BMC Medicine

World TB Day

RESEARCH ARTICLE　　　　　　　　　　　　　　　　　　**Open Access**

CrossMark

A new trial design to accelerate tuberculosis drug development: the Phase IIC Selection Trial with Extended Post-treatment follow-up (STEP)

Patrick P. J. Phillips[1]*, Kelly E. Dooley[2], Stephen H. Gillespie[3], Norbert Heinrich[4,5], Jason E. Stout[6], Payam Nahid[7], Andreas H. Diacon[8,9], Rob E. Aarnoutse[10], Gibson S. Kibiki[11], Martin J. Boeree[12] and Michael Hoelscher[4,5]

Abstract

Background: The standard 6-month four-drug regimen for the treatment of drug-sensitive tuberculosis has remained unchanged for decades and is inadequate to control the epidemic. Shorter, simpler regimens are urgently needed to defeat what is now the world's greatest infectious disease killer.

Methods: We describe the Phase IIC Selection Trial with Extended Post-treatment follow-up (STEP) as a novel hybrid phase II/III trial design to accelerate regimen development. In the Phase IIC STEP trial, the experimental regimen is given for the duration for which it will be studied in phase III (presently 3 or 4 months) and patients are followed for clinical outcomes of treatment failure and relapse for a total of 12 months from randomisation. Operating characteristics of the trial design are explored assuming a classical frequentist framework as well as a Bayesian framework with flat and sceptical priors. A simulation study is conducted using data from the RIFAQUIN phase III trial to illustrate how such a design could be used in practice.

Results: With 80 patients per arm, and two (2.5 %) unfavourable outcomes in the STEP trial, there is a probability of 0.99 that the proportion of unfavourable outcomes in a potential phase III trial would be less than 12 % and a probability of 0.91 that the proportion of unfavourable outcomes would be less than 8 %. With six (7.5 %) unfavourable outcomes, there is a probability of 0.82 that the proportion of unfavourable outcomes in a potential phase III trial would be less than 12 % and a probability of 0.41 that it would be less than 8 %. Simulations using data from the RIFAQUIN trial show that a STEP trial with 80 patients per arm would have correctly shown that the Inferior Regimen should not proceed to phase III and would have had a high chance (0.88) of either showing that the Successful Regimen could proceed to phase III or that it might require further optimisation.

Conclusions: Collection of definitive clinical outcome data in a relatively small number of participants over only 12 months provides valuable information about the likelihood of success in a future phase III trial. We strongly believe that the STEP trial design described herein is an important tool that would allow for more informed decision-making and accelerate regimen development.

Keywords: Tuberculosis, Clinical trials, Middle development, Phase IIC, STEP, Drug development, Regimen development

* Correspondence: patrick.phillips@ucl.ac.uk
[1]MRC Clinical Trials Unit at University College London, Aviation House, 125 Kingsway, London, WC2B 6NH, UK
Full list of author information is available at the end of the article

Phillips *et al. BMC Medicine* (2016) 14:51

Page 2 of 11

Background

New regimens are urgently needed for the treatment of tuberculosis (TB). The current standard of care, a 6-month four-drug regimen, has been in use for decades. Under optimal clinical trial conditions it can achieve a durable cure in 92 % of patients [1]; in practice, approximately 10 % of patients are forced to stop treatment prematurely owing to toxicity or because life circumstances present challenges to prolonged adherence to supervised treatment. Many of these patients will later experience relapse and, under programme settings, completion rates (a proxy for cure) are closer to 80–85 % [2]. Thus, the standard regimen is inadequate to control the epidemic and shorter, simpler regimens are needed to defeat what is now the world's greatest infectious disease killer [2].

Exciting progress has been made, with two compounds in different classes (bedaquiline and delamanid) receiving accelerated approval by regulators for the treatment of multi-drug resistant TB (MDR-TB) since 2012 [2]. In addition, there has been growing interest in optimising old TB drugs [3] and repurposing drugs with other indications that were previously thought to have limited activity against TB [4, 5]. A number of clinical trials of shorter regimens combining new and old drugs are ongoing (ClinicalTrials.gov identifiers NCT02410772, NCT02581527 and NCT02342886). Innovative trial designs originally developed in oncology, including the multi-arm multi-stage design (MAMS) that allows for poorly performing arms to be stopped early, have been implemented in TB phase II trials with some success [6, 7]. There has, however, been no change in the treatment of drug-sensitive TB for more than 40 years. Most recently, three major international multicentre phase III trials demonstrated that four different 4-month regimens did not provide as good a standard of care as the 6-month regimen [1, 8, 9].

Antibiotics, and in particular those used for the treatment of TB, were some of the first drugs evaluated systematically in medicine. While establishing the composition of anti-TB first-line treatment is often thought to be the first example of evidence-based medicine or 'rational therapeutics' [10, 11], this feat was achieved largely empirically through a series of studies that we would now define as phase III trials [12]. Thus, there has been no established pathway for clinical development in TB, and the approach put forward by drug developers has been instituted only in the last 10–15 years. In fact, no new compound has been approved for the treatment of drug-sensitive TB since the disease was declared a global emergency by the World Health Organization in 1993, with the exception of intermittent low-dose rifapentine in the continuation phase of treatment [13], which has been implemented only in a few low-incidence settings. The accelerated approvals of the two new drugs for MDR-TB, bedaquiline and delamanid, were based on data from phase II trials only. This was possible because these trials employed surrogate endpoints that were considered likely to reflect clinical outcomes [14] and because MDR-TB treatment outcomes with the current standard of care are so poor (54 % cure [15]) and the need for new medicines so acute. Full approval will only be achieved once phase III trials are completed successfully. The accelerated approval framework with the US and EU regulators could not be used for drug-sensitive TB because there is an established and highly effective, albeit prolonged, treatment. Furthermore, such accelerated approval mechanisms are not often available for many regulators outside the US and the EU.

Clinical development of new treatments is conventionally split into three phases. Phase I trials are first-in-human studies aimed at exploring safety, pharmacokinetics, and relevant drug-drug interactions [16]. The safety of the participants (often healthy volunteers) is the overwhelming primary concern because incidence and the magnitude of toxicities cannot be readily predicted by preclinical toxicology studies, and unexpected serious adverse events not seen in animal studies do sometimes occur [17]. Phase III trials are large and expensive multicentre randomised studies that are designed to provide confirmatory evidence that a new treatment is safe and efficacious. Thus, the objectives of the phase III trial are to provide adequate evidence to convince regulators to approve the treatment for the disease in question and to change policy and practice. In between these two poles of development is a critical area: phase II or 'middle development' [18], where the treatment is given to individuals with the target disease with the objectives of identifying an optimal therapeutic dose or doses, exploring and establishing safety in a larger group of participants, and generating efficacy data (often on intermediate endpoints) sufficient to provide the confidence needed to justify a phase III trial. Because TB is treated with regimens rather than individual drugs, an additional goal of middle development is to optimise and select those combinations with the highest likelihood of success in phase III treatment-shortening trials.

Middle development for TB drugs typically follows a sequential two-step or three-step pathway. The first step is a 14-day monotherapy trial (a new drug can only be given for a maximum of 14 days as monotherapy owing to the risk of acquiring drug resistance), which is used to demonstrate the microbiological activity of the agent alone at different doses. This is often followed by a 7-day to 14-day study of drugs in combination. These trials are followed by longer-duration studies of drugs in combination (traditionally 8 weeks). An example of a development pathway is that of pretomanid (PA-824), first evaluated as monotherapy, then in combination in 14-day and 8-week studies [19–23]. The 14-day studies

Phillips *et al. BMC Medicine* (2016) 14:51

Page 3 of 11

are considered phase IIA and 8-week studies as phase IIB, and in each case all patients are given standard treatment after completing the experimental therapy so that the total duration of treatment is not less than the standard 6 months. The primary endpoints of such studies are microbiological intermediate endpoints, that is, measuring the speed at which bacilli are killed using serially collected sputum during the first days and weeks of treatment, and include the decline in the number of colony-forming units growing on solid culture media [23], increase in time to positivity in liquid culture media [23], time to stable negative culture conversion [3], or proportion of participants with negative sputum cultures at a specific time during treatment [24]. Killing bacilli during a course of treatment is a complex process that is not well understood. It has been proposed that different drugs act on up to four different sub-populations of bacilli, because drug activity may depend on the metabolic state of the organism and the compartment in which the organism resides in the infected host [25]. Some drugs, such as isoniazid, are thought to act earlier, targeting metabolically active, rapidly multiplying organisms. Others, such as pyrazinamide and rifampicin, are believed to act later, when the hardest-to-kill bugs must be eradicated. Penetration of different drugs into the lesions where the bacilli reside is another area that affects the timing and extent of drug activity [26]. For this reason, there is limited value in only measuring the activity of a dose or a regimen during the first 14 days to identify the combinations to advance from phase IIA to phase IIB studies [27], and the activity may depend on the drug's mechanism of action and pharmacokinetic characteristics. For example, only a small effect was seen with bedaquiline in an initial 7-day study [28], but in a follow-on 14-day trial that included loading doses, clear activity was demonstrated [29]. Clofazimine had no effect over 14 days at tested doses [22]. Nevertheless, both bedaquiline and clofazimine have important roles in combination with other agents in longer duration randomised clinical trials for the treatment of MDR-TB [30, 31]. In practice, the 14-day and 8-week studies merely demonstrate that the drug or regimen has some anti-mycobacterial activity in humans, but both yield incomplete efficacy or safety data. More importantly, even microbiological activity over 8 weeks has a limited role in decision-making for advancing regimens to phase III trials [32–35]. The modest benefits in microbiological activity with moxifloxacin seen in 8-week studies were confirmed in the REMoxTB phase III trial [1], but these benefits did not translate into activity sufficient to cure patients in just 4 months. Stated directly, many regimens with impressive microbiological activity over 2 or 8 weeks of treatment are not sufficiently active to allow for treatment shortening, demonstrated by high rates of relapse [1, 8, 9, 12]. Furthermore, safety data generated in phase II trials are

valuable but limited given the shorter duration of use over only 8 weeks.

The model of clinical development employed by the pharmaceutical industry assumes a limited number of drug candidates and is focused on gathering adequate data to inform go/no-go decisions along the development pathway so that promising compounds are advanced efficiently and unpromising ones are abandoned early. In general, about 60–70 % of drugs that enter phase III trials are successfully licensed, but the success rate depends highly on the disease entity and drug class [36]. This model has informed many of the common approaches in TB regimen development, with doses and regimens selected based on early clinical studies and carried through the development process until they fail. This pathway for developing regimens for TB has not been successful to date. Because TB is a major public health problem with drug development funded largely through public and philanthropic sources [37], the focus of middle development should be the efficient identification and testing of those regimens that are most likely to shorten TB therapy rather than excessive attention to particular drugs or combinations. A barrier to this is the panoply of potential old, repurposed, and new drugs, which results in a large number of possible three-drug and four-drug combination regimens that might lead to shorter, simpler regimens.

To redress this over-reliance on 14-day and 8-week endpoints and to facilitate this broader perspective on the middle development of TB drugs and regimens, we propose a new hybrid phase II/III trial design: the Phase IIC Selection Trial with Extended Post-treatment follow-up (STEP). We believe that the Phase IIC STEP trial design will help to bridge the gap between phase II and phase III by improving the quality and relevance of data generated in middle development. This improved evidence base prior to phase III will allow for informed decisions for selecting novel regimens to proceed to confirmatory trials and will reduce the risk of expensive phase III failures. In this paper we describe the STEP design and explore its operating characteristics under typical scenarios. We have also conducted a simulation study using data from a large, recently completed phase III trial to illustrate how such a design could be used in practice.

Methods

STEP trial design

We define a STEP trial as one with the following features:

1. The novel regimen is given for the duration for which it is intended to be studied in phase III (presently 3 or 4 months).
2. Treatment is stopped after experimental therapy is completed (provided there is sufficient confidence

Phillips *et al. BMC Medicine* (2016) 14:51

Page 4 of 11

that it is safe to do so from evidence collected in the study itself or in previous studies).

3. Patients are followed for clinical outcomes of treatment failure and relapse for a total of 12 months from randomisation.

This can be contrasted with the current phase IIB design (Fig. 1). A STEP trial is designed with a similar sample size to a standard phase IIB study with the primary outcome defined in typical fashion—time to culture conversion or decline in bacillary load over time [23, 38]. However, because the regimen is given for its intended duration, and patients are followed for a total of 12 months from randomisation, data on the definitive clinical endpoint, status at 12 months post-randomisation (a composite outcome of treatment failure and relapse) is also collected. This composite outcome is the same as that used in confirmatory phase III trials [1, 8]. Twelve months of follow-up is considered adequate because more than 75 % of relapses occur within 6 months of stopping

Fig. 1 Comparison of standard phase IIB design (**a**) with the STEP design (**b**)

treatment [39]; 18 months of follow-up would considerably prolong the duration of the trial without significantly enhancing the knowledge gained. While the precision in this composite unfavourable outcome will be low because unfavourable outcomes are rare and phase II trials are relatively small (and it therefore should formally be a secondary rather than primary outcome in the trial), it will provide invaluable information to guide decisions about further clinical development of the regimen. Equally important are the longer-term safety and tolerability data, both during treatment and following completion of treatment. A STEP trial could be considered as a pilot phase III trial, with the collection of similarly rich data on clinically relevant outcomes but at a much lower sample size and cost. Results of a STEP trial may support several potential development strategies: direct advancement for evaluation in a definitive phase III trial, further phase II work to optimise the doses or change the regimen components, pharmacokinetic-pharmacodynamic (PK-PD) assessments to understand underperforming regimens or identify populations that are not seeming to benefit, or cessation of development activities altogether.

Statistical methods

Sample sizes for TB phase IIB trials have varied from 60 to 100 per arm, with this number considered to be adequate to demonstrate important differences in intermediate endpoints such as time to culture conversion or slope of decline in bacillary load [23, 38, 40]. We therefore selected this range for a STEP trial to explore the operating characteristics for the important secondary outcome of an unfavourable outcome. This allows us to show what could be achieved without an increase in sample size as compared to a phase IIB trial. Inference from this secondary outcome is considered first from a more familiar classical frequentist perspective and second from a Bayesian framework, which lends itself more to the sort of decision-making necessary for clinical development of regimens.

In the frequentist framework, hypothesis testing for the secondary outcome of an unfavourable outcome was set up as a 90 % two-sided confidence interval to exclude a specified proportion at the upper bound. The power for a given sample size, the proportion to be excluded at the upper bound, and the assumed true proportion were estimated from standard sample size formulae for a one-sample proportion [41]. A 90 % confidence interval (corresponding to a 5 % one-sided level of significance) was chosen given that this is a phase II trial with the objective of collecting evidence to guide decisions for future confirmatory trials, although other TB phase II trials more often use 95 % confidence intervals [24].
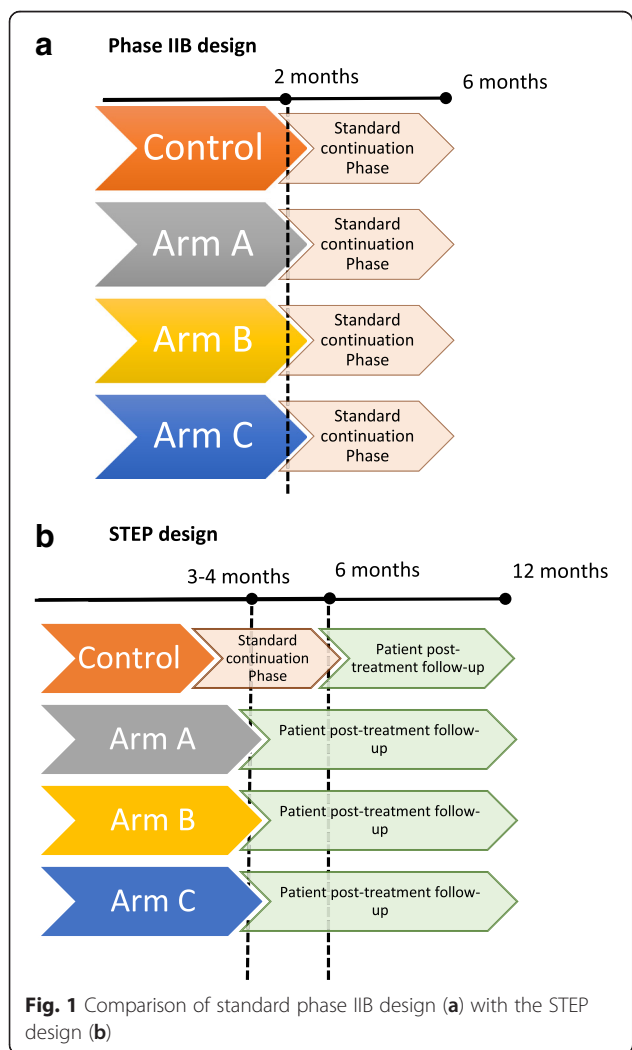
Phillips *et al. BMC Medicine* (2016) 14:51

Page 5 of 11

In the Bayesian framework, the posterior distribution of the proportion of unfavourable outcomes was calculated using a binomial distribution for the observed data and using both flat and sceptical beta conjugate priors. A flat prior assigns equal (low) probability for each unfavourable outcome proportion from 0 % to 100 %. The sceptical prior was chosen as having a fairly wide distribution with a higher density around a 20 % proportion of unfavourable outcomes, which would be an unsuitable regimen for evaluation in a phase III trial, and a lower density for very high proportions of unfavourable outcomes (extremely unlikely for a regimen selected for phase II evaluation). A prediction of the number of unfavourable outcomes for a future phase III trial was calculated from the predictive posterior distribution, which follows a beta-binomial distribution [42]. The predictive probability mimics the clinical decision-making process and provides the probability that a regimen will have an acceptably low proportion of unfavourable outcomes in a future phase III trial. The hypothetical phase III trial is said to have 500 patients per arm, which is not dissimilar to other previous and ongoing TB phase III trials.

In order to further explore the operating characteristics of the STEP trial design, data were repeatedly sampled without replacement from the recently completed RIFAQUIN phase III trial, with the unfavourable outcome as defined in the trial, censored at 12 months from randomisation. RIFAQUIN evaluated two novel regimens for the treatment of TB in seven sites across southern Africa as compared to the standard 6-month control [8]. RIFAQUIN included one arm that was shown to be non-inferior to the control regimen (the Successful Regimen) and one that did not meet criteria for non-inferiority (the Inferior Regimen). The Successful Regimen in RIFAQUIN was actually a 6-month regimen, but for the purposes of illustration, we can assume that both experimental arms were 4-month regimens. This trial is thus useful for exploring what additional information a STEP trial might have added had it been conducted before the phase III trial.

Ethical permissions

The trial data from the RIFAQUIN trial used for the simulation study were fully anonymised for the purposes of these analyses and therefore no further ethical approval was required. The RIFAQUIN trial received ethical approval and all patients gave informed consent, as detailed in the primary publication [8].

Results

Power calculations in a frequentist framework

Figure 2 shows the power curves for the secondary outcome of an unfavourable outcome under various scenarios within the frequentist framework. With a STEP trial of 80 assessable patients per arm (Fig. 2b), there would be 80 % power to exclude a proportion of unfavourable outcomes of 18.3 % at the upper bound of the 90 % confidence interval, or 90 % power to exclude a proportion of 19.8 % assuming a true proportion of unfavourable outcomes of 8 %, similar to that observed with the 6-month control regimen in recent phase III trials [1]. The same trial would give 80 % power to exclude a proportion of unfavourable outcomes of 9.3 % and 90 % power to exclude a proportion of 10.2 % for a superior regimen with a true unfavourable outcome of 2 %.

Bayesian predictive probability

An alternative perspective would be to consider the Bayesian predictive probability of the unfavourable outcome being lower than a pre-specified threshold (p_1) in a future phase III trial with 500 patients per arm, given the observed unfavourable outcome from the STEP trial and incorporating prior information. This approach allows for explicit incorporation of the prior confidence in a regimen before the data are collected in the STEP trial. Figure 3 shows three different prior distributions reflecting a range of prior confidence from sceptical to enthusiastic. Figure 4 shows the Bayesian predictive probability that the proportion of unfavourable outcomes is less than a pre-specified threshold (p_1) in a potential phase III trial given a sceptical or flat prior. With the sceptical prior and 80 patients per arm, two (2.5 %) unfavourable outcomes in the STEP trial would mean there is a probability of 0.99 that the proportion of unfavourable outcomes in the potential phase III would be less than $p_1 = 12$ % and a probability of 0.91 that the proportion of unfavourable outcomes would be less than $p_1 = 8$ %. With six (7.5 %) unfavourable outcomes in the STEP trial, there is a probability of 0.82 that the proportion of unfavourable outcomes in a potential phase III would be less than $p_1 = 12$ % and a probability of 0.41 that it would be less than $p_1 = 8$ %.

Results from simulations employing the RIFAQUIN trial data

Sampling data from the RIFAQUIN trial allows us to consider what the additional benefit of a STEP trial would have been had it been conducted before the phase III trial, showing how such a trial design might be used in practice. Focusing specifically on the Bayesian formulation, Fig. 5 shows the results of 10,000 random draws from the RIFAQUIN trial data using the primary outcome as defined in the trial censored at 12 months [8] using a sceptical prior. Similar results with the flat prior are shown in Fig. 6. The predictive probability is grouped into four categories to show possible decision outcomes: 0.00 to <0.50, 0.50 to <0.75, 0.75 to <0.95, and 0.95 to 1.00 (Fig. 7). A predictive probability of <0.50 might be used to indicate that the development of a regimen be abandoned
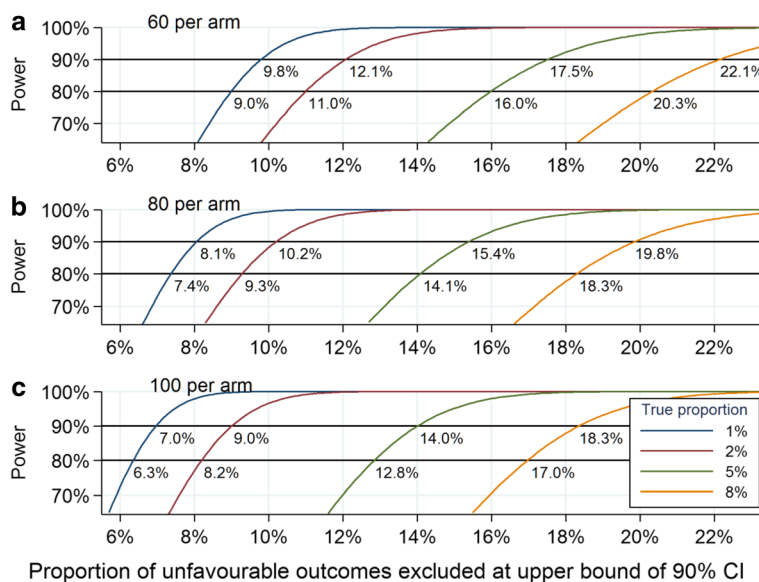
Fig. 2 Power curves for excluding a proportion of unfavourable outcomes under different assumptions of the true proportion, assuming (**a**) 60, (**b**) 80, and (**c**) 100 patients per arm for a STEP trial. Numbers on graphs show the proportion of unfavourable outcomes that could be excluded at the upper bound of the 90 % two-sided confidence interval (95 % one-sided confidence interval, *CI*) with 80 % and 90 % power

(*stop*, red bars) and a predictive probability of >0.95 might be used to indicate strong evidence for continuing the regimen to a phase III trial (*go*, green bars). Predictive probabilities of 0.75 to <0.95 (*proceed with caution*, amber bars) might be used to indicate further optimisation is required before proceeding to evaluation in a phase III trial, with the decision depending on other outcomes from the

STEP trial. Predictive probabilities of 0.50 to <0.75 (blue bars) might be used to indicate changes to the regimen should be considered before possible further re-evaluation in a STEP trial.

Considering a STEP trial of 80 patients per arm, 100 % of the simulated trials for the Inferior Regimen had a predictive probability of less than 0.50 that the unfavourable proportion in a phase III trial would be less than $p_1 = 8$ % (Fig. 5b) indicating this regimen should be abandoned (red bars). In contrast, 28 % of the
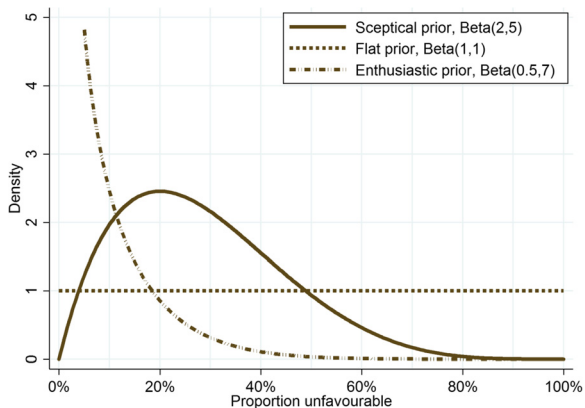

Fig. 3 Distribution functions under various prior assumptions. The flat prior assigns equal (low) probability for each unfavourable outcome proportion from 0 % to 100 %. The sceptical prior assigns a higher probability to a 20 % proportion of unfavourable outcomes (a regimen that would not be suitable for evaluation in a phase III trial), and a lower probability to very high or very low proportions of unfavourable outcomes. The enthusiastic prior assigns a high probability to proportions of unfavourable outcomes less than 10 % and lower probabilities to proportions greater than 10 %. The sceptical prior has a median at 26.4 % and the enthusiastic prior has a median at 3.3 %
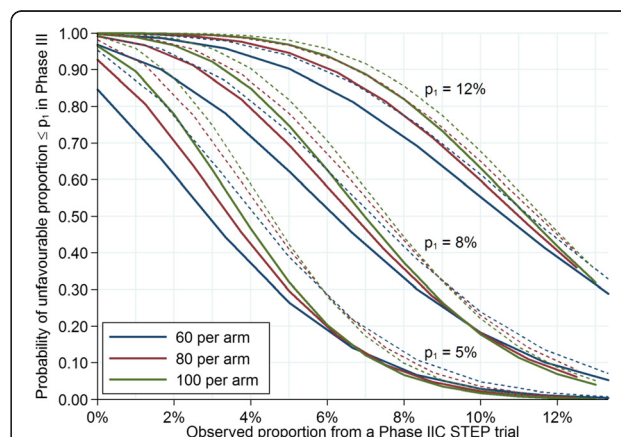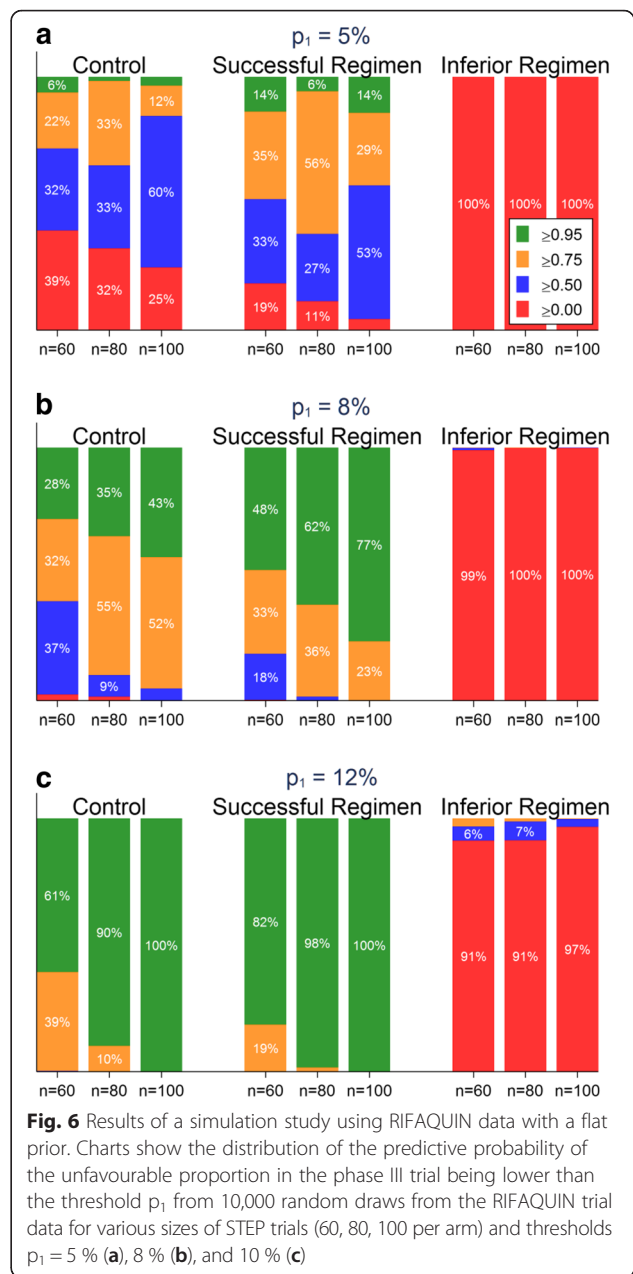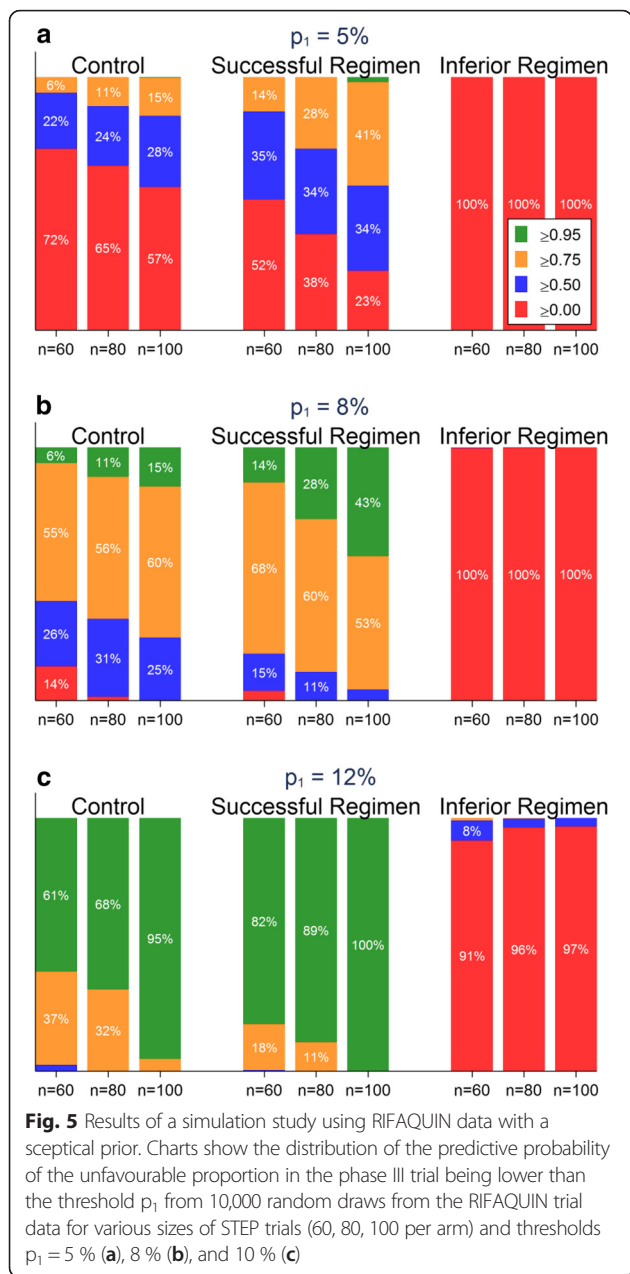

Fig. 4 Bayesian predictive probability of the unfavourable outcome being less than a specific threshold, p_1, in a future phase III trial of 500 patients per arm, following various observed proportions of unfavourable outcomes and patient numbers in a STEP trial of equal treatment duration. *Solid lines* correspond to a sceptical prior and *dashed lines* to the flat prior

Phillips *et al. BMC Medicine* (2016) 14:51

Page 7 of 11

Fig. 5 Results of a simulation study using RIFAQUIN data with a sceptical prior. Charts show the distribution of the predictive probability of the unfavourable proportion in the phase III trial being lower than the threshold p_1 from 10,000 random draws from the RIFAQUIN trial data for various sizes of STEP trials (60, 80, 100 per arm) and thresholds p_1 = 5 % (**a**), 8 % (**b**), and 10 % (**c**)

Fig. 6 Results of a simulation study using RIFAQUIN data with a flat prior. Charts show the distribution of the predictive probability of the unfavourable proportion in the phase III trial being lower than the threshold p_1 from 10,000 random draws from the RIFAQUIN trial data for various sizes of STEP trials (60, 80, 100 per arm) and thresholds p_1 = 5 % (**a**), 8 % (**b**), and 10 % (**c**)
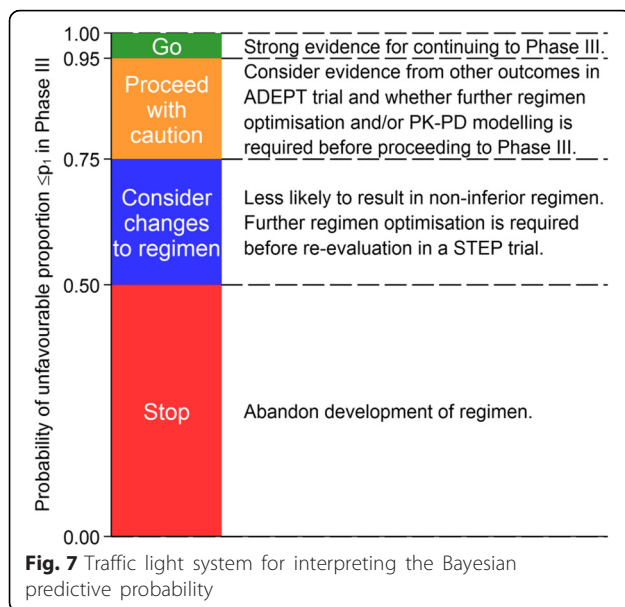
simulated trials for the Successful Regimen had a predictive probability of greater than 0.95 that the unfavourable proportion in a phase III trial would be less than p_1 = 8 % (green bars), and 60 % for the Successful Regimen had a predictive probability between 0.75 and 0.95 (amber bars). This shows a STEP trial would have correctly shown that the Inferior Regimen should not proceed to phase III (this regimen was shown to have significantly more unfavourable outcomes than the control arm in the RIFAQUIN phase III trial). It would also have had a high chance (88 %) of either showing that the Successful Regimen could proceed to phase III or that it might require further optimisation (the Successful Regimen

was shown to be non-inferior to control in the RIFAQUIN phase III trial). Using a larger STEP trial with 100 patients per arm results in a greater chance of a predictive probability of more than 0.95 (43 % of trials), as would using a flat prior instead of a sceptical prior (Fig. 6).

Discussion

We have described how a trial with the novel Phase IIC STEP design and a similar sample size to that commonly used in phase IIB trials for TB regimens could be used to provide important additional information to inform the decision on whether to proceed to evaluation in phase III. Furthermore, a simulation study using data

Phillips *et al. BMC Medicine* (2016) 14:51

Page 8 of 11

Fig. 7 Traffic light system for interpreting the Bayesian predictive probability

from the RIFAQUIN trial demonstrated that conducting a STEP trial first with 80 patients per arm would have led to the correct decision in 100 % of the simulated trials to not proceed with evaluation in a phase III trial with a regimen that was subsequently shown to be inferior, saving cost and reducing risk to participants. Use of such a design could fundamentally improve our ability to choose those regimens most likely to succeed in phase III evaluation and help us avoid costly phase III failures.

We favour a Bayesian approach to the interpretation of a STEP trial, but have also presented the frequentist framework to show how this could be applied. The Bayesian probability is easy to interpret and fits more naturally into a decision framework for deciding whether to move a regimen into phase III evaluation or whether a regimen requires further optimisation. A traffic light system for 'stop', 'go', or 'proceed with caution' might be used to improve understanding of the results. A probability of at least 0.95 for the unfavourable proportion (p_1) being less than 8 % might be considered a reasonable threshold for progressing to a phase III trial, which would be analogous to the 95 % confidence intervals and 5 % type I error rate commonly adopted in frequentist hypothesis testing.

In principle, STEP trial results could be used to support subsequent trials of a different duration than that which was studied in the STEP trial itself. A very high predictive probability might indicate a shorter regimen could be considered for phase III (a 3-month regimen if a 4-month regimen was studied in the STEP trial, for example), and a more modest predictive probability might indicate a longer regimen should be considered for phase III. While this paper focuses on the development of regimens for drug-sensitive TB, a STEP trial would

also be suitable for regimens for MDR-TB where 6-month or 9-month regimens are tested.

The Bayesian framework allows for the incorporation of prior information. We have used and would recommend a slightly sceptical prior rather than a flat, so-called non-informative prior because this more accurately reflects the likelihood of a new regimen failing in clinical development. Our chosen sceptical prior assumes that an unfavourable proportion of around 20 % is most likely, but has a wide distribution and therefore has only limited impact on the interpretation of the results (shown by the closeness of the lines assuming the sceptical and flat priors in Fig. 4). However, more sophisticated priors could be used to formally incorporate promising preclinical as well as any other earlier clinical trial data on efficacy into the decision-making process. The beta prior lends itself to this sort of approach, where the distributional parameters a and b can be interpreted as the strength of real data with a unfavourable outcomes and b favourable outcomes. For example, a relapse murine study of 15 mice with only one relapse and 14 cures could be summarised as a prior with a Beta(0.5, 7) distribution (the enthusiastic prior in Fig. 3), if considering data from a murine study as having half the strength of a human study ($a = 1/2 = 0.5$, and $b = 14/2 = 7$). In such a situation, it would be appropriate to conduct the analysis using both the sceptical and enthusiastic priors and compare the results.

The results regarding the secondary outcome (an unfavourable outcome) must be considered alongside the results of the primary outcome (killing of bacilli) when deciding how to proceed with the regimen under evaluation. If the predictive probability was greater than 0.75, a large increase in killing of bacilli with a new regimen would strengthen the decision to move to phase III whereas only a modest increase would provide evidence that further optimisation was needed. The results regarding the intermediate outcome could also be formally incorporated into the decision process by updating the prior with this information in conjunction with a prediction model that links the intermediate outcome with the long-term clinical outcome (one published model uses 2-month culture results on solid media [43]). The primary intermediate outcome data would likely be available before the long-term clinical unfavourable outcome data, so this updating of the prior could be done after that primary outcome analysis and before the long-term clinical data were available.

There are a number of factors that will determine whether a STEP trial could replace a traditional phase IIB trial in TB regimen development. It would be essential to have reasonable confidence that the drug combination could be used safely for 3 or 4 months, based on the adverse event profile and knowledge about the relationship between the duration of treatment and risk of toxicity. It

Phillips *et al. BMC Medicine* (2016) 14:51

Page 9 of 11

is also necessary to have sufficient confidence in the efficacy of a regimen and the capacity of the trial sites to follow up patients to be secure in stopping treatment after 3 or 4 months; otherwise, trial participants are put at unnecessary risk of relapse, with its resultant morbidity. Where confidence is insufficient to proceed directly to a STEP trial based on available knowledge after a smaller 14-day study, a STEP trial might still be considered after the phase IIB trial owing to the benefits in preventing costly phase III failure. Where the evidence preceding a STEP trial may not be as strong as that preceding a large phase III trial, some arms might be at unusually high risk of failure and early relapse. In such cases, careful patient monitoring (including regular cultures during treatment) and frequent and efficient oversight by an independent data monitoring committee are important features.

We believe that the richness of data from a STEP trial supports a broader philosophy in the middle development of 'iterative optimisation' where the objective of efficiently selecting from among numerous optimised combination regimens is met. As combinations are shown to be effective and safe, newer drugs can be added, doses adjusted (making fullest use of PK-PD modelling), and drugs replaced in sequential STEP evaluation (or phase IIB where necessary) until a regimen is identified that is sufficiently promising to proceed to evaluation in phase III. Indeed, competing combinations can be evaluated side-by-side within the same multi-arm trial. A MAMS design [7] would be appropriate to allow recruitment to unpromising arms to be stopped early, taking into account long-term clinical outcomes and short-term microbiological outcomes, thereby enhancing efficiency. The flexibility of the MAMS design also allows for adding new arms so that sequential evaluations could be managed within the same platform trial, an approach that has been successful in prostate cancer trials [44].

There are a number of limitations with a STEP trial. Additional follow-up visits are required after the end of treatment and these incur costs and increase the duration of the trial by 6 months, thereby delaying the start of the phase III trial. However, a visit at the end of treatment, at 12 months post-randomisation, and a third intermediate visit, with sputum taken for culture at each visit, would likely be adequate to identify treatment failure and relapse and determine 12-month cure. The simulation study using the RIFAQUIN trial data also indicated that a slightly larger sample size of 100 per arm might be preferable for a STEP trial to give more information and greater confidence in decision-making for phase III.

More work is needed to consider further extensions of the methodology and how this might operate in practice. A Bayesian adaptive trial design could also be used where the prior is continually updated as data accrue on both the primary intermediate outcome and

the long-term unfavourable outcome during the STEP trial, with rules for stopping arms early. This would allow for real-time assessment of risks and benefits to participants [42]. The Bayesian framework has been set up around the predictive probability of the unfavourable proportion being lower than a pre-specified threshold (p_1) in a future phase III trial, but a more practical alternative would be to incorporate results on the control arm from the STEP trial and calculate the unconditional probability of actually demonstrating non-inferiority in a future phase III. This is known as 'assurance', which is a Bayesian extension of the concept of the power of a clinical trial that more formally incorporates the uncertainty resulting from a phase II trial [45].

The iterative approach to regimen development, in contrast to the linear approach, may be of most value when there are multiple promising drugs and combinations to be evaluated in middle development so that the most promising regimen or regimens can be taken forward to phase III. In practice, this may be limited by the extent to which individual drugs can be evaluated in combination prior to regulatory approval and the limitations imposed by regulators and the companies which own the intellectual property. These limitations must be overcome in order to improve the treatment of TB and are being addressed in other arenas, an example being the 3P Project [46].

We have proposed a traffic light system for the predictive probability. Our intention was to set a high bar for moving regimens forward to phase III trials, given that the STEP trial design was motivated by the desire to reduce the risk of costly phase III failures. For this reason, the relatively small sample size with the STEP trial design means that there is a higher chance of a false negative result, where a truly effective regimen is not immediately advanced to phase III. Simulations from the RIFAQUIN trial showed that only 28 % of STEP trials with 80 patients per arm would have resulted in a predictive probability of greater than 0.95. Further work is needed to evaluate whether the 0.95 threshold could be relaxed; one approach would be to consider how this impacts frequentist concepts like power and type I error rate [47]. However, the Bayesian approach is particularly useful here because it quantifies the likelihood of phase III success, rather than imposing a dichotomous go/no-go decision. As we have shown, a predictive probability of less than 0.95 need not necessarily lead to a regimen being abandoned, but would suggest that the regimen may require further optimisation before proceeding to phase III. In this setting, critical evaluation of all the trial outcomes (not just the primary endpoint) and of variability in measurement of the primary endpoint would be essential, as would sophisticated PK-PD analyses (to see if higher exposures improve outcomes). A STEP trial may

Phillips *et al. BMC Medicine* (2016) 14:51

Page 10 of 11

have delayed the start of the RIFAQUIN trial, but we submit that this would have been a worthwhile risk given that it would have shown that the Inferior Regimen should not have been advanced.

Conclusions

We have proposed the Phase IIC STEP design that (1) yields rich, highly relevant information about treatment outcomes that is additional and complementary to the intermediate microbiological data typically gathered in phase II trials and (2) facilitates iterative optimisation of combination regimens in middle development. We have shown that collecting definitive clinical outcome data in a relatively small number of participants over only 12 months does provide invaluable information about the likelihood of success in a future phase III trial. For these reasons, we strongly believe that the STEP design described herein is an important tool for use in middle development for new TB regimens that would allow for more informed decision-making and accelerate regimen development.

Availability of data and materials

The data from the RIFAQUIN trial used for these analyses is available for eligible researchers as part of a repository of TB trial data. See http://c-path.org/programs/tb-pacts/ for further details.

Authors' contributions
PPJP wrote the first draft of the manuscript and conducted the statistical analyses. All authors contributed to the formulation of the STEP trial design, provided critical input into the manuscript, and read and approved the final version.

Author details
[1]MRC Clinical Trials Unit at University College London, Aviation House, 125 Kingsway, London, WC2B 6NH, UK. [2]School of Medicine, Johns Hopkins University, Baltimore, Maryland, USA. [3]School of Medicine, University of St. Andrews, St. Andrews, UK. [4]Division of Infectious Diseases and Tropical Medicine, Medical Center of the Ludwig-Maximilians-University (LMU), Munich, Germany. [5]German Center for Infection Research (DZIF), Munich Partner Site, Munich, Germany. [6]School of Medicine, Duke University, Durham, North Carolina, USA. [7]University of California San Francisco, San Francisco, California, USA. [8]Division of Physiology, Department of Medical Biochemistry, Stellenbosch University, Tygerberg, South Africa. [9]TASK Applied Science, Bellville, South Africa. [10]Department of Pharmacy, Radboud University Medical Center, Nijmegen, The Netherlands. [11]East African Health Research Commission (EAHRC), East African Community, Arusha, Tanzania. [12]Department of Lung Diseases, Radboud University Medical Center, UCCZ Dekkerswald, Nijmegen, The Netherlands.

References
1. Gillespie SH, Crook AM, McHugh TD, Mendel CM, Meredith SK, Murray SR, et al. Four-month moxifloxacin-based regimens for drug-sensitive tuberculosis. N Engl J Med. 2014;371(17):1577–87.
2. World Health Organization. Global Tuberculosis Report 2015. Geneva: WHO; 2015.
3. Boeree MJ, Diacon AH, Dawson R, Narunsky K, du Bois J, Venter A, et al. A dose-ranging trial to optimize the dose of rifampin in the treatment of tuberculosis. Am J Respir Crit Care Med. 2015;191(9):1058–65.
4. Dooley KE, Obuku EA, Durakovic N, Belitsky V, Mitnick C, Nuermberger EL. World Health Organization group 5 drugs for the treatment of drug-resistant tuberculosis: unclear efficacy or untapped potential? J Infect Dis. 2012;207(9):1352–8.
5. Dooley KE, Mitnick CD, Ann DeGroote M, Obuku E, Belitsky V, Hamilton CD, et al. Old drugs, new purpose: retooling existing drugs for optimized treatment of resistant tuberculosis. Clin Infect Dis. 2012;55(4):572–81.
6. Dawson R, Narunsky K, Carman D, Gupte N, Whitelaw A, Efron A, et al. Two-stage activity-safety study of daily rifapentine during intensive phase treatment of pulmonary tuberculosis. Int J Tuberc Lung Dis. 2015;19(7):780–6.
7. Phillips PPJ, Gillespie SH, Boeree M, Heinrich N, Aarnoutse R, McHugh T, et al. Innovative trial designs are practical solutions for improving the treatment of tuberculosis. J Infect Dis. 2012;205 Suppl 2:S250–7.
8. Jindani A, Harrison TS, Nunn AJ, Phillips PP, Churchyard GJ, Charalambous S, et al. High-dose rifapentine with moxifloxacin for pulmonary tuberculosis. N Engl J Med. 2014;371(17):1599–608.
9. Merle CS, Fielding K, Sow OB, Gninafon M, Lo MB, Mthiyane T, et al. A four-month gatifloxacin-containing regimen for treating tuberculosis. N Engl J Med. 2014;371(17):1588–98.
10. Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. Br Med J. 1948;2(4582):769–82.
11. Cochrane A. 1931-1971: A critical review, with particular reference to the medical profession. In: Teeling-Smith G, Wells N, editors. Medicines for the Year 2000. London: Office of Health Economics; 1979. p. 1–11.
12. Fox W, Ellard GA, Mitchison DA. Studies on the treatment of tuberculosis undertaken by the British Medical Research Council tuberculosis units, 1946-1986, with relevant subsequent publications. Int J Tuberc Lung Dis. 1999;3(10 Suppl 2):S231–279.
13. US FDA Priftin (Rifapentine) Drug Approval Package. [http://www.accessdata.fda.gov/drugsatfda_docs/nda/2000/21024S5_Priftin.cfm]. Accessed 21 Mar 2016.
14. Cox E, Laessig K. FDA approval of bedaquiline–the benefit-risk balance for drug-resistant tuberculosis. N Engl J Med. 2014;371(8):689–91.
15. Ahuja SD, Ashkin D, Avendano M, Banerjee R, Bauer M, Bayona JN, 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(8):e1001300.
16. Pocock SJ. Clinical trials: a practical approach. Chichester: Wiley; 1983.
17. Suntharalingam G, Perry MR, Ward S, Brett SJ, Castello-Cortes A, Brunner MD, et al. Cytokine storm in a phase 1 trial of the anti-CD28 monoclonal antibody TGN1412. N Engl J Med. 2006;355(10):1018–28.
18. Piantadosi S. Clinical trials: a methodologic perspective. 2nd ed. Hoboken: Wiley-Interscience; 2005.
19. Diacon AH, Dawson R, Hanekom M, Narunsky K, Maritz SJ, Venter A, et al. Early bactericidal activity and pharmacokinetics of PA-824 in smear-positive tuberculosis patients. Antimicrob Agents Chemother. 2010;54(8):3402–7.
20. Diacon AH, Dawson R, du Bois J, Narunsky K, Venter A, Donald PR, et al. Phase II dose-ranging trial of the early bactericidal activity of PA-824. Antimicrob Agents Chemother. 2012;56(6):3027–31.
21. Diacon AH, Dawson R, von Groote-Bidlingmaier F, Symons G, Venter A, Donald PR, et al. 14-day bactericidal activity of PA-824, bedaquiline, pyrazinamide, and moxifloxacin combinations: a randomised trial. Lancet. 2012;380(9846):986–93.
22. Diacon AH, Dawson R, von Groote-Bidlingmaier F, Symons G, Venter A, Donald PR, et al. Bactericidal activity of pyrazinamide and clofazimine alone and in combinations with pretomanid and bedaquiline. Am J Respir Crit Care Med. 2015;191(8):943–53.
23. Dawson R, Diacon AH, Everitt D, van Niekerk C, Donald PR, Burger DA, et al. Efficiency and safety of the combination of moxifloxacin, pretomanid (PA-824), and pyrazinamide during the first 8 weeks of antituberculosis treatment: a phase 2b, open-label, partly randomised trial in patients with drug-susceptible or drug-resistant pulmonary tuberculosis. Lancet. 2015;385(9979):1738–47.

Phillips *et al. BMC Medicine* (2016) 14:51

Page 11 of 11

24. Dorman SE, Goldberg S, Stout JE, Muzanyi G, Johnson JL, Weiner M, et al. Substitution of rifapentine for rifampin during intensive phase treatment of pulmonary tuberculosis: study 29 of the tuberculosis trials consortium. J Infect Dis. 2012;206(7):1030–40.

25. Mitchison DA. The action of antituberculosis drugs in short-course chemotherapy. Tubercle. 1985;66(3):219–25.

26. Prideaux B, Via LE, Zimmerman MD, Eum S, Sarathy J, O'Brien P, Chen C, Kaya F, Weiner DM, Chen PY, et al. The association between sterilizing activity and drug distribution into tuberculosis lesions. Nat Med. 2015;21(10):1223–7.

27. Wallis RS, Nacy C. Early bactericidal activity of new drug regimens for tuberculosis. Lancet. 2013;381(9861):111–2.

28. Rustomjee R, Diacon AH, Allen J, Venter A, Reddy C, Patientia RF, et al. Early bactericidal activity and pharmacokinetics of the diarylquinoline TMC207 in treatment of pulmonary tuberculosis. Antimicrob Agents Chemother. 2008;52(8):2831–5.

29. Diacon AH, Dawson R, Von Groote-Bidlingmaier F, Symons G, Venter A, Donald PR, et al. Randomized dose-ranging study of the 14-day early bactericidal activity of bedaquiline (TMC207) in patients with sputum microscopy smear-positive pulmonary tuberculosis. Antimicrob Agents Chemother. 2013;57(5):2199–203.

30. Diacon AH, Pym A, Grobusch MP, de los Rios JM, Gotuzzo E, Vasilyeva I, et al. Multidrug-resistant tuberculosis and culture conversion with bedaquiline. N Engl J Med. 2014;371(8):723–32.

31. Tang S, Yao L, Hao X, Liu Y, Zeng L, Liu G, et al. Clofazimine for the treatment of multidrug-resistant tuberculosis: prospective, multicenter, randomized controlled study in China. Clin Infect Dis. 2015;60(9):1361–7.

32. Phillips PP, Mendel CM, Burger DA, Crook A, Nunn AJ, Dawson R, et al. Limited role of culture conversion for decision-making in individual patient care and for advancing novel regimens to confirmatory clinical trials. BMC Med. 2016;14(1):19.

33. Phillips PPJ, Fielding K, Nunn AJ. An evaluation of culture results during treatment for tuberculosis as surrogate endpoints for treatment failure and relapse. PLoS One. 2013;8(5):e63840.

34. Wallis RS, Wang C, Doherty TM, Onyebujoh P, Vahedi M, Laang H, et al. Biomarkers for tuberculosis disease activity, cure, and relapse. Lancet Infect Dis. 2010;10(2):68–9.

35. Phillips PP, Davies GR, Mitchison DA. Biomarkers for tuberculosis disease activity, cure, and relapse. [Correspondence]. Lancet Infect Dis. 2010;10(2): 69–70. author reply 70-61.

36. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nat Biotechnol. 2014; 32(1):40–51.

37. Treatment Action Group, Frick M. 2015 report on tuberculosis research funding trends, 2005–2014: a decade of data. 2015. http://www. treatmentactiongroup.org/tbrd2015. Accessed 21 Mar 2016.

38. Boeree M, Hoelscher M; The PanACEA consortium. High-dose rifampin, SQ109 and moxifloxacin for treating tb: The PanACEA MAMS-TB Trial. In: 2015 Conference on retroviruses and opportunistic infection (CROI), 23–26 February 2015. Seattle, Washington: CROI Foundation; 2015.

39. Nunn AJ, Phillips PPJ, Mitchison DA. Timing of relapse in short-course chemotherapy trials for tuberculosis [Short communication]. Int J Tuberc Lung Dis. 2010;14:241–242(242).

40. Dorman SE, Savic RM, Goldberg S, Stout JE, Schluger N, Muzanyi G, et al. Daily rifapentine for treatment of pulmonary tuberculosis. A randomized, dose-ranging trial. Am J Respir Crit Care Med. 2015;191(3):333–43.

41. Machin D. Sample size tables for clinical studies. 3rd ed. Wiley-Blackwell: Chichester; 2008.

42. Berry SM. Bayesian adaptive methods for clinical trials. Boca Raton: CRC Press; 2011.

43. Wallis RS, Peppard T, Hermann D. Month 2 culture status and treatment duration as predictors of recurrence in pulmonary tuberculosis: model validation and update. PLoS One. 2015;10(4):e0125403.

44. Parker CC, Sydes MR, Mason MD, Clarke NW, Aebersold D, de Bono JS, et al. Prostate radiotherapy for men with metastatic disease: a new comparison in the STAMPEDE trial. Clin Oncol (R Coll Radiol). 2013;25(5):318–20.

45. O'Hagan A, Stevens JW, Campbell MJ. Assurance in clinical trial design. Pharm Stat. 2005;4(3):187–201.

46. PUSH, PULL, POOL: Accelerating innovation and access to medicines for tuberculosis. [http://www.msfaccess.org/3Ps-project]. Accessed 21 Mar 2016.

47. Ventz S, Trippa L. Bayesian designs and the control of frequentist characteristics: a practical solution. Biometrics. 2014;71(1):218–26.