Supplementary Material for "Adding Experimental Arms to Platform Clinical Trials: Randomization Procedures and Interim Analyses"

Steffen Ventz[1,2], Matteo Cellamare[2,3,4], Giovanni Parmigiani[2,4] and Lorenzo Trippa[2,4]

April 20, 2017

Contents

S1 Examples and Algorithms for Section 3

Example S1.1. We continue example 2.1 and apply the bootstrap algorithm to the BR design. The application to BAR and DBCD is similar. As before, the study starts as a multi-arm trial with two experimental arm and a planned sample size of $n_1 = 159$ patients. Experimental arms 3 and 4 are added after 12 months and 24 months, $(M_2, M_3) = (72, 144), n_2 = n_3 = 53$. We used the same parameters $q_k = Q_k$ (defined in (2.3)) as in example 2.1. Interim efficacy analyses are performed after 100 and 200 observed outcomes, and a final analysis is conducted after all outcomes are observed, so $J = 3$. For the initial set of arms $a = 1, 2$ the type I error probabilities are $\left(\alpha_a^{(1)}, \alpha_a^{(2)}, \alpha_a^{(3)}\right) = (0.025, 0.025, 0.05)$, whereas for arms $a = 3, 4$ $\left(\alpha_a^{(1)}, \alpha_a^{(2)}, \alpha_a^{(3)}\right) = (0, 0.05, 0.05)$. We consider 4 scenarios. In scenario 1, all arms have identical response rates of 0.3. The remaining Scenarios 2 to 4 are identical as in example 2.1.

For scenario 1, the type I error rate across 5000 simulated trials was $0.10, 0.09, 0.11$ and 0.09, for arms 1 to 4. Arms 1 and 2 were stopped early for futility in 50% of all simulations, whereas ineffective arms 3 and 4 were stopped early for futility in 43% of all simulations. In scenario 2, the initial arm $a = 1$ has 79% power. The probability of rejecting the null hypothesis H_1 in stages 1 to 3 are $0.31, 0.25$ and 0.23, respectively, and the empirical type I error rates for arms $a = 2, 3, 4$ are

[0]Corresponding authors: SV steffen@jimmy.dfci.harvard.edu, LT ltrippa@jimmy.harvard.edu
[1]University of Rhode Island, Kingston, US
[2]Dana-Farber Cancer Institute, Boston, US
[3]Sapienza University of Rome, Rome, Italy
[4]Harvard School of Public Health, Boston, US

0.11, 0.10 and 0.10. Similarly, in scenarios 3 and 4, the 1st added arm $a = 3$ and the 2nd added arm $a = 4$ have 79% and 78% power, respectively, with type I error rates of (0.11, 0.10, 0.10) in scenario 3, and (0.10, 0.11, 0.11) in scenario 4 for the remaining 3 ineffective arms. Effective arms in scenarios 2 to 4 were stopped early for futility in less than 2% of all simulations.

Algorithm 1 bootstrap hypothesis testing (without early stopping for efficacy).

Input 1: Trial design d (BR, BAR or DBCD)
Input 2: M_j, for all arms' sets $\mathcal{A}_j$ added before τ_a
Input 3: Sufficient statistics for all arms added before τ_a
Input 4: The test statistics T_a for arm $a \in \mathcal{A}_k$
Set $\widehat{\theta}_{a'}$ equal to the MLE of $\theta_{a'}$ for all arms $a' \neq a$ added before τ_a
Set $\widehat{\theta}_0$ and $\widehat{\theta}_a$ equal to the MLE under the assumption $\theta_0 = \theta_a$
for $c = 1, \cdots, C$ **do**
    Simulate the study forward from the M_k-th enrolled patient until $\tau_{a,c}$ under d:
        – A group of A_j arms is added at the enrollment of the M_j-th patient.
        – Patients respond to therapy a' with probability $\widehat{\theta}_{a'}$.
        – The response probability for the control and treatment a is equal to $\widehat{\theta}_a = \widehat{\theta}_0$.
    Set $S_{a,c} = 0$ if arm a is dropped for futility and 1 otherwise.
    Compute the statistics $T_{a,c}$ at $\tau_{a,c}$.
end for
Output, estimated p-value: $\widehat{p(T_a)} = \sum_{c=1}^{C} I\{T_{a,c} \geq T_a, S_{a,c} = 1\}/C$

S2 Trends in the patient population during the platform trial

In studies with long accrual periods variations in the composition of the patient population may occur during the study. With population trends the comparison of an added arm to outcomes of patients that have been randomized to the control before the experimental arm was activated should be avoided. This has the goal to prevent type I error rates above the targeted α level and to reduce bias of treatment effect estimates. The primary aim of this subsection is to emphasize possible bias issue and inflated type I error rates that can arise with population trends.

Using the setting of the EndTB trial, first quantify the effect of trends via simulations for the null scenario 1 in Section 5 with two different types of trends. In the first case, the population changes towards patients that are more likely to respond, and the probabilities of response to treatments increase from 0.55 to 0.75 between the first and last enrollment. In the second case, these probabilities decrease from 0.55 to 0.35 over the same period. When all control outcomes are used for inference and randomization, without accounting for the trend, then, with positive trends, the type I error rates increase to 19% for BR, and 14% for both BAR and DBCD (Table S1). With negative trends, the type I error rates decrease to 1% for all designs.

We outline a straightforward strategy to reduce these effects of the population trend. For all designs we restrict the comparison of an experimental arm k to outcome data from patients that have been enrolled during the accrual period of arm k. This restriction is used in the computation of randomization probabilities, early futility stopping decisions, and in the bootstrap for hypothesis

testing. Recall that the bootstrap simulations generate outcome data between the activation of arm k and the time of hypothesis testing as described in Section 3. Additionally one could model the population trend during the platform trial. This additional modeling component could be combined with the computations of randomization probabilities, futility early stopping decisions, as described in the previous sections, and the proposed bootstrap procedure.

With a positive trend, using only concurrent data, the type I error decreases to 5-6% for BR, and 3-5% for both BAR and DBCD. Under negative trends, BR's type I error rates remain close to the nominal value of 5%. For BAR and DBCD, we observe type I error rates up to 6.4% for BAR and 9% for the DBCD.

When trends of the patient population are likely to occur during the study BR allows to use relatively simple approaches (Altman and Royston, 1988; Cohen and Sackrowitz, 1989; Kimani *and others*, 2013; Bowden and Trippa, 2015) and arguments to reduce bias of the treatment effects' estimates and prevent Type I error rates different from the target α. On the other hand a precise and rigorous account of possible population trends with BAR and DBCD appears more challenging. Indeed correction techniques to reduce bias of the treatment effect estimates and to obtain type I error rates close to the target α, are simpler for the BR design (Cohen and Sackrowitz, 1989; Bowden and Glimm, 2008; Kimani *and others*, 2013) than for response-adaptive designs (Bowden and Trippa, 2015). In the endTB trial, based on previous experiences of investigators trends are not expected during the accrual period of 2.5 years and we adopted a Bayesian randomized design (Cellamare *and others*, 2017).

	Positive Trend						Negative Trend					
	use all outcomes			use concurrent outcomes			use all outcomes			use concurrent outcomes		
	BR	BAR	DBCD	BR	BAR	DBCD	BR	BAR	DBCD	BR	BAR	DBCD
Type I error rates												
$\mathcal{A}_1$	0.053	0.043	0.052	0.048	0.054	0.044	0.052	0.040	0.040	0.053	0.049	0.044
$\mathcal{A}_2$	0.134	0.102	0.910	0.045	0.032	0.029	0.021	0.018	0.024	0.045	0.063	0.090
$\mathcal{A}_3$	0.189	0.141	0.136	0.059	0.032	0.028	0.009	0.014	0.015	0.031	0.064	0.074
Bias of the treatment effect												
$\mathcal{A}_1$	-0.033	-0.047	-0.002	-0.000	-0.006	-0.007	-0.029	-0.046	-0.004	0.000	0.002	-0.005
$\mathcal{A}_2$	0.010	-0.006	0.019	-0.038	-0.031	-0.027	-0.025	-0.046	-0.027	-0.008	0.009	0.017
$\mathcal{A}_3$	0.032	0.013	0.031	-0.055	-0.041	-0.038	-0.018	-0.041	-0.025	-0.014	0.012	0.015

Table S1: Type I error rates and bias of the treatment effect estimates $\gamma_a = \widehat{p}_a - \widehat{p}_0$ for each arm $a \in \mathcal{A}_k$ in group $k = 1, 2, 3$ across simulations under trends in the patient population during the trial. We consider two types of trends: (1) The true outcome probabilities increase for all treatment arms from 0.55 top 0.75 between the 1st enrollment and the end of the trial. (2) The true outcome probabilities decrease from from 0.55 top 0.35 over the same period. The trial starts with four initial experimental treatments $\mathcal{A}_1 = \{1, 2, 3, 4\}$ and two experimental arms are added after $M_2 = 200$ enrollments, $\mathcal{A}_2 = \{5, 6\}$, and after $M_3 = 300$ enrollments, $\mathcal{A}_3 = \{7, 8\}$. Results are based on 5000 simulations under balanced randomization (BR), Bayesian adaptive randomization (BAR) or the doubly adaptive biased coin design (DBCD), with an initial planned sample size of $n_0 = 500$ patients and an extension of the overall sample size by 200 patients at time M_2 and M_3. Without patient trends the mean treatment effects across simulations equal $(-0.016, -0.016, -0.016)$ for BR, $(-0.023, -0.021, -0.020)$ and $(-0.004, -0.005, -0.004)$ for BAR and DBCD.

S3 Supplementary Figures for Section 4

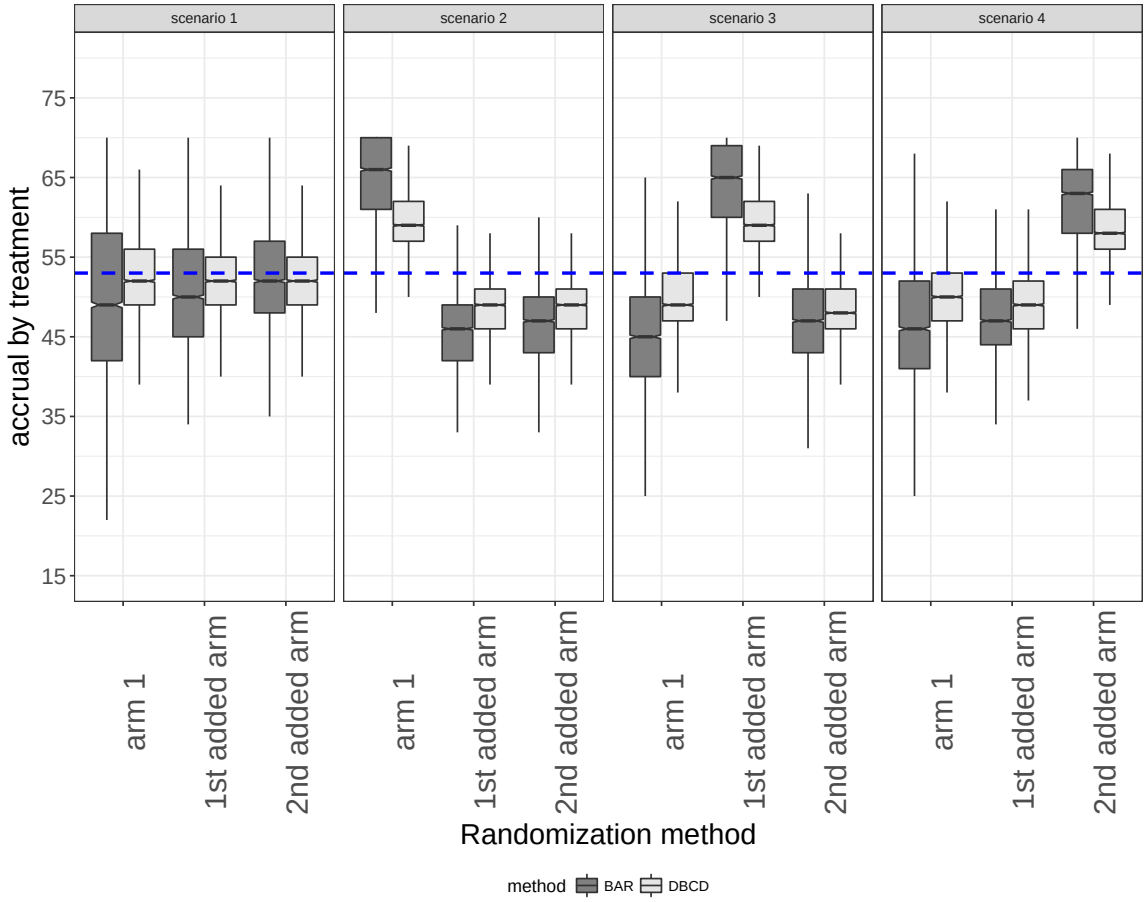

Figure S1: Variability of treatment assignments at the end of the trial for the simulation study of Section 4. Boxplots of the number of patients randomized to each treatment arm under BR, BAR and DBCD across 5000 simulations for a trial with 2 initial arms and two arms that are added after the enrollment of $M_2 = 72$ and $M_3 = 144$ patients. The dashed line shows the number of patients randomized to each arm under BR.

References

Altman, Douglas G and Royston, J Patrick. (1988). The hidden effect of time. *Statistics in medicine* **7**(6), 629–637.

Bowden, Jack and Glimm, Ekkehard. (2008). Unbiased estimation of selected treatment means in two-stage trials. *Biometrical Journal* **50**(4), 515–527.

Bowden, Jack and Trippa, Lorenzo. (2015). Unbiased estimation for response adaptive clinical trials. *Statistical methods in medical research*, 0962280215597716.

Cellamare, Matteo, Ventz, Steffen, Baudin, Elisabeth, Mitnick, Carole D and

Trippa, Lorenzo. (2017). A bayesian response-adaptive trial in tuberculosis: The endtb trial. *Clinical Trials* **14**(1), 17–28.

Cohen, Arthur and Sackrowitz, Harold B. (1989). Two stage conditionally unbiased estimators of the selected mean. *Statistics & Probability Letters* **8**(3), 273–278.

Kimani, Peter K, Todd, Susan and Stallard, Nigel. (2013). Conditionally unbiased estimation in phase ii/iii clinical trials with early stopping for futility. *Statistics in Medicine* **32**(17), 2893–2910.