Design and Evaluation of an External Control Arm Using Prior Clinical Trials and Real-World Data

Steffen Ventz1,2,3, Albert Lai4, Timothy F. Cloughesy4, Patrick Y. Wen5, Lorenzo Trippa1,2,3, and Brian M. Alexander3,5,6

Abstract

Purpose: We discuss designs and interpretable metrics of bias and statistical efficiency of *externally controlled* trials (ECT) and compare ECT performance to randomized and single-arm designs.

Experimental Design: We specify an ECT design that leverages information from real-world data (RWD) and prior clinical trials to reduce bias associated with interstudy variations of the enrolled populations. We then used a collection of clinical studies in glioblastoma (GBM) and RWD from patients treated with the current standard of care to evaluate ECTs. Validation is based on a *'leave one out' scheme, with iterative selection of a single-arm from one of the studies, for which we estimate treatment effects using the remaining studies as external control. This produces interpretable and robust estimates on ECT bias and type I errors.

Results: We developed a model-free approach to evaluate ECTs based on collections of clinical trials and RWD. For GBM, we verified that inflated false positive error rates of standard single-arm trials can be considerably reduced (up to 30%) by using external control data.

Conclusions: The use of ECT designs in GBM, with adjustments for the clinical profiles of the enrolled patients, should be preferred to single-arm studies with fixed efficacy thresholds extracted from published results on the current standard of care.

Introduction

Randomized controlled trials (RCT) have been the gold standard for clinical experimentation since the Medical Research Council trial of streptomycin for tuberculosis in 1948 (1). Randomization is the foundation for many statistical analyses and provides a method for limiting systematic bias related to patient selection and treatment assignment (2). Indeed, many failures in phase III drug development may be attributed to overestimating treatment effects from previous early-stage uncontrolled trials (3). Although RCTs reduce the risk of bias compared with single-arm trials, they tend to require larger sample sizes to achieve the targeted power (4), take longer to complete enrollment, and patients have typically a lower propensity to enroll into a RCT than a single-arm trial (5–7).

Many methods have been suggested as a compromise between uncontrolled trials and RCTs (8–11). Recently, availability of data collected from electronic health records (EHR) at scale has increased the interest in using real-world data (RWD; ref. 12) as a “synthetic” or “external” control (12–14). In addition, data from prior clinical trials can be integrated in the design and analysis of single-arm trials (15) rather than using a single published estimate of the standard of care primary outcome distribution to specify a benchmark. Leveraging RWD and prior clinical trials has the potential for controlling for known prognostic factors that cause intertrial variability of outcome distributions. This can reduce bias in single-arm studies, and ultimately could lead to better decision making by sponsors and regulators.

In this article, we illustrate the design and validation of an externally controlled trial (ECT) design to test for therapeutic impact on overall survival (OS) using both RWD and data from prior clinical trials for patients with newly diagnosed glioblastoma (GBM). We compare the ECT design to single-arm trial designs and RCTs and show the benefits and limitations of the ECT approach.

Materials and Methods

General approach to design and evaluate ECTs

To design an ECT, estimate the sample size for a targeted power, and evaluate relevant operating characteristics, our approach was the following. First, define the patient population for the ECT (in our case GBM). Next, identify a set of prognostic factors associated with the outcome of interest. Finally, specify the control therapy and identify available datasets (trials and RWD).
for the control treatment and extract relevant outcomes and patients’ characteristics.

As described below, to evaluate the ECT design, the control arm of each study is compared (using adjustment methods) to an external control, which is defined by the remaining available data for patients that received the same control treatment. In these comparisons the treatment effect is zero by construction, which facilitates interpretability and produces bias and variability summaries for ECT’s treatment effect estimates, and type I error rates estimates.

If the ECT design maintains (approximately) the targeted type I error rate, we can then determine the sample size required for ECT, single-arm trial, and RCT designs for a targeted probability of treatment discovery at a predefined treatment effect.

The binary variable $A$ indicates the assignment of a patient to the experimental treatment, $A = 1$, or to the control arm, $A = 0$ and $Y$ denotes the outcome. We focus on binary endpoints, such as survival at 12 months from enrollment (OS12) and expand the discussion to time-to-event outcomes in the Supplementary Material. The vector $X$ indicates a set of pretreatment patient characteristics. We evaluate whether characteristics $X$ are sufficient to obtain (nearly) unbiased treatment effect estimates or not.

ECT design

The ECT is a single-arm clinical study that uses the trial data (experimental treatment) and external data (control) to conduct inference on treatment effects. More specifically, for a hypothetical randomized study, we estimate the unknown average treatment effect

$$E = \sum_x [P(Y = 1|A = 1, X) - P(Y = 1|A = 0, X)] P_X(x). \quad (1)$$

which is a weighted average of the conditional outcome probabilities weighted with respect to a distribution $P_X(x)$ of patients characteristics $X$. Possible definitions for $P_X(x)$ used by existing adjustment methods are the distribution of patient characteristics $X$ in the single-arm study, $P_{SAT}(x)$, or the distribution of $X$ in the external (historical) control, $P_{HIC}(x)$. The unknown probabilities $P(Y = 1|A, X)$ do not refer to a particular parametric statistical model but are unknown model-free quantities. We considered four adjustment methods, all based on the usual hypothesis of no unmeasured confounders (16) to estimate the unknown average treatment effect $TE$ (1), direct standardization, matching, inverse probability weighting, and marginal structural methods (Supplementary Material; refs. 16, 17).

Datasets

To develop an ECT design for newly diagnosed GBM, we used data from patients receiving standard temozolomide in combination with radiation (TMZ-RT) from both prior clinical trials and RWD (Table 1). Clinical trial data were from the phase III AVAglio (NCT00943826; ref. 18) trial and two phase II trials (PMID: 22120301 and NCT00441142; refs. 19, 20). RWD was abstracted from patients undergoing treatment for newly diagnosed GBM at the Dana-Farber Cancer Institute (Boston, MA) and UCLA (Los Angeles, CA), and a previously published RWD dataset (21).

Model-free evaluation

We evaluated the ECT design by mimicking the comparison of an ineffective experimental arm to an external control. Hypothetical ECT experimental arms were generated from data of the TMZ-RT arm of one of the studies in Table 1 using the following model-free procedure. For each study, we iterated the following steps:

(i) We randomly selected $n$ patients (without replacement) from the TMZ-RT arm of the study and use the clinical profiles $X$ and outcomes $Y$ of these patients as experimental arm of the ECT. Here, $n$ is the number of enrolled patients.

(ii) TE is estimated for the external control (step ii) using one of the adjustment methods (Supplementary Material), and tested the null hypothesis of no-benefit, $H_0 : TE \leq 0$, at a targeted type I error rate of 10%.

We repeated steps (i–iii) with different sets of $n$ randomly selected patients. Here, $n$ is less or equal to the size of the TMZ-RT arm of each GBM study.

A similar model-free procedure allows one to evaluate the operating characteristics of ECTs in presence of positive treatment effects, by reclassifying in step (i), randomly and with fixed probability, negative individual outcomes $Y$ into positive outcomes. Section S1 of the Supplementary Material presents a detailed description of this procedure.

Comparison, ECT, single-arm, and RCT designs

We compared the ECT to single-arm and RCT designs. We used the following criteria:

(i) bias and variability of treatment effect estimates,

(ii) deviations of type I error rates from targeted control of false positive results, and

(iii) the sample size to achieve a targeted power.

In a single-arm trial, an estimate of the proportion $\pi_E$ of patients surviving, at a specific time point, say 12 months (OS-12), is compared with a historical estimate for the standard-of-care $\pi_{HIC}$, typically the result of a prior trial. Here, $\pi_{HIC}$ can be expressed as a weighted average of the OS-12 probability for a patient with profile $X$, which is averaged over the study-specific distribution $P_{HIC}(x)$ of patient characteristics,

$$\pi_{HIC} = \sum_x P(Y = 1|A = 0, X)P_{HIC}(x). \quad (2)$$

Similarly, the parameter $\pi_E$ can be written as a weighted average of the probability $P(Y = 1|A = 1, X)$ over the distribution $P_{SAT}(x)$ of $X$ in the single-arm trial. If the distributions

Translational Relevance

The use of existing data (real-world data or prior trials) to design and analyze clinical studies has the potential to accelerate drug development processes and can contribute to rigorous examination of new treatments.
Design and Evaluation of an External Control Arm

**Table 1.** Distribution of pretreatment patient characteristics for the TMZ+RT arm of three clinical studies and three RWE studies

<table>
<thead>
<tr>
<th>Study</th>
<th>NCT ID</th>
<th>PubMed ID</th>
<th>DFCI-cohort</th>
<th>UCLA-cohort</th>
<th>NCT0103285</th>
<th>PM2135282</th>
<th>NCT00441142</th>
<th>PM25910950</th>
<th>—</th>
<th>PM22120301</th>
<th>AVAglio</th>
<th>NCT00943826</th>
<th>PM2455231B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arm</td>
<td>TMZ+RT</td>
<td>TMZ+RT</td>
<td>TMZ+RT</td>
<td>TMZ+RT</td>
<td>Phase II</td>
<td>Phase II</td>
<td>Phase II</td>
<td>Phase II</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Enrollments to SOC</td>
<td>378</td>
<td>305</td>
<td>110</td>
<td>29</td>
<td>16</td>
<td>344</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>OS events</td>
<td>265</td>
<td>265</td>
<td>89</td>
<td>24</td>
<td>15</td>
<td>344</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Age</td>
<td>Median</td>
<td>58</td>
<td>57</td>
<td>59</td>
<td>58</td>
<td>59</td>
<td>57</td>
<td>57</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Range</td>
<td>18-91</td>
<td>20-84</td>
<td>20-90</td>
<td>26-73</td>
<td>36-69</td>
<td>18-79</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>SD</td>
<td>13</td>
<td>13</td>
<td>14</td>
<td>11</td>
<td>11</td>
<td>10</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Sex (%)</td>
<td>Females</td>
<td>0.43</td>
<td>0.36</td>
<td>0.36</td>
<td>0.45</td>
<td>0.5</td>
<td>0.36</td>
<td>0.36</td>
<td>0.36</td>
<td>0.36</td>
<td>0.36</td>
<td>0.36</td>
<td>0.36</td>
</tr>
<tr>
<td>Sex (%)</td>
<td>Males</td>
<td>0.57</td>
<td>0.64</td>
<td>0.64</td>
<td>0.55</td>
<td>0.5</td>
<td>0.64</td>
<td>0.64</td>
<td>0.64</td>
<td>0.64</td>
<td>0.64</td>
<td>0.64</td>
<td>0.64</td>
</tr>
<tr>
<td>KPS (%)</td>
<td>&lt;80</td>
<td>0.55</td>
<td>0.39</td>
<td>0.32</td>
<td>0.24</td>
<td>0.44</td>
<td>0.31</td>
<td>0.31</td>
<td>0.31</td>
<td>0.31</td>
<td>0.31</td>
<td>0.31</td>
<td>0.31</td>
</tr>
<tr>
<td>KPS (%)</td>
<td>&gt;80</td>
<td>0.45</td>
<td>0.61</td>
<td>0.68</td>
<td>0.76</td>
<td>0.56</td>
<td>0.69</td>
<td>0.69</td>
<td>0.69</td>
<td>0.69</td>
<td>0.69</td>
<td>0.69</td>
<td>0.69</td>
</tr>
<tr>
<td>Data missing (n)</td>
<td>27</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>RPA (%)</td>
<td>3</td>
<td>NA</td>
<td>0.22</td>
<td>0.25</td>
<td>NA</td>
<td>0.12</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
</tr>
<tr>
<td>RPA (%)</td>
<td>4</td>
<td>NA</td>
<td>0.42</td>
<td>0.41</td>
<td>NA</td>
<td>0.75</td>
<td>0.61</td>
<td>0.61</td>
<td>0.61</td>
<td>0.61</td>
<td>0.61</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td>RPA (%)</td>
<td>5</td>
<td>NA</td>
<td>0.34</td>
<td>0.33</td>
<td>NA</td>
<td>0.13</td>
<td>0.23</td>
<td>0.23</td>
<td>0.23</td>
<td>0.23</td>
<td>0.23</td>
<td>0.23</td>
<td>0.23</td>
</tr>
<tr>
<td>RPA (%)</td>
<td>6</td>
<td>NA</td>
<td>0.02</td>
<td>0.01</td>
<td>NA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RPA (%)</td>
<td>Data missing (n)</td>
<td>378</td>
<td>0</td>
<td>0</td>
<td>29</td>
<td>1</td>
<td>0</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>RPA (%)</td>
<td>Resection (%)</td>
<td>Biopsy</td>
<td>0.14</td>
<td>0.22</td>
<td>0.21</td>
<td>0.21</td>
<td>0</td>
<td>0</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
</tr>
<tr>
<td>RPA (%)</td>
<td>Sub total</td>
<td>0.47</td>
<td>0.47</td>
<td>0.36</td>
<td>0.48</td>
<td>0.31</td>
<td>0.49</td>
<td>0.49</td>
<td>0.49</td>
<td>0.49</td>
<td>0.49</td>
<td>0.49</td>
<td>0.49</td>
</tr>
<tr>
<td>RPA (%)</td>
<td>Gross total</td>
<td>0.39</td>
<td>0.31</td>
<td>0.43</td>
<td>0.31</td>
<td>0.69</td>
<td>0.42</td>
<td>0.42</td>
<td>0.42</td>
<td>0.42</td>
<td>0.42</td>
<td>0.42</td>
<td>0.42</td>
</tr>
<tr>
<td>RPA (%)</td>
<td>Data missing (n)</td>
<td>12</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RPA (%)</td>
<td>MGMT (%)</td>
<td>Unmethylated</td>
<td>0.43</td>
<td>0.71</td>
<td>0.61</td>
<td>0.86</td>
<td>0.43</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
</tr>
<tr>
<td>RPA (%)</td>
<td>Sub total</td>
<td>0.57</td>
<td>0.29</td>
<td>0.40</td>
<td>0.14</td>
<td>0.56</td>
<td>0.32</td>
<td>0.32</td>
<td>0.32</td>
<td>0.32</td>
<td>0.32</td>
<td>0.32</td>
<td>0.32</td>
</tr>
<tr>
<td>RPA (%)</td>
<td>Data missing (n)</td>
<td>194</td>
<td>128</td>
<td>40</td>
<td>7</td>
<td>0</td>
<td>0.23</td>
<td>0.23</td>
<td>0.23</td>
<td>0.23</td>
<td>0.23</td>
<td>0.23</td>
<td>0.23</td>
</tr>
<tr>
<td>RPA (%)</td>
<td>IDH (%)</td>
<td>Wild-type</td>
<td>0.91</td>
<td>0.91</td>
<td>0.98</td>
<td>0.83</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>RPA (%)</td>
<td>Sub total</td>
<td>0.09</td>
<td>0.09</td>
<td>0.02</td>
<td>0.17</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>RPA (%)</td>
<td>Data missing (n)</td>
<td>188</td>
<td>0.46</td>
<td>52</td>
<td>6</td>
<td>16</td>
<td>344</td>
<td>344</td>
<td>344</td>
<td>344</td>
<td>344</td>
<td>344</td>
<td>344</td>
</tr>
</tbody>
</table>

Abbreviations: IDH1, isocitrate dehydrogenase 1; KPS, Karnofsky performance status; MGMT, O6-methylguanine-DNA methyltransferase; RPA, recursive partitioning analysis.

PrSAT(X) and PrHC(X) differ substantially, then πHC and πE are not comparable, treatment effect estimates can be biased, and type I error rates can deviate from the targeted value. If the patient’s prognostic profiles in the single-arm study are favorable compared with the study used as benchmark, then the type I error probability tends to be above the targeted α-level, and vice versa. In the latter case, the power decreases.

In an RCT, patients are randomized to the control and experimental arm, with patient characteristics, on average, equally distributed between arms, reducing the risk of bias compared with single-arm trial designs.

**Results**

**Limitations of the single-arm design**

We illustrate the bias and type I error deviation associated with single-arm trials using an example for a hypothetical ineffective experimental treatment in a disease with one known prognostic biomarker X. We assume, for each patient, identical outcome probabilities under the experimental and control treatment. Figure 1A shows the difference (πE − πHC) when the prevalence of the biomarker varies between PrSAT(X = 1) = 0.1 and 0.9 for different correlation levels between the outcome Y and the biomarker X. Even with a moderate association between the biomarker and the outcome, the differences between the distributions (PrHC − PrSAT) result in bias and departures from the intended 10% type I error rate (Fig. 1B).

**TMZ+RT in newly diagnosed GBM**

The standard of care of TMZ+RT for newly diagnosed GBM was established in 2004 based on results from the EORTC-NCIC CE.3 (22). Subsequently, nine additional trials enrolled patients on TMZ+RT control arms between 2005 and 2013 (Supplementary Table S1). The majority of single-arm studies used the reported results of EORTC-NCIC CE.3 as historical benchmark (Supplementary Table S1). Sample sizes of the TMZ+RT control arms in the RCTs varied between 16 (19) patients and 463 (18) patients. Supplementary Figure S1 shows reported Kaplan–Meier estimates, median OS, and OS-12 for the TMZ+RT arms. Point estimates for OS-12 varied between 0.56 and 0.81 across studies, and between 13.2 and 21.2 months for median OS.

**Prognostic variables**

Through a literature review, we identified prognostic factors associated with survival in newly diagnosed GBM (23–26). A Cox regression analysis, stratified by trial and treatment arm, was used to quantify association of covariates with OS (Table 2). On multivariable analyses, age (HR 1.03; P < 0.001), male gender...
HR 1.15, \( P = 0.012 \), KPS >80 (HR 0.78; \( P < 0.001 \)), gross total resection versus biopsy (HR 0.62; \( P < 0.001 \)), subtotal resection versus biopsy (HR 0.82; \( P = 0.028 \)), MGMT promoter methylation (HR 0.46; \( P < 0.001 \)), and IDH1 (HR = 0.52; \( P = 0.01 \)) showed association with OS.

The prevalence of these factors varied across studies, from 0.5 to 0.64 for male gender, from 0.45 to 0.76 for KPS >80, and 0.14 to 0.57 for MGMT methylated status. Minimum (maximum) age varied between 18 and 36 (68 and 91) years across trials, and also resection showed noticeable variation across trials. We selected all five variables (age, gender, KPS, MGMT, extent of resection) for adjustments in the ECT design.

ECT and inconsistent definitions of outcomes and pretreatment characteristics

We initially generated for each study \( j = 1, 2, \ldots, 6 \) in Table 1, an ECT by selecting all patients on the TMZ+RT arm. This produces a hypothetical experimental arm, which is compared with all TMZ+RT patients in the remaining studies with adjustments for differences in patients' characteristics \( X \) (Supplementary Fig. S2). Treatment effects estimates appeared biased for the NCT01013285 dataset (\( \tilde{T}_{\text{ECT}} = 0.1, 90\% \) confidence interval (CI), 0.01–0.18). Upon further inspection of the definitions of patient characteristics and outcome, we noticed that OS in NCT01013285 was defined as time from diagnosis to death. In contrast, the clinical trials (and DFCI, UCLA cohorts) used time of randomization (beginning of therapy) to death. Unsurprisingly, different definitions of the outcome or prognostic variables can be important sources of bias.

Evaluation of the ECT design

In consideration of the described definitions of the outcome \( Y \) across studies, we removed the NCT01013285 dataset (Figs. 2 and 3).

Model-free ECT evaluation

Figure 2A shows ECT treatment effect estimates for each of the remaining 5 studies. Treatment effect estimates, in all cases,
Table 2. Pretreatment patient characteristics associated with OS estimated
HRs in univariable and multivariable stratified Cox regression models. The baseline hazard rate was stratified by study and treatment arm.

<table>
<thead>
<tr>
<th>Model variables</th>
<th>Univariable</th>
<th>Multivariable</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HR</td>
<td>p</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear</td>
<td>1.02</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Sex (ref. female)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1.17</td>
<td>0.004</td>
</tr>
<tr>
<td>KPS (ref. ≤ 80)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>&gt;80</td>
<td>0.64</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>RPA (ref. class 3)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Class 4</td>
<td>1.50</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Class 5</td>
<td>2.29</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Class 6</td>
<td>7.10</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Resection (ref. biopsy)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sub total</td>
<td>0.78</td>
<td>0.001</td>
</tr>
<tr>
<td>Gross total</td>
<td>0.56</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>MGMT (ref. unmethylated)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Methylated</td>
<td>0.47</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>IDH (ref. wild-type)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Mutant</td>
<td>0.35</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

were close to zero. In comparison, a single-arm trial design, with the EORTC-NCIC CE.3 used as historical benchmark (Supplementary Table S1), would lead to overestimation of treatment efficacy.

Next, we generated ECTs for a fixed sample size of n = 46. The sample size was selected for a targeted power of 80% and 10% type I error rate, for a single-arm trial (one-sided binomial test) with OS-12 improvement from \( \hat{H}_{EC} = 61\% \) to 76%, with \( \hat{H}_{EC} \) from the EORTC-NCIC CE.3 (22) study. Because the TMZ+RT arm of PM22120301 and NCT00441142 had only 16 patients and 29 patients, we could not use these studies to generate ECTs with size n = 46. Figure 2B and C show the results of nearly identical analyses as Fig. 2A across 10,000 subsamples of n = 46 randomly selected patients using four adjustment methods—direct standardization (DSM), matching (M), inverse probability weighting (IPW), and marginal structural model (MSM). For IPW and MSM, we used multiple reference distribution \( P_X(x) \) (see expression 1; refs. 27–29). We used IPW-T in the analyses for Fig. 2A and C. Supplementary Fig. S3 shows ECT treatment effect estimates obtained by adjustment using different sets of prognostic characteristics.

Figures 2C shows the distributions of treatment effect estimates across the generated trials for ECT (in blue) and RCT (in black). The RCT data were obtained by randomly dividing the simulated single-arm trial dataset into two parts of 23 patients, which are labeled as control and experimental arms. With identical sample size (n = 46), the assignment of all patients in ECT to the experimental arm results in lower variability of the treatment effect estimates compared with the RCT. The empirical type I error rate (targeted value 10%) across generated ECTs (model-free analysis with n = 46) was 9.1%, 6.1%, and 8.6% for the TMZ+RT arm of the AVAglio, the DFCI cohort, and the UCLA cohort, compared with 40.7%, 21.9%, and 40.5% for the single-arm trial design [historical benchmark: 0.61 reported in EORTC-NCIC CE.3 (22)] and a targeted value of 10%, respectively. The latter estimates, well above the 10% target, are consistent with different outcome distributions under TMZ+RT observed in these three studies compared with the EORTC-NCIC-CE.3 study. Indeed, underestimation of TMZ+RT’s efficacy in single-arm trial translates into inflated type I error rates.

Model-based ECT evaluation

Figure 3A and B–F report model-based type I error rates and power for hypothetical RCTs, single-arm trials, and ECTs with sample size ranging between 20 and 160. We used the pretreatment characteristics X from the five studies to evaluate designs using a model-based approach (Supplementary Material), which consists in sampling baseline characteristics X from the studies and generating outcomes Y from models \( P(Y|X, A) \). We specify \( P(Y|X, A) \) with a logistic model, obtained by fitting the TMZ+RT data from all five studies combined, with or without the addition of a positive treatment effect.

Both the RCT and ECT have false positive rates, across simulated trials, close to the targeted value of 10% for all five studies (Fig. 3A). The single-arm trial design (historical benchmark: EORTC-NCIC CE.3) overestimates treatment effects, and presents inflated type I error rates, 21%–59% for n = 30 and 30%–83% for n = 60 patients (Fig. 3A).

The reported power in Fig. 3B–D corresponds to a scenario with improvement in OS-12 equal to an odds ratio of 2.6. For example, with X corresponding to a male patient, age 59 (median age in the studies TMZ+RT arms) with biopsy, KPS ≤ 80, and negative MGMT status \( Pr(Y = 1|A = 1, X) \) – \( Pr(Y = 1|A = 0, X) \) = 0.15. The RCT requires more than 139 patients (139, 140, 137, 154, 150 patients, X-distributions of DFCI, UCLA cohorts, AVAglio, PM22120301, and NCT00441142) to achieve a power of 80% at 10% type I error rate. In contrast, the ECT requires between 34 and 40 patients (34, 34, 34, 40, and 37 patients) to achieve the 80% power.

Discussion

Clinical researchers have discussed and debated the relative merits of single-arm versus randomized trial design (4, 30–36). Single-arm trials have obvious attraction for patients, could potentially be smaller, and are logistically easier to employ as pragmatic trials. The associated increased risk for bias, however, could lead to poor therapeutic development and regulatory decision-taking. Overly optimistic analysis and interpretation can result in large negative phase III trials. Negatively biased results can cause discontinuation in the development of therapies with positive effects. This potential for bias is less pronounced for endpoints with minimal variation under the control. For example, single-arm designs for monotherapies using tumor response as an endpoint have low risk for inflated type I error (37). Evaluation of therapeutic combinations and use of endpoints such as PFS and OS are more complicated with increased risk for biased results; however, randomization is the best way to limit this bias. But alternative methodologies could improve on single-arm designs without the limitations of setting up a randomized control.

Historic benchmarks in single-arm trial designs have two major problems. The first problem is ignoring discrepancies of the estimated survival functions across trials due to population differences. Controlling for known prognostic factors has been shown to mitigate this issue somewhat (15). In addition, single-arm trials by design compare a single point of a PFS or OS curve, for example OS-12, to a benchmark. Such approaches do not leverage the power of statistical analyses that incorporate all time-to-event data, including censoring. The use of external control arms can address both these limitations.

Clearly, ECTs can increase power compared with RCTs by leveraging additional information from outside of the trial rather...
than committing resources to an internal control. In our analysis in GBM, the ECT reached nearly the same power as a single-arm trial that specifies the correct historical response rate (zero bias) of the standard of care. This efficacy gain of ECTs will not necessarily be the same in other disease settings, and will depend on the size of the external control, the number of patient characteristics X required to control for confounding, and the variability of these patient characteristics across and within studies. By standard statistical arguments (38–40), in settings that are favorable to the ECT design (large external control cohort, a few relevant covariates, and no unmeasured confounders), the sample size for an RCT to match uncertainty summaries of ECTs such as standard deviation or the length of a CI for TE is approximately four times larger than the sample size of an ECT. The major question is whether this comes with the downside of increasing type I errors. In our analysis in newly diagnosed GBM, the ECT’s type I error rates were comparable to RCTs, whereas single-arm trials showed significantly inflated type I error rates. These results are in concordance with previous findings of a meta-analysis in GBM of Vanderbeek and colleagues, 2018 (41), which associate single-arm phase II trial in GBM during a 10-year period with underestimation of the TMZ+RT efficacy. The majority of single-arm studies used the EORTC-NCIC CE.3 trial as historical benchmark. Underestimation of the control efficacy translates into inflated type I error rates.

Another key question is whether our results were limited to newly diagnosed GBM or generalizable. We can consider three questions when evaluating the use of an ECT for a given experimental treatment: (i) are there time trends in the outcome under the control; (ii) are the available prognostic factors sufficient to explain most of the variation in the outcome distributions across trials; and (iii) is there evidence of significant latent unobserved confounding after controlling for known prognostic factors. Our evaluation of ECTs in GBM required use of an entire collection of datasets, to address these questions, and there is not a simple strategy to determine how other diseases or indications might compare.

Figure 2.
Treatment effect estimates of the ECT design. For each study, the TMZ + RT arm was used as ECT’s experimental arm and (after adjustment for patients’ characteristics) compared with the TMZ + RT arms of the remaining four studies. A, For each of the study, covariate adjusted treatment effect estimates (point estimates and 90% CI, n equals to the arm-specific size). B, Treatment effect estimates (average value, 5th and 95th percentile) across 10,000 subsamples of n = 46 patients using different adjustment methods. We consider direct standardization, matching, inverse-probability weighting, and marginal structural models (DSM, PS-M, IPW, MSM). For IPW and MSM, we use different reference distributions Pr_x(x) (see expression 1) of pretreatment characteristics X. C, The distribution of treatment effect estimates of the ECT (blue line) and RCT (black line) across subsamples of n = 46 patients.
The first step of our validation analyses was the selection of potential confounders. On the basis of previous recommendations, see for example Greenland, 2008 (42), before data collection and analyses, we identified potential confounders through a review of the GBM literature. In selecting the set of patient characteristics, we tried to be as comprehensive as possible, because the exclusion of confounders can compromise the ECT performance. Sensitivity summaries of the validation analysis, similar to Supplementary Fig. S2, can then be computed to illustrate variations of the estimated ECT performances when smaller sets of variables are used for adjustments.

A strength of our evaluation was the use of both data from prior clinical trials and RWD. Most discussions of external controls tend to focus on one or the other, but each has strengths and weaknesses. Clinical trial data are more meticulously collected, resulting in more standardized definitions and entry. This was evident in our dataset where several covariates from RWD datasets were characterized by missing data. This problem can be mitigated through the use of RWD datasets at scale rather than from a single institution. Furthermore, we initially found erroneous treatment effects due to differences in the definitions of the time-to-event variables (index dates) in our RWD compared with the remaining RCTs. Although this is easily correctable in our example, care must be taken to define endpoints in RWD (43). Conversely, RWD has the advantage of being generated during routine clinical care, which is less costly, potentially available at larger scale, and more contemporary. Because each kind of data have benefits and limitations, leveraging both has value for ECT generation.

A limitation of our study is the relatively small number of datasets that we used to evaluate the ECT design. The higher the number of available trials and cohorts, the more precise the estimation of ECT type I error rates and other key operating characteristics can be. The extension of our results in the future for GBM and the generation of ECTs for other diseases will undoubtedly be aided by clinical trial data sharing efforts like Project Data Sphere (44), Vivli (45), YODA (46), and the availability of RWD at scale from groups like Flatiron Health, Tempus, and ASCO through CancerLinQ (47).

The validation procedure that we used builds on clinical studies that included the same control (in our case TMZ+RT) with the aim to evaluate the use of ECT for future trials. A limitation of the procedure is the identical use of all the available studies. Potentially relevant differences, such as the year when each study started the enrollment, are not considered. Nonetheless, a simple and interpretable scheme for validation has the advantage of being robust to selection bias. We included all studies for which we could access patient-level data. However, after the utility of the ECT design has been rigorously evaluated for a specific disease setting based on interpretable and robust procedures, it becomes appropriate

Figure 3.
Model-based evaluation of the type I error and power for RCT, ECT, and single-arm trial (SAT) designs for an overall study sample size of $n = 20, \ldots, 160$ patients. In the model-based approach (Supplementary Material), we sampled baseline characteristics $X$ from the five studies in Table 1, and generated outcomes $Y$ from models $Pr(Y|X, A)$. A, For all studies, the type I error rates of RCT, ECT, and SAT designs at different overall sample sizes. Different line types (solid, dashed, dotted, etc.) indicate different studies (Table 1). B–F, For each study, the power of RCT, SAT, and ECT designs, and sample size to achieve 80% power (dotted vertical lines). In A, the SAT experimental outcomes have been generated as in the ECT simulations, but outcomes $Y$ are directly compared with the EORTC-NCIC CE.3 study estimates, without adjustments for different distributions of patients’ characteristics. For RCTs, half of the randomly selected profiles $X$ are used to define the experimental arm and the remaining half defines the control arm. Two-group (RCT) and single-group (SAT) $z$-tests for proportions were used for testing. To compute the power in B–F of the SAT, we assumed that the historical control benchmark $x_{HC}$ was correctly specified.
to refine the set of studies used as external control. This could be done for example by multistudy analyses to identify studies with a risk of inducing bias in the ECT results if used as external control. Moreover, Bayesian models that incorporate differences across studies (48, 49) could be used to compute treatment effects estimates and credible intervals for ECTs.

An extension of our validation framework could include the evaluation of procedures to use external control data in the analysis and interpretation of RCTs. In some cases, external data may contribute to more accurate treatment effects estimates in RCTs. Statistical methods for the use of external control data in RCTs, for example early-stopping rules, will require, similar to ECT designs, validation studies before their implementation in clinical trials.

Conclusions
ECT designs have the potential to improve the evaluation of novel experimental agents in clinical trial and accelerate the drug development process by leveraging external data. Challenges in the use external data compared with standard RCTs include (i) the identification of a comprehensive list of potential confounders X for adjustments, (ii) access to a large set of RWD datasets or completed RCT to create a library of studies for robust validation analyses, (iii) availability of patient-level data and possible missing data problems, (iv) coherent definitions and consistent measurements of patient characteristics and outcomes across datasets, (v) possible trends in calendar time in the distributions of the outcomes under the control treatment due to improved clinical practice and (vi) the use of robust statistical procedures to evaluate ECT designs in comparison with traditional single-arm and RCT designs.

Here, we introduced a simple algorithm to evaluated operating characteristics such as bias, variability of treatment effect estimates, and type I error rates of ECT designs. We considered different ECT designs that use distinct adjustment methods. Our results indicate that the ECTs constitute a useful alternative to standard single-arms trials and RCTs in GBM, which could significantly reduce the current false positive rates (41) of single-arm phase II GBM trials.

Disclosure of Potential Conflicts of Interest
A. Lai reports receiving commercial research grants from Genentech/Roche and is a consultant/advisory board member for Merck. T. F. Cloughesy holds ownership interest (including patients) in Notable Labs, is listed as a co-inventor on a patent regarding the composition of matter for cancer therapy owned by UCLA, and is a consultant/advisory board member for Bayer, Del Mar Pharmaceuticals, Tocagen, Kryopharm, GW Pharma, Klyceptor, Abbvie, Roehring Ingelheim, VBL, Deacervice, VBL, Agios, Merck, Roche, Genocea, Celgene, Puma, Lilly, Bristol-Myers Squibb, Coricite, Welome Trust, Novacure, Novogen, Boston Biomedical, Sunovion, Human Longevity, Insyxs, ProNaI, Pfizer, Notable Labs, and MedQia. B. M. Alexander is an employee of Foundation Medicine, Inc., and holds ownership interest (including patients) in Hoffman La Roche. No potential conflicts of interest were disclosed by the other authors.

Authors’ Contributions
Conception and design: S. Ventz, P.Y. Wen, L. Trippa, B.M. Alexander
Development of methodology: S. Ventz, L. Trippa, B.M. Alexander
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): S. Ventz, A. Lai, T.F. Cloughesy, P.Y. Wen, B.M. Alexander
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): S. Ventz, A. Lai, T.F. Cloughesy, L. Trippa, B.M. Alexander
Writing, review, and/or revision of the manuscript: S. Ventz, A. Lai, T.F. Cloughesy, P.Y. Wen, L. Trippa, B.M. Alexander
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): S. Ventz, B.M. Alexander
Study supervision: S. Ventz, B.M. Alexander

Acknowledgments
This work was supported by the Burroughs Wellcome Innovations in Regulatory Science Award.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received March 11, 2019; revised April 28, 2019; accepted June 4, 2019; published first June 7, 2019.

References
Design and Evaluation of an External Control Arm Using Prior Clinical Trials and Real-World Data


Updated version
Access the most recent version of this article at:
doi:10.1158/1078-0432.CCR-19-0820

Supplementary Material
Access the most recent supplemental material at:
http://clincancerres.aacrjournals.org/content/suppl/2019/06/07/1078-0432.CCR-19-0820.DC1

Cited articles
This article cites 48 articles, 9 of which you can access for free at:
http://clincancerres.aacrjournals.org/content/25/16/4993.full#ref-list-1

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, use this link
http://clincancerres.aacrjournals.org/content/25/16/4993.
Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.