well as scenarios with unmeasured confounding.

Leave-one-study-out resampling algorithm

To evaluate the operating characteristics of the HT design we used a resampling scheme similar to the one described by Ventz et al.[17] applied to datasets from completed clinical trials and electronic health records in ES-SCLC and GBM (see Figs. 3 and S6). The algorithm provides estimates of the operating characteristics, including type I error rate, power and the average sample size.

ES-SCLC datasets: We used patient-level data available at Project Data Sphere[28] from three randomized Phase III clinical trials: CALGB-9732[11] (NCT00003299), Pirker et al.[13] (NCT00119613), and GALES[25] (NCT00363415). For the Pirker et al. study, a random subsample containing 80% of the original study population was available. We used data from patients who received etoposide in combination with platinum-based cisplatin (CALGB-9732, Pirker et al., GALES) or carboplatin (Pirker et al.) chemotherapy; both treatments were SOC regimens in ES-SCLC. The resampling algorithm to generate *in silico* ECTs and HTs assumes identical conditional outcome distributions, given the available pre-treatment characteristics, for these two SOC regimes. The comparison of cisplatin and carboplatin has been previously discussed[33]. Supplementary analysis based on data for patients randomized to the control arm of NCT00119613, which received either etoposide plus carboplatin or etoposide plus cisplatin supported this assumption (Log-rank test: p-value 0.4). Nonetheless, undetected differences between these two regimes could impact the operating characteristics of trial designs that leverage EC data.

GBM datasets [8,17]: We used patient-level data from a phase III study (Chinot et al.[16] [NCT00943826], 460 patients), two phase II studies (Cho et al. [34] [PMID: 22120301], 16 patients; Lee et al.[35] [NCT00441142], 29 patients) and two real-world datasets[17] (378 and 305 patients) from DFCI and UCLA. We only used data from patients treated with temozolomide and radiation therapy (TMZ + RT), the SOC in GBM[18]. Pre-treatment variables included age, sex, Karnofsky performance status, MGMT methylation status, and extent of tumor resection[36–38] (see Table S4).

Algorithm

For each ES-SCLC (or GBM) study, the algorithm repeatedly samples at random, without replacement, a subset of patients from the control arm. These subsets are used to mimic the data generated during the HTs. Patient-level data from the control arms of the remaining ES-SCLC (or GBM) datasets are used as EC.

Specifically, for each ES-SCLC (or GBM) study *k*, we randomly generated 2,000 trials by repeating the following steps 2,000 times (using different computer-generated random subsamples):

(i) Randomly subsample (without replacement) *n* patient profiles X and the corresponding outcomes Y from the control arm (SOC) of study *k*.

(ii)Use the control arms of the remaining studies as EC data.

(iii)Randomize (without replacement) n_1 of the patients in Step (i) to the experimental and control arms of the *in silico* HT in ratio $r_{1,C}$: $r_{1,E}$ and compute the index W_1.

(iii.a) If $W_1 \leq w_1$, use the ratio $r_{2,C}$: $r_{2,E}$ for the remaining $n_2 = n - n_1$ patients in stage 2.

(iii.b) If $W_1 > w_1$, use the ratio $r_{1,C}$: $r_{1,E}$ for the remaining $n_2 = n - n_1$ patients in stage 2.

(iv) Use the output of Steps (i)-(iii) to generate an *in silico* HT trial, including the futility IA and, if the *in silico* HT is not discontinued, final hypothesis testing (Fig. 1A).

We used the statistical software R[39] to implement the algorithm and generate the random samples in Steps (i)-(iii), which differed across the 2,000 *in silico* HTs.

The n_1 patients (randomly selected) from the control arm of study *k* in Step (iii.a) allowed us to mimic the data of the experimental and IC arms of the HT during the first stage of the study, whereas the remaining n_2 patients in Step (iii.b) mimicked the second stage of the HT. In these *in silico* HTs, the treatment effect is null by construction of the algorithm because the outcome distributions in the two arms of the trial are identical.

To evaluate the power of the HT design, we added a component to Step (iii) of the algorithm (see Figure S6), which allowed us to produce *in silico* HTs with positive treatment effects. For each enrollment *i* to the experimental arm ($A_i = 1$), if the patient had a negative response ($Y_i = 0$), we randomly generate a binary random variable R_i, with $Pr\left(R_i = 1\right) = \pi$, representative of the treatment effect for patient *i*. If $R_i = 1$, then the negative outcome is relabeled as a positive outcome (i.e., we set $Y_i = 1$). If $R_i = 0$, then the outcome remains unchanged ($Y_i = 0$). The computer-generated random variables $\{R_i\}$differed across the 2,000 *in silico* HTs. We used $\pi = 0.4$ for ES-SCLC and $\pi = 0.5$ for GBM analyses reported in the Results, and different values of π for analyses reported in the Supplementary Material.

Data Availability

Simulated datasets were generated in R, version 4.1 with the Supplementary R code. The SCLC datasets (NCT00003299, NCT00119613, NCT00363415) are freely available for download from Project Data Sphere[28].

Code Availability

R code used to generate the HTs, ECTs and CTs are available with the Supplementary R code.

Declarations

Competing interests: P.W is on the advisory board of Agios, Astra Zeneca, Bayer, Black Diamond, Boston Pharmaceuticals, Elevate Bio, Imvax, Karyopharm, Merck, Mundiphrama, Novartis, Novocure, Nuvation Bio, Prelude Therapeutics, Sapience, Vascular Biogenics, VBI Vaccines, Voyager. All other authors declare no competing interests.

References

1. Ratain MJ, Sargent DJ. Optimising the design of phase II oncology trials: The importance of randomisation. *Eur J Cancer*. 2009;45(2):275-280. doi:10.1016/j.ejca.2008.10.029

2. Gan HK, Grothey A, Pond GR, Moore MJ, Siu LL, Sargent D. Randomized phase II trials: Inevitable or inadvisable? *J Clin Oncol*. 2010;28(15):2641-2647. doi:10.1200/JCO.2009.26.3343

3. Unger JM, Hershman DL, Fleury ME, Vaidya R. Association of Patient Comorbid Conditions with Cancer Clinical Trial Participation. *JAMA Oncol*. 2019;5(3):326-333. doi:10.1001/jamaoncol.2018.5953

4. M VA, Ventz S, Rahman R, Fell G, Trippa L, Alexander BM. To randomize, or not to randomize, that is the question: a meta-analytic methodology for determining the context-speci c value of randomization. *Neuro Oncol*. 2019;Submitted.

5. Ventz S, Lai A, Cloughesy TF, Wen PY, Trippa L, Alexander BM. Design and evaluation of an external control arm using prior clinical trials and real-world data. *Clin Cancer Res*. 2019;25(16). doi:10.1158/1078-0432.CCR-19-0820

6. Corrigan-Curay J, Sacks L, Woodcock J. Real-world evidence and real-world data for evaluating drug safety and effectiveness. *JAMA*. 2018;320(9):867-868. doi:10.1001/jama.2018.10136

7. Ventz S, Trippa L, Schoenfeld JD. Lessons learned from deescalation trials in favorable risk HPV-associated squamous cell head and neck cancer–a perspective on future trial designs. *Clin Cancer Res*. 2019;25(24). doi:10.1158/1078-0432.CCR-19-0945

8. Ventz S, Comment L, Louv B, et al. The use of external control data for predictions and futility interim analyses in clinical trials. *Neuro Oncol*. 2021. doi:10.1093/neuonc/noab141

9. Khozin S, Blumenthal GM, Pazdur R. Real-world Data for Clinical Evidence Generation in Oncology. *J Natl Cancer Inst*. 2017;109(11):1-5. doi:10.1093/jnci/djx187

10. Rahman R, Fell G, Ventz S, et al. Deviation from the proportional hazards assumption in randomized phase 3 clinical trials in oncology: Prevalence, associated factors, and implications. In: *Clinical Cancer Research*. Vol 25. ; 2019:6339-6345. doi:10.1158/1078-0432.CCR-18-3999

11. Niell HB, Herndon JE, Miller AA, et al. Randomized phase III intergroup trial of etoposide and cisplatin with or without paclitaxel and granulocyte colony-stimulating factor in patients with extensive-stage small-cell lung cancer: Cancer and Leukemia Group B trial 9732. *J Clin Oncol*. 2005;23(16):3752-3759. doi:10.1200/JCO.2005.09.071

12. Socinski MA, Smit EF, Lorigan P, et al. Phase III study of pemetrexed plus carboplatin compared with etoposide plus carboplatin in chemotherapy-naive patients with extensive-stage small-cell lung cancer. *J Clin Oncol*. 2009;27(28):4787-4792. doi:10.1200/JCO.2009.23.1548

13. Pirker R, Ramlau RA, Schuette W, et al. Safety and efficacy of darbepoetin alfa in previously untreated extensive-stage small-cell lung cancer treated with platinum plus etoposide. *J Clin Oncol*. 2008;26(14):2342-2349. doi:10.1200/JCO.2007.15.0748

14. Cox DR. Regression Models and Life-Tables. *J R Stat Soc Ser B*. 1972;34(2):187-202. doi:10.1111/j.2517-6161.1972.tb00899.x

15. Rahman R, Ventz S, McDunn J, et al. Leveraging external data in the design and analysis of clinical trials in neuro-oncology. *Lancet Oncol*. 2021;22(10). doi:10.1016/s1470-2045(21)00488-5

16. Chinot OL, Wick W, Mason W, et al. Bevacizumab plus Radiotherapy–Temozolomide for Newly Diagnosed Glioblastoma. *N Engl J Med*. 2014;370(8):709-722. doi:10.1056/NEJMoa1308345

17. Ventz S, Lai A, Cloughesy TF, Wen PY, Trippa L, Alexander BM. Design and Evaluation of an External Control Arm Using Prior Clinical Trials and Real-World Data. *Clin Cancer Res*. 2019;25(16):4993-5001. doi:10.1158/1078-0432.ccr-19-0820

18. Stupp R, Mason WP, van den Bent MJ, et al. Radiotherapy plus Concomitant and Adjuvant Temozolomide for Glioblastoma. *N Engl J Med*. 2005;352(10):987-996. doi:10.1056/NEJMoa043330

19. Viele K, Berry S, Neuenschwander B, et al. Use of historical control data for assessing treatment effects in clinical trials. *Pharm Stat*. 2014;13(1):41-54. doi:10.1002/pst.1589

20. Carrigan G, Whipple S, Taylor MD, et al. An evaluation of the impact of missing deaths on overall survival analyses of advanced non–small cell lung cancer patients conducted in an electronic health records database. *Pharmacoepidemiol Drug Saf*. 2019;28(5):572-581. doi:10.1002/pds.4758

21. Kanapuru B, Gong Y, Mishra-Kalyani PS, et al. Project Switch: Lenalidomide and dexamethasone (Len-Dex) as a potential synthetic control arm (SCA) in relapsed or refractory multiple myeloma (rrMM). *J Clin Oncol*. 2019;37(15_suppl). doi:10.1200/jco.2019.37.15_suppl.8047

22. Robins JM, Hernán MÁ, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11(5):550-560. doi:10.1097/00001648-200009000-00011

23. Imbens GW, Rubin DB. *Causal Inference: For Statistics, Social, and Biomedical Sciences an Introduction*. Cambridge University Press; 2015. doi:10.1017/CBO9781139025751

24. Hirano K, Imbens GW. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Heal Serv Outcomes Res Methodol*. 2001;2(3-4):259-278. doi:10.1023/A:1020371312283

25. Thatcher N, Hirsch FR, Luft A V., et al. Necitumumab plus gemcitabine and cisplatin versus gemcitabine and cisplatin alone as first-line therapy in patients with stage IV squamous non-small-cell lung cancer (SQUIRE): An open-label, randomised, controlled phase 3 trial. *Lancet Oncol*. 2015;16(7):763-774. doi:10.1016/S1470-2045(15)00021-2

26. Ready NE, Pang HH, Gu L, et al. Chemotherapy with or without maintenance sunitinib for untreated extensive-stage small-cell lung cancer: A randomized, double-blind, placebo-controlled phase II study -

CALGB 30504 (Alliance). *J Clin Oncol*. 2015;33(15):1660-1665. doi:10.1200/JCO.2014.57.3105

27. Bertagnolli MM, Sartor O, Chabner BA, et al. Advantages of a Truly Open-Access Data-Sharing Model. *N Engl J Med*. 2017;376(12):1178-1181. doi:10.1056/NEJMsb1702054

28. Green AK, Reeder-Hayes KE, Corty RW, et al. The Project Data Sphere Initiative: Accelerating Cancer Research by Sharing Data. *Oncologist*. 2015;20(5):464. doi:10.1634/theoncologist.2014-0431

29. Krumholz HM, Waldstreicher J. The Yale Open Data Access (YODA) Project — A Mechanism for Data Sharing. *N Engl J Med*. 2016;375(5):403-405. doi:10.1056/NEJMp1607342

30. Bierer BE, Li R, Barnes M, Sim I. A Global, Neutral Platform for Sharing Trial Data. *N Engl J Med*. 2016;374(25):2411-2413. doi:10.1056/NEJMp1605348

31. Rubinstein SM, Warner JL. CancerLinQ: Origins, Implementation, and Future Directions. *JCO Clin Cancer Informatics*. 2018;(2):1-7. doi:10.1200/cci.17.00060

32. Agresti A. *An Introduction to Categorical Data Analysis: Second Edition*.; 2006. doi:10.1002/0470114754

33. Rossi A, Di Maio M, Chiodini P, et al. Carboplatin- or cisplatin-based chemotherapy in first-line treatment of small-cell lung cancer: The COCIS meta-analysis of individual patient data. *J Clin Oncol*. 2012;30(14):1692-1698. doi:10.1200/JCO.2011.40.4905

34. Cho DY, Yang WK, Lee HC, et al. Adjuvant immunotherapy with whole-cell lysate dendritic cells vaccine for glioblastoma multiforme: A phase II clinical trial. *World Neurosurg*. 2012;77(5-6):736-744. doi:10.1016/j.wneu.2011.08.020

35. Lee EQ, Kaley TJ, Duda DG, et al. A multicenter, phase II, randomized, noncomparative clinical trial of radiation and temozolomide with or without vandetanib in newly diagnosed glioblastoma patients. *Clin Cancer Res*. 2015;21(16):3610-3618. doi:10.1158/1078-0432.CCR-14-3220

36. Thakkar JP, Dolecek TA, Horbinski C, et al. Epidemiologic and molecular prognostic review of glioblastoma. *Cancer Epidemiol Biomarkers Prev*. 2014;23(10):1985-1996. doi:10.1158/1055-9965.EPI-14-0275

37. Curran WJ, Scott CB, Horton J, et al. Recursive partitioning analysis of prognostic factors in three Radiation Therapy Oncology Group malignant glioma trials. *J Natl Cancer Inst*. 1993;85(9):704-710. doi:10.1093/jnci/85.9.704

38. Lamborn KR. Prognostic factors for survival of patients with glioblastoma: Recursive partitioning analysis. *Neuro Oncol*. 2004;6(3):227-235. doi:10.1215/S1152851703000620

39. Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. *J Comput Graph Stat*. 1996;5(3):299-314. doi:10.1080/10618600.1996.10474713

Tables 1-3

Tables 1-3 are available in the Supplementary Files section.

Figures

Figure 1

See image above for figure legend.

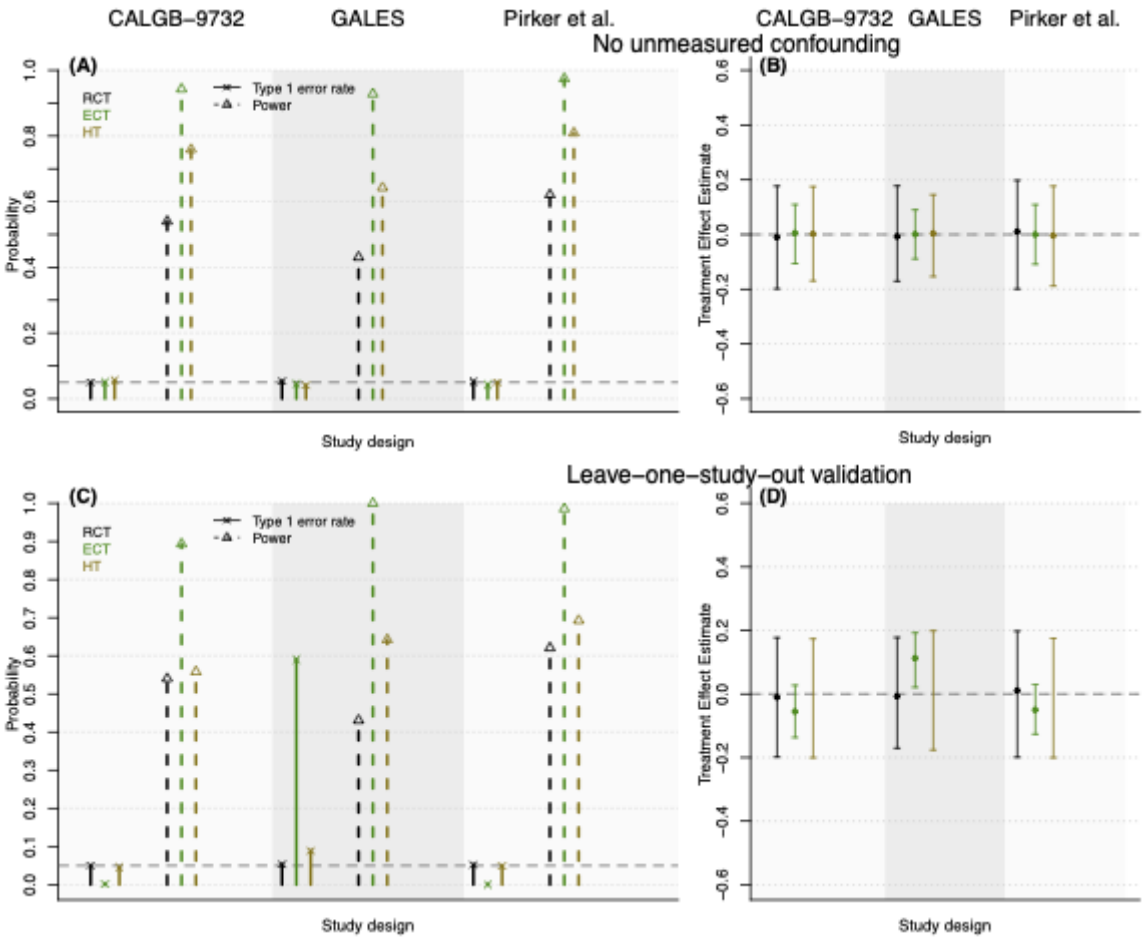

Figure 2

Operating characteristics of *in silico* ECT, HT, and RCT trials generated by resampling the control arms of the ES-SCLC studies. The top row shows type I error rates (Panel A, solid vertical lines with a cross), power (Panel A dotted vertical lines with an arrow), and the variability/bias of the treatment effect estimates (Panel B). In panel B, the dots indicate the average treatment effect estimates across *in silico* trials and

the vertical bars indicate the 5% and 95% quantiles. Panels A and B are representative of an ideal setting, without unmeasured confounders, and identical conditional outcome distributions of the SOC across studies. The bottom row (Panels C and D) shows the same operating characteristics when we used the leave-one-study-out resampling algorithm.

Figure 3

See image above for figure legend.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Tables13.pdf
- 2740741supp0r32prk.pdf