A systematic review of endpoint definitions in late phase tuberculosis therapeutic trials

Nancy Kathryn Hills (nancy.hills@ucsf.edu)
University of California San Francisco https://orcid.org/0000-0002-1041-8182
Johnson Lyimo
Ministry of Health, Dodoma, Tanzania
Payam Nahid
University of California San Francisco
Rada Savic
University of California San Francisco
Patrick P.J. Phillips
University of California San Francisco https://orcid.org/0000-0002-6336-7024

Research Article

Keywords: Estimand, Tuberculosis, Phase III, Intercurrent events

Posted Date: April 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-397643/v1

Version of Record: A version of this preprint was published at Trials on August 3rd, 2021. See the published version at https://doi.org/10.1186/s13063-021-05388-1.

Abstract

Background. Safe, more efficacious treatments are needed to address the considerable morbidity and mortality associated with tuberculosis (TB). However, the current practice in TB therapeutics trials is to use composite binary outcomes, which in the absence of standardization may inflate false positive and negative errors in evaluating regimens. The lack of standardization of outcomes is a barrier to the identification of highly efficacious regimens and the introduction of innovative methodologies

Methods. We conducted a systematic review of trials designed to advance new TB drugs or regimens for regulatory approval and inform practice guidelines. Trials were primarily identified from the WHO International Clinical Trial Registry Platform (ICTRP). Only trials that collected post-treatment follow-up data and enrolled at least 100 patients were included. Protocols and Statistical Analysis Plans (SAP) for eligible trials from 1995 to the present were obtained from trial investigators. Details of outcome data, both explicit and implied, were abstracted and organized into three broad categories: Favorable, Unfavorable, and Not Assessable. Within these categories, individual trial definitions were recorded and collated, and areas of broad consensus and disagreement were identified and described.

Results. From 2205 TB-related trials, 51 were selected for protocol and SAP review, from which 31 were both eligible and had accessible documentation. Within the three designated categories, we found broad consensus in the definitions of Favorable and Unfavorable outcomes, although specific details were not always provided, and when explicitly addressed, were heterogeneous. Favorable outcomes were handled the most consistently but were widely variable with respect to specification. In some cases, the same events were defined differently by different protocols, particularly in distinguishing Unfavorable from Not Assessable events. Death was often interpreted conditional on cause. Patients who did not complete the study because of withdrawal or loss to follow-up presented a particular challenge to consistent interpretation and analytic treatment of outcomes.

Conclusions. In a review of 31 clinical trials, we found that outcome definitions were heterogeneous, highlighting the need to establish clearer specification and a move towards universal standardization of outcomes across TB trials. The ICH E9 (R1) addendum provides guidelines for undertaking and achieving this goal.

Registration PROSPERO 2020 CRD42020197993

Introduction

Tuberculosis (TB) kills more people globally than any other single pathogen(1), with mortality and morbidity likely to increase as a result of the COVID-19 pandemic and the many ensuing challenges posed to national TB control programs(2). New shorter, safer and more efficacious treatments are urgently needed(3). In response to this need, more than a dozen new compounds are in early or middle clinical development (https://www.newtbdrugs.org/pipeline/clinical) with numerous late-phase

randomized controlled trials expected in the near future, conducted either by individual pharmaceutical companies, or as part of publicly or philanthropically funded networks.

Most recent and ongoing late-phase TB therapeutics trials have used a composite binary outcome that combines bacteriological failure and relapse, death, treatment changes, and loss to follow-up as the primary efficacy outcome. Multiple analysis populations are usually proposed as co-primary. These include an intention-to-treat analysis (ITT) population including all patients randomized, classifying as unfavorable any participants with substantial missing data; a modified ITT (mITT) analysis population excluding some losses to follow-up from the analysis, and a per protocol (PP) analysis population excluding participants who had a protocol violation or did not complete a sufficient proportion of treatment. This approach has a number of limitations:

- **Not standardized.** Outcome definitions are not standardized across phase III TB treatment trials. This leads to considerable challenges in combining data, interpreting results, assessing comparative efficacy, implementing predictive modelling, and conducting necessary meta-analyses (as exemplified in the TB-ReFLECT project(4)).
- **Outdated**. The emphasis on simple, unadjusted per protocol analyses (not considering causal inference methods(5)) and even modified intention-to-treat analyses with post-randomization exclusions is at odds with best practice in other disease areas(5, 6) and regulatory guidance(7). The draft version of the FDA guidance document for non-inferiority trials (2010) initially accommodated an “as-treated” analysis, but this was removed in the final guidance document (2016)(7).
- **May inflate Type I and II errors**. Classifying the outcome of participants lost to follow-up as unfavorable (i.e., defining “missing” as “failure”) is likely to result in conservative estimates in superiority trials by diluting any treatment effect (and is therefore often favored by regulators). This is not necessarily conservative in a non-inferiority trial, can inflate type I and type II errors, and also results in mis-leading decisions in the context of adaptive platform trial designs.
- **A barrier to identifying highly efficacious regimens.** Including events that are less likely to be related to treatment (including loss to follow-up and non-TB mortality) in a composite outcome increases variability in treatment effect estimates and therefore necessitates an increased sample size. This added “noise” also makes it challenging to identify interventions (like stratified medicine approaches(4)) that may result in very high cure rates (97%-100%) without requiring prohibitively large sample sizes(8).
- **At odds with policy makers and guideline developers.** WHO guidelines generally rely on WHO programmatic outcomes definitions(9) when considering evidence (the 2018 DR-TB guidelines is a case in point(10)). The “catch-all” nature of the composite outcome currently used in phase III trials is likely to have contributed to this disconnect between trials and the approach taken for guidelines.
- **Mixes efficacy and safety events**. Including treatment changes due to adverse events during treatment in the composite outcome conflates safety and tolerability with efficacy.
- **Impedes progress in prediction modelling.** A phase III outcome defined by composite events does not allow for efficient and predictive linkage with phase IIB endpoints, such as time to culture conversion,

that are essential for bridging the gap between phase II and phase III trials(11), and that will be increasingly important as new biomarkers of TB treatment response are identified(12). Similarly, translational modelling across the species (NHP, mice, rabbit) is limited due to discordance in outcomes, enabling translational errors and suboptimal decision-making which regimens to advance in clinical development.

Furthermore, regulatory guidance is changing with the ICH E9 (R1) addendum on Estimands and Sensitivity Analyses (finalized November 2019), which formalizes a new approach to specifying trial objectives, endpoints and analysis populations (collectively called the Estimand).

With new late-phase trials expected on the horizon, it is therefore vital to carefully consider and refine the phase III primary efficacy outcome, making use of the language proposed in the ICH E9 (R1) addendum. The objective of this systematic review was to first catalogue phase III outcome definitions (including analysis populations and primary objectives) from recent phase III trials for new regimens for drug susceptible (DS) and drug resistant (DR) tuberculosis, and then to conduct a thematic analysis on these outcomes to identify areas of consensus and disagreement. The overarching goal of this work is to use these results to develop standardized consensus estimands for phase IIC and III TB therapeutics trials.

Methods

The protocol for the systematic review was prospectively registered on the PROSPERO registry (PROSPERO 2020 CRD42020197993)(13) and is provided as an online supplement, along with the PRISMA checklist(14)

Briefly, this systematic review sought to identify trials that have been designed to advance a new drug or regimen for regulatory approval and therefore inform and impact practice guidelines. The focus was on phase III and other late-phase randomized controlled trials (RCTs), or non-randomized trials of new drugs intended specifically for regulatory approval. Trials of treatment for latent TB, prevention of TB, diagnosis of TB, extrapulmonary TB, adjuvant nutritional supplements or immune therapies, ART initiation among TB patients, and trials of TB vaccines and programmatic interventions looking at adherence interventions (DOT or mHealth initiatives) were excluded as endpoints in these trials are defined differently. Trials that did not collect outcome data on post-treatment follow-up (for relapse) or that enrolled fewer than 100 patients were excluded since these were clearly not designed to change guidelines and practice.

The WHO International Clinical Trial Registry Platform (ICTRP) was the primary database searched to identify relevant trials. To increase the likelihood that no trials were missed, we also contacted experts in the field of TB trials to identify other trials and reviewed the excellent list of DR-TB clinical trials maintained by RESIST-TB (www.resisttb.org).

Two individuals (PPJP and JJL) independently reviewed the list of trials identified from the search strategy using titles and other fields from the ICTR platform to determine whether they met the inclusion criteria. Investigators or sponsor representatives of final selected studies were contacted to access the

study statistical analysis plans (SAPs) and study protocols; these were downloaded from the public domain when available. Two individuals (NKH and JJL) reviewed all protocols and Statistical Analysis Plans and abstracted relevant information.

Qualitative data from primary endpoint definitions of different studies were analyzed using thematic analysis in the five stages outlined by Braun and Clarke(15) Qualitative analyses and summaries were done by NKH. The final draft of the manuscript was circulated to PIs of all completed trials for their comments, approval and edits. Our objective was to describe areas of consensus and disagreement as drawn from protocols and SAPs across trials, rather than to critique individual trials. For this reason, we do not discuss nuances in definitions in specific trials, but rather aim to provide a summary of broad trends in outcomes definitions and analyses used in recent TB treatment trials.

Results

Due to heavy traffic generated by the COVID-19 pandemic during early 2020 and limited ability to use the on-line search portal for the WHO ICTR, we downloaded the full ICTR database (3.5GB, 19 May 2020) and used it for this systematic analysis. From 632,787 clinical trials registered, we identified 2205 with condition containing ‘tb’ or ‘tubercul’ and selected 510 for independent registry review by two reviewers. All registry information was available in English. From these, we identified 51 trials that were highly likely to be relevant and eligible for inclusion (See Figure 1 for PRISMA flow diagram(14)). We then contacted Principal Investigators of the selected trials to request the most current versions of their protocols and, when possible, SAPs. We received protocols from 31 studies, and SAPs from 18 (58%). Many trials were listed on more than one trial registry; the majority (27, 87%) were listed at least on clinicaltrials.gov; two of the studies was only listed on ISRCTN registry (isrctn.com), and two trials were only listed on Clinical Trials Registry of India (ctri.nic.in). Registration across all trials was finalized between the years of 2001 and 2020 (although in some early cases, trials were not registered until after completion; the earliest trial began enrolling in 1995 but was not registered until 2001), with 21 (68%) of the trials registered during or after 2010. All protocols were available in English. Twenty-six of the trials (84%) were phase III, either with (n=29) or without (n=2) internal controls; one trial was described as phase IIB/III, two were listed as phase IIC, and two as phase IV (see Table 1).

Ten of the trials targeted patients with drug-resistant TB (DR-TB) and the remaining 21 trials enrolled patients whose TB was drug susceptible (DS-TB). Two protocols included patients diagnosed with either DS- or DR-TB, although in each case those with DR-TB were enrolled as a non-randomized interventional cohort that was not ”statistically analyzed.” Five (17%) of the trials included participants enrolled in African sites, eight (26%) included participants enrolled in Asian sites, and 16 (52%) included participants enrolled on both continents. Seven (24%) included subjects in South American sites, two (7%) in Latin America and five (17%) in North America. Proposed subjects were as young as 12 (one trial), 14 (two trials) and 15 (four trials), although one trial did not impose a lower age limit; however, most trials included patients aged 18 years and older. Two protocols capped the age of participants at 60, five at 65, one at 70, and another at age 75; in the remaining trials, an upper age limit was not specified. Only one

trial exclusively conducted in children and adolescents was included in the 51 trials for protocol and SAP review, but the protocol was not made available for inclusion in our review.

The primary objective uniformly across all but one study was to investigate whether a novel treatment regimen had non-inferior or superior efficacy in terms of a “long-term durable cure extending through post-treatment follow-up.” In the remaining study, efficacy outcomes were secondary to safety outcomes. Novel interventions varied across trials, and included shortening treatment, evaluating the efficacy of new combination regimens, utilizing oral medications exclusively, testing different doses and durations of treatment, testing fixed dose combination formulations, and simplifying treatment by utilizing intermittent dosing. A non-inferiority analysis comparing a new treatment regimen to standard treatment was specified in 18 (58%) protocols, with margins of non-inferiority ranging from 4% to 12%. Other techniques used included equivalence testing (n=6), superiority testing (n=3) and logistic regression to compare differences in proportions of participants achieving a Favorable outcome (or, conversely, an Unfavorable outcome). In 15 (48%) protocols, an intention-to-treat (ITT) or modified intention-to-treat (mITT) analysis was defined as primary, while per protocol (PP) analyses were also planned as secondary or confirmatory analyses. In 14 (48%) studies, the mITT and PP analyses were considered co-primary. In only one of the protocols we reviewed was the PP analysis considered primary; in one other, no specification was made (although in this case we did not have access to the trial SAP).

The duration of experimental treatment regimens ranged from 13 weeks to 26 weeks for DS-TB trials, and from 24 to 44 weeks for DR-TB trials. Duration of post-treatment follow-up was of varying lengths, these might be measured as time post-randomization or post-treatment, sometimes in weeks, at others in months. Some protocols specified “time windows” around evaluation dates, while others cited only the week or month representing the end of follow-up without explanation as to how much time before or after defined the follow-up “window.” Total trial duration time from randomization to end of follow-up ranged from 78 to 130 week for DS-TB trials, and from 104 to 132 weeks for DR-TB trials. In general, the primary trial outcomes were measured at the end of follow-up. At the time of writing, 8 (26%) trials were still open to enrollment. Seven trials (19%) were complete with study findings not yet available or in follow-up, and 2 (6%) were completed and had results posted on clinicaltrials.gov. For 14 (45%) trials, the primary results of the trial had been published in a peer-reviewed journal or presented at an international conference.

Outcome Definitions

Outcomes across study protocols were assigned to one of three broad categories: Favorable, Unfavorable, or Not Assessable. Protocols generally defined an outcome as Favorable in terms of timing of culture conversion and required number of negative cultures at the end of the follow-up period. Similarly, determination of an outcome as Unfavorable primarily involved the observation of a specific number of positive cultures with or without reference to a time frame for the samples. All protocols specified these bacteriological conditions to some degree, although the circumstances under which determinations were made, and the granularity with which these were defined in individual protocols, varied considerably (see Supplemental Table 1 for a listing of outcome definitions found in protocols).

Protocols from recent studies were more likely to allow for categorization of an outcome as Not Assessable if it could not be clearly classified as Favorable or Unfavorable, e.g., deaths unrelated to TB, recurrence due to re-infection with a different strain, and loss-to-follow-up with last culture negative. However, in some cases identical outcome-determining events were categorized as Not Assessable in some cases and Unfavorable in others. Protocols from earlier trials seldom specifically labeled an outcome Not Assessable, although this designation sometimes could be inferred from descriptions of patients excluded from analyses. In others, however, this possibility was neither explicitly nor implicitly addressed. Outcomes determined to be Not Assessable will be discussed simultaneously with Unfavorable outcomes, since the same event could be interpreted as one or the other by different trial protocols. Table 2 summarizes the range of outcome definitions and the frequency of their occurrence across protocols.

Protocols additionally addressed issues around treatment and adherence with respect to categorization of outcome. These will be considered last, as they often coincide with or contribute to other reasons of categorization of outcomes as either Unfavorable or Not Assessable.

Favorable Outcomes

In contrast to Unfavorable and Not Assessable outcomes, Favorable outcomes received the most consistent treatment across protocols. In all protocols that we reviewed, a patient with a Favorable outcome was defined as one who tested negative on a varying number of cultures, with reference to the end of treatment and/or follow-up. Nonetheless, this seemingly straightforward outcome underwent a multitude of permutations across trials. Some trials required only that a patient be “culture negative;” others defined an outcome as Favorable based on a single negative culture. The majority of trials required at least two negative cultures, and in a small number of trials, a patient was required to have three negative cultures to achieve negative status. In addition to the variability in number of negative cultures required, Favorable status was conditional on a variety of restrictions in terms of timing (with reference to either the end of treatment, the end of follow-up or both), spacing (amount of time between the negative cultures that ranged from occurrence on different days to requiring at least four intervening weeks between negative cultures), and culture medium type (solid or liquid).

Spontaneous sputum production usually decreases or resolves during successful treatment and follow-up for TB and most such patients are culture-negative for *M.tb*(16). A smaller number of studies addressed a patient’s potential inability to produce sputum at various points in the trial as indicative of a Favorable outcome. One protocol interpreted a patient’s inability to ever produce sputum as a Favorable outcome; another further stipulated that never producing sputum would be considered Favorable even if the patient never achieved culture negative status but completed follow-up without clinical or microbiological relapse. Others defined circumstances under which failure to produce sputum at the end of the follow-up period could be classified as negative, e.g., provided this coincided with a patient having prior culture negative status or lacking clinical symptoms. In only one trial was failure to produce sputum at the end of follow-up categorized as an Unfavorable outcome. While generally classified as a Not

Assessable outcome (see below), one study classified patients who developed an infection with a strain different from that with which they had originally been infected (an exogenous reinfection) as having a Favorable outcome if the original strain was shown to have been cured. In another, a contaminated culture result or one which could not be evaluated was categorized as Favorable, provided there were no positive cultures at the end of follow-up. Two studies allowed for a patient to have had a Favorable outcome even with a culture at the end of follow-up that was inconclusive, if clinical and radiological symptoms were supportive of the assessment.

Unfavorable Outcomes vs. Not Assessable Outcomes

In the broadest sense, we found that all the reviewed protocols deemed that a patient's outcome would be considered Unfavorable primarily based on positive sputum cultures. However, the level of detail attached to culture positivity varied from the most general ("Failure at end of treatment") to the bewilderingly complex: in one trial, for example, the outcome of a patient not attending the final visit could not be categorized as Unfavorable until all of four specified conditions were met, and two additional conditions had been taken into account.

Categories of unfavorable/not assessable outcomes

Failure to ever achieve negative culture conversion. A patient's failure to respond successfully, as defined bacteriologically, to the prescribed regimen by the end of the treatment period constituted the most straightforward type of unfavorable outcome. In some protocols, however, the treatment duration could be extended if necessary or if some limited number of treatments had been missed, thus lengthening the time a patient was given to achieve culture conversion or culture negative status.

Relapse and Re-infection. Recurrence of bacterial infection can occur as an endogenous relapse, defined as a patient's recurrence with positive culture status with the originally diagnosed strain, having previously attained negative status, or as an exogenous re-infection, i.e., a new infection with a different strain. Not all protocols specifically addressed an analytical approach to both. One protocol did not address either relapse or reinfection; some categorized the status of relapse but not re-infection, and several addressed re-infection but not relapse. Other protocols addressed and categorized both.

Relapse. In all studies, relapse was considered an Unfavorable outcome in terms of its analytical treatment. Although some studies provided specific definitions of relapse, others included it as either part of a composite outcome or (in a few cases where patients were required to have been previously treated and cured prior to the study) as the primary outcome. Definitions, when provided, varied as to when and how relapse was defined, and with what level of detail, however. Some studies defined a relapse as occurring in patients who were culture-negative at the end of treatment, but with different constraints on the conversion to culture-positive. These included diagnosing relapse in a patient who tested positive twice with no intervening negatives, whose two positive tests occurred at least one day apart, who had positive sputum cultures during four consecutive monthly exams (at least one with 20 or more colonies), or who had a subsequent diagnosis and treatment for the same or another DR strain (in a study targeting

DR infections). Similarly, two additional DR studies defined relapse as having occurred when a patient was prescribed a new DR regimen after treatment and before the end of follow-up. Another study specified that a patient's conversion to negative status had to occur over at least four weeks, with subsequent positive status (on solid medium) confirmed by a second positive culture on a different day. Other studies offered less specific criteria, including simply "recurrence by the end of the study," "after cure, single culture positive," and "one culture positive and clinical features suggestive of recurrent disease."

Reinfection. Unlike relapse, patients who acquired an infection with a different type of TB were regarded by most DR-TB and DS-TB studies as having outcomes that were Not Assessable. Only one protocol viewed re-infection with a different strain as Unfavorable; one study targeting patients with DR-TB categorized re-infection with a *different* DR strain as Unfavorable, but with a DS strain as Not Assessable. As previously mentioned, one protocol categorized a patient's re-infection as Favorable if occurring after a confirmed conversion to negative status with respect to the original strain.

Death. With varying degrees of granularity, most protocols addressed death as an outcome, whether occurring during treatment, after treatment during the follow-up period, or during either. One protocol did not mention death in relation to outcome, and another mentioned death only in that it precluded a Favorable outcome; we were unable to obtain SAPs for either of these studies. The death of a patient was generally categorized as an Unfavorable outcome, although under certain specified circumstances, deaths could also result in study outcomes being considered Not Assessable.

Death during treatment. A patient's death during treatment could fall into one of the following categories: (1) death due to any cause, (2) death directly related to TB, and (3) death due to causes unrelated to TB. Non-TB deaths were categorized differently across studies; some considered these to be Not Assessable, while more frequently, studies treated them as Unfavorable, with the exception of deaths due to accident, violence, trauma or suicide (with the exception of suicide, these latter were generally classified as Not Assessable). Death by suicide was specifically addressed in a third of the protocols, but was considered by some as Unfavorable, and by others as Not Assessable. An additional protocol specified that the outcome of a patient whose death during treatment was unrelated to TB, but whose culture status at the time of death was unknown, would be classified as Not Assessable.

Death during post-treatment follow-up. During the post-treatment follow-up phase, "all cause deaths" (without further differentiation) were regarded as Unfavorable outcomes in some studies, while in others, deaths were only considered Unfavorable if TB-related. A small number of studies considered a generalized category of non-TB deaths to be Not Assessable for purposes of analysis. In several studies, the treatment of death during follow-up was determined with respect to bacteriological status. Several trial protocols classified the outcomes of patients who died with their last culture negative as Not Assessable. Additional studies more specifically proposed that deaths be considered Not Assessable only if a patient died while culture negative, under the condition that the last positive culture had been followed by two negative cultures at least seven days apart. In addition, one study specified that a patient

who died from extrapulmonary TB would be considered as having an Unfavorable outcome; another classified patients whose deaths were due to an infection other than with the originally diagnosed strain to have outcomes that were Not Assessable.

Withdrawal of consent/ lost-to-follow up. Across study protocols, outcomes of patients who were lost to follow-up or who withdrew from the study appeared to be the most challenging to categorize. These patients were variously noted as having been lost to follow-up or withdrawn: (1) while still being treated; (2) at any point, during follow-up, (3) after being cured at the end of treatment, during follow-up, or (4) "when last seen."

During the treatment phase. With respect to patients lost or withdrawn while treatment was still ongoing (without further caveats), a quarter of the protocols classified their outcomes as Unfavorable; one study alone categorized them as Not Assessable. Other protocols determined categorization based on the reason for the patient's withdrawal. In one protocol, patients who withdrew or were lost due to clinical reasons were considered to have an Unfavorable outcome. More frequently, patients who exited the study during the treatment phase were considered to have outcomes that were Not Assessable, including those whose withdrawal was either unrelated to TB or was due to protocol violation, pregnancy, or moving away and/or becoming untraceable at any point.

After treatment completion. In addressing patients who were lost to follow-up or who withdrew after completing treatment, Unfavorable outcomes could include those who exited the study under any circumstances (although one protocol classified such a patient as having an outcome that was Not Assessable); those whose last positive culture was not followed by at least two negative cultures ≥7 days apart; those who terminated the study early, but were known to be alive at last contact, or who were lost to follow-up with vital status unknown; patients who had not achieved culture negative status or who had been classified as having an Unfavorable outcome before their withdrawal; patients who could not be contacted for some specified period of time prior to the last study visit; and those who had no culture results within a specified window of time prior to the study endpoint. As specified by two protocols, it was also necessary for these latter patients to be either culture positive when last tested, have no other post-baseline results, or have a negative culture at their most recent result, but with radiological or clinical symptoms that were inconclusive.

Alternatively, the following patients were categorized with varying frequency as having outcomes that were Not Assessable: those whose last culture before study exit was negative; patients whose last two culture results prior to exit were negative, who had not otherwise been deemed Unfavorable; patients whose last culture was negative and whose last positive culture was followed by at least two negative cultures at different visits ≥7 days apart, without an intervening positive culture; and patients not otherwise classified as Unfavorable prior to exit from study.

Patients who withdrew or were lost to follow-up after having been cured at the end of treatment were specifically addressed by one study; those who either did so with their most recent culture positive or who moved away with their most recent culture positive were considered to have Unfavorable outcomes, while

those who under the same circumstances were culture negative or whose most recent culture was contaminated were categorized as Not Assessable.

In some protocols, outcomes were defined at the time when patients “were last seen.” Detailed events included being culture positive, being culture positive with the same type (whether confirmed or not), culture positive not followed by two negatives, or simply not having achieved or maintained culture negative status at the time of their last visit (prior to study endpoint).

Treatment-Related Issues (including treatment changes for adverse events)

Most protocols addressed to some extent their analysis plans regarding treatment issues, including extension, restart, change, and discontinuation of the medications which comprised the specific study regimens. Although in most cases patients who experienced treatment disruptions were considered to have unfavorable outcomes, details varied considerably from study to study. A patient whose treatment was extended for any reason was considered to have had an unfavorable outcome by one study. More commonly, however, the outcomes of patients whose treatment was extended were considered Unfavorable but with exceptions that were considered Not Assessable, including: temporary drug re-challenge, over-treatment with assigned drugs, ≤21 days non-study anti-TB meds for active TB, secondary isoniazid preventive therapy in HIV+ patients, re-infection, pregnancy, making up missed doses, and remaining on treatment at the end of the study without having been declared a treatment failure.

Some protocols categorized patients whose treatment had to be restarted as experiencing an Unfavorable outcome, again with the exceptions that they either had been infected with a different TB type in some cases or had become pregnant in others; another protocol limited designation of an Unfavorable outcome to the period after completion of treatment but before the study’s end.

A change in treatment can take many forms, and this was reflected across protocols. A patient who had any change of medication frequency or dose (except in the case of re-infection) was usually considered as having an Unfavorable outcome, although two protocols made exceptions for patients with a single drug replacement, or those whose drug replacement was due to a guideline change in the standard of care group (neither affected outcome classification). Patients whose treatment was changed due to clinical or radiological deterioration, or because of non-response or poor adherence, were considered to have an Unfavorable outcome by one study each, respectively. Several studies considered as Unfavorable outcomes those of patients for whom one drug was replaced or added, while other studies required that a patient’s treatment involve the replacement or addition of at least two drugs. Such categorization based on number of drug changes ranged from the simple to the overly complex: one study placed further conditions on a two-drug change, declaring that it defined a patient’s outcome as Unfavorable if this occurred because the patient (1) had not converted by the end of the first (more intense) phase of treatment, (2) had bacteriologically reverted during the second treatment phase after having converted to negative in the first, (3) had evidence of additional acquired resistance to fluoroquinolones or 2nd line injectables, or (4) had not converted their sputum cultures to negative status and had two positive

cultures during a specific time period, with the caveat that if one or more of the samples were unavailable or contaminated this would be considered culture positive if the patient displayed deteriorating clinical symptoms.

A patient whose treatment was discontinued was considered by various protocols as having an Unfavorable outcome if study treatment was halted for reasons including the following: experiencing a serious adverse event; starting a different DR-TB regimen; failing to convert after the first phase of a trial where the treatment regimen occurred in two phases, or because the trial regimen needed to be significantly modified for some (unspecified) reason. In most trials, study treatment was discontinued in patients who became pregnant during therapy, who were then treated with standard therapy. In some trials, patients who discontinued treatment because they became pregnant were considered to have an outcome that was Not Assessable, while in others a patient's outcome was considered Not Assessable if the patient's last culture was negative, but Unfavorable if it were positive.

Incomplete treatment in patients whose culture status could not be evaluated at the end of follow-up was considered Unfavorable in several studies; an additional protocol defined a patient's outcome as Unfavorable if, in addition to incomplete treatment, a patient had not attained culture negative status by the end of follow-up. The effect of a patient's missing drugs during the treatment phase was addressed by one protocol that considered this to be Unfavorable if some or all drugs were missed regularly, or if all drugs were missed for more than two consecutive weeks.

Patients who took TB-related but off protocol drugs, or who started TB treatment outside of the study with the most recent culture positive, were considered by one study to have an Unfavorable outcome, while off-protocol drugs not related to TB rendered the outcome Not Assessable. In two other studies, only patients taking specific off-protocol drugs were categorized as having Unfavorable outcomes.

Discussion

In our review of primary efficacy outcomes as defined in the protocols (and SAPs, if available), in 31 confirmatory clinical TB trials for the treatment of active TB, we found broad *conceptual* agreement. A patient's outcome was classified as Favorable or Unfavorable based on the number and timing of negative/positive cultures, and most protocols explicitly acknowledged that outcomes were Not Assessable under certain circumstances (in other cases, this was implicit in descriptions of inclusions and exclusions from given analyses). However, even though they achieved compliance with decidedly broad guidelines for trial sponsors provided by stringent regulatory authorities(17, 18), we found a considerable degree of heterogeneity in outcome definition across trials. In addition, outcomes were for the most part comprised of composite events, and inconsistencies abounded with respect to the ways in which outcome definition was determined by such factors as deviation from treatment regimens, patient withdrawal/loss to follow-up, relapse or reinfection, and even death; the contributions of the individual components to the composite outcome were in all cases unweighted. These outcomes then dictated inclusions and exclusions from different target populations for analysis—ITT, mITT, and per protocol (PP)

—which in turn were variously considered to be of primary, secondary, or equal importance; sensitivity analyses were also sometimes variously used.

Nonetheless, we found that certain areas were treated consistently across protocols, indicating implicit areas of consensus that would facilitate standardization of endpoint definitions. The fairly straightforward criteria for determining a Favorable outcome allowed this diagnosis to be more easily reached, as compared to an Unfavorable or Not Assessable one. While details differed in terms of the number and timing of cultures indicating conversion, and the duration required to validly declare it a durable, long-term cure, it is probable that a consensus definition of a patient with a Favorable outcome would neither be difficult to reach, nor particularly controversial, across investigators and trials. Similarly, a patient who suffered a relapse with the originally diagnosed infecting strain of *M.tb* after reaching confirmed culture negative status was universally classified as having had an Unfavorable outcome, although here, too, small variations in definition are needed to standardize this event across trials.

Understandably, standardizing Unfavorable outcomes presents a far greater challenge. In a systematic review of outcomes reported in 248 peer-reviewed and published phase III TB studies, Bonnett *et al*. found substantial differences in the way Unfavorable outcomes were defined and implemented across numerous dimensions(19). That review was limited to data derived from trial publications and included TB trials from 1950 to 2017 (only 18% of which occurred after 1995), yet Bonnett reported inconsistencies as to what constituted an Unfavorable outcome even in the most recent trials. In our review, with the granular data obtained from the more necessarily detailed study protocols and SAPs (all of which had been registered since 1995), we likewise found little consensus in the specific details attached to endpoint definitions for Unfavorable.

As a result, combining data, interpreting and comparing results, and performing individual patient data (IPD) meta-analyses across trials *that essentially are all working towards the same goal* is at best highly challenging (as experienced with the largest such analysis of TB clinical trial data(4)) and at worst, impossible. The concept of a “Favorable” outcome can be more difficult to define for patients who fail to produce sputum, do not complete the trial, or require treatment changes. Distinguishing between Unfavorable and Not Assessable outcomes, as we have shown, presents even greater potential for discordance, and more differing opinions about what event constitutes each. Adding to the confusion is the fact that some patients inevitably exit the study prematurely, and their reasons for not completing the study, along with their culture status at the time of their exit, are inconsistently used to classify their outcomes as Unfavorable or Not Assessable. Even death, an undeniable and immutable outcome, is cause for dissension; while all protocols considered a patient who died from TB to have had an unfavorable outcome, deaths that were not related to TB could be interpreted as Unfavorable or Not Assessable, depending on circumstances. The conventional use of composite outcomes, which may or may not be directly related to treatment, and that involve numerous assumptions about each “piece” of the outcome, further clouds evaluations of efficacy.

The ICH E9 (R1) addendum, in providing a framework and language for defining clinical trial estimands and outcomes, directly addresses these problems. While no guidelines can cover all circumstances that may arise in a trial, and it is unlikely that one estimand will satisfy the interests of all categories of trial stakeholders, interpretation and comparison across trials would be greatly facilitated by a standardization of the elements of a trial used to evaluate the efficacy of its intervention. In our review of protocols, we found a wide range of granularity of definitions. While some protocols defined outcomes in the most general terms, others complicated definitions by attempting to cover every possible eventuality; in the latter case, the outcome definitions were clouded with minutia, making consensus with other trials extremely unlikely. On the other hand, the absence of precise definitions of outcomes in the protocol or SAP means that some classification decisions are left to the data analyst; these may have only been ”documented” in the analysis code, which is rarely reviewed by study investigators. The ICH E9 (R1) addendum has taken an instructive approach in addressing these problems, separating from the definition of the primary outcome, called an estimand, the many events that occur and either preclude or affect observation of an outcome (referred to by the ICH E9 as “intercurrent events”). This allows not only for a consistent definition of the primary efficacy outcome across trials but also gives a structure for specifying how intercurrent events will be handled in the analysis, thus reducing potential inflation of Type I and Type II errors. Events that have until now been viewed as rendering an outcome “Not Assessable” can rather be categorized as intercurrent events, with decisions about how they will be dealt with in analyses made prior to the beginning of the trial, based on the defined estimands. Thus, within the same trial, an intercurrent event may be treated one way for one estimand, and another way for a second estimand, dependent on the needs of particular stakeholders.

Bringing such events to the forefront therefore would allow for standardization in the way they are classified and how they are treated in the analysis, resulting in the reporting of outcomes that are comparable across trials. Even if *standardization* is not possible between different trials, at the very least the ICH E9 (R1) addendum provides a lingua franca for *specification* to support clear interpretation and translation into clinical practice guidelines. Some preliminary work has been done in this area in the context of individual trials(20, 21).

Our study has several limitations. We were not able to obtain an SAP for every trial we reviewed, and in these cases lacked the more detailed descriptions of how events would be dealt with in analyses than are often provided in the protocol. The sheer length of the protocols themselves made locating specific pieces of information difficult (and in some cases, particularly in the case of older studies, it was not present or was treated in general terms, with specifics purportedly left to the SAP). While our search for clinical trials was as thorough as possible, we cannot be sure that all recent clinical trials were included. We did not receive responses from investigators for 17 of the 51 trials selected for protocol review. These were necessarily excluded, although many may not have been within the scope of our review (notably some of the country-specific trial registries had limited data to assess whether trials were within scope). We did not distinguish trials of unlicensed drugs conducted under stringent regulatory oversight authorities and often sponsored by the pharmaceutical industry from the investigator-initiated trials of

licensed drugs. Both are designed to inform policy and practice, and better specification and standardization of endpoint definitions is relevant to all future TB treatment trials.

While no estimand can include all possible trial occurrences, the standardization of definitions and of the treatment of intercurrent events that occur most frequently will enhance comparability across trials, while allowing for interpretation of rare or unanticipated events. The approach outlined by the ICH E9 addendum can also be used to develop different estimands to address the concerns of specific audiences. It is therefore important, following the recommendations of the ICH E9 addendum, to prioritize both the specification and standardization of outcomes across TB trials. As new drugs and treatment regimens are discovered and tested in trials, the ability to make valid comparisons to old treatments and regimens is also essential if researchers are to effectively collaborate towards our common goals of developing shorter, simpler, and more effective and safe treatment to cure patients with TB. Following this review, our next step will be to produce recommendations for estimands and methods of estimation for TB treatment trials. At a time when the world has begun to establish large adaptive platforms with core protocols for the search of active treatment of COVID-19, it is past time that we, as a TB community, move towards better standardization and harmonization of trial methods.

This work was supported, in whole or in part, by the Bill & Melinda Gates Foundation [Grant #INV-002039]. Under the grant conditions of the Foundation, a Creative Commons Attribution 3.0 Generic License has already been assigned to the Author Accepted Manuscript version that might arise from this submission.

Declarations

Ethics approval and consent to participate:

Not applicable

Consent for publication:

Not applicable

Availability of data and materials:

Not applicable

Competing interests:

The authors declare that they have no competing interests.

Funding:

This work was supported, in whole or in part, by the Bill & Melinda Gates Foundation (INV-002039). Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 Generic License has already

been assigned to the Author Accepted Manuscript version that might arise from this submission.

Authors' contributions:

PP was responsible for conception and design of the work. PP and JL substantially contributed to identification of studies and review for eligibility and inclusion. PP was responsible for obtaining protocols and SAPs from study investigators. All authors contributed to the interpretation of the data and revised the work. All authors have approved the submitted version. All authors have agreed both to be personally accountable for the authors' own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature.

Acknowledgements:

We thank the study participants, the staff at clinical sites, and all investigators for generous sharing of protocols and statistical analysis plans. We also thank Gopalan Narendran, Christian Lienhardt, John Johnson, Angela Crook, Lorenzo Guglielmetti, and Bern-Thomas Nyang'wa for their review of the manuscript.

References

1. World Health Organization. Global Tuberculosis Report 2020. Geneva: World Health Organization; 2020.
2. Mandavilli A. 'The Biggest Monster' Is Spreading. And It's Not the Coronavirus. The New York Times. 2020 Aug. 3.
3. Lienhardt C, Vernon AA, Cavaleri M, Nambiar S, Nahid P. Development of new TB regimens: Harmonizing trial design, product registration requirements, and public health guidance. PLoS Med. 2019;16(9):e1002915.
4. Hernan MA, Robins JM. Per-Protocol Analyses of Pragmatic Trials. N Engl J Med. 2017;377(14):1391-8.
5. Mauri L, D'Agostino RB, Sr. Challenges in the Design and Interpretation of Noninferiority Trials. N Engl J Med. 2017;377(14):1357-67.
6. U.S. Department of Health and Human Services FaDA, Center for Drug Evaluation and Research (CDER),. Guidance for Industry. Non-inferiority clinical trials to establish effectiveness: U.S. Department of Health and Human Services , Food and Drug Administration, Center for Drug Evaluation and Research (CDER); 2016.
7. Imperial MZ, Nahid P, Phillips PPJ, Davies GR, Fielding K, Hanna D, et al. A patient-level pooled analysis of treatment-shortening regimens for drug-susceptible pulmonary tuberculosis. Nature Medicine. 2018;24(11):1708-15.

8. Collaborative Group for the Meta-Analysis of Individual Patient Data in MDR-TB treatment, Ahmad N, Ahuja SD, Akkerman OW, Alffenaar JC, Anderson LF, et al. Treatment correlates of successful outcomes in pulmonary multidrug-resistant tuberculosis: an individual patient data meta-analysis. Lancet. 2018;392(10150):821-34.
9. 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.
10. Walzl G, McNerney R, du Plessis N, Bates M, McHugh TD, Chegou NN, et al. Tuberculosis: advances and challenges in development of new diagnostics and biomarkers. Lancet Infect Dis. 2018;18(7):e199-e210.
11. John J, Phillips P, Hills N. Primary efficacy outcomes of TB phase IIc and phase III clinical trials: A systematic review. PROSPERO 2020 CRD42020197993 2020 [Available from: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42020197993.
12. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative research in psychology. 2006;3(2):77-101.
13. Perrin FM, Breen RA, McHugh TD, Gillespie SH, Lipman MC. Are patients on treatment for pulmonary TB who stop expectorating sputum genuinely culture negative? Thorax. 2009;64(11):1009-10.
14. U.S. Department of Health and Human Services FaDA, Center for Drug Evaluation and Research (CDER),. Guidance for Industry. Pulmonary Tuberculosis: Developing Drugs for Treatment, Draft Guidance: U.S. Department of Health and Human Services , Food and Drug Administration, Center for Drug Evaluation and Research (CDER); 2013.
15. European Medicines Agency. Addendum to the guideline on the evaluation of medicinal products indicated for treatment of bacterial infections to address the clinical development of new agents to treat pulmonary disease due to Mycobacterium tuberculosis: European Medicines Agency; 2017.
16. Bonnett LJ, Ken-Dror G, Davies GR. Quality of reporting of outcomes in phase III studies of pulmonary tuberculosis: a systematic review. Trials. 2018;19(1):134.
17. Phillips PPJ, Van Deun A, Ahmed S, Goodall RL, Meredith SK, Conradie F, et al. Investigation of the efficacy of the short regimen for rifampicin-resistant TB from the STREAM trial. BMC Med. 2020;18(1):314.
18. Meyvisch P, Alonso A, Van der Elst W, Molenberghs G. On the relationship between association and surrogacy when both the surrogate and true endpoint are binary outcomes. Stat Med. 2020;39(26):3867-78.
19. Benator D, Bhattacharya M, Bozeman L, Burman W, Cantazaro A, Chaisson R, et al. Rifapentine and isoniazid once a week versus rifampicin and isoniazid twice a week for treatment of drug-susceptible pulmonary tuberculosis in HIV-negative patients: a randomised clinical trial. Lancet. 2002;360(9332):528-34.
20. el-Sadr WM, Perlman DC, Matts JP, Nelson ET, Cohn DL, Salomon N, et al. Evaluation of an intensive intermittent-induction regimen and duration of short-course treatment for human immunodeficiency

virus-related pulmonary tuberculosis. Terry Beirn Community Programs for Clinical Research on AIDS (CPCRA) and the AIDS Clinical Trials Group (ACTG). Clin Infect Dis. 1998;26(5):1148-58.

21. Lienhardt C, Cook SV, Burgos M, Yorke-Edwards V, Rigouts L, Anyo G, et al. Efficacy and Safety of a 4-Drug Fixed-Dose Combination Regimen Compared With Separate Drugs for Treatment of Pulmonary Tuberculosis. JAMA: The Journal of the American Medical Association. 2011;305(14):1415-23.
22. 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.
23. Johnson JL, Hadad DJ, Dietze R, Maciel EL, Sewali B, Gitta P, et al. Shortening treatment in adults with noncavitary tuberculosis and 2-month culture conversion. Am J Respir Crit Care Med. 2009;180(6):558-63.
24. Velayutham B, Jawahar MS, Nair D, Navaneethapandian P, Ponnuraja C, Chandrasekaran K, et al. 4-month moxifloxacin containing regimens in the treatment of patients with sputum positive pulmonary tuberculosis in South India - a randomized clinical trial. Trop Med Int Health. 2020.
25. 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.
26. 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.
27. Gopalan N, Santhanakrishnan RK, Palaniappan AN, Menon PA, Lakshman S, Chandrasekaran P, et al. Daily vs Intermittent Antituberculosis Therapy for Pulmonary Tuberculosis in Patients With HIV: A Randomized Clinical Trial. JAMA Intern Med. 2018;178(4):485-93.
28. Jindani A, Nunn AJ, Enarson DA. Two 8-month regimens of chemotherapy for treatment of newly diagnosed pulmonary tuberculosis: international multicentre randomised trial. Lancet. 2004;364(9441):1244-51.
29. Jawahar MS, Banurekha VV, Paramasivan CN, Rahman F, Ramachandran R, Venkatesan P, et al. Randomized clinical trial of thrice-weekly 4-month moxifloxacin or gatifloxacin containing regimens in the treatment of new sputum positive pulmonary tuberculosis patients. PLoS One. 2013;8(7):e67030.
30. Decroo T, de Jong BC, Piubello A, Souleymane MB, Lynen L, Van Deun A. High-Dose First-Line Treatment Regimen for Recurrent Rifampicin-Susceptible Tuberculosis. Am J Respir Crit Care Med. 2020;201(12):1578-9.
31. Nunn AJ, Phillips PPJ, Meredith SK, Chiang CY, Conradie F, Dalai D, et al. A Trial of a Shorter Regimen for Rifampin-Resistant Tuberculosis. N Engl J Med. 2019;380(13):1201-13.
32. von Groote-Bidlingmaier F, Patientia R, Sanchez E, Balanag V, Jr., Ticona E, Segura P, et al. Efficacy and safety of delamanid in combination with an optimised background regimen for treatment of multidrug-resistant tuberculosis: a multicentre, randomised, double-blind, placebo-controlled, parallel group phase 3 trial. Lancet Respir Med. 2019;7(3):249-59.
33. Conradie F, Diacon AH, Ngubane N, Howell P, Everitt D, Crook AM, et al. Treatment of Highly Drug-Resistant Pulmonary Tuberculosis. N Engl J Med. 2020;382(10):893-902.

Tables

Tables 1-2 and S1 are available in the Supplmentary Files.

Figures

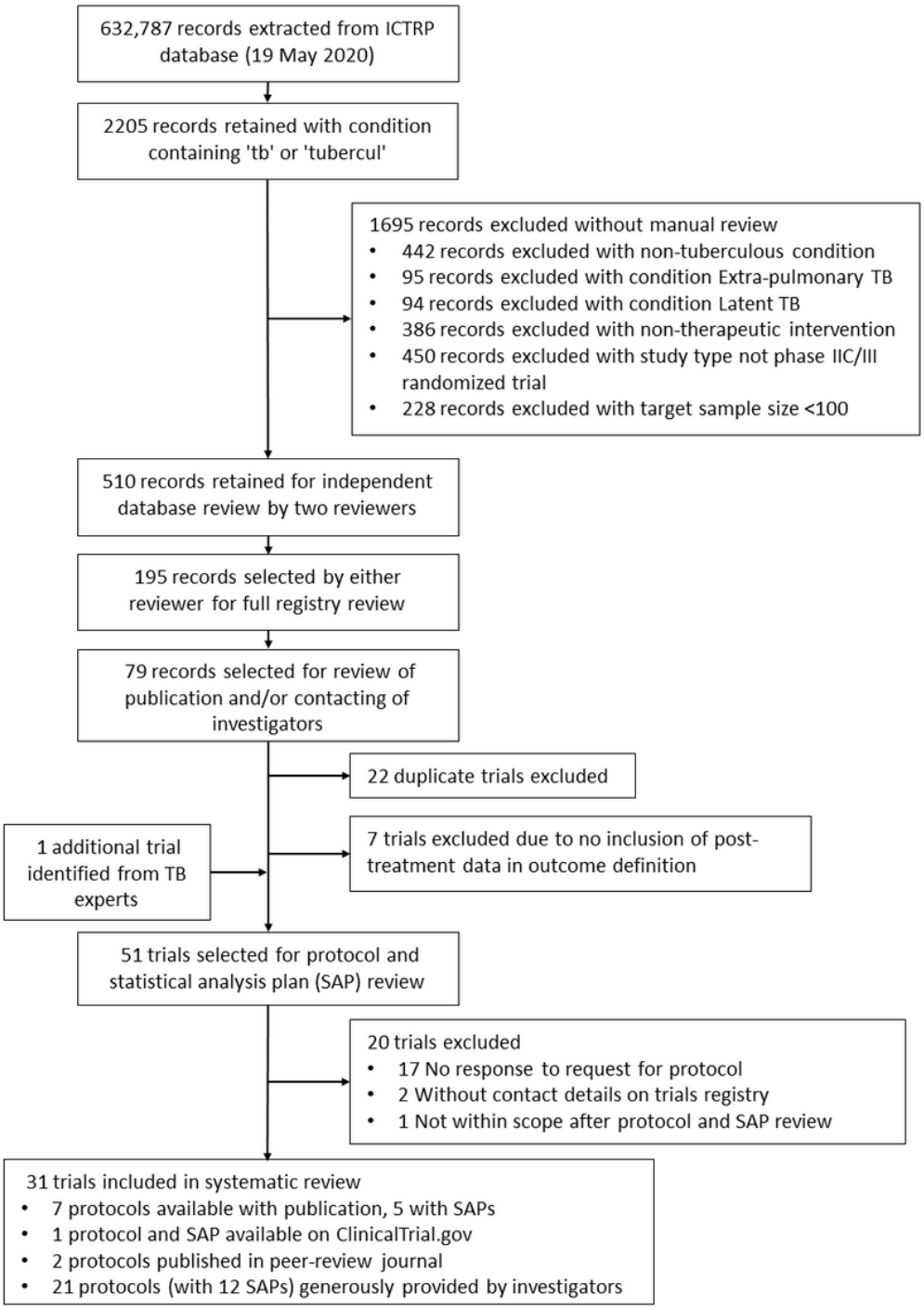

Figure 1

PRISMA flow diagram

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Tables.docx
- PRISMAchecklistTBendpoints.docx