Published in final edited form as:

Int J Tuberc Lung Dis. 2018 May 01; 22(5): 34–39. doi:10.5588/ijtld.17.0358.

Statistical considerations for pediatric multidrug-resistant tuberculosis efficacy trials

Soyeon Kim¹, James A. Seddon², Anthony J. Garcia-Prats³, and Grace Montepiedra¹

¹Frontier Science Foundation, Brookline, Massachusetts, USA

²Centre for International Child Health, Department of Paediatrics, Imperial College London, London, United Kingdom

³Desmond Tutu TB Centre, Department of Paediatrics and Child Health, Faculty of Medicine and Health Sciences, Stellenbosch University, South Africa.

Summary

Inclusion of newly licensed or repurposed drugs in regimens to treat children for multidrug-resistant-tuberculosis may lead to therapy that is shorter than traditional regimens and composed only of oral medications. As an all-oral regimen may be more acceptable and have a better safety profile than current regimens, demonstrating non-inferiority may be satisfactory. Demonstrating non-inferior efficacy requires setting a non-inferiority margin and safeguarding study assay sensitivity. Multi-arm multi-stage designs may currently not be appropriate in pediatric trials because of the lack of sensitive and specific intermediate outcomes. However, including an arm with an agent to ameliorate toxicity would be efficient. Covariates can be used to stratify randomization, define subgroups, and improve efficiency of analysis. Enriching the sample for the confirmed-TB subgroup to ensure that they are well represented may be important. Primary outcomes using a fixed time point from randomization for all study arms will result in variation in post-treatment duration but may be the best choice. While blinding of site personnel and patients may not be possible when regimens differ substantially in drugs and modes of administration, blinding should be maintained for independent endpoint review groups and other personnel. Type I error and family-wise error rates should be tightly controlled.

Keywords

MDR-TB; tuberculosis; children; resistant; clinical trials design

Introduction

Newly licensed or repurposed drugs are promising candidates for inclusion in all-oral, shortened duration regimens for treating multi-drug resistant (MDR)-tuberculosis (TB). Traditionally, efficacy of anti-tuberculosis drugs has been extrapolated for use in children

from adult studies, and pediatric efficacy trials have not been pursued. The priority in children has been trials of drug safety and pharmacokinetics, which permits an evaluation of the dosage required to achieve similar drug exposure as compared to adults, and age-specific toxicity.^{1,2} However, as the spectrum of TB disease in children differs from that in adults, with children tending to have less severe TB, extrapolating efficacy from adult studies may result in the dismissal of regimens that could benefit children. Use of the second-line injectable anti-tuberculosis drugs (amikacin, kanamycin and capreomycin, hereafter injectables), key components of existing regimens for MDR-TB, result in ototoxicity in greater than 20% of children.³ As children may be more affected by hearing loss if it occurs at a crucial time in their neurodevelopment, the risk-benefit profiles of new regimens may differ, as compared to adults, with exclusion of injectable drugs being a priority for pediatric MDR-TB treatment.

TB in children can have differing clinical characteristics and pathophysiology from adults.⁴ The spectrum of TB disease in children is wide, but is frequently sputum smear-negative (80–90%), culture-negative (50–60%), and paucibacillary, as opposed to the sputum smear-positive, culture-confirmed, cavitary disease with high bacillary load and necrotic lesions typical of adults with TB. These factors are correlated, co-vary with age, and may predict treatment outcome. Drug susceptibility testing (DST) profile will also likely predict treatment outcome. Effective treatments are needed for all presentations of TB disease and DST pattern.

Confirming the diagnosis of TB in children is often challenging, as disease is frequently paucibacillary, compounded by the difficulty of obtaining respiratory samples. Ultimately, the majority of children treated for TB lack bacteriological confirmation and are treated based on clinical signs or symptoms, radiography consistent with TB, and exposure to a case of MDR-TB. As mycobacterial culture is more sensitive than the rapid genotypic tests for confirming the diagnosis in children⁵, when disease is confirmed, the results and DST may be delayed. As the initiation of MDR-TB treatment as early as possible is expected to achieve optimal outcomes³, treatment is typically started while awaiting culture and DST results and modified later, if necessary. Clinically-diagnosed TB (i.e., not bacteriologically-confirmed) precludes the use of culture-conversion as a treatment outcome. As an intermediate outcome, early clinical improvements may have low sensitivity for treatment outcome.

These features of pediatric TB suggest that a pragmatic attitude is required in designing a trial to evaluate a shortened, all-oral treatment for children with MDR-TB. A pragmatic approach is defined by Schwartz and Lellouch⁶ to distinguish trials that evaluate treatments under usual conditions vs explanatory trials that evaluate treatments under optimal conditions. In this paper, we discuss statistical considerations for a pragmatic approach to designing high quality trials to evaluate efficacy of MDR-TB treatments for pediatric TB disease.

SUPERIORITY VS NON-INFERIORITY OBJECTIVES

In the setting of MDR-TB disease, use of an active control group is incontrovertible. If a trial's objective is to show that experimental treatment is better than the control, a superiority study is ideal. When well designed and conducted, a superiority trial will have internal validity, and, if a statistically significant difference is found, can be used to declare an improvement over the control. When treatment outcomes are expected to be good, as for children with MDR-TB⁷, the anticipated superiority of the experimental regimen over the control may be small and would require a large sample size for high power. For example, when 15% of the control arm and 12.5% (2.5% lower) of the experimental arm have unfavorable treatment outcomes, 3983 evaluable participants per arm are required for 90% power using a two-sided 5% significance level test. A sample size of this magnitude may be prohibitive.

With the currently recommended treatment regimens for MDR-TB, which are long (often 18 months or longer) and require daily painful intramuscular injections, resulting in hearing loss in a substantial proportion of children, an all-oral treatment regimen with small improvements in or slightly inferior efficacy but of shorter duration with less toxicity and improved acceptability may be welcome. The public health impact with a shortened, simpler regimen may also be considerable.

If the experimental regimen is expected to have the same or worse efficacy than the control, a superiority trial cannot be designed with acceptable Type I and II errors. However, a non-inferiority trial can identify regimens that are expected to be better or not unacceptably worse than the control, and, because of the potential improvements in risk-benefit profile, may be useful in evaluating a shortened, all-oral MDR-TB regimen in children.

A study with a non-inferiority objective requires attention to several crucial methodological details. A non-inferiority margin, δ , must be set that defines the difference in the experimental and control arms for which the experimental arm would not be considered unacceptably worse. International Conference on Harmonisation (ICH) Guidelines⁸ suggests that establishing δ must be based on "historical data and relevant clinical and statistical considerations," and should preserve an acceptably conservative "fraction of the control drug effect". Although clinical consideration for δ is important, it is typically overlooked.⁹ It may be appropriate to use a wider δ if the experimental arm is "associated with fewer serious adverse effects or better tolerability," or has other advantages.¹⁰ Hence it may be more appropriate to use a wider δ for pediatric trials testing an injectable-sparing regimen than for adult trials for all its advantages. In addition, the trial must show assay sensitivity, which is "the ability to distinguish an effective treatment from a less effective or ineffective treatment." "[I]f a trial intended to demonstrate efficacy by showing a test treatment to be non-inferior to an active control, but lacks assay sensitivity, the trial may find an ineffective treatment to be non-inferior and could lead to an erroneous conclusion of efficacy." Assay sensitivity requires that "similarly designed trials in the past regularly distinguished effective treatment from less effective or ineffective treatments."⁸ Similarity of trial design ensures that the constancy assumption holds. High quality study conduct prevents erosion of assay sensitivity.⁸

Lack of prior placebo-controlled MDR-TB trials with demonstrated assay sensitivity is a challenge. While the above ICH criteria are ideal, not all non-inferiority trials may require an antecedent trial that demonstrates efficacy of the active-control in a similar trial. Indeed that requirement would preclude the conduct of non-inferiority studies of treatment for MDR-TB, because no such trials have been conducted or would be ethical to conduct. Observational studies, primarily in adults, using historical or non-concurrent controls have demonstrated the effectiveness of a 9–12 month regimen for treating MDR-TB. On this basis, World Health Organization (WHO) has provided a conditional recommendation for its use, with very low certainty in the evidence.³ In MDR-TB, a non-inferiority trial with a reasonably sized α may be able to demonstrate non-inferiority of an experimental treatment provided that treatment success in the control arm is high and the study is well conducted. In fact, the STREAM trial evaluating a shortened regimen for MDR-TB in adults using a non-inferiority design with α of 10% began enrolling in 2012 with Stage 1 results anticipated in early 2018.

The trial sample size will depend on the true difference between arms and δ . Figure 1 shows the effect of the non-inferiority margin on the required evaluable sample size assuming that 15% of control participants have an unfavorable outcome (e.g. absence of treatment failure or relapse) for 90% power using a one-sided 2.5% significance level test. Sample sizes are higher when the experimental regimen is expected to perform worse or if δ is smaller. In the scenario when the experimental regimen is 2.5% better than the control, a δ of 10% requires a sample size of 160 evaluable participants per arm, while a δ of 5% would require 443 per arm. Both would be substantially smaller than for a superiority trial with a two-sided 5% significance level test (3983 per arm).

When safety improvement is the primary motivation for a proposed experimental regimen, such as an all-oral shortened regimen, a non-inferiority efficacy comparison and a superiority test for safety could be co-primary objectives. Sample size calculations could be performed for each objective and the trial would enroll the larger sample size to ensure adequate power for both objectives. The actual Type I error of the study would be conservative since both null hypotheses would need to be rejected to declare the experimental regimen better.

Non-inferiority studies should be analyzed in both intention-to-treat (ITT) and per protocol (PP) populations. If a modified-ITT (mITT) population is defined, we recommend that exclusions be based on information, or samples obtained, antecedent to randomization. For example, the mITT population might exclude individuals with DST indicating resistance to second-line drugs based on samples taken just prior to randomization. D'Agostino et al⁹ provide an excellent discussion of many important issues for non-inferiority trials.

Montepiedra and colleagues¹¹ discussed use of the Desirability of Outcome Ranking (DOOR) approach¹² that combines both efficacy, safety, and potentially other factors in a composite outcome that may allow the use of a superiority design to evaluate MDR-TB treatments in children aimed at reducing adverse events without compromising efficacy. Cautionary advice regarding the DOOR approach is offered by Phillips et al,¹³ and we agree

that additional research evaluating the performance of DOOR outcomes in the context of TB trials is needed.

MULTI-ARM MULTI-STAGE ADAPTIVE DESIGNS

Sequential pairwise testing of regimens might require a long duration and many participants before a beneficial regimen is identified. Multi-arm multi-stage (MAMS) designs^{14,15} are appealing for use in MDR-TB as multiple drugs can be combined into treatment regimens requiring phase 2 testing. Initially proposed to speed the development of new agents for use in cancer,¹⁶ Phillips et al¹⁷ outlined how this class of adaptive designs could be extended in which multiple experimental arms and a control arm are initiated and then “[p]lanned repeated interim analyses that are pre-specified in the protocol allow for poorly performing regimens to be dropped early and resources switched to evaluation of other potentially effective regimens. The trial continues until there are sufficient data for the best regimen(s) to be identified and taken forward to phase 3.” Regimen dropping criteria during interim analyses balance false-negative (discontinuation of an efficacious regimen) and false-positive (continuation of a non-efficacious regimen) decisions.

MAMS designs used in seamless Phase 2/3 trials, where, at the final stage, remaining arms are evaluated on a definitive endpoint are also of interest.¹⁷ This requires a sensitive and specific intermediate outcome for screening arms to avoid continuing ineffective or discarding effective arms. If disease status can be culture-confirmed, as in most adult trials, conversion status at six months might be considered for use as an intermediate outcome. However, data from adults with MDR-TB show high sensitivity (92%) but poor specificity (58%) for the six month culture-conversion status to predict patient-level long term treatment outcome and show substantial variability in sensitivity and by participant characteristics and geography.¹⁸ This issue limits their usefulness at the screen-out phase of seamless trials. At a trial level, because of the critical need for efficacious, safe, and tolerable treatment, decision rules based on the intermediate outcome should have high sensitivity for continuing useful treatment arms which is paramount in order to avoid discarding critically needed beneficial regimens, whereas continuing regimens that are not useful (lower specificity) would only result in wasted resources and patient time. If estimates and behavior of these performance characteristics can be extrapolated to pediatric MDR-TB, then the MAMS design would have the same value for MDR-TB pediatric trials. However, since a minority of children with clinical TB disease have confirmed diagnoses, intermediate bacteriologic outcomes are even less useful in children. As respiratory sampling methods for children and rapid diagnostic tests that are highly sensitive and specific for the diagnosis of disease and for treatment outcome become available however, MAMS designs may become more realistic for MDR-TB trials in children. For a MAMS trial which has multiple comparisons, regulatory guidance suggests controlling for family-wise error rate, which is the probability of making a Type I error for all study comparisons.^{19,20}

FACTORIAL DESIGNS

Using a factorial design to answer multiple questions is appealing. In addition to the anti-TB regimens, there are opportunities to evaluate non-anti-mycobacterial adjuvant therapies,

drugs to ameliorate toxicity, different health systems, or nutritional interventions. When no interactions between factors exist, a factorial design is an efficient use of resources. When interactions are present, treatment effects can be evaluated within subgroups but may be underpowered.

If novel drugs are evaluated by simply adding or substituting the novel drug in existing regimens then a factorial design might work. However, this is increasingly felt to be inefficient and most TB trials going forward seek to evaluate completely novel regimens which include novel drugs. This makes a standard factorial design inappropriate. For example, because increased ototoxicity is exclusively expected in an injectable-containing regimen, evaluating the addition of the potentially otoprotective agent N-acetylcysteine (NAC), in an injectable-sparing regimen is not relevant. However, an incomplete factorial design could include: 1) a control using the WHO-approved injectable-containing regimen, 2) a novel all-oral regimen, and 3) a WHO-approved injectable-containing regimen plus NAC. Arms 1 and 3 would be compared to evaluate the reduction of ototoxicity with NAC and because the reduction is expected to be substantial, this third arm could be small. The primary study efficacy analysis should not combine Arms 1 and 3 when evaluating the novel anti-TB regimen, because the addition of NAC may affect treatment outcomes through improved tolerability and adherence.

SUBGROUPS, STRATIFICATION, AND COVARIATE ADJUSTMENT

Subgroups will be of interest in pediatric MDR-TB trials, based on culture-confirmation, age, disease severity, DST profile, HIV status, and other factors, due to the potential for heterogeneity of both safety and efficacy outcomes. Stratification may serve administrative purposes, such as stratification by study center to address differences in clinical or laboratory practices. More importantly, stratification may serve statistical purposes, such as to define distinct study populations in which treatment comparisons will be separately powered or to ensure balance of important prognostic factors. To be used as a stratification factor in the randomization, that factor must be evaluated at study entry. Subgroup enrichment can be used to ensure sufficient representation of subgroups of interest.²¹

Bacteriologic-confirmation status would be a desirable randomization stratification factor if available at study entry. Unfortunately, many children will enter the trial prior to confirmation. A factor which is not used in the randomization may still be used to define distinct subgroups or for covariate adjustment during analysis. Providing the specimen is obtained prior to randomization, comparisons within subgroups defined using bacteriologic-confirmation status maintain the integrity of the randomization. Because participants with bacteriologic confirmation have unambiguous disease and can be assessed on microbiologically-based treatment outcomes, it may be important to ensure that this subgroup is well represented in the study population and, at the extreme, the study could be powered to allow separate evaluation of primary outcomes among those with confirmed TB. In the latter case, clinically diagnosed MDR-TB patients should not be excluded from study participation as they represent an important population as discussed earlier. Since proportionally more children will have clinically-diagnosed MDR-TB disease, the study

should be able to enroll sufficient clinically–diagnosed cases in the same time frame that confirmed cases are enrolled.

Subgroup analyses are used to verify robustness of results or identify heterogeneity. These analyses should be pre–specified as secondary or exploratory to guard Type I error. It is uncommon for a study to be adequately powered to detect heterogeneity, and observed effects are likely to vary due to chance even if they are homogeneous. When heterogeneity is anticipated, the study design should take it into account. If a non–inferiority trial design is used when heterogeneity is expected, it may be important to consider distinct subgroup–specific non–inferiority margins when assessing the consistency of an overall conclusion of non–inferiority. Pre–specification of tests to be performed and adjustment for any multiplicity to control Type I error is essential. Subgroup analyses in clinical trials are addressed by Alosch and colleagues.^{21,22}

OUTCOMES

The decision to evaluate primary outcomes at a fixed time point post–randomization versus at a fixed time post–treatment completion needs to be made with consideration to the respective risks. The former may favor longer duration regimens, whereas the latter, consistent with most programmatic guidelines, may favor shorter duration regimens. We argue that evaluating the primary endpoint for a MDR–TB clinical trial at a fixed time point post–randomization should allow for the greatest consistency of ascertainment and, in the context of treatment shortening trials, is more conservative. For longer duration regimens, a longer treatment period may protect from including cases of reinfection but some relapses may be missed during the shorter post–treatment period. From a biological efficacy standpoint, investigators may not want to capture differences in protection from reinfection, but from a pragmatic standpoint this may be a benefit of longer treatment. Follow–up duration for the primary outcome should balance between allowing sufficient time post–treatment completion for the longest regimen to capture relapses versus capturing reinfection. This may differ from the length of follow–up advised by WHO or other guidelines, but it does not preclude longer follow–up, for consistency with guidelines or other clinical considerations. Longer duration follow–up can be a secondary outcome. Whole genome sequencing or high–throughput typing methods used in adult studies to distinguish reinfection from relapse may play less of a role in pediatric TB studies since proportionally fewer participants will have bacterial isolates.

BLINDING

Blinding plays an important role in reducing bias in clinical trials. In studies that evaluate treatment regimens composed of fewer drugs, test the substitution of a single drug in a complex regimen, or test shortening the duration of treatment, blinding may be useful. However, in a study comparing regimens that require different drugs, modes of administration and treatment durations in combination, as is likely in pediatric MDR–TB, blinding may not be possible or acceptable. Employing placebos for each drug that differs between regimens may result in a high pill burden which may negatively affect adherence. Improved treatment outcomes for an all–oral regimen may be attributed in part to improved

adherence; employing a placebo requiring an injection would obviate this benefit and may not be acceptable to families or IRBs. This does not preclude the use of blinding, where possible, such as for the independent endpoint review committee evaluating primary efficacy and/or safety outcomes, or investigators reviewing radiographs or recording diagnostic test or other results. Since a trial will likely be open-label, designs utilizing response-adaptive randomization,²³ for their potential merits, may not be ideal in this setting. An astute site investigator in such a trial may not maintain a state of equipoise as the proportions of participants assigned to specific treatments change over the course of the study.

CONCLUSION

We have highlighted key issues in designing an efficacy trial to evaluate novel treatment regimens for MDR-TB in children with pragmatic approaches to some of the important challenges posed by childhood TB. Because current efficacy of MDR-TB treatment in children is good but there is potential for better safety and acceptability with an all-oral, shortened duration regimen, a non-inferiority trial design is likely the best approach. Such a trial would not blind study participants and site personnel, but should blind independent endpoint review committees and other study personnel. The study will likely enroll bacteriologically-confirmed and clinically-diagnosed cases and may be enriched to ensure bacteriologically-confirmed cases are well represented or be powered for each subgroup. Other subgroup analyses should be designated in the protocol as secondary or exploratory or will require testing that maintains tight control of Type I error.

Acknowledgements:

We acknowledge RESIST-TB, IMPAACT, Vital Strategies, and New Ventures for hosting the Pediatric Multidrug Resistant Tuberculosis Clinical Trials Landscape Meeting in Arlington, Virginia, USA on June 17, 2016 that was the impetus for this work. SK and GM received funding support from the National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health (NIH) under award number UM1 AI068616 (IMPAACT SDMC); JAS is supported by an NIHR Academic Clinical Lectureship and through a grant from the Academy of Medical Sciences. We are grateful to the reviewers for their thoughtful comments that helped us to substantially improve this manuscript.

REFERENCES

1. Food, Drug Administration HHS. Guidance for industry, General considerations for pediatric pharmacokinetic studies for drugs and biological products. 1998.
2. Food, Drug Administration HHS. Guidance for industry, Pulmonary tuberculosis: Developing drugs for treatment. 2013.
3. WHO Treatment Guidelines for Drug-Resistant Tuberculosis, 2016 Update. 2016 at <http://apps.who.int/iris/bitstream/10665/250125/1/9789241549639-eng.pdf?ua=1>.
4. Perez-Velez CM , Marais BJ . Tuberculosis in children. *N Engl J Med* 2012;367:348–61.22830465
5. Detjen AK , DiNardo AR , Leyden J , et al. Xpert MTB/RIF assay for the diagnosis of pulmonary tuberculosis in children: a systematic review and meta-analysis. *Lancet Respir Med* 2015;3:451–61.25812968
6. Schwartz D , Lellouch J . Explanatory and pragmatic attitudes in therapeutical trials. *J Chronic Dis* 1967;20:637–48.4860352
7. Ettehad D , Schaaf HS , Seddon JA , Cooke GS , Ford N . Treatment outcomes for children with multidrug-resistant tuberculosis: a systematic review and meta-analysis. *Lancet Infect Dis* 2012;12:449–56.22373593

8. Food, Drug Administration HHS. International Conference on Harmonisation; choice of control group and related issues in clinical trials; availability. Notice. Fed Regist2001;24390–1.12356096
9. D'Agostino RB , Massaro JM , Sullivan LM . Non-inferiority trials: design concepts and issues – the encounters of academic consultants in statistics. Stat Med 2003;22:169–86.12520555
10. Food, Drug Administration HHS. Non-inferiority Clinical Trials to Establish Effectiveness, Guidance for industry In: Office of Communications DoDI, Office of Communication OaD, Research CfDEa, eds.2016.
11. Montepiedra G , Yuen CM , Rich ML , Evans SR . Totality of outcomes: A different paradigm in assessing interventions for treatment of tuberculosis. J Clin Tuberc Other Mycobact Dis 2016;4:9–13.28042610
12. Evans SR , Rubin D , Follmann D , et al. Desirability of Outcome Ranking (DOOR) and Response Adjusted for Duration of Antibiotic Risk (RADAR). Clin Infect Dis 2015;61:800–6.26113652
13. Phillips PP , Morris TP , Walker AS . DOOR/RADAR: A Gateway Into the Unknown? Clin Infect Dis 2016;62:814–5.26658302
14. Bratton DJ , Phillips PP , Parmar MK . A multi-arm multi-stage clinical trial design for binary outcomes with application to tuberculosis. BMC Med Res Methodol 2013;13:139.24229079
15. Royston P , Barthel FM , Parmar MK , Choodari-Oskoei B , Isham V . Designs for clinical trials with time-to-event outcomes based on stopping guidelines for lack of benefit. Trials 2011;12:81.21418571
16. Parmar MK , Barthel FM , Sydes M , et al. Speeding up the evaluation of new agents in cancer. J Natl Cancer Inst 2008;100:1204–14.18728279
17. Phillips PP , Gillespie SH , Boeree M , et al. Innovative trial designs are practical solutions for improving the treatment of tuberculosis. J Infect Dis 2012;205 Suppl 2:S250–7.22448027
18. Kurbatova EV , Cegielski JP , Lienhardt C , et al. Sputum culture conversion as a prognostic marker for end-of-treatment outcome in patients with multidrug-resistant tuberculosis: a secondary analysis of data from two observational cohort studies. Lancet Respir Med 2015;3:201–9.25726085
19. The European Agency for the Evaluation of Medicinal Products, Evaluation of Medicines for Human Use, Committee for Proprietary Medicinal Products (CPMP). Points to Consider on Multiplicity Issues in Clinical Trials. 2002.
20. Food, Drug Administration HHS. Guidance for industry, Adaptive Design Clinical Trials for Drugs and Biologics, Draft guidance. 1998.
21. Alosch M , Fritsch K , Huque M , et al. Statistical considerations on subgroup analysis in clinical trials. Statistics in Biopharmaceutical Research 2015;7:286–303.
22. Alosch M , Huque MF , Bretz F , D'Agostino RB Tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials. Stat Med 2016.
23. Rosenberger WF , Sverdlov O , Hu F . Adaptive randomization for clinical trials. J Biopharm Stat 2012;22:719–36.22651111

Figure 1:

Sample size requirements (evaluable participants per arm) versus non-inferiority margin, assuming 15% of participants with unfavorable outcome in the control arm for 90% power using a one-sided 2.5% significance level test for various assumed differences between experimental and control regimens