Combining Bayesian Experimental Designs and Frequentist Data Analyses: Motivations and Examples

Lorenzo Trippa, Giovanni Parmigiani, Steffen Ventz

September 8, 2016

Abstract

Recent developments in experimental designs for clinical trials are stimulated by advances in personalized medicine. Clinical trials today are very different from traditional studies, and typically seek to answer several research questions for multiple patient subgroups. Bayesian designs, which enable the use of sound utilities and prior information, can be tailored to these settings. On the other hand, frequentist concepts of data analysis remain pivotal; for example type I/II error rates are the accepted standards for reporting trial results and are required by regulatory agencies. Bayesian designs are often perceived as incompatible with these established concepts, which hinders widespread clinical applications. We discuss a pragmatic framework for combining Bayesian experimental designs with frequentists analyses. The approach wishes to facilitate a more widespread application of Bayesian experimental designs, and ultimately analysis, in clinical trials. We discuss several applications of this framework in different clinical settings, including bridging trials and multi-arm trials in infectious diseases and Glioblastoma. We also outline computational algorithms for implementing the

proposed approach.

1 Introduction

What is a Bayesian Clinical Trial Design? The Bayesian design of a clinical trial is characterized by the collection and subsequent formalization of available information through a prior distribution. Previous clinical trials, data from epidemiological studies or disease models are standard examples of relevant information used to specify the prior. In summary, the design of the trial starts from a prior distribution π over a set of unknown parameters

$$\theta \sim \pi.$$

Throughout our discussion π will be a genuine representation of the investigators beliefs and uncertainties on key parameters θ. Typically in medicine θ includes response probabilities, survival curves or toxicity rates of different treatments. These parameters will be estimated and compared using the data generated by the clinical trial.

The Bayesian Designer and the use of the prior π. The information embedded in the prior π can be used in several contexts and for different purposes. Examples are (i) the choice of the sample size for a two-arm or a multi-arm study [41], (ii) the definition of a two stage design, with stage-specific samples sizes selected using the prior π [38], and (iii) Bayesian adaptive randomization, with reinforcement of the randomization probabilities during the trial toward the most promising arms [35, 36, 26].

Decision Theory. Some of these designs, for example two arm studies, can be optimized by a direct application of the decision theoretic paradigm. The design is selected by the prior π and the utility function u, which is representative of the investigators preferences. In general, the solution of the decision problem coincides with the design d that maximizes the expected value

$$\mathbb{E}_{\pi,d}(u)$$

of the utility generated by the experiment [13]. Here the utility $u = u(X, d)$ is a random quantity, and it is a function of the data X collected during the study with design d. Additionally, in some cases, it is convenient to include the unknown parameter $\theta \sim \pi$ in the definition of $u = u(X, \theta, d)$ to simplify its interpretation. For instance, the sample size for a two-arm study d can be selected by specifying a utility function that captures the trade off between marginal costs associated with the enrollment of each patient and the likelihood to correctly identify and recommend the best available treatment. In this example

$$u(X,\theta,d) = \texttt{data support recommendation of the experimental treatment} \\ - \texttt{costant} \times \texttt{sample size} \tag{1}$$

In other contexts the selected design d is not the solution of a maximization problem. The use of a prior distribution π is combined with less explicit utility criteria. Examples include the use of adaptive randomization probabilities in multi-arm trials, with randomization probabilities proportional to the posterior probabilities of positive treatment effects [6, 35]. In this case the utility criteria is not explicitly stated, but the intention is explicitly to increase the accrual toward the most promising arms. This type of studies use the prior π and the data generated during the trial for interpretable decisions, such as variations of the randomization probabilities, or to drop arms during the trial. We refer to Berry and Fristedt [7] for discussions of the decision theoretic framework to define adaptive randomization probabilities which illustrate computational complexities and justify the use of alternative heuristic algorithms.

Prior information and utility criteria in non Bayesian designs. For trials designed without the explicit used of a utility functions u and a prior distributions π, sample sizes and interim decision rules are often selected using substitutes of (π, u), such as tables which report operating characteristics under a list of simulation scenarios. These evaluations typically involve several candidate designs. The list of scenarios, similarly to π, is representative of prior beliefs and predictions of the investigator. Symmetrically, the choice of the operating characteristics to be compared across potential designs mirrors the investigators preferences. We often have a one-to-one correspondence between the key components of the decision theoretic framework (π, u) and those of a simulation study, scenarios and operating characteristics [21].

What is the motivation for the study of Bayesian designs? We list a few closely related advantages for using prior distributions and utility functions. First a pragmatic aspect. The selection of a design based on examining tables and summaries across simulation scenarios, candidate designs and competing operating characteristics can be quite challenging and time consuming. Second, the use of the decision theoretic approach forces investigators to think through and explicitly state goals and assumptions via a prior π and and a utility function. In routine tasks, for example selection of futility stopping boundaries, it is easier to interpret and subsequently agree or disagree on the choice of π and u, than having a debate over large tables of operating characteristics. Additionally, a clinical trial design selected based on decision theoretic arguments can always – and in most cases should – be scrutinized through interpretable summaries of the resulting operating characteristics. Still, skepticism can be an appropriate reaction towards attempts to declare exhaustive the evaluation of a design through simulations and tables of operating characteristics. These tables can be necessary but not sufficient for a solid evaluation the trial design. Third, in complex trials it is difficult to replace prediction and posterior probabilities with alternative data summaries with comparable level of interpretability. In particular, prediction and posterior probabilities are useful and interpretable to specify interim analyses during the the study. For instance, in studies with biomarker-treatment interactions, posterior probabilities can be used to modify arm-specific eligibility criteria base on accumulating data [23, 3].

Beyond Pros and Cons of the Bayesian approach. Limitations of the Bayesian framework that prevent a more widespread use in clinical trials, including computational demand, prior elicitation and the acceptance of a single utility function from several stakeholders, have been discussed in the literature [22, 17]. The goal of the sections that follow is complementary to these discussions of the pros and cons of the Bayesian framework. We discuss a possible strategy to facilitate the use of Bayesian foundations in clinical trials. Most clinical investigators and scientific review panels are not against the use of Bayesian designs. But there are barriers to a rapid diffusion of Bayesian methods in clinical trial designs. Here we only focus on one of them, perhaps an important one, by illustrating that the results reported at completion of a Bayesian trial do not necessary need to be linked and influenced by the choice of the prior π.

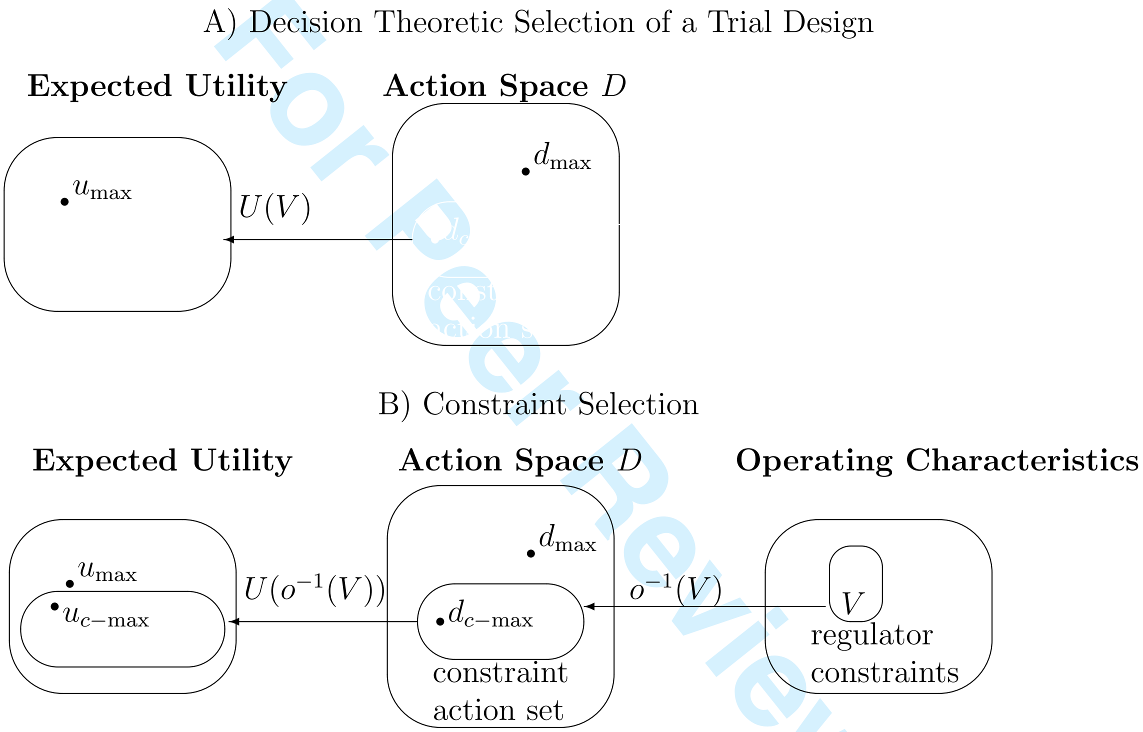
Reporting results from a Bayesian trial. Clinicians, scientific review panels, and other stakeholders in the clinical trials arena, in most cases, are familiar with key statistical concepts from the frequentist literature; type I error rates, hypothesis testing and confidence intervals to name a few. These concepts are accepted standards for reporting results in clinical trials and to communicate evidence of positive effects or futility of novel treatments. Bayesian designs are often perceived as incompatible with these established metrics for reporting results, in particular p-values and hypotheses testing. This is the perceived barrier that we will discuss. We illustrate the use of methods to combine the use of Bayesian models π, utility functions u and frequentist analyses, including the control of type I errors rates and confidence intervals.

How do we formalize the goal of combining π, u, and frequentist analyeses? We include frequentists constraints into a Bayesian decision theoretic framework [38]. These constraints reflect desiderata from collaborators and other stakeholders. Example include control of type I error below 0.05, or minimal bias in the effect estimates. The first panel of Figure 1 illustrates graphically the application of the decision theoretic framework. The action space D, i.e. the set of all trial designs, is shown on the right. A point d in D is a candidate trial design and, typically it includes sample size, stopping rules, and also a plan on how to analyze the data and communicate the final results of the trial. Estimators and procedures to report evidence of treatment effects or futility are components of the trial design d. Importantly the plan for final analyses can vary substantially across candidate designs in D. The space D is mapped to the range of expected utilities $U(D)$. The Bayesian statistician selects the trial design $d_{\max}$ that maximizes the expected utility. In the first panel of Figure 1 $u_{\max}$ is the maximum of the expected utility surface, which is achieved by the design $d_{\max}$.

We can now describe the strategy of our Bayesian biostatistician to select a trial design, by including its interactions with the scientific community, clinicians, editors of scientific journals and review committees. We model these interactions by adding constraints to the operating characteristics of the trial (see Figure 1, Panel B). Examples, as we mentioned, include the requirement to bound Type I/II error rates below explicit thresholds, or to limit the expected enrollment below a pre-specified threshold under the hypothesis of a detrimental or toxic treatment. These are well defined frequentist constraints, and a candidate design d can satisfy the requirements or not, irrespective of the prior distibution π. In Figure 1 we indicate these constraints through the set V. The subset of designs that satisfy them $o^{-1}(V)$ is identified by the map o which links designs d to their operating characteristics $o(d)$. The choice is now constrained to the selection of a possibly suboptimal design within $o^{-1}(V)$. Our

Bayesian decision maker selects the design $d_{c-\max}$ that optimizes the expected utility surface within the subset $o^{-1}(V)$.

Figure 1: Graphical representation of the optimal Bayesian design $d_{\max}$ and of the constrained optimal Bayesian design $d_{c-\max}$. In this diagram V, $o^{-1}(V)$ and $U(o^{-1}(V))$ denote the regulator constraints, the subset of designs with operating characteristics in V and the corresponding expected utilities. The expected utilities of $d_{\max}$ and $d_{c-\max}$ are $u_{\max}$ and $u_{c-\max}$ respectively.



Why should one follow this framework? We list a few properties of our constrained decision theoretic (CDT) framework:

- It includes an explicit unambiguous utility function u as the primary criterion to select candidate designs.

- It is straightforward to extend the approach to prediction-based adaptive strategies and algorithms that remain similar in spirit, and share a similar interpretation.

- It makes effective use of prior estimates and scientific knowledge.
- It allows dissemination of the major findings of trial s using an established scientific language, including frequentist concepts such as hypothesis testing and power.
- It facilitates communication of the design characteristics with multiple stakeholders and non-statisticians.

Roadmap. In the sections that follow we first provide an example of a direct application the CDT framework. We will then move to examples of more complex sequential designs where, similarly to the standard decision-theoretic framework, exact solutions become computationally unfeasible, and it is necessary to relax the optimization strategy with heuristic algorithms.

2 Constrained Optimal Designs

2.1 Constrained Optimal Bridging Trials

In Ventz and Trippa [38] we previously explored the use of the CDT framework for the design of bridging trials [12]. Here we provide a summary of the results obtained by applying the CDT framework. A bridging trial assesses whether a drug recently approved in a region A, say Europe, can be marketed in a different region B, for instance Japan. Clinical data from region A should guarantee that the drug is effective and safe, and the bridging trial is a supplementary study to test whether the drug has a similar treatment effect and safety profile in population B [12]. In this setting we have historical data from randomized trials and information, which the investigator can incorporate in the prior π. The use of the CDT framework requires the specification of three components π, u and V. One can argue that the available data allow straightforward specification of the first component π [12]. Additionally, the investigator can specify sound utility functions based on estimates of relevant utility parameters, such as the potential number of prescriptions per year in region B. Regulators and other stakeholders, such as patient representatives, can express the need for controlling type I error rates and/or other characteristics of the study from a frequentist viewpoint. We indicate the constraints by V as before.

Problem setting. For each patient i, the primary endpoint Y_i, say the reduction of blood pressure, conditional on the treatment $C_i = 1$ or the placebo $C_i = 0$, is assumed normal distibuted with mean θ_k, $k = 0, 1$. The company has to test $H_0 : \gamma_B = 0$ versus $H_1 : \gamma_B > 0$, where $\gamma_B = \theta_0 - \theta_1$. A group-sequential trial with a possible early termination at interim analyses $t = 1, \cdots, T-1$ in favor of H_1 is used, and N patients will be randomized to each arm between consecutive interim analyses. We use the summary $Z_t = (\hat{\theta}_{0,t} - \hat{\theta}_{1,t})/\sqrt{(2\sigma^2)/(Nt)}$ and, without loss of generality, assume a common variance σ^2 for the two arms. Here $\hat{\theta}$ denotes maximum likelihood estimates. The vector $Z_{1:T} = (Z_1, \dots, Z_T)$ is Gaussian with mean $\mu = (\gamma_B \sqrt{Nt/2\sigma^2})_{t \le T}$ and covariance matrix $W = (W_{t,t'})_{1 \le t,t' \le T}$, where $W_{t,t'} = \sqrt{t/t'}$ for $t \le t'$.

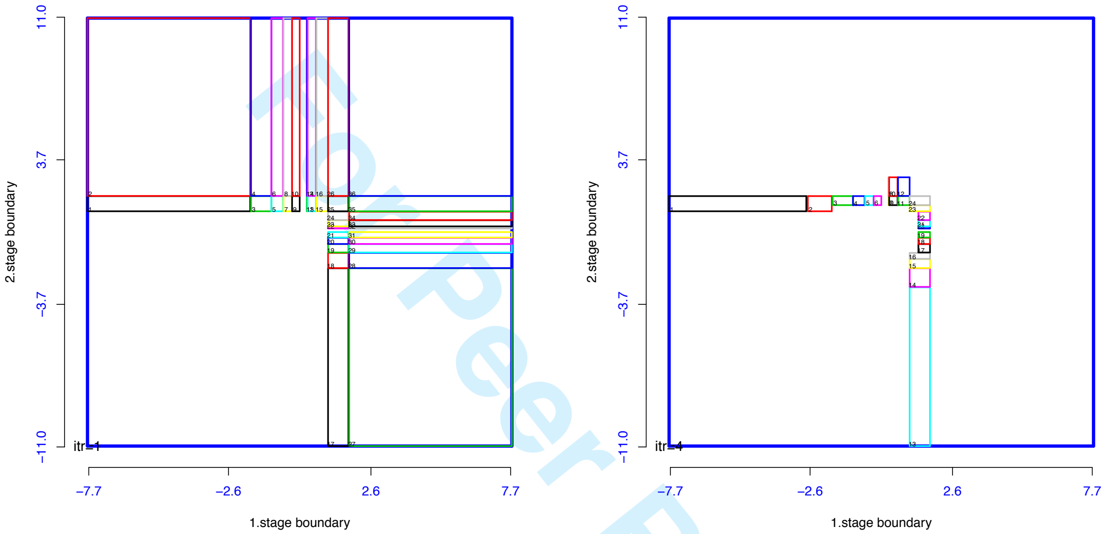
Specification of V, π and u. Low power could delay patients' access to an effective drug. We therefore assume that the regulator requires type I and II error rates, at H_0 and $\gamma_B = \gamma_B^* > 0$, to be controlled at suitably chosen α and β levels respectively. The information from region A can be summarized by a Gaussian prior for the parameter γ_B. Power priors [11], for example, are directly applicable to specify π based on information from region A. Finally we use an interpretable utility function similar to (1), with costs linear in the number of randomized patients and, in case of a true positive finding, a fixed payoff at termination of the trial [38].

Decision problem. A design is characterized by the sample size N and stopping boundaries $z_{1:T} = (z_1, \cdots, z_T)$ at interim and final analyses $t = 1, \cdots, T$.

Results: Z thresholds characteristics. We solved the constrained optimization problem and computed optimal thresholds $z_{1:T}$ for the summary statistics $Z_{1:T}$ to allow early termination of the study with several variations on the prior π and utility function u. The solution showed negligible departures from linear thresholds. As expected, by varying utility parameters and prior distribution, we obtained considerably different thresholds.

Algorithm 1: Cut and Zoom-in To compute the optimal stopping rules $z_{1:T}$ we leverage on monotonicity properties [38]. First, the expectation of the utility function (1) can be written as the sum of two monotone functions $U(d) = U_1(z_{1:T}) - U_2(z_{1:T})$. Here $U_1(z_{1:T})$ is equal to the probability of the intersection of two events (i) presence of a positive treatment effects, and (ii) reporting evidence of treatment effects at completion of the study, while $U_2(z_{1:T})$ is proportional to the expected sample size. Second, assuming that N is fixed, the operating characteristic $o(d(z_{1:T})) = \sup_{\{\gamma_k:\gamma_k \leq \Delta_0\}} \mathbb{P}_{\gamma_k}[Z_{1:T,k} \geq z_{1:T}]$ is monotone in the thresholds $z_{1:T}$. Similarly, also the indicator function $1(o(d) \in V)$, which indicates if the candidate design satisfies the operating characteristics requirements V or not, is monotone in $z_{1:T}$. The algorithm used to compute the optimal design partitions the space of designs and computed lower and upper expected utility bounds for each partition set. Figure 2 is a graphical representation of the optimization algorithm. In this case we have one interim analysis, it is therefore necessary to compute $T = 2$ thresholds. The two panels show the current status of the algorithm at different iterations, which is a collection of rectangles that could potentially harbor the constrained optimum. At each iteration a single rectangle is either (i) removed from the list because it does not contain $d_{c-\max}$ or (ii) divided in two sub-rectangles. By computations of the expected utilities and operating characteristics at the extremes of the rectangles, and exploiting monotonicity, the algorithm progressively and iteratively remove candidate designs d and zooms into regions of the action space with comparable operating characteristics that include the constrained optimum.

Figure 2: Cut-&-Zoom-in algorithm for computing the contained optimal design $d_{c-\max}$. The left and right panel show the second and fifth iteration of the algorithm. At each iteration the algorithm (i) either removes rectangles for which the utility can be bound by an other rectangle, (ii) or splits the rectangle into two disjoined rectangles.



2.2 A multi-arm response-adaptive design in Glioblastoma

Glioblastoma and motivations to look beyond standard designs. Glioblastoma is a brain cancer associated with a poor prognosis. Numerous treatments, in recent years, have shown promise in preclinical models, but translation into tangible treatment effects and survival improvement has been slow and nearly negligible [28]. Current trial designs and more generally pipelines for developing new treatments have been severely criticized for being very inefficent [5]. Most of the current early-phase trials for patients with glioblastoma are single-arm studies. In contrast, we proposed and evaluated potential benefits of using controlled, response-adaptive multi-arm trials in this context [36].

Response adaptive randomization Adaptive randomization schemes are designed to obtain a more desirable assignment of patients in the trial to competing treatments compared to balanced designs. Several contributions considered two-arm and multi-arm controlled trials and provide motivations for adaptively tuning the randomization probabilities during the study on the basis of the accumulation outcome data [35, 25, 3, 23]. Response adaptive randomization can be defined as the application of a map, used each time a patient is enrolled in the trial, which transforms the available data into suitable randomization probabilities. Frequentist approaches are direct, in that, intuitive and heuristic rules are used to map the available data into randomization probabilities [40, 14, 32, 31]. These maps have been assessed using asymptotic theoretical analysis in and simulation studies [32, 19, 20, 43]. In contrast Bayesian randomization methods are indirect, model based and exploit Bayesian predictions during the trial. The prior distribution π models jointly the primary outcome distributions $\theta_0, \theta_1, \ldots, \theta_K$ for control and experimental treatments. Most Bayesian adaptive strategies map posterior probabilities of treatment effects, say $(\theta_k - \theta_0)$, into randomization probabilities [35, 24, 42].

Missing utility. Most Bayesian adaptive randomization procedures do not maximize an explicit utility function. The computational burden to optimize a sequential multi-arm study motivates the use of heuristic procedures. In different words, we will discuss procedures that relax the decision theoretic paradigm. Zhang et al. [44] compare heuristics and decision theoretic optimal designs within the context of biomarker-subgroup trials. The development of nearly optimal assignment procedures tailored to explicit utility functions u remains an attractive area of research.

Randomization. In Trippa et al. [36] we consider ed a controlled four-arm trial. The response to treatments is evaluated using progression-free survive endpoints and $(S_0, S_1, \cdots, S_3)$ denote the unknown time to event distributions for the control arm $k = 0$ and experimental thearpies $k = 1, 2$ and 3. These are assume to follow a proportional hazards model, with unknown positive hazard ratios $\theta = (\theta_1, \theta_2, \theta_3)$, such that the equalities $S_k(t) = [S_0(t)]^{\theta_k}$ hold, for every $t \geq 0$ and $k = 1, 2, 3$. We use identical symmetric prior distributions with mean zero for the log-hazard ratios $\log(\theta_1), \log(\theta_2)$ and $\log(\theta_3)$. In different words, the prior π assigns symmetric probabilities to scenarios where treatment k has a positive or negative effect compared to the control. We consider time varying randomization probabilities

$$R_i^k = p(i\text{-th enrolled patient is randomized to treatment } k|\text{available DATA})$$

defined by the following expressions:

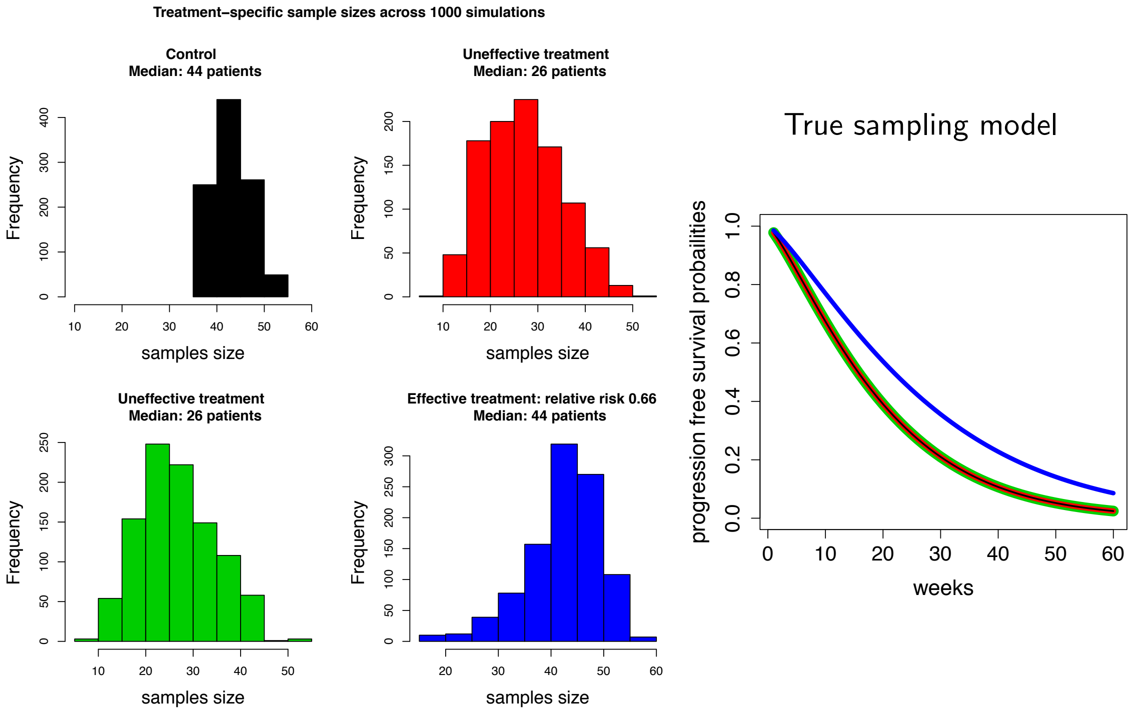
$$R_i^k \propto \frac{p\left(\theta_k < 1 \mid \text{ available DATA}\right)^{\gamma(i)}}{\sum\limits_{\ell=1,2,3} p\left(\theta_\ell < 1 \mid \text{ available DATA}\right)^{\gamma(i)}} \quad \text{if } k = 1, 2, 3, \text{ and}$$

$$R_i^0 \propto \exp\left\{\eta(i) \times \left(\max_{\ell=1,2,3} \#[\text{assignments to arm } \ell] - \#[\text{assignments to control}]\right\}/3.$$

The above two expressions have a clear interpretation. The first one shows that for any choice of the tuning function $\gamma(i) > 0$ the algorithm assigns patients with higher probabilities to experimental arms with evidence of a positive treatment effect $\theta_k < 1$. Natural candidates for the tuning parameters $\gamma(i)$ are non decreasing functions with values close to zero during the initial stage of the trial. The second expression aims at approximately match patient accrual to the control treatment with the number of patients on the experimental are with the highest patient accrual. In our experience values of η close to 0.25 during the final stage of the trial suffices to obtain the desired balance without making treatment assignment highly predictable.

Figure 3 shows the distribution of the arm specific sample sizes under a fixed scenario across 10,000 simulated trials. The above formulation of the adaptive randomization probabilities can be straightforwardly extended and used for binary or continuous outcomes as in the next example.

Figure 3: Distribution of patients accrual to the control and experimental arms across simulated trials using Bayesian adaptive randomization. The left panel shows the distribution of patients accrual for each therapy. The right panel shows the true progression-free survival for the control and experimental arms.



2.3 The endTB trial: An adaptive Study in Tuberculosis

In 2010 there were an estimated 650,000 prevalent cases of multi-drug resistant tuberculosis (MDR-TB), and nearly 500,000 new cases emerge annually through acquisition of resistance during treatment and through airborne transmission [27]. The need for new regimens is therefore indisputable. The recent conditional approval by regulatory authorities of two new anti-TB drugs, bedaquiline and delamanid, presents the first opportunity of a significant improvement in the treatment of MDR-TB since half a centory.

The endTB study is a Phase III trial that seeks to evaluate five novel treatments for

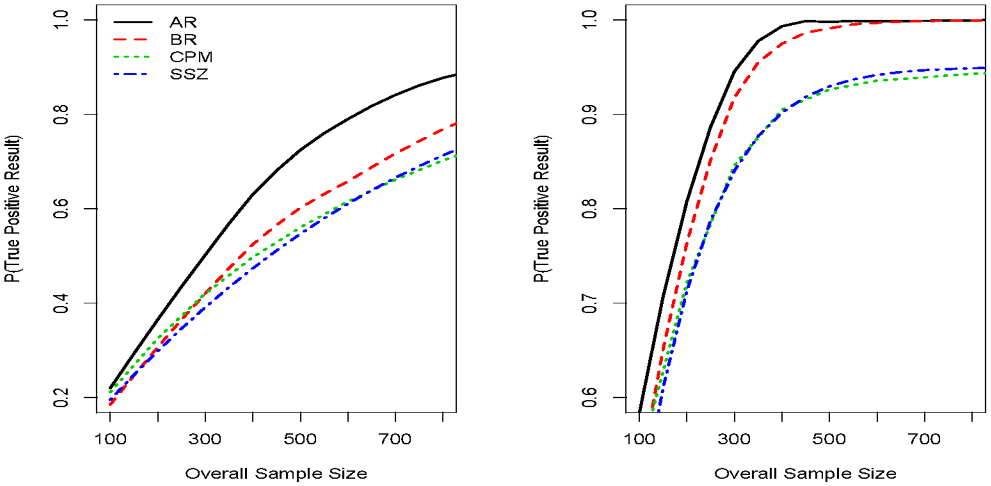
MDR-TB. The study is sponsored by Médecins sans Frontiéres, planned in conjunction with Partners In Health, Harvard Medical School, Epicentre, and the Institute for Tropical Medicine, and supported by UNITAID. It will generate evidence on efficacy and recommendations for those arms that will show treatment effects. The endTB is estimated to have a final enrollment of 750 patients. We designed the trial using Bayesian outcome-adaptive randomization [10], adapting on surrogate and primary endpoints based on joint modeling of the binary culture conversion outcome after 8 and 39 weeks of treatment

$$\theta_{39,k} = \theta_{39,PR,k} \times \theta_{8,k} + \theta_{39,NR,k} \times (1 - \theta_{8,k}).$$

Here $\theta_{8,k}$ denotes the probability of a positive early response to therapy k after 8 weeks of treatment, while $\theta_{39,PR,k}$ and $\theta_{39,NR,k}$ are the response rates after 39 weeks given a positive (PR) or negative (NR) 8-weeks response. We include interim analyses at regular intervals after a total of 100, 200, and so on, primary outcomes become available. Arms are dropped for futility if the available data and posterior probabilities suggest no treatment effect on the primary outcome. The endTB trial uses outcome adaptive randomization, followed, at the end of the trial by frequentists analysis using a strong control of pre-specified targeted type I error rates. In Section 3.2 and 3.3 below we discuss algorithms for the control of type I error rates of adaptive trials.

A detailed study of the design is provided in [10], here we provide a brief summary of the results. When we compare the statistical power under adaptive and non-adaptive designs, under several hypothetical scenarios, see Figure 4 for two examples, we observe that Bayesian outcome-adaptive randomization requires fewer patients than non-adaptive designs to achieve the targeted power. Moreover, adaptive randomization consistently allocates more participants to effective arms compared to alternative non-adaptive designs.

Figure 4: Power comparison under Bayesian adaptive-randomization (AR) and alternative balanced designs with (CPM and SSZ) and without (BR) interim analyses for a trial with two effective and three ineffective experimental arms for different overall sample sizes. Th panels correspond to scenarios with treatment effects for the primary outcome equal to 0.65 (left penal) and 0.75 (right panel) for both effective therapies compare to 0.5 for the control.



2.4 Combining Progression-Free and Overall-Survival outcome in Glioblastoma

We recently considered the use of a surrogate progression-free survival (PFS) outcome jointly with the primary outcome, overall survival (OS), also in GBM [37]. One potential way to shorten the time from trial initiation to early results of efficacy is to use imaging-based assessments of progression, such as PFS, with earlier times to event than OS [9]. Furthermore, since experimental therapies most directly influence the time until progression, it can be easier to detect effects on PFS, especially if there is long and heterogeneous treatment post progression [33]. There has been some concern, however, regarding the use of progression-based endpoints for clinical trials in neuro-oncology. While outcomes, such as OS, may have clear clinical relevance, endpoints based on imaging assessments, such as response or progression status, are not as clearly linked to patients benefit [1]. It is not trivial to anticipate how positive effects on overall response rates or PFS translate to effects in OS [2]. The approach that we followed is similar to the one illustrated for the endTB trial.

Trippa et al. [37] defined an adaptive randomization procedure for multi-arm trials based on a joint Bayesian model for PFS and OS outcomes. The model includes $(K + 1)$ PFS distributions $(S_{PSF,0}, \cdots, S_{PSF,K})$ and $(K + 1)$ OS distributions $(S_{OS,0}, \cdots, S_{OS,K})$, one for each of the K experimental arms and the control arm $k = 0$. Survival distributions are assumed to follow a proportional hazard model $S_{x,k} = S_{x,0}^{\theta_{x,k}}$ for both $x = PFS, OS$ and $k = 1, 2, 3$ with joint prior distribution for the unknown hazard ratios $\pi(\theta_{PFS}, \theta_{OS})$. Adaptation based on OS leverages on early PFS information through the joined model $\pi(\theta_{PFS}, \theta_{OS})$. At each patients' enrollment the posterior distributions of θ_{OS} given available PFD and OS outcomes is translated into randomization probabilities.

Advantages of joint modeling in this setting can be summarized by two properties. First, when treatment effects $\theta_{PFS,k}$ and $\theta_{OS,k}$ for PFS and OS are concordant, the proposed approach results in efficiency gains compared with randomization based on OS alone while sacrificing minimal efficiency compared with using PFS as the primary endpoint. Second, if treatment effects are limited to PFS, our approach provides randomization probabilities that are close to those based on OS alone. The alternative to our composite model would be to use OS only. Results in Trippa et al. [37] showed that the OS-only adaptive design still results in efficiency gains over a balanced randomization and, as expected, is not sensitive to randomizations driven by PFS effects that do not translate into OS improvements.

3 Computational Methods

We discuss computational approaches which helped us to evaluate and monitor frequentist operating characteristics for Bayesian designs. We illustrate (i) a stochastic search algorithm for the optimal constrained design $d_{c-\max}$, followed by (ii) a bootstrap procedure and (iii) an importance sampling algorithm, which we used for the endTB and Glioblastoma trial designs to control frequentist operating characteristics.

3.1 Simulated Annealing for Constrained Optimal Designs

Finding the constrained optimal designs analytically is infeasible in most cases. Stochastic search procedures can be used to find $d_{c-\max}$. We describe a simulated annealing algorithm for finding $d_{c-\max}$ [38, 29]. The procedure is summarized in Algorithm 2. The algorithm approximately identifies the constrained optimum within a compact set of candidate designs. The procedure starts from a candidate design d_1. For instance by generating random designs d from the set of designs $o(V)$, which satisfy the desired regulatory constraints V, and then selecting the design with the highest expected utility as starting value d_1. In the Bridging trial, in Section 2.1, a design is represented by the efficacy thresholds $z_{1:T}$, while V specifies a bound on type II/I error rates for these thresholds under fixed hypotheses.

The simulated annealing algorithm generates a Markov sequence of designs d_t from the set of designs which satisfy V [38]. At each iteration, the algorithm generates a design $d^\star \in o^{-1}(V)$ from a proposal distribution g_t with values of $d^\star$ in a neighborhood of the current state of the Markov chain d_t. The chain selects $d_{t+1} = d^\star$ with probability w_t or otherwise sets $d_{t+1} = d_t$ with probability $1 - w_t$. The acceptance probability w_t is an increasing function of the difference of expected utility $(U(d^\star) - U(d_t)) * \lambda_t$, where λ_t is an increasing multiplier. The optimal design is approximated by the design which attained the highest expected utility. Details are outlined in Algorithm 2.

Algorithm 2 Simulated annealing for constrained optimal designs

```
set $t = 1$
Select an initial design $d_t \in D$ with operating characteristics $o(d_t)$ in $V$
Compute the expected utility $U(d_t)$ of the design $d_t$
for $t$ equal to $1, \cdots, T$ do
    Generate a design $d^\star \sim g_t$ from a neighborhood $B(d_t, \epsilon_t) \cap V$ of $d_t$
    Compute $\Delta_t = U(d^\star) - U(d_t)$
    Compute the acceptance probability $w_t = \min(1, \exp\{\Delta_t \lambda_t\})$
    Generate $U \sim U(0, 1)$ and select $d_{t+1} = d^\star$ if $\Delta_t \leq U$ and $d_{t+1} = d_t$ otherwise.
end for
Return: $\widehat{d}_{c-\max} = \arg\max_{d \in \{d_1, \ldots, d_T\}} U(d)$
```

3.2 A Bootstrap Scheme for Controlling Type I Error Rates

We describe a bootstrap algorithm for combining Bayesian designs with frequentist analyses. The algorithm is a variation of the bootstrap scheme proposed in [30] for computing confidence intervals and is summarized below in Algorithm 3. The procedure is implemented separately for each treatment arm k, and tests the presence of a treatment effect for experimental arm k, with null hypothesis H_k. First, based on the data generated by the adaptive trial $\mathcal{T}$, and for a fixed arm of interest k, a test statistics Z_k is computed. In the TB trial [10] we use the standardized difference between the culture conversions proportions of experimental arm k and control therapy. Second, we compute for each arm k' consistent estimators of the outcomes distributions $\widehat{F}_{k'}$ under the null hypothesis H_k of no treatment effect for arm k. In the TB study for example this includes estimation of (a) the response probabilities for the surrogate endpoint; and (b) the conditional response probabilities for the primary endpoint, given a positive and negative early outcome. A consistent estimator of the accrual rate is also computed. In the TB trial response probabilities and the accrual rate will be estimated by sample averages and observed accrual rates. Third, to test H_k, we simulate $t = 1, \cdots, T$ adaptive trials $\mathcal{T}_t$, with the same stopping rules and tuning parameters as used in the actual trial. For the t-th simulated trial, patients enter the trial according to the estimated arrival rate, and each patient assigned to the control or experimental arms responds to therapies with probability identical to the estimates $\widehat{F}$. [1] Note that patients respond therefore to treatment k across the T simulations with probabilities that might be different from those observed in the trial because simulations have to be consistent with the null hypothesis H_k that we test. For each simulated trial $\mathcal{T}_t$ we then obtain a statistics $Z_k^{(t)}$, which represents approximately a draw from the null distribution under H_k. We finally estimate the p-value as the proportion of simulated trials with statistic $Z_k^{(t)}$ larger than the observed statistics Z_k. Lastly, the null hypothesis for arm k is rejected when the p-value is below the pre-specified level α. See Algorithm 3 for details.

Algorithm 3 A Bootstrap algorithm for testing treatment efficacy of therapy k.

```
Input: A design $d$ and a trial $\mathcal{T}$
Input: The experimental arm $k$ and hypothesis $H_k$ which should be tested
Compute the statistics $Z_k$ for arm $k$
Estimate the accrual rate of the trial by $\widehat{\lambda}$
Estimates of the outcome distributions for each arm $k'$ under $H_k$ by $\widehat{F}_{k'}$
for $t$ in 1 to $T$ do
    Simulate a trial $\mathcal{T}_t$ under $d$ with accrual rate $\widehat{\lambda}$ and outcome distributions $\widehat{F}_{k'}$
    Compute the statistics $Z_k^{(t)} = Z_k(\mathcal{T}_t)$
end for
reject $H_k$ at level $\alpha$ if $\widehat{p}_k = \frac{1}{T}\sum_1^T I(Z_k^{(t)} > Z_k) \leq \alpha$
```

3.3 Control of Type I Error Rates with Importance Sampling

Importance sampling has been recently used as an alternative approach to control the type I error under a pre-specified threshold α in Wason and Trippa [39]. To simplify the presentation we assume binary outcomes with response probabilities $\theta = (\theta_0, \ldots, \theta_K)$,

[1] There was a flaw in the sentence

one for each of the $K + 1$ therapies. Let Z be a summary statistics that, similarly to a p-value, evaluates evidence against a generic null hypothesis H_0, with large values indicating strong evidence against it. The approach is applicable to both Bayesian and non-Bayesian adaptive randomization schemes, and is summarized in Algorithm 4. The algorithm iteratively simulates $t = 1, \cdots, T$ adaptive clinical trials varying $\theta^{(t)}$ at each iteration.The response probabilities $\theta^{(t)}$ at each simulation t are generated independently from a continuous distribution g, for instance a beta distribution. Each simulated trial $\mathcal{T}_t$ is based on a different set of response probabilities $\theta^{(t)} \sim g$, where g is a conveniently selected distribution. Let $\mathcal{L}_\theta(\mathcal{T})$ be the likelihood of a trial $\mathcal{T}$ under the adaptive scheme; this is the probability of a specific sequence of outcomes and treatment assignments at a fixed value of the vector $\theta = (\theta_0, \ldots, \theta_K)$. We chose the distribution g so that for each generated trial, the importance weights

$$w(\mathcal{T}_t; \theta) = \frac{\mathcal{L}_\theta(\mathcal{T}_t)}{\int \mathcal{L}_{\theta'}(\mathcal{T}_t) g(\theta') d\theta'}$$

can be straightforwardly computed. Standard importance sampling, using the above weights, enables us to use the same draws $\{\mathcal{T}_t\}$ to approximate the distribution of Z at any desired θ value (see step 5 of Algorithm 4). The second part of the algorithm estimates the cut-off point z with the constraint that for every value θ consistent with the null hypothesis H_0 the inequality $p_\theta(Z > z) < \alpha$ holds. That is the cut-off point z controls the type I error at the α level. The algorithm use a grid of values for θ and selects z such that the estimated type I error rate – obtained by importance sampling – across possible probabilities θ is bounded by a pre-specified α value.

Algorithm 4 Importance Sampling for the control of type I error rates

```
Simulate $T$ response probabilities $\theta^{(t)} = (\theta_0^{(t)}, \cdots, \theta_K^{(t)}) \sim g(\theta), t = 1, \cdots, T$
Generate a trial $\mathcal{T}_t$ under design $d$ with patients response rates $\theta^{(t)}$ for each $t = 1, \cdots, T$
Compute the statistics $Z^{(t)}, t = 1, \cdots, T$
For each trial $t$ compute the importance weight
    $$w(\mathcal{T}_t; \theta) = \frac{\mathcal{L}_\theta(\mathcal{T}_t)}{\int \mathcal{L}_s(T) g(s) ds}$$
Approximate the type I error for the threshold $z$ at $\theta$ by
    $$\widehat{p}_\theta(Z > z) = \sum_{t=1}^{T} \frac{w(\mathcal{T}_t; \theta)}{\sum_\ell w(\mathcal{T}_\ell; \theta)} \times I(Z^{(t)} > z)$$
Compute $\widehat{z}_\alpha = \min\{z : \widehat{p}_\theta(Z > z) \leq \alpha$ for all $\theta$ in $H_0\}$
```

4 Summary

Clinical trials are evolving from traditional two-arm studies in large heterogeneous patient populations towards studies with many subpopulations, multiple research questions and substantial correlative analyses [18]. Traditional frequentist and Bayesian designs are often challenged by these new directions, which demand designs which are applicable in a variety of settings, and can be tailored towards specific research questions[15, 8]. Bayesian designs, which enable the use of explicit or implicit utilities u and prior probabilities π to incorporate existing information in the design, can be tailored to specific study purposes [4, 34]. Clinical investigators and medical journals are typically familiar with frequentists measures of evidence. Bayesian testing using Bayesian factors or posterior probabilities, while based on coherent foundational axioms, can be difficult to communicate to these audiences. In addition regulatory agencies, for instance the US Food and Drug Administration, continue to make systematic use of frequentists testing principles for drug approval and practice changing recommendations.

We outlined several applications to clinical trial designs and presented algorithms for the implementation of the proposed approach. The hybrid consists of a Bayesian design followed by a frequentist analysis Etzioni and Kadane [16]. The use of Bayesian designs is motivated by the desire to optimize the acquisition of information about the clinical utility of therapies by incorporating available prior knowledge and using response-adaptive assignments rules. The use of frequentists analysis is motivated by the desire to communicate results of clinical trials to the medical community, pharmaceutical companies and regulatory authorities using widely accepted frequentists metrics.

References

[1] B. M. Alexander and L. Trippa. Progression-free survival: too much risk, not enough reward? *Neuro-oncology*, 16(5):615–616, 2014.

[2] B. M. Alexander, E. Galanis, W. A. Yung, K. V. Ballman, J. M. Boyett, T. F. Cloughesy, J. F. Degroot, J. T. Huse, B. Mann, W. Mason, et al. Brain malignancy steering committee clinical trials planning workshop: report from the targeted therapies working group. *Neuro-oncology*, page nou154, 2014.

[3] A. Barker, C. Sigman, G. Kelloff, N. Hylton, D. Berry, and L. Esserman. I-spy 2: An adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clinical Pharmacology & Therapeutics*, 86(1):97–100, 2009.

[4] D. A. Berry. Bayesian statistics and the efficiency and ethics of clinical trials. *Statist. Sci.*, 19(1):175–187, 02 2004. doi: 10.1214/088342304000000044. URL http://dx.doi.org/10.1214/088342304000000044.

[5] D. A. Berry. Adaptive clinical trials: the promise and the caution. *Journal of Clinical Oncology*, 29(6):606–609, 2011.

[6] D. A. Berry and S. G. Eick. Adaptive assignment versus balanced randomization in clinical trials: a decision analysis. *Statistics in medicine*, 14(3):231–246, 1995.

[7] D. A. Berry and B. Fristedt. *Bandit problems: sequential allocation of experiments (Monographs on statistics and applied probability)*. Springer, 1985.

[8] S. M. Berry, J. T. Connor, and R. J. Lewis. The platform trial: an efficient strategy for evaluating multiple treatments. *JAMA*, 313(16):1619–1620, 2015.

[9] K. R. Broglio and D. A. Berry. Detecting an overall survival benefit that is derived from progression-free survival. *Journal of the National Cancer Institute*, 2009.

[10] M. Cellamare, S. Ventz, E. Boudin, C. Mitnick, and L. Trippa. Bayesian adaptive randomization in a clinical trial to identify new regimens for multidrug-resistant tuberculosis. *Clinical Trials*, in press, 2016.

[11] M.-H. Chen, J. G. Ibrahim, et al. The relationship between the power prior and hierarchical models. *Bayesian Analysis*, 1(3):551–574, 2006.

[12] S.-C. Chow, C. Chiang, J.-p. Liu, and C.-F. Hsiao. Statistical methods for bridging studies. *Journal of biopharmaceutical statistics*, 22(5):903–915, 2012.

[13] M. H. DeGroot. *Optimal statistical decisions*, volume 82. John Wiley & Sons, 1970.

[14] J. R. Eisele. The doubly adaptive biased coin design for sequential clinical trials. *Journal of Statistical Planning and Inference*, 38(2):249–261, 1994.

[15] L. J. Esserman and J. Woodcock. Accelerating identification and regulatory approval of investigational cancer drugs. *Jama*, 306(23):2608–2609, 2011.

[16] R. Etzioni and J. B. Kadane. Optimal experimental design for another's analysis. *Journal of the American Statistical Association*, 88(424):1404–1411, 1993.

[17] B. Freidlin and E. L. Korn. Biomarker-adaptive clinical trial designs. *Pharmacogenomics*, 11(12):1679–1682, 2010.

[18] D. Harrington and G. Parmigiani. I-spy 2 a glimpse of the future of phase 2 drug development? *New England Journal of Medicine*, 375(1):7–9, 2016. doi: 10.1056/NEJMp1602256. URL http://dx.doi.org/10.1056/NEJMp1602256. PMID: 27406345.

[19] F. Hu and W. F. Rosenberger. Optimality, variability, power: evaluating response-adaptive randomization procedures for treatment comparisons. *Journal of the American Statistical Association*, 98(463):671–678, 2003.

[20] F. Hu and L.-X. Zhang. Asymptotic properties of doubly adaptive biased coin designs for multitreatment clinical trials. *Annals of Statistics*, pages 268–301, 2004.

[21] L. Y. T. Inoue, D. A. Berry, and G. Parmigiani. Relationship Between Bayesian and Frequentist Sample Size Determination. *The American Statistician*, 59(1): 79–87, Feb. 2005.

[22] J. Jack Lee and C. T. Chu. Bayesian clinical trials in action. *Statistics in medicine*, 31(25):2955–2972, 2012.

[23] E. S. Kim, R. S. Herbst, I. I. Wistuba, J. J. Lee, G. R. Blumenschein, A. Tsao, D. J. Stewart, M. E. Hicks, J. Erasmus, S. Gupta, et al. The battle trial: personalizing therapy for lung cancer. *Cancer discovery*, 1(1):44–53, 2011.

[24] J. J. Lee and D. D. Liu. A predictive probability design for phase ii cancer clinical trials. *Clinical Trials*, 5(2):93–106, 2008.

[25] J. J. Lee, Xuemin Gu, and Suyu Liu. Bayesian adaptive randomization designs for targeted agent development. *Clinical Trials*, 7:584 – 596, 2010.

[26] J. J. Lee, N. Chen, and G. Yin. Worth adapting? revisiting the usefulness of outcome-adaptive randomization. *Clinical Cancer Research*, 18(17):4498–4507, 2012.

[27] W. H. Organization. *Global tuberculosis report 2015*. World Health Organization, 2015.

[28] E. C. Quant, J. Drappatz, P. Y. Wen, and A. D. Norden. Recurrent high-grade glioma. *Current treatment options in neurology*, 12(4):321–333, 2010.

[29] C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2004.

[30] W. F. Rosenberger and F. Hu. Bootstrap methods for adaptive designs. *Statistics in medicine*, 18(14):1757–1767, 1999.

[31] W. F. Rosenberger and J. M. Lachin. *Randomization in clinical trials: theory and practice*. John Wiley & Sons, 2016.

[32] W. F. Rosenberger, N. Stallard, A. Ivanova, C. N. Harper, and M. L. Ricks. Optimal adaptive designs for binary response trials. *Biometrics*, 57(3):909–913, 2001.

[33] M. Terasaki, K. Murotani, Y. Narita, R. Nishikawa, T. Sasada, A. Y. K. Itoh, and M. Morioka. Controversies in clinical trials of cancer vaccines for glioblastoma. *Journal of Vaccines & Vaccination*, 2013, 2013.

[34] P. F. Thall. Bayesian models and decision algorithms for complex early phase clinical trials. *Statist. Sci.*, 25(2):227–244, 05 2010. doi: 10.1214/09-STS315. URL http://dx.doi.org/10.1214/09-STS315.

[35] P. F. Thall and K. J. Wathen. Practical Bayesian adaptive randomisation in clinical trials. *European Journal of Cancer*, 5:859–866, 2007.

[36] L. Trippa, E. Q. Lee, P. Y. Wen, T. T. Batchelor, T. Cloughesy, G. Parmigiani, and B. M. Alexander. Bayesian adaptive randomized trial design for patients with recurrent glioblastoma. *Journal of Clinical Oncology*, 30:3258–3263, 2012.

[37] L. Trippa, P. Y. Wen, G. Parmigiani, D. A. Berry, and B. M. Alexander. Combining progression-free survival and overall survival as a novel composite endpoint for glioblastoma trials. *Neuro-oncology*, page nou345, 2015.

[38] S. Ventz and L. Trippa. Bayesian designs and the control of frequentist characteristics: a practical solution. *Biometrics*, 71(1):218–226, 2015.

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

[40] L. J. Wei and S. Durham. The randomized play-the-winner rule in medical trials. *Journal of the American Statistical Association*, 73(364):840–843, 1978.

[41] D. A. B. Yi Cheng, Fusheng Su. Choosing sample size for a clinical trial using decision analysis. *Biometrika*, 90(4):923–936, 2003.

[42] G. Yin, N. Chen, and J. Jack Lee. Phase ii trial design with bayesian adaptive randomization and predictive probability. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(2):219–235, 2012.

[43] L.-X. Zhang, F. Hu, S. H. Cheung, and W. S. Chan. Asymptotic properties of covariate-adjusted response-adaptive designs. *The Annals of Statistics*, pages 1166–1182, 2007.

[44] Y. Zhang, L. Trippa, and G. Parmigiani. Optimal bayesian adaptive trials when treatment efficacy depends on biomarkers. *Biometrics*, 2015.