Supplementary Material for "Design and validation of an external control arm using prior clinical trials and real-world data"

Steffen Ventz PhD, Albert Lai, Timothy F. Cloughesy, Patrick Y. Wen MD,
Lorenzo Trippa PhD, Brian M. Alexander MD MPH

April 28, 2019

Contents

S1 Model-free evaluation of ECTs in presence of positive treatment effects

Hypothetical ECTs can be generated using the TMZ+RT arm of one of the studies in Table 1. For each study, the algorithm iterates the following steps.

(a) Select randomly (without replacement) n patients with clinical profiles X_i and outcomes Y_i from the TMZ+RT arm of the study

(b) For all selected patients (step-a) with negative outcome $Y_i = 0$, change randomly and with a fixed probability Δ ($0 < \Delta < 1$) the value of Y_i from 0 to 1. Then use the clinical profiles X_i and (modified) outcomes Y_i of all n patients as experimental arm of the ECT. [1]

If $\Delta = 0$ the whole procedure becomes identical to the one discussed in the main manuscript.

(c) Use the TMZ+RT arms of the remaining studies (Table 1) as external control.

(d) Estimate the treatment effect TE comparing the experimental arm (step a+b) and the external control (step c) using one of the adjustment methods (Section S3), and test the null hypothesis of no-benefit, $H_0 : TE \leq 0$, at a targeted type I error rate of 10%.

Repeat steps (a-c) 10,000 times and compute for each study the proportion of ECT tests that rejected the null hypothesis at 10% type I error level.

S2 Model-based evaluation of ECTs

We also used a model-based approach to evaluate the ECT. Based on studies in Table 1, we estimated a logistic model for the response to TMZ+RT given patient characteristics, $\widehat{Pr}(Y = 1|A = 0, X) = F(X'\widehat{\beta})$, where $F(x) = 1/(1+\exp\{-x\})$. We fitted the model assuming identical regression coefficients $\widehat{\beta}$ across all studies. A positive treatment effect γ (regression parameter) is added to specify a probability model $\widehat{Pr}(Y = 1|A = 1, X) = F(X'\widehat{\beta} + \gamma)$ for an effective experimental treatment. We

[1] The parameter Δ can be consideblack the magnitude of the experimental treatment effect compablack to the control. Using counterfactual notation $\Delta = Pr(Y_i(T) = 1|Y_i(C) = 0)$, where $Y_i(T)$ and $Y_i(C)$ are the potential outcomes if the patient i receives the experimental or the control therapy. This is the probability of a positive outcome $Y_i(T) = 1$ under the experimental treatment conditionally on the fact that the outcome would have been negative if the patient was treated with the control $Y_i(C) = 0$. We assume $Pr(Y_i(T) = 1|Y_i(C) = 1) = 1$.

then generate for each study in Table 1 a hypothetical ECT with effective experimental arm and fixed sample size n:

(a) Select n patient profiles X (with replacement) from the study and generate the corresponding outcomes Y using $\widehat{Pr}(Y = 1|A = 1, X) = F(X'\widehat{\beta} + \gamma)$ for the experimental arm.

(b) Randomly select N patient profiles X from the remaining studies (where N is the sample size of external control data) and generate outcomes Y using $\widehat{Pr}(Y = 1|A = 0, X) = F(X'\widehat{\beta})$ for TMZ+RT.

(c) Conduct a covariate-adjusted ECT test using experimental arm data (step a) and the external control (step b).

We repeated steps (a-c) 10,000 times and computed for each study the proportion of ECT tests that rejected the null hypothesis at $\alpha = 0.1$. By repeating this calculation over a grid of sample sizes we determine the smallest size that achieve an 80% power. We repeated this analysis for all 5 studies in Table 1.

S3 Statistical Details for the ECT design

We summarize the four methods (direct standardization, matching, inverse probability weighting, marginal structural models) that we used to estimate treatment effects

$$TE = E_X\Big\{Pr[Y = 1|X, A = 1] - Pr[Y = 1|X, A = 0]\Big\}$$

in the ECT design. The first method uses a probit regression model to estimate the unknown probabilities $Pr(Y = 1|A, X)$ and the unknown treatment effect TE. Whereas the remaining methods use matched pairs of outcomes Y_i (matching) and weighted samples Y_i of patients in the experimental and control arm to estimate the unknown TE, without directly estimating the conditional probabilities $Pr(Y = 1|A, X)$.

All adjustment methods use assumptions that are difficult to test, for instance the absence of unmeasublack confounders, and use modeling assumptions that may be violated in practice. This motivates the use of a model-free procedure (see the method section of the manuscript) to evaluate bias, type I error rates and other operating characteristics.

(i) With **direct standardization** [1, 2] the average treatment effect TE,

$$TE = E_X\Big\{Pr[Y=1|X,A=1] - Pr[Y=1|X,A=0]\Big\}$$

is estimated by first fitting a regression mode for the response to treatment $\widehat{Pr}[Y=1|X,A=a]$ given the pre-treatment characteristics vector X and treatment assignment $a=0,1$. We use a logistic regression model. Then, for each patient i with characteristics X_i, irrespective of the actual treatment A_i, the difference between $\widehat{Pr}[Y=1|X_i,A_i=1]$ and $\widehat{Pr}[Y=1|X_i,A_i=0]$ is computed, conditioning on the hypothetical events that the patient had been assigned to arm $A_i=1$ or $A_i=0$. Lastly, the difference is averaged over patients $i=1,\ldots,n$,

$$\widehat{TE}_{DS} = \sum_i \Big\{\widehat{Pr}[Y=1|X_i,A_i=1] - \widehat{Pr}[Y=1|X_i,A_i=0]\Big\}\Big/ n \ .$$

(ii) We used a **Matching** algorithm based on estimates $\widehat{e}_i$ of the patient's propensity scores $e_i = Pr(A=1|X_i)$. Under standard (non-verifiable) assumptions [3] each individual pair of potential/counterfactual outcomes, under the control and experimental treatment, is independent of the treatment assignment A_i given e_i. The propensity score can be used to match patients i in the experimental arm and patients j in the control arm with similar propensity scores $e_i \approx e_j$. If exact matching on e_i can be achieved, then the distribution of pre-treatment variables would be identical on both arms [4]. For each patient i in the experimental arm, $A_i=1$, we indicate with $j(i)$ the index of the patients on the control arm with estimate propensity score closest to patient i ($|\widehat{e}_i - \widehat{e}_{j(i)}| \leq |\widehat{e}_i - \widehat{e}_{j'}|$ for all j' with $A_{j'}=0$). Then the average treatment effect with $P_X = P_{SAT}$ is estimated by

$$\widehat{TE}_M = 1/n_m \sum_{i:A_i=1} \Big\{Y_i - Y_{j(i)}\Big\} I(|\widehat{e}_i - \widehat{e}_{j(i)}| < \epsilon),$$

where ϵ (see [3]) restricts matching to patients with low propensity score dissimilarity and n_m indicates the number of patients i with $A_i=1$ that are matched, $|\widehat{e}_i - \widehat{e}_{j(i)}| < \epsilon$.

(iii) **Inverse probability weighting** (IPW) methods [5, 6, 7] estimate the average treatment effect in a population with reference distribution P_X by contrasting weighted averages of outcomes Y_i of patients in the control and experimental arms. More specifically, the IPW estimator $\widehat{TE}_{IPW}$ is

defined as

$$\widehat{TE}_{IPW} = \sum_i A_i \times w_{i,1} \times Y_i - \sum_i (1 - A_i) \times w_{i,0} \times Y_i,$$

where for both groups $a = 1$ and $a = 0$ the weights $w_{i,a} \in [0, 1]$ sum to $1 = \sum_{i:A_i=a} w_{i,a}$ [7]. These weights $w_{i,a}$ are used to re-weight the outcomes Y_i of patients in both treatment groups to produce pseudo-samples of control and experimental patients with (approximately) common distribution of pre-treatment characteristics $X_i \sim P_X$, and are defined as

$$w_{i,a} \propto a\frac{h(e_i)}{e_i} + (1 - a)\frac{h(e_i)}{1 - e_i}. \tag{S1}$$

The function $h(\cdot)$ defines the particular reference distribution P_X [6, 7], and Table S1 summarizes the functions $h(\cdot)$ that we used [7] .

$h(x)$	Target population X
1	Combined population
e_i	SAT population
$1 - e_i$	External control population
$e_i(1 - e_i)$	Overlap population
$I\{e_i \in [\alpha, 1 - \alpha]\}$	Truncated combined population
$\min\{e_i, 1 - e_i\}$	Matching

Table S1: Reference Distributions used in IPWs and MSMs

(iii) **Marginal structural (regression) models** (MSMs) [8, 9] estimate the average treatment effect TE by first estimating a marginal regression functions $E_{P_X}[E[Y|A = a, X]] = g(\beta_0 + \beta_{TE}a)$ for the outcome Y given experimental and control treatment $a = 0, 1$ and then setting $\widehat{TE}_{ave} = g(\beta_0) - g(\beta_0 + \beta_{TE}a)$ [9]. The marginal regression function is estimate by maximizing the weighted log-likelihood $l(\beta) = \sum_{i=1}^{n} w_{i,A_i} \log Pr(Y_i|A_i, \beta)$ [9] with weights w_{i,A_i} defines as in IPW (see Table S1 and equation (S1)) .

For all four methods, confidence intervals for the average treatment effect have been generated by a bootstrap algorithm [3, 10, 11]. Let n_E and n_C be the sample size of the experimental and control arm. The algorithm draws n_E patients (with replacement) from the original set of experimental patients $(X_i, A_i = 1, Y_i), i = 1, \cdots, n_E$ and n_C patients from the observed control arm $(X_i, A_i = 0, Y_i), i = 1, \cdots, n_C$. We then apply one of the above causal inference methods to the resampled data and estimate the average treatment effect $\widehat{TE}^{(boot)}$. We repeat these two steps 10000 times to

obtain resampled bootstrap estimates $\widehat{TE}^{(boot,c)}, c = 1, \cdots, 10000$ and use the empirical 5% and 95% percentiles of these estimates to obtain approximate 90% confidence intervals.

S3.1 ECT design with OS outcomes

Various methods to control for confounding (imbalance in the distribution of pre-treatment variables) with time-to-event data have been proposed; for instance MSMs and direct standardization for Cox and accelerated failure time models [12, 13, 14], and IPW methods to estimating survival functions [15, 16]. We use an IPW method for survival functions proposed by [15]. Let $t_j, j = 1, \ldots, J$ indicate the ordeblack failure times in the sample, $C_i \in \{0, 1\}$ indicate if the patient outcome Y_i is censoblack ($C_i = 0$), $w_{i,a}$ is defined as above in (S1), and $d_{w,a}(t_j) = \sum_i I(Y_i = t_j, C_i = 1, A_i = a) w_{i,a}$ and $N_{w,a}(t_j) = \sum_i I(Y_i \geq t_j, A_i = a) w_{i,a}$ indicate the weighed number of deaths and patients at risk at time t_j for patients in the experimental arm $a = 1$ and external control arm $a = 0$. Then the IPW estimate of the survival function for arm $a = 0, 1$ equals

$$\widehat{S}_a(t) = \prod_{j:t_j \leq t} \left(1 - \frac{d_{w,a}(t_j)}{N_{w,a}(t_j)}\right). \tag{S2}$$

S3.2 Prior work on integration of external data into clinical trials

Most RCTs and single arm trial designs to evaluate experimental therapies utilize either a study-specific control arm or a single historical benchmark value for the control therapy. Pocock [17], to the best of our knowledge has been the first to discuss statistical methods to incorporate an external control data into a RCT with binary endpoints by modeling inter-study variability with random effects. Thall and Simon [18] consideblack an RCT design that leverages external control data and selects the randomization parameters with the aim to minimizes the variance of the treatment effect estimate, again utilizing a random effects model. We previously discussed methods specific for Bayesian multi-arm studies and platforms [19, 20, 21, 22] that evaluate multiple treatment, in some cases studied during different periods [23, 24], sharing a control arm [25, 26, 27]. Bayesian models to incorporate external control data in the evaluation of new treatments, based on power priors, commensurate priors, and meta-analysis techniques, have been discussed in Neuenschwander et al. [28], Schmidli et al. [29], Viele et al. [30], van Rosmalen et al. [31], Kaizer et al. [32] and references therein. Most of these methods

focus on study-specific marginal probabilities $Pr(Y|A)$, without modeling individual profiles $Pr(x)$ and conditional distributions $Pr(Y|A,x)$ across studies. Hobbs et al. [33] and Murray et al. [34] used Bayesian regression models to incorporate historical control data for normal and time-to-event outcome data. These approaches are based on hierarchical models and commensurate priors that allows estimation of conditional treatment effects $TE(x) = Pr(Y|A=1,x) - Pr(Y|A=0,x)$. The ECT design that we evaluated in this manuscript builds on established methods from causal inference, which estimate marginal treatment effects. These methods correct, similar to Hobbs et al. [33], Murray et al. [34], for differences in the patient populations across studies, but estimate marginal effects $TE = E_X[Pr(Y|A=1,X) - Pr(Y|A=0,X)]$. Operating characteristics of both types of treatment effect estimates relative to RCTs and single arm studies can be evaluated using the proposed model-free validation algorithm.

S4 Supplementary Material Figures and Tables

	PubMed ID	NCT ID	Treatment		Enrollment Period	Primary Endpoint	Sample Size	OS Events
			Control	Experimental				
Historical Control	15758009	NCT00006353		TMZ+RT	8/2000-3/2002	OS	305	254
Phase II RCT	25910950	NCT00441142	TMZ+RT		2/2009-6/2011	OS	29	-
	26481741	-	TMZ+RT		6/2008-6/2012	OS	52	-
	26843484	NCT00589875	TMZ+RT		1/2006-1/2010	OS	134	133
	29126203	NCT01062399	TMZ+RT		12/2012-9/2013	PFS	83	44
	28142059	NCT00190424	TMZ+RT		10/2005-10/2008	OS-24	42	42
	21135282	NCT01013285	TMZ+RT		8/2006-11/2008	OS	110	48
	22120301	-	TMZ+RT		8/2005-2/2011	OS	16	15
Phase IIII RCT	24552318	NCT00943826	TMZ+RT		6/2009-3/2011	OS & PFS	463	458
	24101040	NCT00304031	TMZ+RT		1/2006-6/2008	OS	411	320
		NCT00884741	TMZ+RT		4/2009-5/2011	OS & PFS	309	198
Phase II SAT	20564147	NCT00544817	NCT00006353: TMZ+RT		4/2007-7/2008	PFS	54	-
	20615924	NCT00262730	NCT00006353: TMZ+RT		1/2006-1/2007	OS	97	75
	21531816	NCT00597402	NCT00006353: TMZ+RT		4/2007-9/2008	OS-16	75	-
	22706484	NCT00805961	NCT00006353: TMZ+RT		2/2009-10/2009	PFS	68	-
	25586468	NCT00458601	NCT00006353: TMZ+RT		8/2007-11/2009	PFS-5.5	65	56

Table S1: Single Arm trials (SATs) and randomized controlled trials (RCTs) in newly diagnosed Glioblastoma during the period if 2000-2016. For SATs the "control treatment" column indicates the historical trial (and treatment arm is the trial) that was used in the SAT to define a benchmark value for the SATs' experimental arm.

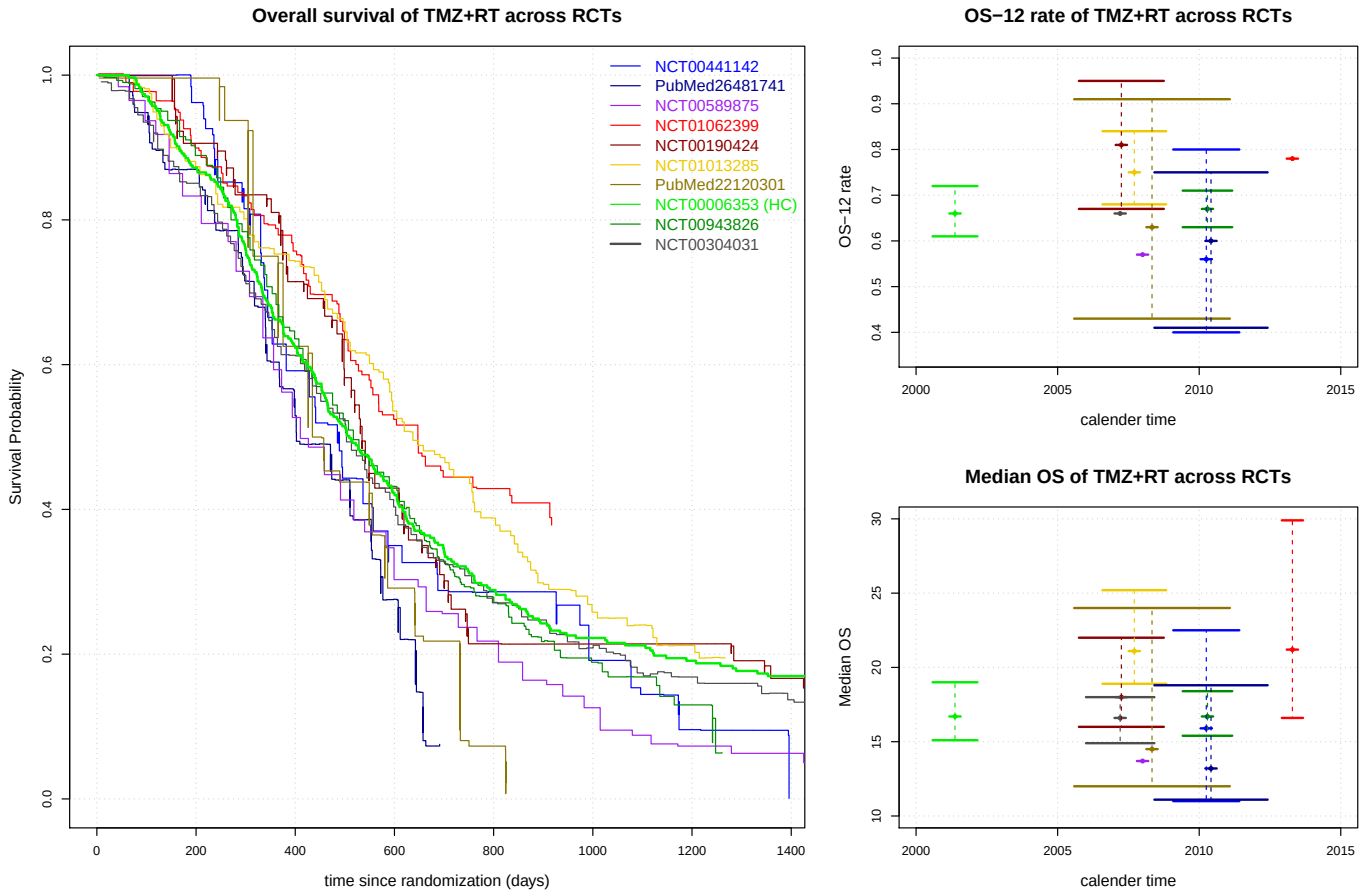

Figure S1: Reported estimates (point estimates and 95% confidence intervals (error bars)) of the overall survival (OS) functions, OS proportion at 12 months (OS-12) from randomization and median OS for the TMZ+RT arm in 10 RCTs that enrolled patients during the years 2000 – 2014. For median OS and OS-12 the error bars cover to the RCTs enrollment period.

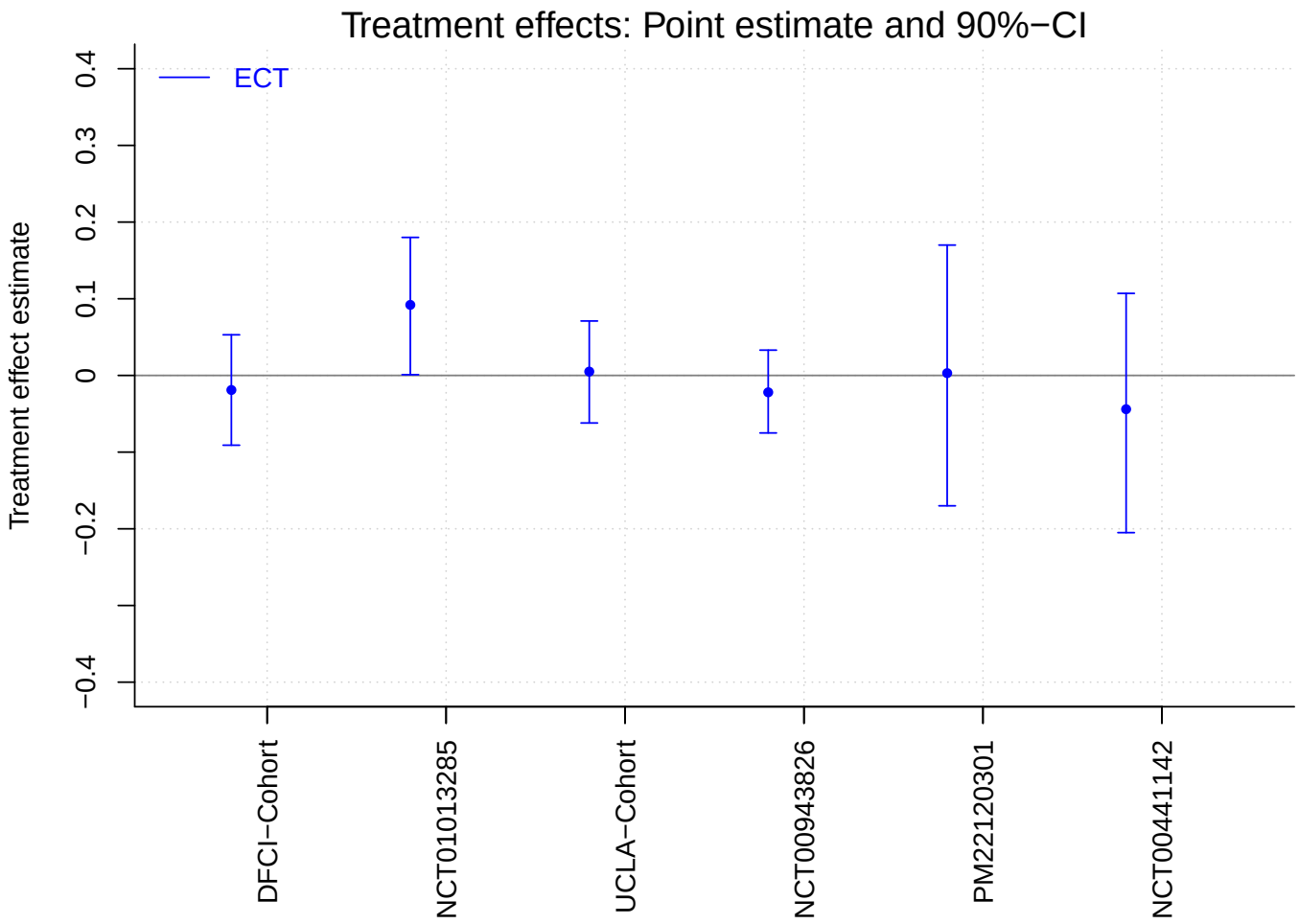

Figure S2: Treatment effect estimates of the ECT design. For each study the RT+TZM arm was used as ECT's experimental arm and (after adjustments for patent characteristics) compablack to the RT+TZM arms of the remaining five studies. The figures shows covariate adjusted treatment effects estimates $\widehat{TA}_{Ave}$ (point estimates and 90% confidence interval) for each of the 6 studies.

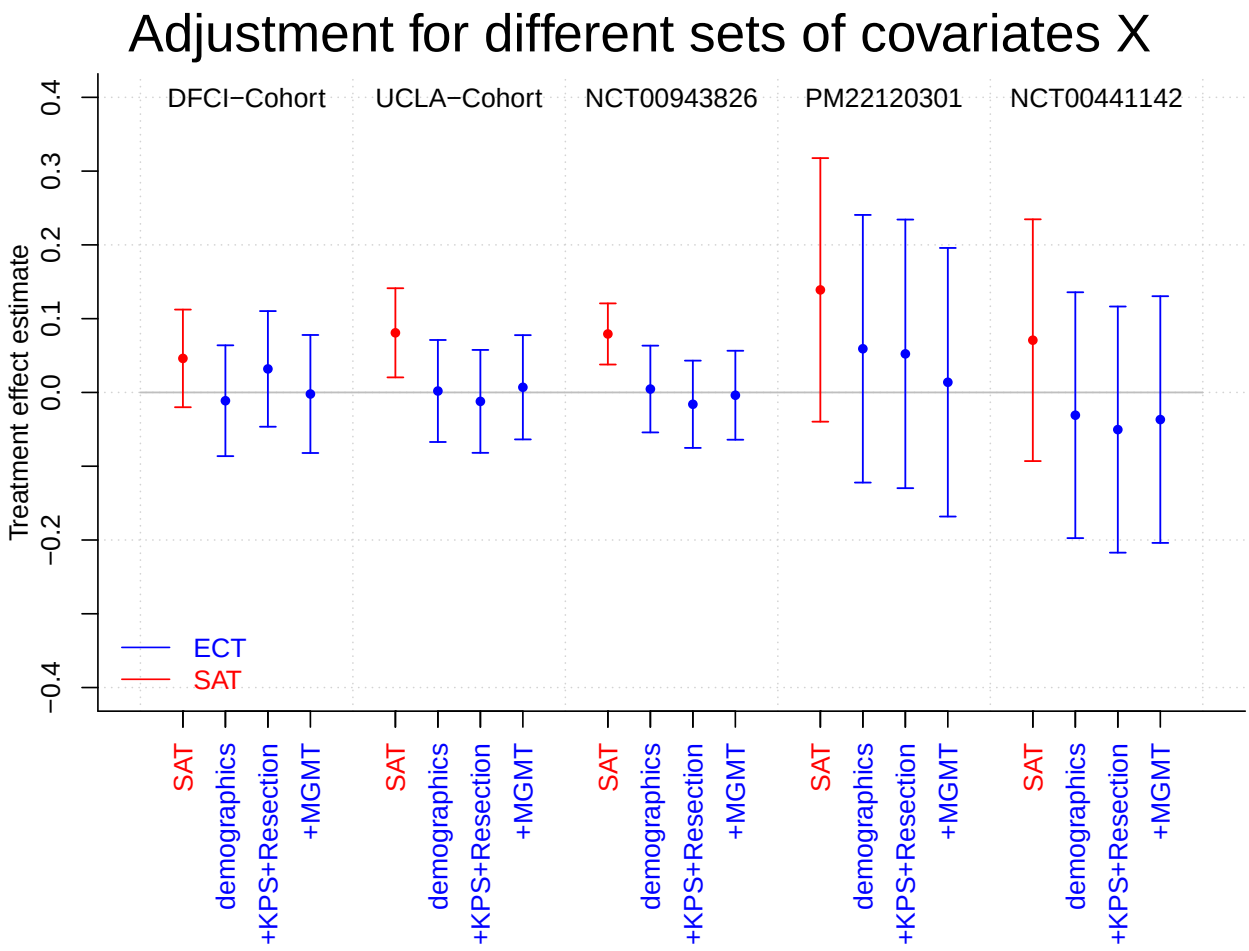

Figure S3: Adjusting ECT's treatment effect estimates for different sets of patients' covariates. The figure shows covariate-adjusted treatment effect estimates (point estimates and 90% confidence interval) for each of the five studies (DFCI and UCLA cohorts, NCT00943826, PM22120301 and NCT00441142) RT+TZM arm, when we consider a SAT using the RT+TZM arm of EORTC-NCIC as historical control arm, and a ECT with adjustments only for demographic pre-treatment variables, demographics in combination with Karnofsky Performance Status (KPS) and resection, and all five pre-treatment variables.

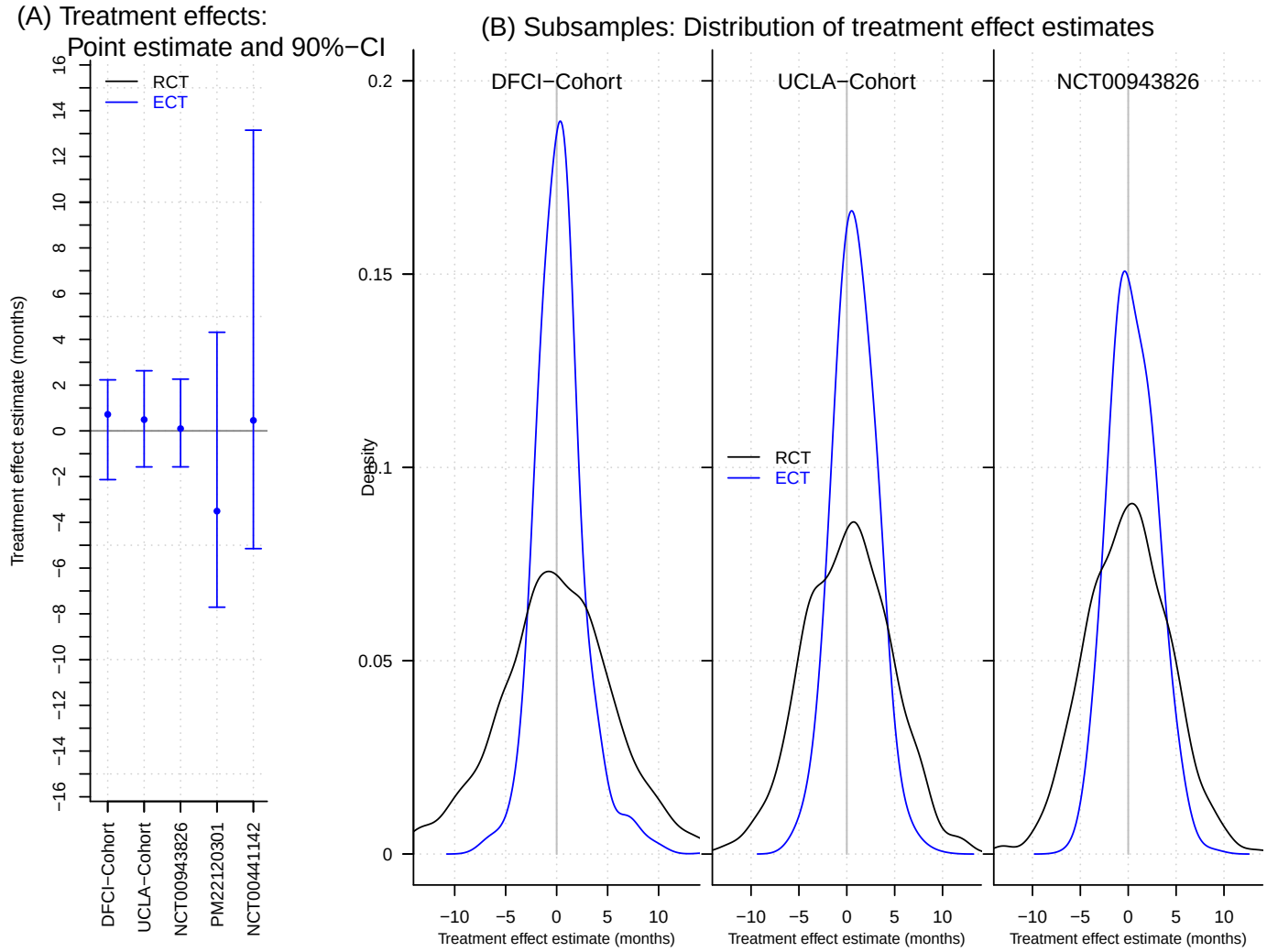

Figure S4: Treatment effect estimates of the ECT design (difference in median OS between the ECTs experimental arm and the external control arm) . For each study (DFCI and UCLA cohorts, NCT00943826, PM22120301 and NCT00441142) the RT+TZM arm was used as ECT's experimental arm and (after adjustment for patents characteristics) compablack to the RT+TZM arms of the remaining four studies. Panel (A) shows covariate adjusted treatment effect estimates $\widehat{TA}_{Ave}$ - difference on median OS - (point estimates and 90% confidence interval) for each of the five studies. Panel (B) shows the distribution of treatment effect estimates of the ECT (blue line) and RCT (black line) across subsamples of $n = 46$ patients.

References

[1] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Mathematical modelling*, 7 (9-12):1393–1512, 1986.

[2] Jonathan M Snowden, Sherri Rose, and Kathleen M Mortimer. Implementation of g-computation on a simulated data set: demonstration of a causal inference technique. *American Journal of Epidemiology*, 173(7):731–738, 2011.

[3] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[4] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[5] Paul R Rosenbaum. Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394, 1987.

[6] Liang Li and Tom Greene. A weighting analogue to pair matching in propensity score analysis. *The international journal of biostatistics*, 9(2):215–234, 2013.

[7] Fan Li, Kari Lock Morgan, and Alan M Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

[8] James M Robins. Correction for non-compliance in equivalence trials. *Statistics in medicine*, 17 (3):269–302, 1998.

[9] James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

[10] Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

[11] Peter C Austin and Dylan S Small. The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Statistics in medicine*, 33(24):4306–4319, 2014.

[12] Peter C Austin. Absolute risk reductions and numbers needed to treat can be obtained from adjusted survival models for time-to-event outcomes. *Journal of clinical epidemiology*, 63(1):46–55, 2010.

[13] Per K Andersen, Elisavet Syriopoulou, and Erik T Parner. Causal inference in survival analysis using pseudo-observations. *Statistics in medicine*, 36(17):2669–2681, 2017.

[14] Peter C Austin. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Statistics in medicine*, 33(7):1242–1258, 2014.

[15] Jun Xie and Chaofeng Liu. Adjusted kaplan–meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Statistics in medicine*, 24(20):3089–3110, 2005.

[16] Stephen R Cole and Miguel A Hernán. Adjusted survival curves with inverse probability weights. *Computer methods and programs in biomedicine*, 75(1):45–49, 2004.

[17] Stuart J Pocock. The combination of randomized and historical controls in clinical trials. *Journal of chronic diseases*, 29(3):175–188, 1976.

[18] Peter F Thall and Richard Simon. Incorporating historical control data in planning phase ii clinical trials. *Statistics in medicine*, 9(3):215–228, 1990.

[19] James MS Wason and Lorenzo Trippa. A comparison of bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials. *Statistics in medicine*, 33(13):2206–2221, 2014.

[20] Brian M Alexander, Patrick Y Wen, Lorenzo Trippa, David A Reardon, Wai-Kwan Alfred Yung, Giovanni Parmigiani, and Donald A Berry. Biomarker-based adaptive trials for patients with glioblastoma?lessons from i-spy 2. *Neuro-oncology*, 15(8):972–978, 2013.

[21] Steffen Ventz, William T Barry, Giovanni Parmigiani, and Lorenzo Trippa. Bayesian response-adaptive designs for basket trials. *Biometrics*, 73(3):905–915, 2017.

[22] Brian M Alexander, Lorenzo Trippa, Sarah Gaffey, Isabel C Arrillaga-Romany, Eudocia Q Lee, Mikael L Rinne, Manmeet S Ahluwalia, Howard Colman, Geoffrey Fell, Evanthia Galanis, et al. Individualized screening trial of innovative glioblastoma therapy (insight): A bayesian adaptive

platform trial to develop precision medicines for patients with glioblastoma. *JCO Precision Oncology*, 3:1–13, 2019.

[23] Steffen Ventz, Matteo Cellamare, Giovanni Parmigiani, and Lorenzo Trippa. Adding experimental arms to platform clinical trials: randomization procedures and interim analyses. *Biostatistics*, 19 (2):199–215, 2017.

[24] Steffen Ventz, Brian M Alexander, Giovanni Parmigiani, Richard D Gelber, and Lorenzo Trippa. Designing clinical trials that accept new arms: an example in metastatic breast cancer. *Journal of Clinical Oncology*, 35(27):3160–3168, 2017.

[25] Steffen Ventz and Lorenzo Trippa. Bayesian designs and the control of frequentist characteristics: A practical solution. *Biometrics*, 71(1):218–226, 2015.

[26] Steffen Ventz, Giovanni Parmigiani, and Lorenzo Trippa. Combining bayesian experimental designs and frequentist data analyses: motivations and examples. *Applied Stochastic Models in Business and Industry*, 33(3):302–313, 2017.

[27] Steffen Ventz, Matteo Cellamare, Sergio Bacallado, and Lorenzo Trippa. Bayesian uncertainty directed trial designs. *Journal of the American Statistical Association*, pages 1–13, 2018.

[28] Beat Neuenschwander, Gorana Capkun-Niggli, Michael Branson, and David J Spiegelhalter. Summarizing historical information on controls in clinical trials. *Clinical Trials*, 7(1):5–18, 2010.

[29] Heinz Schmidli, Sandro Gsteiger, Satrajit Roychoudhury, Anthony O'Hagan, David Spiegelhalter, and Beat Neuenschwander. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4):1023–1032, 2014.

[30] Kert Viele, Scott Berry, Beat Neuenschwander, Billy Amzal, Fang Chen, Nathan Enas, Brian Hobbs, Joseph G Ibrahim, Nelson Kinnersley, Stacy Lindborg, et al. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical statistics*, 13(1):41–54, 2014.

[31] Joost van Rosmalen, David Dejardin, Yvette van Norden, Bob Löwenberg, and Emmanuel Lesaffre. Including historical data in the analysis of clinical trials: Is it worth the effort? *Statistical methods in medical research*, 27(10):3167–3182, 2018.

[32] Alexander M Kaizer, Joseph S Koopmeiners, and Brian P Hobbs. Bayesian hierarchical modeling based on multisource exchangeability. *Biostatistics*, 19(2):169–184, 2018.

[33] Brian P. Hobbs, Bradley P. Carlin, Sumithra J. Mandrekar, and Daniel J. Sargent. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, 67(3):1047–1056, 2011.

[34] Thomas A. Murray, Brian P. Hobbs, Theodore C. Lystig, and Bradley P. Carlin. Semiparametric bayesian commensurate survival model for post-market medical device surveillance with non-exchangeable historical data. *Biometrics*, 70(1):185–191. doi: 10.1111/biom.12115. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12115.