CLINICAL TRIALS

Clinical Trials
1–12

DOI: 10.1177/1740774516665090
ctj.sagepub.com
SAGE

A Bayesian response-adaptive trial in tuberculosis: The *endTB* trial

Matteo Cellamare[1,2], Steffen Ventz[1,3], Elisabeth Baudin[4], Carole Mitnick[5,6] and Lorenzo Trippa[1]

Abstract

Purpose: To evaluate the use of Bayesian adaptive randomization for clinical trials of new treatments for multidrug-resistant tuberculosis.

Methods: We built a response-adaptive randomization procedure, adapting on two preliminary outcomes for tuberculosis patients in a trial with five experimental regimens and a control arm. The primary study outcome is treatment success after 73 weeks from randomization; preliminary responses are culture conversion at 8 weeks and treatment success at 39 weeks. We compared the adaptive randomization design with balanced randomization using hypothetical scenarios.

Results: When we compare the statistical power under adaptive randomization and non-adaptive designs, under several hypothetical scenarios we observe that adaptive randomization requires fewer patients than non-adaptive designs. Moreover, adaptive randomization consistently allocates more participants to effective arm(s). We also show that these advantages are limited to scenarios consistent with the assumptions used to develop the adaptive randomization algorithm.

Conclusion: Given the objective of evaluating several new therapeutic regimens in a timely fashion, Bayesian response-adaptive designs are attractive for tuberculosis trials. This approach tends to increase allocation to the effective regimens.

Keywords

Tuberculosis, clinical trial, Bayesian analysis, adaptive designs, preliminary outcomes

Introduction

In 2010, there were an estimated 650,000 prevalent cases of multidrug-resistant tuberculosis. Nearly 500,000 new cases emerge annually through acquisition of resistance during treatment and through airborne transmission.[1] Over 80% of the annual burden of drug-resistant tuberculosis worldwide occurs in 27 low- and middle-income countries; at the end of 2014, only 23% of the world's drug-resistant tuberculosis patients were receiving treatment.

One of the current recommended treatment regimens for patients is long (usually 20–24 months), expensive, difficult to implement, and associated with substantial adverse effects. Globally, success is achieved in approximately 50%.[1] The need for new regimens is, therefore, indisputable. The recent conditional approval by stringent regulatory authorities of two new anti-TB drugs, *bedaquiline* and *delamanid*, presents the first opportunity in more than 50 years to revolutionize treatment for tuberculosis. Regimens that meet most or all of the criteria for new regimens, as laid out in the 2014 article by Brigden et al.,[2] would have the widest applicability which include $\geq$1 new drug class, 3–5 likely effective drugs, effective against tuberculosis and extensively drug-resistant tuberculosis strains, 6-month duration, all oral, simple dosing profile, good side-effect profile, and minimal interaction with antiretroviral therapy.

[1]Department of Biostatistics, Dana-Farber Cancer Institute, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[2]Department of Statistical Sciences, Sapienza University of Rome, Rome, Italy
[3]Department of Computer Science and Statistics, The University of Rhode Island, Kingston, RI, USA
[4]Department of Clinical Research, Epicentre–MSF, Paris, France
[5]Harvard Medical School, Boston, MA, USA
[6]Partners In Health, Boston, MA, USA

Corresponding author:
Matteo Cellamare, Department of Statistical Sciences, Sapienza University of Rome, Rome 00185, Italy.
Email: matteo.cellamare@uniroma1.it

Availability of multiple effective treatments would be beneficial for several reasons. These include possible supply-chain issues, heterogeneity in distribution of anti-TB resistance among populations, and individual patient characteristics such as comorbidities requiring concomitant medications. A single regimen is unlikely to be sufficient to address the global burden of tuberculosis. Rather, identifying multiple, improved regimens would be extremely beneficial to scale up efforts.

Identifying new regimens is of the utmost urgency. The pace of development of new treatments to date is appallingly slow. Strategies to accelerate regimen development have, to date, focused on the regulatory pathway.[3] Advances in methodology may also help expedite late-stage regimen trials.[4,5] Bayesian adaptive randomized trial designs, which have gained attention in oncology and cardiovascular diseases,[6–9] represent a possible advance. At its core, the adaptive randomization approach uses the data generated during the trial—in this case, response data—to randomize more patients to the most promising regimens.[10]

The *endTB* study is an investigator-initiated Phase III trial that seeks to evaluate five novel treatments. It will generate evidence on efficacy and recommendations limited to those arms that will show treatment effects, for introduction into clinical routine practice. In the context of the *endTB* trial, we explore whether Bayesian outcome-adaptive randomization[11,12] based on preliminary outcomes may accelerate the development of tuberculosis regimens.

Methods

Regimen selection

The *endTB* trial will examine five standardized regimens (Table 1) for fluoroquinolone-susceptible tuberculosis that meet as many of the criteria for new regimens as possible.[2] Regimens all contain at least one new antituberculosis drug. Regimens contain four to five oral drugs. Treatment duration will be 9 months. Dosing will mostly be once daily.

The side-effect profile will be enhanced by limiting the number of strong QT interval–prolonging drugs to two (among five drugs that are used in treatment of tuberculosis despite potential cardiotoxicity). We expect no unmanageable interaction with antiretroviral therapy.[13–16] Priority is given to drugs that have had limited use in most settings (i.e. *bedaquiline, delamanid, clofazimine*, and *linezolid*) to avoid resistance due to prior exposure. Also favored are drugs whose role in tuberculosis treatment is thought to be so important that in the absence of definitive evidence of resistance, they are often included in regimens (*fluoroquinolones* and *pyrazinamide*).[17,18]

Preliminary and final end points

The primary outcome is treatment success at 73 weeks (TS-73) after randomization. Two preliminary binary end points will be measured at 8 (culture conversion, TS-8) and 39 weeks (treatment success, TS-39) after randomization. A 6-week incubation is required to assess these end points, so these results will be available, respectively, 14 and 45 weeks after randomization. The TS-39 response rate for the control arm in most of our simulations is set at 60%, the upper limit of the range of responses from published observational studies, meta-analyses, and placebo arms of randomized trials, from 32% to 62%.[19–23] The TS-73 response rate for the control arm in most of our simulations is set to 55%, reflecting an expected 5% relapse between 39 and 73 weeks from randomization.[24] Since we expect strong correlation between the 39-week and the 73-week end points, randomization will utilize the preliminary end points.

Adaptive randomization

For each arm k, our randomization algorithm estimates the TS-39 response

$$P(\text{TS} - 39 \text{ for treatment } k) = \psi_1(k)\phi(k) + \psi_0(k)(1 - \phi(k))$$

where $\phi(k)$ is the TS-8 probability, and $\psi_i(k)$ denotes the conditional probability of a TS-39 given a positive ($i = 1$) or negative ($i = 0$) 8-week outcome.

A 2-month (8-week) culture conversion has been accepted as predictive for the end-of-treatment

Table 1. Six regimens proposed for testing in *endTB* trial.

#	Bdq	Dlm	Cfz	Lzd	FQ	Z
1	Bdq			Lzd	Mfx	Z
2	Bdq		Cfz	Lzd	Lfx	Z
3	Bdq	Dlm		Lzd	Lfx	Z
4		Dlm	Cfz	Lzd	Lfx	Z
5		Dlm	Cfz		Lfx	Z
6	Conventional control, composed according to WHO Guidelines, including the possible use of delamanid or bedaquiline					

Bdq: bedaquiline; Dlm: delamanid; Cfz: clofazimine; Lzd: linezolid; FQ: fluoroquinolone; Mfx: moxifloxacin; Lfx: levofloxacin; Z: pyrazinamide.

outcome by the Food and Drug Administration (FDA).[25] Similarly, our model a priori assumes positive correlation between TS-8 and TS-39 outcomes. The trial will confirm or contradict this relation. If contradicted, the model a posteriori will not rely on the initial assumption. Bayesian joint modeling of preliminary outcomes has been previously discussed.[11,26] Our analyses for $\theta_k = (\phi(k), \psi_0(k), \psi_1(k))$ are based on standard Bayesian computations: we use beta priors for the unknown parameters. In particular, the prior probability for the TS-8 rate is a uniform distribution. For the conditional TS-39 rates, given a positive or negative TS-8 outcome, we use beta prior distributions with parameters 1 and 10, and 10 and 1.

Initially, participants will be allocated equally to each arm, and then, after sufficient preliminary data become available, adaptive randomization increases the probability of randomization to the most promising arms. Similar to previous applications of adaptive randomization,[11,27–29] we define the probability of randomizing the ith patient to the kth arm, $w_i(k)$, evaluating the evidence of positive treatment effects at 39 weeks. For experimental arm k, $w_i(k)$ is proportional to

$$\frac{\text{'S} - 39 \text{ for treatment } k > \text{TS} - 39 \text{ for the control given the data)}^{h(i)}}{\sum_{j=1} \text{'S} - 39 \text{ for treatment } j > \text{TS} - 39 \text{ for the control given the data)}^{h(i)}}$$

In other words, if before the enrollment of the ith patient, some regimens accumulate evidence of TS-39 rates higher than the control, then randomization probabilities to those regimens will be higher than to the others. For the control, the randomization probability is defined to approximately match the control sample size and the experimental arm with the highest number of enrolled patients,[27,30] $w_i(0)$ is proportional to

$$\frac{\exp\left(\max_{k=1,\ldots,5}\{\#\text{ enrollment to arm } k\} - \#\text{ enrollment to control}\right)}{5}$$

The function $h(i)$ is defined to enforce balanced randomization in the early stage of the trial; $h(i)$ is equal to 0 until 30 TS-8 outcomes become available for each treatment arm. Then, $h(i)$ increases with the total number of observed TS-8 outcomes, to a maximum value γ, and the randomization probabilities become progressively unbalanced, favoring allocation to the most promising regimens. Specifically, we used

$$h(i) = \gamma^* \left(\frac{N_8(i)}{N}\right)^{\eta} I(i)$$

where $I(i) = 1$ if, at the enrollment of the ith patient, at least 30 patients have been assigned to each arm, and $I(i) = 0$ otherwise. Here, N denotes the (maximum) total number of patients in the trial, and $N_8(i)$ equals the number of observed TS-8 outcomes until the ith enrollment. The parameters γ, $\eta > 0$ were selected similarly to the article by Trippa et al.,[27] through simulations of trials with (γ, η) parameters over a grid of plausible values. We then selected parameter values with a limited coefficient of variation for the final arm-specific sample sizes (values of γ and η equal to 2.5 and 3, respectively).

We also explored the idea of using an additional parameter $N_{\max}$ in the adaptive randomization design to limit the number of patients assigned to a single arm below a pre-specified maximum $N_{\max}$ (Table S4 in the online supplementary file).

Stopping rules

We include interim analyses at regular intervals after a total of 100, 200, and so on, overall TS-73 outcomes become available. Arms are dropped for futility if the available data suggest no treatment effect on the TS-73 probability; more formally, if at the jth interim analysis, the posterior probability of a TS-73 rate larger than the control TS-73 rate falls below the boundary b_j at stage j, then arm k is dropped. The computation of these posterior probabilities involves only the TS-73 outcome data and a uniform prior for the TS-73 outcome rates.

We select an increasing stopping boundary $b_j = b^*(j/J)^p$, where b denotes the boundary value at the end of the trial. The parameters (p, b) impact on the proportion of patients randomized to ineffective treatments, as well as on the design power of detecting effective treatments. We explored different combinations of parameter values b and p across simulation scenarios with different numbers of effective arms and various effect sizes. Values of p around 2, with properly tuned b, provide good operating characteristics, with a low probability of early stopping for effective arms and a high probability of dropping ineffective treatments. To preserve the power, b was selected such that the probability of dropping an arm with strong treatment effect (see scenarios A3 and A4 of Table 2) is bounded below 0.01. Other monotone boundaries could be considered. In our experience, it is important to choose the futility boundary by considering important aspects of the trial, including the number of arms and the maximum number of enrolled patients.

The *endTB* trial will not use early stopping rules for efficacy. We evaluated whether, given a relatively short accrual period of approximately 2.5 years, early stopping for efficacy based on primary outcome data would result in a substantial gain in terms of time necessary to

AQ1 **Table 2.** Simulation scenarios.

	TS-8						TS-39						TS-73					
	Arm						Arm						Arm					
Scenario	C	1	2	3	4	5	C	1	2	3	4	5	C	1	2	3	4	5
N1	0.3	0.3	0.3	0.3	0.3	0.3	0.6	0.6	0.6	0.6	0.6	0.6	0.55	0.55	0.55	0.55	0.55	0.55
N2	0.3	0.25	0.25	0.25	0.25	0.25	0.6	0.55	0.55	0.55	0.55	0.55	0.55	0.5	0.5	0.5	0.5	0.5
N3	0.3	0.45	0.3	0.3	0.3	0.3	0.6	0.6	0.6	0.6	0.6	0.6	0.55	0.55	0.55	0.55	0.55	0.55
N4	0.3	0.3	0.3	0.3	0.3	0.3	0.6	**0.75**	0.6	0.6	0.6	0.6	0.55	0.55	0.55	0.55	0.55	0.55
N5	0.3	**0.45**	0.3	0.3	0.3	0.3	0.6	**0.75**	0.6	0.6	0.6	0.6	0.55	0.55	0.55	0.55	0.55	0.55
N6	0.3	**0.45**	**0.45**	**0.45**	0.3	0.3	0.6	**0.75**	**0.75**	**0.75**	0.6	0.6	0.55	0.55	0.55	0.55	0.55	0.55
N7	0.3	**0.45**	0.3	0.3	0.3	0.3	0.6	**0.75**	0.6	0.6	0.6	0.6	0.55	0.45	0.55	0.55	0.55	0.55
A1	0.3	**0.45**	0.3	0.3	0.3	0.3	0.6	**0.75**	0.6	0.6	0.6	0.6	0.55	**0.7**	0.55	0.55	0.55	0.55
A2	0.3	**0.45**	**0.45**	0.3	0.3	0.3	0.6	**0.75**	**0.75**	0.6	0.6	0.6	0.55	**0.7**	**0.7**	0.55	0.55	0.55
A3	0.3	**0.55**	0.3	0.3	0.3	0.3	0.6	**0.85**	0.6	0.6	0.6	0.6	0.55	**0.8**	0.55	0.55	0.55	0.55
A4	0.3	**0.55**	0.55	0.3	0.3	0.3	0.6	**0.85**	**0.85**	0.6	0.6	0.6	0.55	**0.8**	**0.8**	0.55	0.55	0.55
A5	0.3	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**	0.6	**0.75**	**0.75**	**0.75**	**0.75**	**0.75**	0.55	**0.7**	**0.7**	**0.7**	**0.7**	**0.7**
A6	0.3	**0.45**	0.45	0.3	0.25	0.25	0.6	**0.75**	**0.75**	0.6	0.55	0.55	0.55	**0.7**	**0.7**	0.55	0.5	0.5
A7	0.3	**0.45**	**0.55**	**0.6**	0.3	0.3	0.6	**0.75**	**0.85**	**0.9**	0.6	0.6	0.55	**0.7**	**0.8**	0.85	0.55	0.55
A8	0.3	0.3	0.3	0.3	0.3	0.3	0.6	**0.75**	0.6	0.6	0.6	0.6	0.55	**0.7**	0.55	0.55	0.55	0.55
A9	0.3	0.25	0.3	0.3	0.3	0.3	0.6	**0.75**	0.6	0.6	0.6	0.6	0.55	**0.7**	0.55	0.55	0.55	0.55
A10	0.3	0.3	0.3	0.3	0.3	0.3	0.6	0.6	0.6	0.6	0.6	0.6	0.55	**0.7**	0.55	0.55	0.55	0.55
A11	0.3	0.3	0.3	0.3	0.3	0.3	0.6	0.6	0.6	0.6	0.6	0.6	0.55	**0.7**	**0.7**	0.55	0.55	0.55
A12	0.3	0.3	0.3	0.3	0.3	0.3	0.6	0.6	0.6	0.6	0.6	0.6	0.55	**0.7**	**0.7**	**0.7**	0.55	0.55

TS-8, TS-39, and TS-73 are the response probabilities for the five experimental arms and the control arm under the null scenarios N1–N7 and alternative scenarios A

declare significant results. Since this would not be the case and it would lead to a reduction in power (between 2% and 20% across the scenarios that we considered), we opted against early stopping for efficacy. Patient accrual will therefore proceed until the planned overall sample size, unless all experimental arms are dropped for futility.

Hypothesis testing

We used Bayesian assignment rules, followed at study completion by frequentist analyses.[31] This includes rigorous hypothesis testing of the one-sided null hypothesis that arm $k = 1, \ldots, 5$ has a TS-73 response probability lesser than or equal to the TS-73 control rate, that is, H_k: TS-73 rate for arm k is lesser than or equal to the TS-73 control rate. type I error is controlled at the target level $\alpha = 0$. he p-values for the TS-73 end point, accounting for adaptive randomization and stopping rules, are computed using bootstrap methodology proposed by Trippa et al.[27] and Rosenberger and Hu[32] and described in the supplementary material. When the conclusions of the trial are reported to medical and epidemiology communities, p-values are de facto an accepted standard to communicate results. This is the main motivation for reporting at completion of the study frequentist analyses, after Bayesian randomization. Additional examples of Bayesian designs combined with frequentist hypothesis testing are discussed by Ventz and Trippa,[31] Bowden and Trippa,[33] and Trippa et al.[34]

Comparison to fixed randomization

We compare adaptive randomization to a multi-arm, multi-stage design. A multi-arm, multi-stage design would be a natural alternative design for the *endTB* trial. The multi-stage design assigns a fixed maximum number of patients to each arm using balanced randomization. For each experimental arm, interim analyses for futility, which are performed after 25%, 50%, and 75% of the TS-73 outcomes under the control and experimental arms, have been observed. The stopping boundaries were selected to target cumulative probabilities of 0.25, 0.50, and 0.75 to stop an arm without a TS-73 treatment effect after the first, second, and third interim analyses, respectively.[35] At completion of the study, efficacy is tested at 5% significance level. The sample size was selected to obtain 80% power with TS-73 probabilities at 0.7 and 0.55 under the experimental and control arms, respectively. Subsequently, to facilitate comparisons of different trial designs, for each scenario in Table 2, we tuned the sample size of the adaptive randomization design to match the average sample size of multi-stage designs across simulations and fixed scenarios and estimate power.

Simulation scenarios

We used a number of scenarios. Fixed parameters are the number of regimens tested (5) and the monthly enrollment rate (16 participants). We explored the effect of varying the number of effective arms on power. Table 2 describes 19 simulation scenarios. We considered seven null scenarios N1–N7 and 12 alternative scenarios A1–A12. In the first two null scenarios, the response probabilities for all end points are equal (N1) or inferior (N2) to the control response. In scenarios N3, N4, and N5, one arm has a positive TS-8 effect (N3), a positive TS-39 effect (N4), or a positive TS-8 and TS-39 effect (N5) without an effect for the primary TS-73 end point. In scenario N6, three arms have a positive TS-8 and TS-39 effect but no TS-73 effect. Finally, in scenario N7, one arm has a positive TS-8 and TS-39 effect but is inferior to the control for the TS-73.

Scenarios A1–A12 in Table 2 correspond to different alternative scenarios. In the first four scenarios, either one arm (A1, A3) or two arms (A2, A4) are effective for all three end points, with identical effect sizes. In scenario A5, all experimental arms are effective with identical treatment effects for all three end points; while in scenarios A6 and A7, two arms (A6) or three arms (A7) have positive effects for all three end points, with different arm-specific effect sizes. In the last five scenarios, we explore cases where the early end points are not good surrogates of the primary TS-73 end point. In scenarios A8 and A9, one arm has no TS-8 effect, but a positive TS-39 and TS-73 effect. Finally, in scenarios A10–A12, no arm has a positive TS-8 or TS-39 effect, but one arm (A10), two arms (A11), or three arms (A12) have positive TS-73 effects.

An R package is available at http://bcb.dfci.harvard.edu/~steffen/software.html.

Results

The box plots in Figure 1 show the arm-specific sample sizes at completion of the *endTB* trial across simulations, both under adaptive randomization and multi-stage designs, for scenarios A3 and A4. The mean overall sample size of multi-stage designs and adaptive randomization in these two scenarios equals 715 (A3) and 737 (A4) patients. Each box plot illustrates variability across 10,000 simulations. For scenario A3, with one effective regimen, the median patient accrual to the effective arm using adaptive randomization is 154 with interquartile range 140–170 compared to 128 for multi-stage designs (interquartile range: 128–128). The median assignment to the control is 146 patients (interquartile range 136–159) for adaptive randomization compared to 128 (interquartile range: 128–128) for multi-stage designs, respectively. For scenario A4, with two effective treatments, in more than 50% of the simulations, adaptive randomization assigns at least AQ2

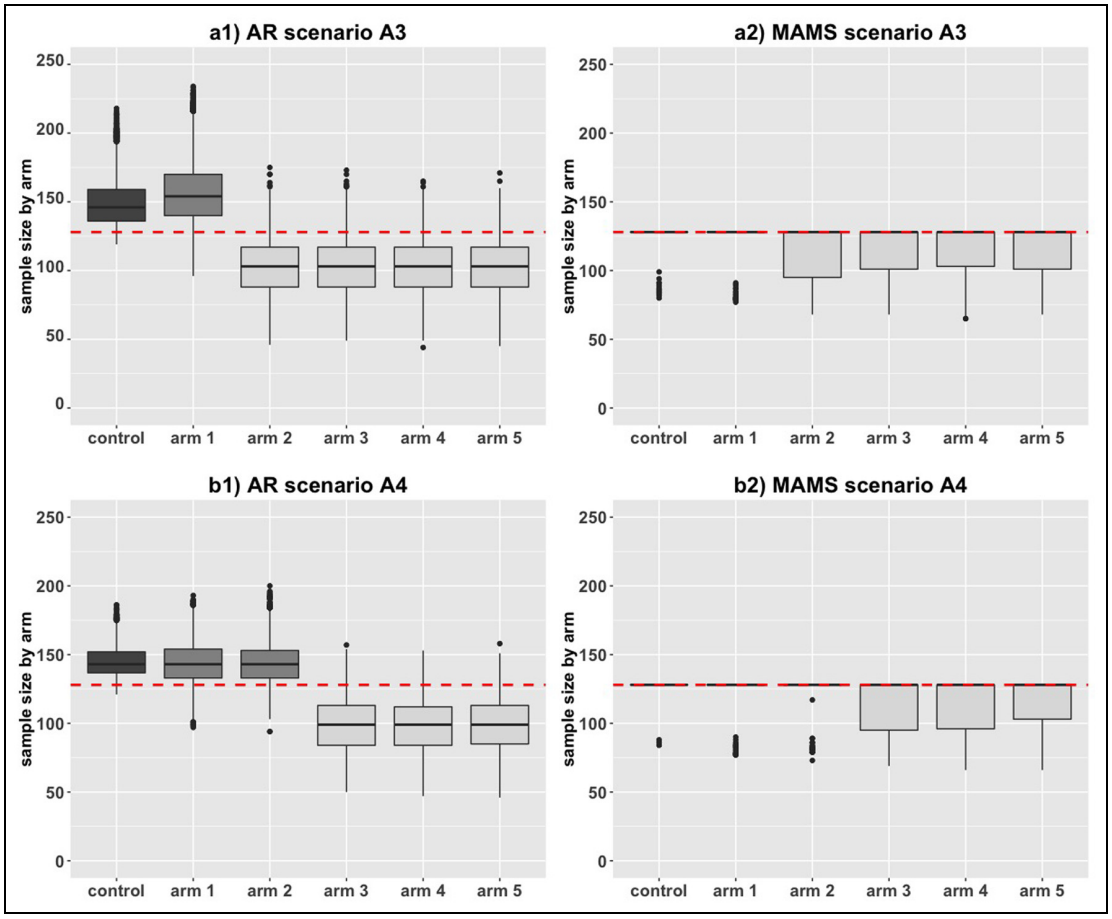

Figure 1. Box plots of the number of patients randomized to each arm at the end of the trial across 10,000 simulations for scenarios A3 and A4 with adaptive randomization and multi-stage designs. For both scenarios, A3 and A4, the sample size for adaptive randomization was selected to match the expected sample size of multi-stage designs (715 and 737 for scenarios A3 and A4, respectively). The dashed line indicates the median sample size for each arm under balanced randomization without early stopping.

143 patients to each effective arm, with interquartile range (133, 153). Under multi-stage designs, the median is 128 with interquartile range (128, 128). Moreover, for the non-effective arms, adaptive randomization and multi-stage designs have a median accrual of 99 and 128 patients with interquartile range of (84, 113) and (95, 128) patients, respectively. For both adaptive randomization and multi-stage designs, the outliers in Figure 1 are a result of early stopping for futility. Due to the relatively late availability of the primary TS-73 end point, the multi-stage designs trial enrolled approximately 65% of the overall sample size before the first interim analysis when 25% of the arm-specific TS-73 outcomes are observed.

For each scenario, Table 3 shows the mean enrollment to each arm and the power to declare a significant TS-73 result of multi-stage designs and adaptive randomization when both designs have an identical average sample size. For instance, for scenario A1, the multi-stage designs trial has a maximum and mean sample size of 768 and 713 patients and 80% power to detect superiority for arm 1. We compare multi-stage designs in this scenario to an adaptive randomization trial with sample size of 713 patients. Adaptive randomization has 85.9% power to detect the treatment effect of arm 1. The second column in Table 2 shows the average sample size under multi-stage designs. For adaptive randomization, the second column indicates the sample size necessary to match the power of multi-stage designs.

The last two columns of Table 2 show the expected number of positive TS-73 outcomes, and the probability of unbalancing the design in the wrong direction. The latter is defined as the probability of assigning less than E_{BR} patients to the effective arms, where E_{BR} is the expected number of patients assigned to effective arms under multi-stage designs without early stopping, $E_{BR} = N/6$*number of effective arms.

Under scenario A7, effective regimens have different effect sizes. The sample size for adaptive randomization was set to match the 80% power level for the arm with the smallest treatment effect. For scenarios with one

AQ3

AQ4

Table 3. Mean enrollment to each arm and the power to declare a significant TS-73 result of multi-stage designs and adaptive randomization.

Multi-arm, multi-stage designs

		Control	Arm 1		Arm 2		Arm 3		Arm 4		Arm 5			
	EN	ESS(SD)	ESS(SD)	Pow	ESS(SD)	Pow	ESS(SD)	Pow	ESS(SD)	Pow	ESS(SD)	Pow	ENO	PU
N1	700	126(8)	115(19)	5.6	115(19)	5.5	115(19)	5.1	115(19)	5.1	115(19)	5.1	302	–
N2	664	124(12)	108(21)	0.8	108(21)	0.9	108(21)	0.9	108(21)	0.8	108(21)	0.8	294	–
N3	700	126(8)	115(19)	5.9	115(19)	5.7	115(19)	5.7	115(19)	5.2	115(19)	5.3	301	–
N4	701	126(8)	115(19)	5.1	115(19)	5.7	115(19)	5.7	115(19)	5.1	115(19)	5.3	302	–
N5	700	126(8)	115(19)	5.6	115(19)	5.9	115(19)	5.9	115(19)	5.5	115(19)	5.3	301	–
N6	701	126(8)	115(19)	5.1	115(19)	5.7	115(19)	5.1	115(19)	5.1	115(19)	5.1	301	–
N7	686	126(9)	100(20)	0.1	115(19)	5.2	115(19)	5.2	115(19)	5.1	115(19)	5.8	304	–
A1	713	128(4)	**127(7)**	**80.4**	114(19)	5.1	115(19)	5.8	115(19)	5.5	115(19)	5.7	299	3.1
A2	735	128(2)	**127(7)**	**79.7**	**127(7)**	**79.5**	115(19)	5.5	114(19)	5.7	115(19)	5.6	287	6.4
A3	715	128(2)	**128(2)**	**99.6**	115(19)	5.2	115(19)	5.6	115(19)	5.5	115(19)	5.1	288	0.2
A4	737	128(1)	**128(2)**	**99.5**	**128(2)**	**99.5**	114(19)	4.9	114(19)	5.0	115(19)	5.6	263	0.3
A5	761	128(1)	**127(7)**	**79.5**	**127(7)**	**79.9**	**127(7)**	**80.2**	**127(7)**	**79.8**	**127(7)**	**79.9**	247	1.4
A6	711	128(3)	**127(7)**	**80.0**	**127(7)**	**80.2**	115(19)	5.3	107(21)	0.9	107(21)	1.0	291	5.4
A7	738	128(0)	**127(7)**	**79.3**	128(2)	99.5	128(0)	100	114(19)	5.8	114(19)	5.5	243	3.7
A8	714	128(4)	**127(7)**	**80.0**	115(19)	5.5	115(19)	5.1	115(19)	5.2	115(19)	5.5	300	3.2
A9	714	128(4)	**127(7)**	**80.2**	115(19)	5.6	115(19)	5.9	115(19)	5.5	115(19)	5.6	299	3.4
A10	713	128(4)	**127(7)**	**79.3**	115(19)	5.5	115(19)	5.6	115(19)	5.3	115(19)	5.1	299	3.5
A11	735	128(2)	**127(7)**	**79.7**	127(7)	79.6	114(19)	5.8	115(19)	5.1	115(19)	5.6	287	6.2
A12	736	128(1)	**127(7)**	**80.2**	**127(7)**	**79.9**	**127(7)**	**79.8**	114(19)	5.7	114(19)	5.7	274	9.3

Bayesian adaptive randomization

		Control	Arm 1		Arm 2		Arm 3		Arm 4		Arm 5			
	N_{Pow}	ESS(SD)	ESS(SD)	Pow	ESS(SD)	Pow	ESS(SD)	Pow	ESS(SD)	Pow	ESS(SD)	Pow	ENO	PU
N1	–	139(13)	112(24)	4.9	112(24)	4.8	112(24)	5.1	112(24)	4.7	112(24)	4.9	314	–
N2	–	135(14)	105(25)	0.7	105(25)	0.6	105(25)	0.8	105(25)	0.7	105(25)	0.7	323	–
N3	–	139(13)	118(24)	5.1	111(24)	4.8	110(24)	4.8	110(23)	5.0	110(24)	5.1	314	–
N4	–	142(15)	138(23)	5.2	105(21)	5.1	105(21)	4.8	105(21)	5.2	105(21)	5.1	314	–
N5	–	142(14)	139(23)	5.0	104(21)	4.8	104(21)	4.7	104(21)	4.8	105(21)	4.8	314	–
N6	–	137(12)	124(18)	4.8	124(18)	4.7	124(18)	5.1	96(18)	5.0	96(18)	4.8	314	–
N7	–	139(14)	121(24)	0.0	106(22)	4.9	106(22)	5.1	107(22)	5.2	106(22)	4.9	320	–
A1	645	146(15)	**151(21)**	**85.9**	104(20)	5.1	104(20)	4.9	104(20)	5.2	104(20)	5.0	298	9.6
A2	675	144(12)	**141(16)**	**82.9**	**141(16)**	**82.9**	100(19)	4.7	100(19)	4.8	99(19)	4.8	284	10
A3	620	149(17)	**156(22)**	**100**	103(20)	5.1	102(20)	4.8	102(19)	4.7	102(20)	5.1	282	6.5
A4	635	145(11)	**143(15)**	**99.9**	**143(15)**	**99.9**	98(19)	5.0	98(19)	4.8	99(18)	4.8	255	7.6
A5	780	138(6)	**125(12)**	**78.9**	**125(12)**	**78.7**	**124(12)**	**78.9**	**125(12)**	**79.1**	**125(12)**	**78.7**	249	0.0
A6	650	145(12)	**143(16)**	**84.3**	**143(16)**	**85.1**	102(19)	4.8	89(17)	0.8	89(17)	0.7	286	7.1
A7	700	142(8)	**130(12)**	**81.1**	**137(12)**	**99.8**	**138(12)**	**100**	96(18)	4.8	96(18)	5.0	236	10.6
A8	650	146(15)	**150(21)**	**85.5**	104(20)	4.8	104(20)	4.9	105(20)	5.1	105(20)	4.9	298	10.3
A9	660	145(15)	**149(21)**	**84.5**	105(20)	4.7	105(20)	4.8	105(20)	5.1	105(20)	4.7	298	10.8
A10	785	142(14)	**116(25)**	**76.9**	113(25)	4.9	114(25)	5.2	114(25)	5.2	114(25)	4.9	303	53.3
A11	775	145(14)	**118(26)**	**77.2**	**117(26)**	**78.8**	115(26)	4.8	115(26)	4.9	115(26)	4.8	290	60.4
A12	765	148(13)	**119(26)**	**77.6**	**119(26)**	**78.9**	119(26)	77.7	116(27)	5.1	116(27)	4.8	277	62.5

Power (Pow), expected sample size (ESS), and standard deviation (SD) for each arm across 10,000 simulations under scenarios N1–N7 and A1–A12. The type I error is controlled at 5%. The last two columns display the expected number of negative TS-73 outcomes (ENO) and the probability of unbalancing the design in the wrong direction (PU). For multi-stage designs, the maximum overall sample size is 768 patients, and the first column shows the expected sample size (EN) under each scenario. For each scenario, the sample size for adaptive randomization was set equal to the expected sample size of multi-stage designs, and the first column N_{Pow} shows the sample size of adaptive randomization that is required to obtain the same power as multi-stage designs.

and two arms equally superior to the control (scenarios A1, A2, A3, and A4), the adaptive randomization design has higher power than multi-stage designs. Under scenario A1, adaptive randomization requires a sample size of 645 patients to detect the effective regimen with 80% power, while a multi-stage designs trial requires an average of 713 randomized patients for 80% power. With two effective regimens, in scenario A2, an average of 675 patients are required to detect the treatment effects with 80% power, compared to 735 patients for multi-stage designs.

When all five arms have identical treatment effects, adaptive randomization randomizes on average an equal number of patients to each experimental arm, but adaptive assignment has a higher variability compared to multi-stage designs. Consequently, under scenario A5, with five effective treatments, adaptive randomization requires 20 patients more than multi-stage designs to detect treatment effects with 80% power (780 vs 761).

In scenario A7, where each of the three effective regimens has a different effect size, adaptive randomization randomizes a relevant proportion of patients to the most effective arm. In this case, adaptive randomization requires 747 patients to detect the smallest treatment effect with 80% power, compared to 738 patients with multi-stage designs, respectively. With 747 patients, adaptive randomization has a power of 99% to detect the remaining two effective regimens. For multi-stage designs, with 738 patients, the power for the remaining two effective arms is 98% and >99%. In scenarios A8 and A9, where arm 1 has no (A8) or a negative (A9) TS-8 effect, but a positive TS-39 and TS-73 effect, adaptive randomization requires 650 and 660 patients for 80% power, compared to 714 for multi-stage designs, respectively.

When treatments have no positive TS-8 or TS-39 effect but a positive TS-73 effect, then adaptive randomization has no preliminary evidence for superiority and assigns patients on average in equal proportions to each experimental arm. In scenarios A10–A12, where one (A10), two (A11), or three (A12) arms have a positive TS-73 effect, but no preliminary TS-8 or TS-39 effect, adaptive randomization assigns on average a smaller proportion patients to the effective arms than multi-stage designs and requires 62 (A10), 50 (A11), and 29 (A12) more patients to achieve the same power as multi-stage designs.

When we compared adaptive randomization with multi-stage designs, under identical average, overall sample sizes, we observed that in scenarios N1 and N2, where the TS-8, TS-39, and TS-73 of all experimental arms, are equal or inferior to the control, both adaptive randomization and multi-stage designs randomize on average approximately the same number of patients to each experimental arm. In scenarios N3–N5, where arm 1 has a positive effect for the TS-8, the TS-39 or both early end points, but no effect for the primary TS-73 end point, adaptive randomization randomizes on average 3 (N3), 23 (N4), and 24 (N5) more patients to arm 1 compared to multi-stage designs trial. In scenario N6, three arms have a strong TS-8 and TS-39 effect, but no TS-73 effect, in this case, adaptive randomization randomizes on average nine more patients to each of the three arms compared to multi-stage designs. In scenario N7, arm 1 has a strong positive effect for the TS-8 and TS-39, but is inferior compared to the control. In this case, adaptation based on the early outcomes assigns significantly more patients to an ineffective arm compared to multi-stage designs (120 vs 100), respectively.

Similarly, in scenarios A8 and A9, one regimen has a positive TS-73 effect, but no TS-8 effect (A8) or an inferior TS-8 response probability (A9). In this case, the power of adaptive randomization declines from 86% under scenario A1 to 84% under A9, compared to 80% for multi-stage designs. If no regimen is effective for any of the preliminary outcomes, then adaptive randomization randomizes on average the same number of patients to each regimen (A10–A12). In this case, adaptive randomization has a lower power than multi-stage designs (77%–78% compared to 80%) due to its larger variability of patient assignment. The last three scenarios illustrate the operating characteristics of adaptive randomization if the preliminary outcomes are not good surrogates for the primary TS-73 outcome.

Finally, we explored the effect of changing the accrual rate and the time at which the early outcomes are available. In particular, we explored the performance of adaptive randomization and multi-stage designs when we increased the planned accrual rate from 16 patients per month to 32 and 64 patients per month. Table S3 in the online supplementary material shows the average enrollment and power for each arm under adaptive randomization for scenarios A1–A12 when the accrual rate is 16, 32, or 64 patients per month. Faster accrual gives less time for adaptive randomization and multi-stage designs to accumulate outcome information. For instance, for scenario A2, the power to detect the two effective arms changes from 83% for 16 enrollments per month to 81% and 80% for 32 and 64 enrollments per month, respectively. In the latter case, the power of adaptive randomization is identical to multi-stage designs. For scenario A7, where three arms are effective, with different effect sizes, adaptive randomization's power remains nearly identical under faster accrual rates, with most notable changes for arm 1 with the weakest treatment effect (81.1%, 80.8%, and 80.1%).

Adaptive randomization is relatively robust to delays in observing the two preliminary outcomes. For example, if outcomes are observed 4 weeks later than expected, in scenarios A1 and A2, the power to detect treatment effect of the first arm remains nearly unchanged (85.4% instead of 85.9%, Table S4).

To mirror the discussion during the design of the *endTB* study on whether to evaluate superiority or non-inferiority of the experimental regimens, we also conducted simulations for adaptive randomization and multi-stage designs testing non-inferiority. In this case, we considered for each experimental arm the null hypothesis that the TS-73 probability of the experimental arm is lesser than or equal to the control TS-73 minus 0.12, that is, we set the non-inferiority margin to 12%. Modifications for testing non-inferiority using adaptive randomization and multi-stage designs are outlined in the supplementary material.

We considered again scenarios N1–N7 and A1–A12, where the TS-8, TS-39, and TS-73 rates of the control remain identical to Table 2. For the experimental arms, the TS-8, TS-39, and TS-73 rates are equal to the corresponding values in Table 2 minus the non-inferiority margin of 12%. The general performance of adaptive randomization compared to multi-stage designs remains identical when testing superiority or non-inferiority, as shown in Table S8. When up to three of the five experimental arms are non-inferior and the TS-39 is predictive of the final TS-73 end points, adaptive randomization has a similar or higher power as multi-stage designs. If the TS-8 and TS-39 are not good surrogate outcomes for the TS-73, then adaptive randomization is slightly outperformed by multi-stage designs.

Discussion

In the context of *endTB*, an investigator-initiated Phase III trial examining five tuberculosis regimens, Bayesian adaptive randomization could expose larger proportions of study participants to effective regimens than a multi-arm, multi-stage study. The gain of adaptive randomization in our simulation study becomes more pronounced with small effect sizes and with strong positive correlations between the surrogate and primary end points. In scenarios with a positive correlation between the TS-39 and TS-73 outcomes, adaptive randomization randomizes more patients to effective arms and has higher power to detect effective arms compared to a multi-arm, multi-stage study. In our simulations, a weak or negative correlation between the TS-8 and TS-73 had only moderate impact on adaptive randomization's operating characteristics when there is a positive relation between the TS-39 and TS-73 outcomes. Nevertheless, increasing numbers of effective arms, with equivalent assumptions on response rates and desired power, requires increased overall sample sizes for both adaptive randomization and multi-stage designs. The gains of adaptive randomization rely on positive relation between the preliminary and primary outcomes and the availability of the outcomes data in a timely fashion. As shown in several simulation scenarios, if the surrogate outcomes are negatively correlated with the final outcome, then the operating characteristics of adaptive randomization are compromised. We expect a positive association between the TS-39 and the TS-73 outcomes, which led to the decision of using an adaptive randomization design in the *endTB* trial.

Both adaptive randomization and multi-stage designs require additional resources and coordination across participation sites compared to independent two-arm studies, that is, one two-arm trial for each experimental arm. But testing all five experimental regimes in separate two-arm studies would require up to 1280 patients, compared to a maximum of 768 patients under adaptive randomization or multi-stage designs in realistic trial scenarios.

Adaptive randomization represents a departure from methods traditionally used for tuberculosis trials, but several factors facilitate exploration and evaluation of this approach. First, the code for the simulations performed for this article is publicly available and can be tested. Second, despite the multiple looks at the data for adaptation and interim analysis, rigorous frequentist hypothesis testing can be performed. We used a bootstrap procedure, applicable in either case, for testing superiority or non-inferiority.[32] Third, these methods have been previously used in different contexts, including oncology.[36] So, while these methods will be new for tuberculosis, they are not untested.

There has been an extensive debate about the merits and drawbacks of outcome-adaptive randomization. We mention some critiques from the literature: (1) adaptation can lead to a larger overall sample size,[37,38] (2) the relative number of patients who receive the best treatment option increases, but at the same time, the number of patients treated with an inferior arm might also increase due to the larger overall sample size,[38] and (3) moreover, treatment allocation can be more variable compared to balanced designs, and time trends in the data may lead to biased effect size estimates.[37–39] Adaptive randomization requires additional resources for planning and coordination,[37,38,40] compared to balanced randomization. Ethical concerns have been recently discussed in by Berry[41] and in subsequent letters.[10,41–46] The benefits of adaptive randomization are small in the two-arm settings and more clearly seen in the multi-arm setting.[10,27] Wason and Trippa[30] compared balanced randomization and adaptive randomization and estimated power differences when multiple superior treatments exist.

The adaptive randomization design relies on early end points that predict later response. In multidrug-resistant tuberculosis, the evidence for validated surrogate end points is limited but growing. The proportion of 2-month sputum culture conversion, as a proxy for later outcomes, was sufficient to result in marketing authorization for the two new anti-tuberculosis drugs, *delamanid* and *bedaquiline*. In the case of treatment shortening, however, it has not been established

whether 2-month conversion, or any other interim end point, has true surrogacy for final treatment response. The recent studies REMox, RIFAQUIN, and OFLOTUB[47–49] revealed that culture conversion at week 8 (similar to *endTB* TS-8) among patients receiving shortened treatment for drug-susceptible tuberculosis was not a sufficiently robust marker to accurately predict treatment effect at the end of the follow-up period. If the correspondence between TS-39 and TS-73 outcomes is poor, this could result in large proportions of participants being assigned to regimens that have encouraging early response rate without a TS-73 treatment effect. Nonetheless, the adaptive algorithm re-estimates continuously with accumulated information the relationship between the TS-8 and TS-39 end points.

Drifts and variations in the population during the accrual period are a general concern across clinical trials.[50] The operating characteristics of outcome-adaptive designs are known to be sensitive to changes in the patient population during the trial.[38] Outcome-adaptive designs use the observed data over the whole trial for adaptation. In the presence of time trends in the patient population, adaptive randomization may produce biased estimates of treatment effects, and the type I error rates may deviate from the nominal target. We conducted additional simulations with trends in the population. Specifically, we consider a linear increase in all response probabilities by 0.2 between the first enrollment and the end of the trial. Table S7 in the supplementary material summarizes the effect these time trends on the type I error/power and treatment effect estimates for adaptive randomization and multi-stage designs under scenarios N1–N7 and A1, A2, A5, and A10. For multi-stage designs, type I error rates are close to the nominal values of 5%, whereas for adaptive randomization, type I error rates are biased downward with values between 5% and 0.2%. The power of adaptive randomization and multi-stage designs and the estimated treatment effects do not show substantial variations due to time trends (see Table S7). Overall, these results show the sensitivity of the operating characteristics of adaptive randomization to time trends, which support the use of balanced randomized designs in settings where one can expect trends in the population and outcome distributions. A secondary analysis will be conducted at the end of the *endTB* study to detect potential seasonal variations and time trends of treatment effects[50] and to prevent potential biased conclusions.

Although two new drugs have recently received conditional approval from regulatory authorities for tuberculosis treatment, these approvals are based on Phase II studies that demonstrated improved treatment efficacy by adding the new drug to the existing regimen.[19,51] Neither efforts reduced the complexity, toxicity, and cost of treatment. Use of these new drugs needs to be promptly optimized for delivery in settings of need: a very small fraction of patients with the disease are estimated to receive effective treatment each year.[52] The design explored for the *endTB* trial allows relatively rapid evaluation of five new, shorter, experimental regimens, enrolling fewer than 1000 patients. For the *endTB* trial, in which we may detect up to three superior treatments, adaptive randomization is an attractive alternative to standard multi-stage designs. Adaptive randomization may reduce the trial duration and hence more rapidly detect multiple effective arms.

Acknowledgements

M.C. and S.V. are co-first authors; C.M. and L.T. are co-last authors.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This work has been partially funded by Burroughs Wellcome Fund—Regulatory Science and UNITAID.

References

1. World Health Organization (WHO). *Global tuberculosis report 2015*. Geneva: WHO, 2015.
2. Brigden G, Nyang'wa B-T, du Cros P, et al. Principles for designing future regimens for multidrug-resistant tuberculosis. *Bull World Health Organ* 2014; 92: 68–74.
3. Breitstein J and Spigelman M.The role of product development partnerships in advancing the development of drugs for unmet needs. *Clin Pharmacol Ther* 2013; 93: 468–470.
4. Bratton DJ, Phillips PP and Parmar MK.A multi-arm multi-stage clinical trial design for binary outcomes with application to tuberculosis. *BMC Med Res Methodol* 2013; 13: 139.
5. Horsburgh CR, Shea KM, Phillips P, et al. Randomized clinical trials to identify optimal antibiotic treatment duration. *Trials* 2013; 14: 88.
6. Barker A, Sigman C, Kelloff G, et al. I-SPY 2: An adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clin Pharmacol Ther* 2009; 86: 97–100.
7. Zhou X, Liu S, Kim ES, et al. Bayesian adaptive design for targeted therapy development in lung cancer—a step toward personalized medicine. *Clin Trials* 2008; 5: 181–193.
8. Connor JT, Elm JJ and Broglio KR.ESETT and ADAPT-IT investigators. Bayesian adaptive trials offer advantages in comparative effectiveness trials: an example in status epilepticus. *J Clin Epidemiol* 2013; 66: S130–S137.
9. Collins SP, Lindsell CJ, Pang PS, et al. Bayesian adaptive trial design in acute heart failure syndromes: moving beyond the mega trial. *Am Heart J* 2012; 164: 138–145.

10. Berry DA.Adaptive clinical trials: the promise and the caution. *J Clin Oncol* 2011; 29: 606–609.
11. Trippa L, Wen PY, Parmigiani G, et al. Combining progression-free survival and overall survival as a novel composite endpoint for glioblastoma trials. *Neuro Oncol* 2015; 17: 1106–1113.
12. Berry SM, Carlin BP, Lee JJ, et al. *Bayesian adaptive methods for clinical trials*. Boca Raton, FL: CRC Press, 2010.
13. Svensson EM, Aweeka F, Park JG, et al. Model-based estimates of the effects of efavirenz on bedaquiline pharmacokinetics and suggested dose adjustments for patients coinfected with HIV and tuberculosis. *Antimicrob Agents Chemother* 2013; 57: 2780–2787.
14. Svensson EM, Dooley KE and Karlsson MO.Impact of lopinavir-ritonavir or nevirapine on bedaquiline exposures and potential implications for patients with tuberculosis-HIV coinfection. *Antimicrob Agents Chemother* 2014; 58: 6406–6412.
15. Van Heeswijk R, Dannemann B and Hoetelmans R. Bedaquiline: a review of human pharmacokinetics and drug–drug interactions. *J Antimicrob Chemother* 2014; 69: 2310–2318.
16. Dooley KE, Park JG, Swindells S, et al. Safety, tolerability, and pharmacokinetic interactions of the antituberculous agent TMC207 (bedaquiline) with efavirenz in healthy volunteers: AIDS Clinical Trials Group Study A5267. *J Acquir Immune Defic Syndr* 2012; 59: 455–462.
17. Falzon D, Jaramillo E, Schünemann H, et al. WHO guidelines for the programmatic management of drug-resistant tuberculosis: 2011 update. *Eur Respir J* 2011; 38: 516–528.
18. European Medicines Agency. *Summary of product characteristics*. London: European Medicines Agency.
19. Diacon AH, Pym A, Grobusch MP, et al. Multidrug-resistant tuberculosis and culture conversion with bedaquiline. *New Engl J Med* 2014; 371: 723–732.
20. Johnston JC, Shahidi NC, Sadatsafavi M, et al. Treatment outcomes of multidrug-resistant tuberculosis: a systematic review and meta-analysis. *PLoS ONE* 2009; 4: e6914.
21. Ahuja SD, Ashkin D, Avendano M, et al. Multidrug resistant pulmonary tuberculosis treatment regimens and patient outcomes: an individual patient data meta-analysis of 9,153 patients. *PLoS Med* 2012; 9: e1001300.
22. Orenstein EW, Basu S, Shah NS, et al. Treatment outcomes among patients with multidrug-resistant tuberculosis: systematic review and meta-analysis. *Lancet Infect Dis* 2009; 9: 153–161.
23. Skripconoka V, Danilovits M, Pehme L, et al. Delamanid improves outcomes and reduces mortality in multidrug-resistant tuberculosis. *Eur Respir J* 2013; 41: 1393–1400.
24. Becerra MC, Appleton SC, Franke MF, et al. Recurrence after treatment for pulmonary multidrug-resistant tuberculosis. *Clin Infect Dis* 2010; 51: 709–711.
25. Navarro E.*Cross discipline team leader review*. Food amnd Drug Administration, 2012.
26. Yin G, Chen N and Lee JJ.Phase II trial design with Bayesian adaptive randomization and predictive probability. *J R Stat Soc Ser C Appl Stat* 2012; 61: 219–235.
27. Trippa L, Lee EQ, Wen PY, et al. Bayesian adaptive randomized trial design for patients with recurrent glioblastoma. *J Clin Oncol* 2012; 30: 3258–3263.
28. Thall PF and Wathen JK.Practical Bayesian adaptive randomisation in clinical trials. *Eur J Cancer* 2007; 43: 859–866.
29. Alexander BM, Wen PY, Trippa L, et al. Biomarker-based adaptive trials for patients with glioblastoma—lessons from I-SPY 2. *Neuro Oncol* 2013; 15: 972–978.
30. Wason J and Trippa L.A comparison of Bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials. *Stat Med* 2014; 33: 2206–2221.
31. Ventz S and Trippa L.Bayesian designs and the control of frequentist characteristics: A practical solution. *Biometrics* 2015; 71: 218–226.
32. Rosenberger WF and Hu F.Bootstrap methods for adaptive designs. *Stat Med* 1999; 18: 1757–1767.
33. Bowden J and Trippa L.Unbiased estimation for response adaptive clinical trials. *Stat Methods Med Res*. Epub ahead of print 11August2015. DOI: 10.1177/0962280215597716.
34. Trippa L, Rosner GL and Müller P.Bayesian enrichment strategies for randomized discontinuation trials. *Biometrics* 2012; 68: 203–211.
35. Anderson JR and High R.Alternatives to the standard Fleming, Harrington, and O'Brien futility boundary. *Clin Trials* 2011; 8: 270–276.
36. Food and Drug Administration. *Adaptive design clinical trials for drugs and biologics*. FDA, 2010.
37. Simon R.Adaptive treatment assignment methods and clinical trials. *Biometrics* 1977; 33: 743–749.
38. Korn EL and Freidlin B.Outcome-adaptive randomization: Is it useful?*J Clin Oncol* 2011; 29: 771–776.
39. Thall P, Fox P and Wathen J.Statistical controversies in clinical research: scientific and ethical problems with adaptive randomization in comparative clinical trials. *Ann Oncol* 2015; 26: 1621–1628.
40. Hey SP and Kimmelman J.Are outcome-adaptive allocation trials ethical?*Clin Trials* 2015; 12: 102–106.
41. Berry DA.Commentary on Hey and Kimmelman. *Clin Trials* 2015; 12: 107–109.
42. Buyse M.Commentary on Hey and Kimmelman. *Clin Trials* 2015; 12: 119–121.
43. Joffe S and Ellenberg SS.Commentary on Hey and Kimmelman. *Clin Trials* 2015; 12: 116–118.
44. Korn EL and Freidlin B.Commentary on Hey and Kimmelman. *Clin Trials* 2015; 12: 122–124.
45. Lee JJ.Commentary on Hey and Kimmelman. *Clin Trials* 2015; 12: 110–112.
46. Saxman SB.Commentary on Hey and Kimmelman. *Clin Trials* 2015; 12: 113–115.
47. Jindani A, Hatherill M, Charalambous S, et al. *A multicentre randomized clinical trial to evaluate high-dose rifapentine with a quinolone for treatment of pulmonary TB: The RIFAQUIN trial*. Atlanta, GA: Program and abstracts of the 20th Conference on Retroviruses and Opportunistic Infections, 2013.
48. Jindani A, Harrison TS, Nunn AJ, et al. High-dose rifapentine with moxifloxacin for pulmonary tuberculosis. *New Engl J Med* 2014; 371: 1599–1608.
49. Rustomjee R, Lienhardt C, Kanyok T, et al. A Phase II study of the sterilising activities of ofloxacin, gatifloxacin

and moxifloxacin in pulmonary tuberculosis. *Int J Tuberc Lung Dis* 2008; 12: 128–138.
50. Altman DG and Royston JP.The hidden effect of time. *Stat Med* 1988; 7: 629–637.
51. Gler MT, Skripconoka V, Sanchez-Garavito E, et al. Delamanid for multidrug-resistant pulmonary tuberculosis. *New Engl J Med* 2012; 366: 2151–2160.
52. World Health Organization. *Global tuberculosis report 2014*. Geneva: WHO, 2014.