Supplement Article

Innovative Trial Designs Are Practical Solutions for Improving the Treatment of Tuberculosis

Patrick P. J. Phillips,1 Stephen H. Gillespie,2 Martin Boeree,3 Norbert Heinrich,4 Rob Aarnoutse,3 Tim McHugh,5 Michel Pletschette,4 Christian Lienhardt,6 Richard Hafner,7 Charles Mgone,8 Alimuddin Zumla,5 Andrew J. Nunn,1 and Michael Hoelscher4,9

1Medical Research Council (MRC) Clinical Trials Unit, London, United Kingdom; 2University of St Andrews School of Medicine North Haugh, St Andrews, Scotland; 3Radboud University Nijmegen Medical Centre, the Netherlands; 4Division of Infectious Diseases and Tropical Medicine, Medical Center of the University of Munich (LMU), Germany; 5University College London Medical School, United Kingdom; 6Stop TB Partnership and Stop TB Department, World Health Organization, Geneva, Switzerland; 7Division of AIDS, National Institute of Allergy and Infectious Diseases, Bethesda, Maryland; 8European and Developing Countries Clinical Trials Partnership, The Hague, the Netherlands; and 9German Centre for Infection Research, University of Munich (LMU), Germany

A growing number of new drugs for the treatment of tuberculosis are in clinical development. Confirmatory phase 3 trials are expensive and time-consuming and the question of whether one particular drug combination can be used to treat tuberculosis is less important from a public health perspective than the question of which are the shortest, simplest, most effective, and safest regimens. While preclinical and phase 1 studies provide some guidance in the selection of combinations for clinical evaluation, a large number of combinations will require phase 2 testing to ensure that only the best regimens advance to phase 3. The multi-arm multi-stage trial design is an example of a treatment selection–adaptive design where multiple experimental arms are each simultaneously compared with a common control and interim analyses allow for poor performing arms to be dropped early. Such designs, if designed and implemented correctly, require fewer patients, can be completed in a shorter time frame, and answer more relevant questions without any loss in statistical validity or scientific integrity. There are, however, practical issues that must be considered in applying this in tuberculosis treatment trials. More innovative trials designs should be considered to speed drug and regimen development for the treatment of tuberculosis.

The emergence and rapid spread of multidrug-resistant and extensively drug-resistant tuberculosis have given an increased priority to the development and evaluation of novel drug regimens that are expected to be more effective, less toxic, and increase adherence. The tuberculosis research community has identified a number of compounds in new classes with promising in vitro and animal model data now in the clinical phase of evaluation (www.newtbdrugs.org/pipeline.php, the drug development pipeline maintained by the Working Group on New TB Drugs of the Stop TB partnership).

The current state-of-the-art in tuberculosis clinical trials to evaluate new regimens for the treatment of tuberculosis is to follow patients for at least 1 year after completion of treatment and to compare the proportion of patients assigned to each regimen with treatment failure or bacteriological relapse at the end of the follow-up period. Given the high efficacy of currently recommended regimens, it is necessary to adopt a noninferiority design [1]. Such confirmatory phase 3 trials are expensive and time consuming even for evaluating only 1 or 2 new shorter regimens. Alternative endpoints such as time to culture conversion (TCC), or nonlinear modeling of the decline in colony-forming units (CFUs) in sputum during treatment have gained currency in the research community as effective methods to determine treatment response in phase 2 trials [2–4]. However, regulatory bodies have not yet accepted these methods as valid endpoints for phase 3 trials because of the concern that they do not fully
capture the treatment effect on the true clinical endpoint (relapse-free cure), a necessary requirement for a reliable surrogate endpoint [5]. Development of a biomarker that will reduce the number of patients required and shorten the duration of studies [6] is a pressing need. With the growing number of novel drug candidates and potential dosages, evaluation of a very large number of combinations in phase 2 is necessary to have the greatest confidence that the new regimens selected for the phase 3 trial are those with the highest probability of success [7]. Although preclinical evaluation of drug combination in vitro and in animal models and phase 1 results will help to establish some priorities in selection of specific combinations of drugs for clinical evaluation, a large number of combinations will require phase 2 testing, and this number increases substantially each time a new drug is introduced for evaluation in combination therapy.

With restricted time, funds, suitable trial sites, and patient populations for enrollment into clinical trials, the traditional approach of conducting multiple phase 2 parallel-group randomized controlled trials for every potential new drug combination before moving to phase 3 is a critical bottleneck for tuberculosis drug combination regimen development. In addition, in the context of the global tuberculosis epidemic, the question of whether one particular drug regimen can be used to treat tuberculosis is less important from a public health perspective than the question of which are the shortest, simplest, most effective, and safest regimens.

THE MULTI-ARM MULTI-STAGE DESIGN

Group sequential trials that have been used widely in other disease areas for more than 30 years [8] and requiring fewer patients can be completed in a shorter time frame and, if designed and implemented correctly, answer the questions without any loss in statistical validity or scientific integrity. With this accelerated program, it is important to distinguish between “false positive” and “false negative” outcomes of phase 2 trials, and to consider the relative “costs” of each. A false-negative outcome corresponds to a regimen that would be truly effective but shows no benefit in the intermediate outcome analysis and is not taken forward to phase 3. In the absence of other similarly effective regimens, a false-negative outcome is of a high cost to the global tuberculosis community as it is unlikely that the regimen will ever be evaluated. A false-positive outcome corresponds to a regimen that shows benefit on the intermediate outcome in phase 2, is taken forward to phase 3, but is shown to be inferior to the control. The false-positive result wastes time and resources for those conducting the futile phase 3 trial and will delay the evaluation of another possibly effective regimen in a confirmatory phase 3 trial. How then should these risks be balanced? In the context of tuberculosis treatment, and the limited funds and patient populations available for phase 3 trials, the cost of a false negative might be considered to be small compared with the cost of a false positive, in light of the large number of new potential treatments in clinical development. This setting is ideal for trials with so-called treatment selection or screening-adaptive designs comparing several new treatments to a common control [9, 10]. Planned repeated interim analyses that are prespecified in the protocol allow for poorly performing regimens to be dropped early and resources switched to evaluation of other potentially effective regimens. At each interim analysis, the independent data monitoring committee (IDMC) reviews the data and makes recommendations to the trial steering committee about which arms should continue or should stop, so that new patients recruited to the trial are no longer randomized to what is likely to prove to be an ineffective arm. The trial continues until there are sufficient data for the best regimen(s) to be identified and taken forward to phase 3. An illustration of a possible 3-stage design is shown in Figure 1.

Such a design is efficient and enhances patient safety as a smaller number of patients are randomized to ineffective regimens before they are dropped. There is a moderate risk that any one effective regimen may be dropped early in the trial purely by chance (a false negative), but the operating characteristics can be set up so that there is a high probability of identifying at least one truly effective regimen.

A suitable application of such a design is the multi-arm multi-stage (MAMS) design [9, 11]. Although care is needed to calculate the operating characteristics of the design, it is straightforward to implement once the critical elements are agreed. A major benefit is that standard statistical techniques can be used to compare the control with those regimens that are not eliminated at interim analyses without any need for

---

Figure 1. Example of a 5-arm phase 2 trial with a 3-stage multi-arm multi-stage design. At the first interim analysis, novel regimen 4 is considered to lack sufficient benefit compared with the control and is not taken forward to stage 2. At the second interim analysis, recruitment to novel regimens 1 and 3 are stopped, and only the control regimen and novel regimen 2 is continued to the end of trial. All patients are followed up for time to culture conversion for the same duration (usually 8 or 10 weeks).
complex correction for bias. This is in contrast to a "pick-the-winner" approach [12] where only the best-performing arm (or arms) is selected at the interim analysis to be taken forward. The process of selecting the arm with the maximum effect at the interim analysis in this design introduces bias because the treatment estimate of this effect is artificially inflated. This bias increases with the number of treatment arms in the study, and a correction must be used to yield unbiased estimates. Because the MAMS design involves the dropping of poorly performing arms rather than the selecting of the best arms, the estimates in the arms that are continued remain unbiased. The estimates in arms that are dropped are not unbiased. There is work currently under way at the MRC Clinical Trials Unit to formally describe the sources of bias in the MAMS design. The MAMS design has been developed largely for a time-to-event endpoint suitable for the tuberculosis context where time to culture conversion currently is probably the most appropriate phase 2 endpoint. A recent implementation of the MAMS design in prostate cancer is the currently ongoing phase 2/3 STAMPEDE trial [11] (see www.stampedetrial.org for the trial protocol and recent trial updates). At the second meeting of the IDMC, recruitment to 2 of the 5 experimental arms was stopped due to lack of sufficient benefit.

The benefit of the MAMS design in quickly identifying the most effective regimen(s) for taking forward for phase 3 evaluation must be considered alongside a number of practical issues specific for tuberculosis (discussed below). Table 1 summarizes some of the differences between the Multi-Arm Trial With Static Design and the MAMS trial design.

### SELECTING THE MINIMUM TREATMENT EFFECT

An important parameter required for any sample size calculation is the minimum treatment effect that could be detected by the clinical trial. A trial designed to detect a small treatment effect will be very large, but a smaller trial runs the risk of not being able to detect a smaller treatment effect that may still be of clinical benefit. This is an important issue for any trial and is therefore not limited to those with a MAMS design. This problem is particularly acute in phase 2 trials for new tuberculosis treatment regimens because the relationship between time to culture conversion and long-term relapse-free cure is not well understood.

Consideration of the fourth-generation fluoroquinolones gatifloxacin and moxifloxacin as treatment-shortening agents being evaluated in current phase 3 trials (OFLOTUB and REMoxTB) is informative to determine what evidence from phase 2 trials led the investigators to initiate and the funders to resource these phase 3 trials. Of the 4 randomized phase 2 trials evaluating regimens that included moxifloxacin, 3 included time to culture conversion as a secondary endpoint [3, 13], but only 1 reported a hazard ratio for the difference with control (the other 2 reporting only P values from nonparametric tests) [2]. In this trial, the replacement of ethambutol with either gatifloxacin or moxifloxacin resulted in an increase in the hazard of culture conversion by a ratio of 1.5 or 1.7, respectively, in the adjusted analyses. This was considered to be sufficient evidence (alongside other clinical and preclinical data) for evaluating these 4-month regimens

### Table 1. Comparison of a Multi-Arm Trial With a Static Design With a Trial With a Multi-Arm Multi-Stage (MAMS) Design

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Traditional Trial With Static Design</th>
<th>MAMS Trial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td>Fixed and known before the start of the trial</td>
<td>Not fixed and therefore not known before the start of the trial</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The maximum sample size is known and is likely to be slightly higher, but the expected sample size will be considerably lower than for the static design</td>
</tr>
<tr>
<td>Accruing trial data</td>
<td>Not used. Restricted until after database lock at trial completion</td>
<td>Planned interim analyses for early dropping of poorly performing arms based on safety or efficacy comparisons</td>
</tr>
<tr>
<td>Randomization</td>
<td>Fixed and known before the start of the trial</td>
<td>Initial randomization list known before start of trial but may be changed during the trial as arms are dropped</td>
</tr>
<tr>
<td>Timelines of data entry and data cleaning</td>
<td>Must be completed by database lock at completion</td>
<td>Must be completed in time for each interim analysis</td>
</tr>
<tr>
<td>Overall type I error (the probability of a false positive)</td>
<td>Fixed at the prespecified α</td>
<td>Slightly less than α if number of stages is large</td>
</tr>
<tr>
<td>Overall power</td>
<td>Fixed at the prespecified power of 1-β</td>
<td>Slightly reduced below 1-β if number of stages is large. Sample size can be increased to account for this.</td>
</tr>
<tr>
<td>Final data analysis</td>
<td>Standard techniques</td>
<td>Standard techniques for arms that are not dropped at interim analyses</td>
</tr>
</tbody>
</table>
in two phase 2 trials (OFLOTUB and REMoxTB). To select regimens for testing that will need to be shorter or more effective than this 4-month regimen, whether or not it is found to be noninferior to the standard control, it will be necessary to find a larger hazard ratio, say, 1.8 or 1.9. The relationship between a hazard ratio for time to culture conversion from a phase 2 trial with a difference in rates of relapse-free cure from a phase 3 trial is not yet well understood. Because of this and the large number of possible novel combination regimens, the minimum treatment effect could further be increased to 2.0 to reduce the chance of taking a regimen forward to phase 3 (a false positive) that will not significantly shorten treatment beyond what might be hoped for instance with the REMoxTB trial. A larger minimum treatment effect will also mean a smaller trial. Figure 2 shows the total trial duration under various scenarios. This is an example of one way to approach the important decision of the choice of hazard ratio and there may be other strategies to balance the risks of false positive and false negatives. Ideally, if the pool of potential new regimens was inexhaustible, the hazard ratio would be set at an ambitious figure so that only a truly effective shorter regimen would emerge after the investment of time, money, and patient effort.

A hazard ratio of 1.7 means that the (instantaneous) probability of culture converting on the experimental arm at any point during the first 2 months of treatment is 1.7 times that on the control arm. The median time to culture conversion on the experimental arm is reduced by a factor of 1/1.7 as compared with the control arm. If the proportion still not cultured converted at 8 weeks on the control arm is 0.35, then the proportion still not cultured converted at 8 weeks on the experimental arm will be $0.35^{(1.7)} = 0.17$. Similarly, a hazard ratio of 2.0 would be interpreted as a reduction in the proportion not culture converted by 8 weeks from 0.35 to $0.35^{(2.0)} = 0.12$.

**THE LACK OF USEFUL SURROGATE ENDPOINTS**

Using time to culture conversion as an endpoint makes defining the minimum treatment effect difficult but is also problematic in other areas. To benefit most from the MAMS design, it is necessary for efficient links in the data flow from a participant attending a study visit to the report being reviewed by the IDMC tasked with recommending which arms should continue and which should be stopped. These links include, but are not limited to, transport from clinics to...
laboratories, sample processing and reporting of results, data entry, data cleaning, and statistical analysis, not to mention the possible delays in coordinating a date for the IDMC meeting where each of the expert members are available to attend. Every week that the IDMC meeting is delayed due to delays in data entry, for example, reduces the benefit of the adaptive design because some patients will be random- ized to a regimen that may subsequently be stopped the following week by the IDMC. Study processes must be carefully thought through before the trial starts, and features such as Web-based remote-data entry, “digital pen” technol- ogy, real-time data cleaning and management, and pre-written statistical programs would all be considered crucial to this end.

A crucial bottleneck in this process is the slow growth of *Mycobacterium tuberculosis* and the time necessary to de- termine that a patient is truly culture negative. Although a negative culture is identified sooner in the MGIT system than on solid media, a wait of 42 days is still required. Thus, it will always be no less than 6 weeks between the last parti- cipant visit necessary for the interim analysis and the last piece of data entered even before data cleaning and analysis. Better—more rapidly available and quantitative—biomarkers are needed that give results in days or hours rather than weeks, and with better understood relationships with long- term relapse-free cure. This latter criterion can only be satis- fied in the context of phase 3 trials.

Improving the speed of enumeration of mycobacteria could have a significant effect on our ability to deliver the MAMS trials in a timely way. Multiple molecular methods to detect and quantify mycobacteria nucleic acids have been reported, and their use for monitoring treatment has been evaluated. Because previous studies have demonstrated that DNA can persist well beyond the time points that cultures are positive, such assays are unlikely to assist in defining culture negativity accurately. In contrast, messenger RNA (mRNA) has a much shorter half life [14] but is present in significantly smaller concentrations and assays become negative rapidly. More abundant mRNA species might, however, be suitable and studies have been performed. These studies have used *M. tuberculosis* fbpB (encoding fibronectin- binding protein, antigen85B) and *hspX* (encoding alpha- crystalline homologue protein) icl isocitrate lyase mRNA showing that they decline rapidly in conjunction with *M. tuberculosis* colony counts after initiation of a rifampin- based standard drug therapy [14–17]. Isocitrate lyase mRNA and *rrnA*-P1 (noncoding ribosomal promoter region) cor- related with CFUs in sputum prior to therapy and during 7 days of monotherapy and was detectable in sputum culture-positive tuberculosis patients for 1 month. Concentrations of *icl* mRNA were detectable following 1 and 2 months of a rifampicin-based drug regimen.

Ribosomal RNA (rRNA) has a longer half-life than mRNA and is present in greater abundance. Some publications have suggested that the longer half-life might interfere with its value as a measure of treatment response [18]. A recent study demonstrated that, after further development of this assay and the introduction of a robust internal extraction control, rRNA measurement may predict culture status down to 100 CFU and the data correlate well with 56-day culture positivity. More importantly, the mean decline in 16rRNA during the first 3 days of standard tuberculosis therapy is 0.99 log_{10} and this is comparable to similar data from solid agar [2]. Data generated using this assay mirrors the results of culture, and results are available in 24 hours [15]. Another method to be evaluated for correlation with clinical treatment outcomes might be imaging with positron emission tomography/computed tomography scanning, which gives results in nearly real time and may show significant changes in the reduction of pulmonary infiltrates within 4 weeks of treatment [19]. Imaging has the added advantage of not depending on sputum production or collection. Dis- advantages of this methodology include high expense, spe- cialized equipment and expertise, and radiation exposure. However, if it proves to be a useful surrogate for treatment response, imaging may be a valuable tool for phase 2 tu- berculosis drug combination development and to facilitate real-time adaptive trial decision making in the future.

**THE PARADOXICAL BENEFITS OF A SLOW RECRUITMENT RATE**

An important drawback of a MAMS trial is that the total duration and the required number of patients is not fixed in advance. All experimental arms could be dropped at the first interim analysis at which point the trial would stop with the conclusion that no arms are effective, but all experimental arms could also be retained at every interim analysis with the conclusion that all arms are effective. It is, however, possible to calculate the maximum and minimum trial duration and the expected duration based on the expected number of truly effective regimens. The trial duration depends on a number of parameters described in Table 2. An important parameter, the recruitment rate, is often very difficult to estimate before the start of the trial and must be considered carefully. Slow recruitment is a major problem in clinical trials in all disease areas. Even when only sites in high endemic areas are in- cluded, recruitment can be unexpectedly low. Based on a recent Cochrane review [21], it is likely that <50% of clinical trials meet their target even with an extension to recruitment. However, the benefits of the MAMS design are increased relative to the overall sample size when re- cruitment is slow and the design can actually mitigate some of the consequences of slow recruitment. The timing of the
interim analyses and meetings of the IDMC will be defined based on recruitment rates. If 10 weeks are necessary, for example, between the last participant visit contributing data toward an interim analysis and the meeting of the IDMC, a slower recruitment rate will mean that fewer patients are randomized during these 10 weeks to a regimen subsequently recommended to be stopped by the IDMC. Figure 2 shows the total trial duration under various scenarios. If, however, rapid molecular or other response evaluation methods are used, this effect is reduced.

In addition, slower than anticipated recruitment prolongs the duration of the trial allowing for more interim analyses and therefore more opportunities to stop recruitment to poorly performing regimens. Suppose that, in planning a trial with a MAMS design, only 2 interim analyses were considered possible based on the chosen design and the resulting expected trial duration. It would be possible to plan for third interim analysis in advance as a precaution against slow recruitment. If the trial progressed as planned, this would not be needed; but if recruitment was slower than expected, the third interim analysis would become feasible and would allow for another opportunity to stop recruitment to poorly performing arms, thereby reducing the final overall sample size (as compared with not having another interim review), despite slow recruitment.

**USING ALL DATA FROM PARTICIPANTS HAVING NOT COMPLETED FOLLOW-UP**

In a conventional phase 2 trial, the analysis of time to culture conversion would begin after all patients had a culture result at a predefined time point (usually 8 or 10 weeks from randomization). In a MAMS trial, such a procedure would result in additional delay of the interim analyses. After the enrollment of the last patient necessary for the interim analysis, it would be necessary to wait for 8 weeks for follow-up and then 6 weeks for the culture of the last sample to grow even before data entry and data cleaning. It is therefore proposed to define a minimum follow-up period that patients must have reached before they are included in the interim analyses, whereby advancing the interim analyses by several weeks. Such a window should take into account that certain tuberculosis drugs might have a pronounced killing effect in the first few days—for example, INH (isonicotinylhydrazine) that “levels out” after 2–3 weeks. Patients should therefore have completed at least 3 weeks of treatment to be included in the time to culture conversion analysis.

**RECRUITMENT STRATEGIES TO IMPROVE EFFICIENCY**

The total duration of the STAMPEDE trial is estimated to be about 7 years including recruitment and follow-up [11]. Although this includes a final phase 3 stage in a seamless phase 2/3 design, it is substantially longer than a phase 2 trial in tuberculosis that is commonly expected to last 1–2 years. One reason for this is the difference in rates of event-free “survival” on the control arms: 50% failure-free survival after 2 years in STAMPEDE and 50%–60% still culture positive on MGIT after only 8 weeks. The relative benefits of the MAMS design in a phase 2 tuberculosis treatment trial are less than those in STAMPEDE, and it is therefore unlikely

---

**Table 2. Design Parameters Required for the Sample Size Calculations**

<table>
<thead>
<tr>
<th>Design Parameter</th>
<th>Impact of Parameter on Overall Trial Duration and Strategies for Choosing a Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of interim analyses/stages</td>
<td>A large number of interim analyses allows for more opportunities to drop arms and therefore results in a shorter expected trial duration. The number of interim analyses is limited by the overall trial duration.</td>
</tr>
<tr>
<td>Stagewise power</td>
<td>The probability that a truly effective arm will not be dropped at an interim analysis. This is kept high (.95) to reduce as much as possible the rate of false negatives</td>
</tr>
<tr>
<td>Stagewise level of significance</td>
<td>This is the probability of not dropping an ineffective regimen at any interim analysis. It is recommended that this is set to .5 at the first interim which is high, but allows for an early look to drop grossly inferior regimens. This is subsequently set to .5 for the ith interim analysis. This means that this will be .25 at the second interim analysis and .125 at the third, such that probability of continuing with an ineffective regimen gets smaller and smaller after each interim.</td>
</tr>
<tr>
<td>Distribution of time to culture conversion in the control group</td>
<td>The nature of the distribution and the parameters must be determined from data from previous studies on patients on the control arm. Experience from trials suggests that a Weibull rather than exponential distribution is more appropriate and a published study on multidrug-resistant tuberculosis concluded that this distribution could be generalized γ or log-normal distribution [20].</td>
</tr>
<tr>
<td>Treatment effect that is being looked for</td>
<td>For time to culture conversion assuming proportional hazards, this will be the hazard ratio. Given the large number of potential new treatment regimens, it would be sensible to look for a large treatment effect, which will reduce the overall expected trial duration.</td>
</tr>
<tr>
<td>Number of experimental treatment regimens</td>
<td>This will in part be dictated by the number of new compounds available to be included in new regimens, and expert opinion of any possible drug interaction in terms of efficacy and safety.</td>
</tr>
</tbody>
</table>
that more than 1–2 interim analyses will be possible in a phase 2 tuberculosis trial unless there are a very large number of novel experimental regimens (more than 6–7).

One way to improve efficiency would be to stop recruitment to the study between the final patient visit contributing data for an interim analysis and the IDMC meeting where the results are reviewed. This would mean that no participants were unnecessarily randomized to an arm that was subsequently stopped. Repeatedly stopping and starting recruitment at a site delays the overall trial duration and could cause a number of problems. This strategy is therefore not recommended.

Another strategy is to alternate recruitment between experimental arms. Under this strategy, participants would be randomized to one of half of the experimental arms or the control arm until enough patients have completed follow-up for the first interim analysis. The randomization is then switched so that new patients are randomized to one of the other half of the experimental arms or the control arm until enough patients have completed follow-up for the first interim analysis for these arms. At this point the IDMC reviewing data from the first set of experimental arms will have met and made recommendations about which arms to continue and which to stop, and the randomization is then switched back so that new patients are randomized to one of these arms that are continued or the control. The randomization switches back and forth in this way for each interim analysis until enough patients are recruited in those arms that are continued. In this approach, no patients are unnecessarily randomized to arms that are subsequently stopped, and there is no interruption in recruitment at sites. The control arm is always included as one of the options in randomization, and so the total number of patients randomized to the control arm is double, resulting in a slightly larger trial. There are also clear logistic complexities in implementing such a design. In reality, a large number of experimental arms (probably >8) are required for this strategy to be feasible and convey any meaningful benefits. Other novel recruitment strategies should, however, be considered to optimize the design.

**ADDING IN NEW EXPERIMENTAL ARMS**

It is very unlikely that all new compounds that might be included in a phase 2 treatment selection trial would be available at the same time. The MAMS design does allow for the possibility of including new experimental treatment regimens during the course of the trial. This would be particularly relevant if multiple arms are stopped early, meaning that spare capacity can be used to include new arms. The analysis of a combination regimen included midway through the trial should only include data from patients on the control regimen randomized concurrently, not from patients randomized before the introduction of the new arm. The criteria used to determine whether this regimen should stop or continue at the next interim analysis will likely be different to the rest of the regimens because fewer data have accrued. This clearly adds more layers of complexity. A new arm was added to the STAMPEDE trial in November 2011, several years after the first patient was randomized.

**SEAMLESS PHASE 2/3 TRIALS**

This article has focused on the use of the MAMS adaptive design for a phase 2 trial for selecting the most promising regimens to take forward to a phase 3 trial. Rather than the traditional stricter stopping guidelines (such as the Haybittle-Peto or Pocock boundaries) that require overwhelming evidence of lack of benefit before an arm is stopped [22], more liberal stopping rules are used to allow for poorly performing arms to be stopped early so that the superior treatment arms can be evaluated efficiently in a multi-arm study. The MAMS design can also be utilized in a seamless phase 2/3 trial, as has been done in the STAMPEDE trial. At each interim analysis, decisions about which arms to continue are based on an intermediate endpoint (as described above). However, after the final interim analysis, additional patients are randomized to the remaining treatment arms that have not been dropped until the required total is reached and all patients are followed up for the definitive endpoint of survival on which the final analysis is based. The trial combines the treatment selection based on an intermediate endpoint (phase 2) with the confirmatory randomized comparison based on the definitive endpoint (phase 3) in a seamless design [11]. A similar design could be considered for a seamless phase 2/3 tuberculosis trial, where early decisions about which arms to drop are made on the basis of time to culture conversion, but patients on the treatment arms that are not dropped remain in long-term follow-up for relapse to confirm efficacy. This would be an efficient way to evaluate new regimens but would require careful planning and preparation and cofunding from multiple partners for such a large undertaking.

**CONCLUSIONS**

In the next few years at least 4 new drug classes will need evaluation in combination with each other and in combination with the standard drugs to define the best possible treatment. When evaluated in different dosages, the number of different studies required will result in a major bottleneck. There is a real risk that the decisions about which combinations to evaluate in expensive phase 3 randomized controlled trials will not be based on rigorous scientific evidence but either on convenience due to commercial interests, existing...
research collaborations, or worse, chance. Even though an enormous effort will be required, we advocate for a rational selection of potentially effective treatment combinations that is inclusive rather than exclusive. The value of early bactericidal activity (EBA) or extended EBA has been extensively and critically discussed recently [23], and animal models can, at best, only give an indication of which drugs could form highly effective combinations. The MAMS design offers a real practical solution to overcome some of the financial and logistical problems that emerge when multiple drug regimens would be evaluated simultaneously; these problems are often written about, but very few solutions are offered. The European and Developing Countries Clinical Trials Partnership (EDCTP)–funded Pan African Consortium for the Evaluation of Antituberculosis Antibiotics (PanACEA) is proposing a MAMS study design for a phase 2 randomized controlled trial to identify a promising regimen to take forward to phase 3 from different combinations of high-dose rifampicin, moxifloxacin, and SQ109. This will give valuable experience with such an adaptive design in a trial with a relatively small number of treatment combinations with a view to conducting larger trials in the future harnessing the full potential of this design. MAMS designs should be also specifically considered for multidrug-resistant tuberculosis trials, where patient recruitment is slow and multiple drug combinations will have to be considered.

The MAMS design will not always be the most appropriate design for phase 2 trials for the treatment of tuberculosis. Other adaptive designs that convey different benefits may be more suitable in other situations. We therefore encourage other trialists and researchers to consider this and other innovative adaptive trial designs to identify optimal regimens that are short, safe, and efficacious and can be used to treat patients rapidly.

Notes

Financial support. Part of this work has been funded by the EDCTP to the Pan African Consortium for the Evaluation of Antituberculosis Antibiotics (EDCTP grant number IP 2007 32011 013).

Potential conflicts of interest. All authors: No reported conflicts.

All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References


Innovative Designs for Tuberculosis Treatment Trials • JID 2012:205 (Suppl 2) • S257