STREAM

The evaluation of a standard streatment regimen of anti-tuberculosis drugs for patients with MDR-TB

ISRCTN 18148631 (Stage 2)

CORE STAGE 2 STATISTICAL ANALYSIS PLAN

Version 6.0 | 9th September 2022

Authorised by:

Name:	Prof. Andrew Nunn
Role:	Co-Chief Investigator
Signature:	Date:
Name:	Dr. Sarah Meredith
Role:	Co-Chief Investigator
Signature:	Date:
Name:	Dr. Ruth Goodall
Role:	Trial Statistician
Signature:	Date:
Name:	Dr. Katleen Callewaert
Role:	Janssen Statistician
Signature:	Date:

Version history:

Version	Date	Who	Comments
0.1	30th Sep 2014	Patrick Phillips	Based on protocol version 5.30
0.2	15th Oct 2014	Patrick Phillips Tom Godec	Expanded first draft
0.3	28th Nov 2014	Ben Van Baelen Els De Paepe	Comments from Janssen
0.4	28th Nov 2014	Patrick Phillips	Incorporation of comments from Janssen
0.5	4th Dec 2014	Ben Van Baelen Els De Paepe	Comments from Janssen
1.0	12th Dec 2014	Patrick Phillips	Up-versioned as final
1.1	20th Nov 2020	Ruth Goodall	Update on basis of protocol version 11.0 - Update trial description, objectives, sample size, definition of favourable, in line with protocol - Add randomisation protocol as a stratification factor - Extension of wk76/132 analysis windows for COVID - Use of Ogawa (or LJ if missing) culture results in analysis; unable to produce sputum=negative culture result - Primary outcome: add subgroups in line with stage 1, sensitivity analyses, Bayesian analysis, estimands notation - Add/update analyses: time to failure or relapse, favourable outcome at wk132, PK analysis in line with IDMC report, time to cessation of clinical symptoms, WHO classification of outcome, TB outcome irrespective of treatment changes - Statement on multiple testing
1.2	17th Dec 2020	Ruth Goodall	Combined comments from Co-PIs and statistician
1.3	18th Dec 2020	Ruth Goodall	Responses and changes based on comments from Co-PIs and statistician after joint meeting - Clarified window extension for COVID - Add sensitivity analysis for missing COVID results at wk76 - Clarified identification done at ITM - Independent review of FoR classification - Events with missing event date allocated to earliest study phase only - Independent review of cause of death
1.4	22nd Dec 2020	Ruth Goodall	Cleaned version circulated to Angela Crook and Janssen
1.5	5th Feb 2021	Ruth Goodall	Combined comments from Angela Crook and Janssen, with responses from Ruth Goodall - Clarification that stratification will prioritise randomisation protocol if small numbers - Update to estimands notation for primary endpoint - Culture with no identification and <14 days incubation will be excluded (was 7) - mPP population added for sensitivity analysis of primary endpoint

			- mortality further sensitivity analyses and PK/PD analyses removed, will be in extensive SAP only
1.6	1st Apr 2021	Ruth Goodall	Final comments from Janssen, with responses from Ruth Goodall after joint meeting. - changes to definition of TEAEs to align with Extensive SAP.
1.7	13th May 2021	Ruth Goodall	Changes after discussion of both SAPs with Janssen. - Clarification of time point for FoR analysis - Not excluding non-assessables from time to unfavourable analysis at Wk132 - New WHO definition of treatment outcomes - Clarification of analysis populations
2.0	14th May 2021	Ruth Goodall	Up-versioned as final
2.1	22nd Sept 2021	Ruth Goodall	Clarification on wording after programming of primary analyses - Word change : Failure or *recurrence* (not relapse) - Word change: *race* not ethnicity - Clarification that randomisation date = day 1 when assigning analysis windows - Clarification that genotypic DST will be used for determination of the mITT population when phenotypic DST is unavailable.
3.0	24th Sept 2021	Ruth Goodall	Up-versioned as final (after review by Janssen)
3.1	19th Nov 2021	Ruth Goodall	Clarification of wording after double programming.
4.0	22nd Nov 2021	Ruth Goodall	Up-versioned as final (after review by Janssen)
4.1	8th April 2022	Ruth Goodall	Clarifications in response to feedback from FDA and PPD
5.0	21st April 2022	Ruth Goodall	Up-versioned as final (after review by Janssen and CIs)
5.1	19th August 2022	Ruth Goodall	Clarifications in response to Janssen double programming exercise plus addition of comparison of long-term efficacy and safety for regimen C vs regimen D.
5.2	6th September 2022	Ruth Goodall	Incorporating comments from MRC CTU (co-PIs and statisticians) for circulation to Janssen for review.
6.0	9th September 2022	Ruth Goodall	Incorporating additional comments from Janssen.

GENERAL INFORMATION

This document describes and substantiates the statistical principles and methods used for the analysis of data from Stage 2 of the STREAM trial. This document is designed to support the STREAM protocol but this Statistical Analysis Plan (SAP) supersedes the SAP as written in the protocol. Every care was taken in the drafting of this SAP, but corrections or amendments may be necessary. A pre-final version (v1.0) of this SAP was signed off before first patient first visit. The final version of the SAP will be signed off before database lock for the primary Week 76 Stage 2 analysis.

The STREAM trial consists of two stages. Stage 1 involves the comparison of two treatment regimens: Regimen A and Regimen B. Stage 2 involves two additional regimens, Regimen C and Regimen D, and makes the comparison between Regimen B and Regimen C for the analysis of the primary endpoint. All treatment regimens are described in detail in the STREAM protocol, Section 2.1.3. Stage 1 and Stage 2 of the STREAM trial each have separate SAPs listed below. Each SAP has differences, but the fundamental statistical principles will be consistent across all SAPs.

Document	Description
Stage 1 SAP	All analyses relating to stage 1
Core Stage 2 SAP	Core analyses for stage 2
Extensive Stage 2 SAP	Expanded analyses for stage 2 supporting CSR

Compliance:
The trial will be conducted in accordance with the ethical principles that have their origin in the Declaration of Helsinki, in accordance with the principles of Good Clinical Practice (GCP) as laid down by the ICH topic E6 (Note for Guidance on GCP), and the applicable regulatory requirements in the participating countries.

Sponsor:
Vital Strategies, f/k/a The International Union Against Tuberculosis and Lung Disease, Inc.
100 Broadway, Fourth Floor, New York, NY 10005 USA
Tel (main): +1 212 500 5720
Fax: +1 212 480 6040
Email: STREAM@vitalstrategies.org

Funders:
The United States Agency for International Development (USAID)
UK Medical Research Council (MRC) / Department for International Development (DFID)
Janssen Research & Development, LLP (Stage 2 only)

Main Contacts:

Co-Chief Investigator
Prof. Andrew Nunn
MRC Clinical Trials Unit at UCL
Institute of Clinical Trials and Methodology
90 High Holborn, London, WC1V 6LJ
Tel: +44 (0)20 7670 4703
Fax: +44 (0)20 7670 4829
Email: a.nunn@ucl.ac.uk

Co-Chief Investigator
Prof. Sarah Meredith
MRC Clinical Trials Unit at UCL
Institute of Clinical Trials and Methodology
90 High Holborn, London, WC1V 6LJ
Tel: +44 (0)20 7670 4703
Fax: +44 (0)20 7670 4829
Email: s.meredith@ucl.ac.uk

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS AND GLOSSARY

AE	Adverse Event
AFB	Acid Fast Bacilli
AR	Adverse Reaction
AST	Aspartate aminotransferase
ALT	Alanine aminotransferase
BDQ	Bedaquiline
CI	Chief Investigator
DAIDS	Division of Acquired Immunodeficiency Syndrome
DOT	Directly Observed Treatment
DST	Drug Susceptibility Test
ECG	Electrocardiogram
FDA	Fluorescein diacetate staining
GCP	Good Clinical Practice
HIV	Human Immunodeficiency Virus
IDMC	Independent Data Monitoring Committee
ISRCTN	International Standard Randomised Controlled Trial Number
ITM	Institute of Tropical Medicine
ITT	Intention To Treat
LFX	Levofloxacin
LPA	Line Probe Assay
M2	Metabolite 2
MDR	Multi-Drug Resistant
MIC	Minimal Inhibitory Concentration
MRC CTU	Medical Research Council Clinical Trials Unit
NE	Notable Event
NTP	National Tuberculosis Programme
PK	Pharmacokinetics
PI	Principal Investigator
QT Interval	A measure of time between the start of the Q wave and the end of the T wave in the ECG complex
QTc	QT interval corrected for heart rate
QTcF	QT interval corrected for heart rate using the Fridericia correction
SAE	Serious Adverse Event
SAR	Serious Adverse Reaction
STREAM	The Evaluation of a Standardised Treatment Regimen of Anti-Tuberculosis Drugs for Patients with MDR-TB
SUSAR	Suspected Unexpected Serious Adverse Reaction
TB	Tuberculosis
ULN	Upper limit of normal
USAID	United States Agency For International Development
WHO	World Health Organisation
XDR	Extensively Drug Resistant
ZN	Ziehl-Neelsen

Note. In this statistical analysis plan, time (in weeks) refers to the time from randomisation, e.g. Week 76 refers to 76 weeks from randomisation.

1 TRIAL OVERVIEW

1.1 Study design

The STREAM study is an international, multi-centre, parallel-group, open-label, randomised, controlled trial studying participants with rifampicin-resistant tuberculosis (MDR-TB).

1.2 Trial objectives

The primary objective of Stage 2 of the STREAM trial is:

> To assess whether the proportion of participants with a favourable efficacy outcome on Regimen C, the fully oral regimen, is non-inferior to that on Regimen B at Week 76, using a 10% margin of non-inferiority

The secondary objectives of the Stage 2 comparison of the STREAM trial are:

1. To assess whether Regimen C is non-inferior to Regimen B with regards to the proportion of participants with a favourable efficacy outcome at Week 132
2. To assess whether Regimen C is superior to Regimen B with regards to the proportion of participants with a favourable efficacy outcome at Week 76 and Week 132 (if non-inferiority is demonstrated at either time-point)
3. To compare the efficacy of 40 weeks of bedaquiline in combination with the other drugs of Regimen C with Regimen B during treatment and follow-up
4. To compare the efficacy of 28 weeks of bedaquiline in combination with the other drugs of Regimen D with Regimen B during treatment and follow-up
5. To estimate the difference between Regimen C and Regimen B_{mox} in the proportion of participants with a favourable efficacy outcome at Week 76 and Week 132
6. To estimate the difference between Regimen C and Regimen B_{lev} in the proportion of participants with a favourable efficacy outcome at Week 76 and Week 132
7. To estimate the difference between Regimen D and Regimen B in the proportion of participants with a favourable efficacy outcome at Week 76 and Week 132
8. To estimate the difference between Regimen B and Regimen A in the proportion of participants with a favourable efficacy outcome at Week 132
9. To estimate the difference between Regimen C and Regimen A in the proportion of participants with a favourable efficacy outcome at Week 132
10. To estimate the difference between Regimen D and Regimen A in the proportion of participants with a favourable efficacy outcome at Week 132
11. To investigate the safety, including the effect on mortality and tolerability of 40 weeks of bedaquiline in combination with the other drugs of Regimen C compared to Regimen B during treatment and follow-up
12. To investigate the safety, including the effect on mortality and tolerability of 28 weeks of bedaquiline in combination with the other drugs of Regimen D compared to Regimen B during treatment and follow-up
13. To compare the proportion of patients who experience grade 3 or greater adverse events during treatment or follow-up in Regimen B as compared to Regimen A
14. To compare the proportion of patients who experience grade 3 or greater adverse events during treatment or follow-up in Regimen C as compared to Regimen B
15. To compare the proportion of patients who experience grade 3 or greater adverse events during treatment or follow-up in Regimen C as compared to Regimen B_{mox}

16. To compare the proportion of patients who experience grade 3 or greater adverse events during treatment or follow-up in Regimen C as compared to Regimen B_{lev}
17. To compare the proportion of patients who experience grade 3 or greater adverse events during treatment or follow-up in Regimen D as compared to Regimen B
18. To investigate the safety, including the effect on mortality and tolerability, of bedaquiline-containing regimens compared to Regimen B during treatment and follow-up.
19. To investigate the effect on mortality of bedaquiline-containing regimens compared to non-bedaquiline containing regimens
20. To evaluate the pharmacokinetics of bedaquiline and M2 in all participants randomised to Regimen C or Regimen D at sites selected for the PK study and assess pharmacokinetic/pharmacodynamics relationships of bedaquiline for safety and efficacy
21. To evaluate the pharmacokinetics of bedaquiline and M2 in a subset of HIV co-infected patients on Regimen C or Regimen D receiving antiretroviral treatment
22. To evaluate the 4β-hydroxycholesterol/cholesterol ratio as a measure of cytochrome P450 3A (CYP3A) activity in patients on Regimen C and Regimen D at sites selected for the PK study.
23. To compare the economic costs incurred during treatment by patients (transport and food costs for attending DOTs and patient assessment visits, food supplements costs and income loss) and by the health system (inpatient stay, laboratory tests, medication, staff, consumables and serious adverse events costs) in Regimen B and C. To calculate economic costs associated with regimen D, and compare these with regimen B, for those sites where this is possible
24. To compare the proportions of patients having undergone lung surgery (resection or pneumonectomy) by Week 76 and Week 132 in Regimen C and Regimen D as compared to Regimen B
25. To compare the development of resistance to background drugs, especially resistance leading to the development of pre-XDR or XDR strains of TB in Regimen C and Regimen D as compared to Regimen B
26. To investigate the development of increased MIC to bedaquiline in Regimen C and Regimen D

1.3 Patient eligibility criteria

Patient eligibility criteria are listed in Section 5 of the protocol.

1.4 Study interventions

In Stage 2 of the STREAM trial, the primary comparison being made is between Regimen C and Regimen B, with secondary comparisons between Regimen D and Regimen B. In addition, a limited number of secondary analyses will make comparisons between Regimen A and the other regimens.

Regimen A: The locally-used MDR-TB regimen in accordance with 2011 WHO MDR-TB treatment guidelines[1].

Regimen B: Regimen B forms the control regimen for the primary and secondary comparisons, and is based on the regimen described by Van Deun 2010[2] (updated results[3]) consisting at the start of STREAM of clofazimine, ethambutol, moxifloxacin, and pyrazinamide given for 40 weeks, supplemented by isoniazid, kanamycin, and prothionamide for the first 16 weeks; this combination is referred to as Regimen B_{mox}. With Version 8.0 of the protocol Regimen B is modified by replacement of moxifloxacin with levofloxacin (referred to as Regimen B_{lev}). Regimen B without specification of which fluoroquinolone is in the regimen

refers to either (B_{mox} or B_{lev}). With Version 10.0 of the protocol, Regimen B is further modified to replace kanamycin with amikacin in sites where:

- the national TB program has replaced kanamycin with amikacin in line with revised WHO guidance[4] and
- amikacin is available for the trial.

Kanamycin will continue to be used until the transition takes place within each country programme.

Regimen C: Regimen C is an all-oral modified version of Regimen B, consisting of bedaquiline, clofazimine, ethambutol, levofloxacin, and pyrazinamide given for 40 weeks, supplemented by isoniazid, and prothionamide for the first 16 weeks.

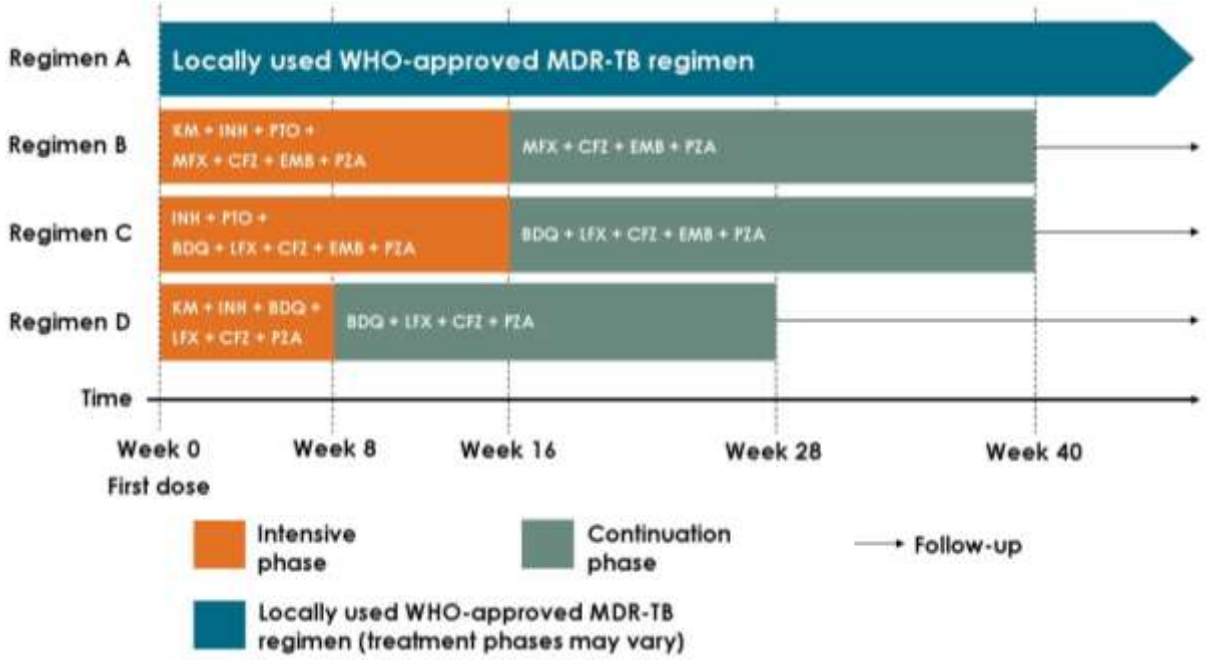
Regimen D: Regimen D a shortened modified version of Regimen B, consisting of bedaquiline, clofazimine, levofloxacin, and pyrazinamide given for 28 weeks, supplemented by isoniazid, and kanamycin for the first 8 weeks.

All patients in Stage 2 of the study will be followed up to Week 132. The primary analysis will be based on data accrued to Week 76; the data accrued to the date when the last patient is projected to have reached Week 96 will be used for the secondary long-term efficacy analysis; all data accrued to Week 132 will be used in the long-term safety analysis.

1.5 Treatment phases

Regimen B, Regimen C, and Regimen D each consist of 2 phases; an intense phase followed by a continuation phase, as shown in Figure 1.

Figure 1: Regimen treatment phases



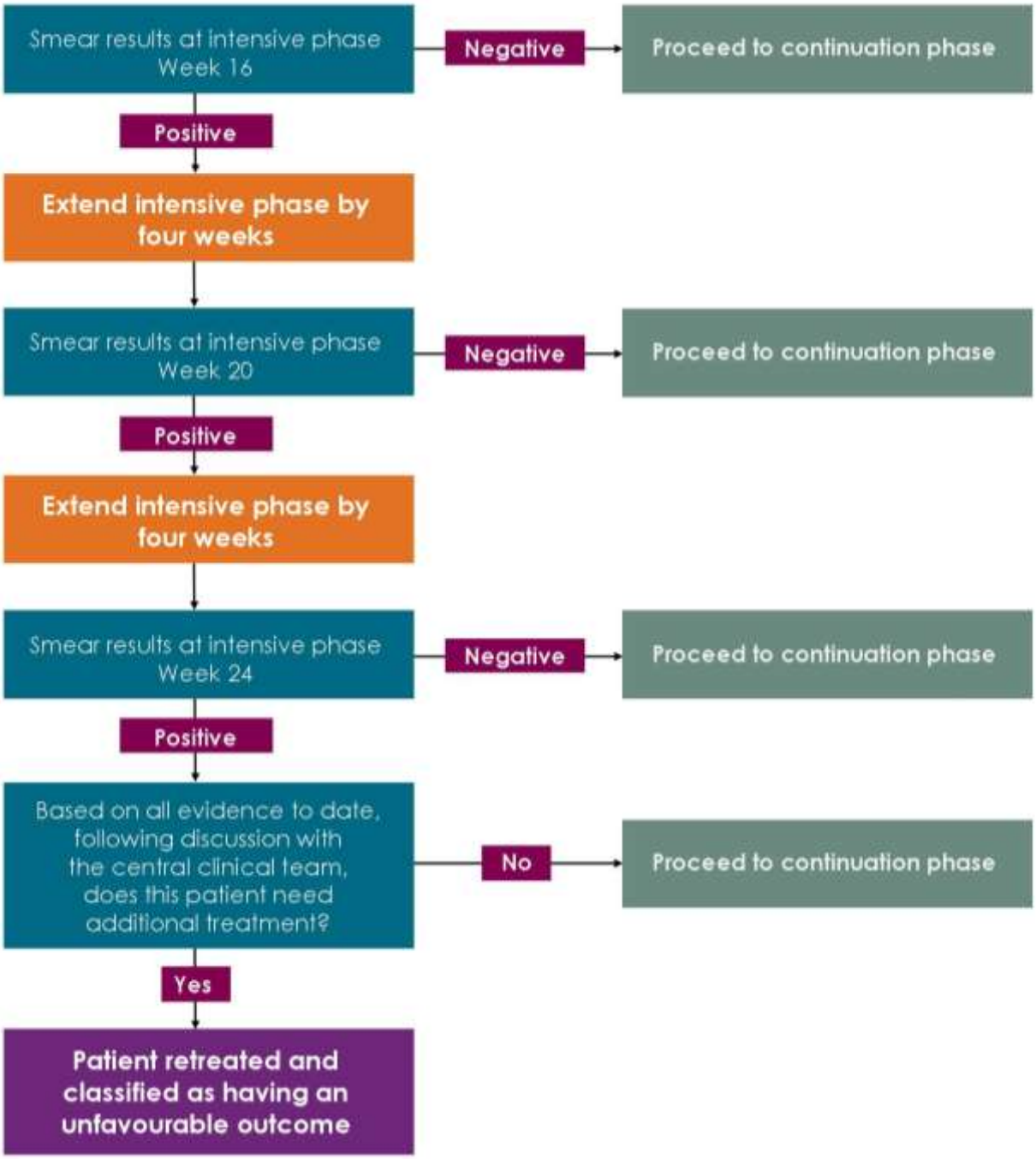
For patients randomised to Regimen B or Regimen C the algorithm described in Figure 2 will be used to determine when a patient can proceed from the intensive to the continuation phase.

Patients randomised to Regimen B or Regimen C will receive 40 weeks of treatment (16 weeks intensive phase plus 24 weeks continuation phase). In the event of positive (at least "scanty" on the IUATLD/WHO scale) AFB smear, the drugs in the intensive phase of these regimens may be extended by 4 weeks twice, allowing a maximum total duration of 48 weeks treatment (except for bedaquiline, which will be given for a maximum duration of 40 weeks regardless of whether the regimen is extended).

Patients randomised to Regimen D will receive 28 weeks of treatment (eight weeks intensive phase plus 20 weeks continuation phase). As Regimen D has shorter intensive phase duration than Regimen B and Regimen C, it is expected that more patients would have a smear positive result at the end of the intensive phase due to the shorter time that patients will have been on treatment to that point. Therefore, the less stringent criterion of a smear positive result of 2+ is sufficient for patients to require an extension of the intense period, i.e. patients can have a positive smear of 1+ and still advance on to the continuation phase, as opposed to Regimen B and Regimen C for which any positive smear result would result in an extension to the intensive phase. In the event of a 2+ or more positive smear, the drugs in the intensive phase of the regimen may be extended by 4 weeks twice, allowing a maximum total duration of 36 weeks treatment (Figure 3).

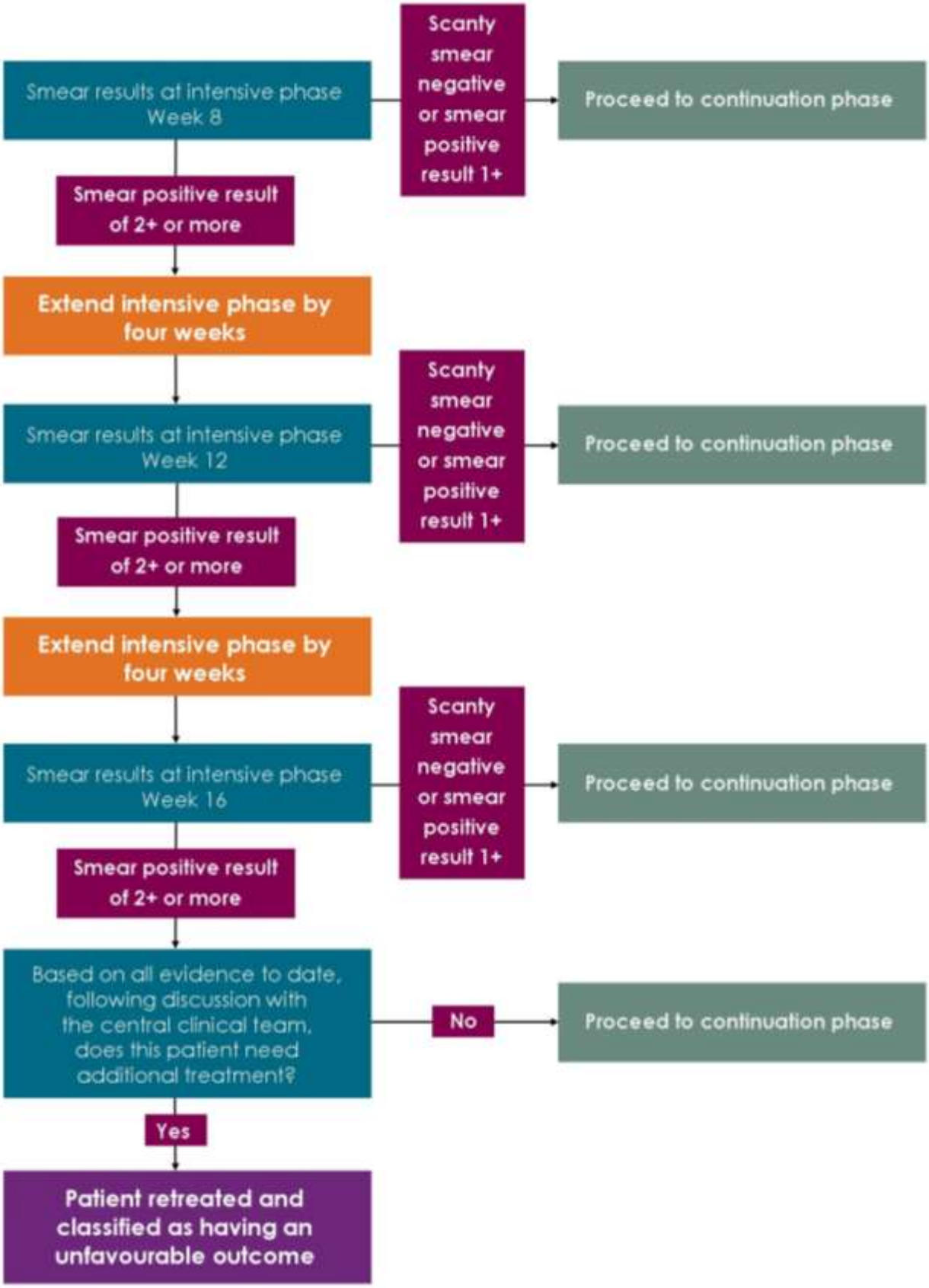
The procedure for transition from the intensive to the continuation phase in Regimen A will be according to local policy.

Figure 2: Transition from intensive to continuation phase for patients on Regimen B and Regimen C



Note: smear results based on regular AFB ZN or auramine staining and not FDA vital staining.

Figure 3: Transition from intensive to continuation phase for patients on Regimen D



1.6 Randomisation procedure

Patients were initially randomised to Regimen A, Regimen B, Regimen C, or Regimen D, in a ratio 1:2:2:2, respectively. From protocol version 6.3 onwards randomisation to Regimen A and/or Regimen D were restricted or stopped (Table 1).

Randomisation will be stratified by (1) site, (2) HIV status (split into three strata: HIV negative, HIV positive with CD4 count ≥ 350 cells/mm^3, and HIV positive with CD4 count < 350 cells/mm^3).

Table 1: Randomisation according to protocol version

Protocol version	Allocation ratio	Regimens
6.2	1:2:2:2	A:B:C:D
6.3 or 7.0	1:1:1	B:C:D in countries where the local NTP adopted or has imminent plans to adopt the WHO 2016 short regimen; as for v6.0 elsewhere
8.0 onwards	1:1	B:C

All analyses will be stratified by randomisation protocol to ensure comparisons are made to concurrent controls:

- comparisons of Regimen C to Regimen B will include all participants allocated to these regimens but be stratified by a 3 level variable : v6.2 vs v6.3/7.0 vs 8.0/10.0
- comparisons of Regimen D to Regimen B will be restricted to participants randomised under protocol v6.2/6.3/7.0 and be stratified by a 2 level variable : v6.2 vs v6.3/7.0
- comparisons to Regimen A will be restricted to participants randomised under v6.2 of the protocol

Analyses of the primary endpoint will also be stratified by HIV status as defined above, unless there is insufficient data within strata. If this is the case initially the two HIV positive groups will be combined, but if there is still insufficient data within strata then stratification by randomisation protocol will be prioritised.

2 SAMPLE SIZE

2.1 Power to demonstrate non-inferiority in the primary efficacy outcome (primary objective relating to Regimens B and C)

Stage 2 aimed to randomise at least 200 patients to each of Regimen B (B_{mox} or B_{lev}) and Regimen C. This revised sample size in protocol version 8.0 was determined based on the assumption that the proportion of patients with a favourable efficacy outcome at Week 76 is 80% for Regimen B (estimated based on preliminary Stage 1 results) and 82% for Regimen C (based on an anticipated minimum benefit in efficacy of using 40 weeks treatment with bedaquiline compared to 16 weeks treatment with kanamycin). Using a non-inferiority margin of 10% and a one-sided significance level of 2.5%, 172 evaluable patients will be required in each of the two regimens to demonstrate non-inferiority of Regimen C to Regimen B with 80% power. To account for 14% of patients excluded from the primary per protocol efficacy analysis population, a total of 400 patients will need to be enrolled across the 2 regimens.

A sample size of 400 patients enrolled to the comparison of Regimens B and C is necessary to rule out an 8% increased mortality in Regimen C compared to the 8% observed on Regimen B in the preliminary analysis of STREAM Stage 1. This will provide 80% power with a one-sided 0.025 type I error.

3 PRIMARY OUTCOME

3.1 Primary analysis Week 76 window

The Week 76 window is defined as the time period from six weeks before 76 weeks since randomisation to six weeks after 76 weeks since randomisation, i.e. from Week 70 to Week 82.

3.1.1 Extension of visit window due to the COVID-19 pandemic

The end dates of the Week 76 window will be extended for any Stage 2 patients whose Week 76 appointment is scheduled to occur during the COVID-19 pandemic and did not occur due to restrictions on movement, unacceptable risk of exposure to COVID-19 in connection with the scheduled visit, or any other reason related to the pandemic. For those patients, the Week 76 sputum samples must be taken in a window beginning six weeks prior to the scheduled visit date and ending within the Week 84 visit window i.e. within 14 weeks of the scheduled Week 76 visit date.

3.2 Primary efficacy outcome

The primary efficacy outcome measure is the proportion of patients with a favourable outcome (as defined below) at Week 76.

Only data from events occurring before the end of the Week 76 window will be included in the primary analysis of the primary efficacy outcome.

Culture results obtained using acidified Ogawa (Kudoh medium) will be used in the primary analysis, unless missing, when results from Löwenstein-Jensen (LJ) media will be used.

A positive culture is defined as at least one colony and a negative culture is defined as absence of growth (no colonies).

Favourable

A patient's outcome will be classified as **favourable** if their last two culture results are negative unless they have previously been classified as unfavourable. These two cultures must be taken on separate visits (on different days); the latest of which being within the Week 76 window.

Patients that don't have a culture result within the Week 76 window because they were unable to produce sputum or their sample was contaminated, will be classified as favourable if their last two cultures before the Week 76 window are negative and they have not previously been classified as unfavourable; such patients will be identified separately in tables.

Participants with no sputum sample available at Week 76 due to Covid-19 restrictions who are not otherwise classified as unfavourable will be considered unfavourable. Sensitivity analyses will reclassify these participants as i) non-assessable and excluded from the primary efficacy analysis, and ii) favourable if they meet the definition of favourable above, with the latest of the 2 negative culture results being within the Week 68 window, and unfavourable otherwise.

Unfavourable
A patient's outcome will be classified as **unfavourable** in Stage 2 if:

1. They are discontinued from their allocated study treatment and subsequently restarted on a different MDR-TB regimen.
2. Treatment is extended beyond the scheduled end of treatment for any reason other than making up days when no treatment was given (missed treatment) for a maximum of eight weeks. A maximum of 14 days of extra treatment (irrespective of reason) is acceptable before it is classified as treatment extension.
3. They are restarted on any MDR-TB treatment after the scheduled end of treatment, but before 76 weeks after randomisation.
4. They change their allocated study treatment for any reason other than the replacement of a single drug.
5. Bedaquiline is started where the allocated regimen did not originally contain that drug (Regimen A or B).
6. A second line injectable agent is started in Regimen C.
7. A drug from the class of nitroimidazoles (delamanid or pretomanid) or linezolid is started.
8. They die at any point during treatment or follow-up.
9. At least one of their last two culture results, from specimens taken on separate occasions, is positive.
10. They do not have a culture result within the Week 76 window.

Starting a single drug other than bedaquiline (in Regimen A or B) or from the class of nitroimidazoles (delamanid or pretomanid) or linezolid (in any regimen) is not considered to be a substantial change to the regimen and therefore does not result in an unfavourable outcome, providing none of the other criteria above are met.

An extension of the intensive phase of treatment in any study arm does not constitute an unfavourable outcome, as long as the extension is in accordance with the algorithms described in Figure 5 and Figure 6 in Section 1.5 for patients on Regimen B, Regimen C, and Regimen D. Similarly, the discontinuation of drugs that are not replaced does not constitute an unfavourable outcome.

These definitions of favourable and unfavourable apply to all four treatment regimens and for analyses of both the mITT, PP and mPP analysis populations.

3.3 Primary outcome defined using estimands notation

We define 2 co-primary estimands that are the same in all components other than the handling of intercurrent events. Information on the estimator and estimation procedure are given in Section 6.

The estimands for the primary objective are defined as follows:
(a) the comparison is between Regimen C and Regimen B as defined in Section 1.4;
(b) the patient population of interest are patients with confirmed RR-TB that meet the inclusion/exclusion criteria as given in the trial protocol, Section 5 (see Section 4.4);
(c) the endpoint is favourable outcome at Week 76 as defined in Section 3.2;

(d) the population level summary is the difference in proportions of favourable outcome between treatment arms (See Section 6.1.1). Non-inferiority will be shown if the upper bound of the 2-sided 95% confidence interval of the difference in proportion of favourable outcomes at Week 76 between Regimens B and C is less than the 10% margin of non-inferiority;
(e) the intercurrent events will be handled as described below.

Intercurrent Event	**Strategy for Addressing Intercurrent Event for Primary Estimand 1**	**Strategy for Addressing Intercurrent Events for Primary Estimand 2**
Change of 1 non-major drug in the allocated treatment regimen	Treatment policy strategy: use primary variable regardless of whether or not this intercurrent event had occurred	Treatment policy strategy: use primary variable regardless of whether or not this intercurrent event had occurred
Treatment extension to make up for days when no treatment was given (up to 8 weeks)	Treatment policy strategy: use primary variable regardless of whether or not this intercurrent event had occurred	Treatment policy strategy: use primary variable regardless of whether or not this intercurrent event had occurred
Other changes to allocated treatment regimen (drug substitutions, other treatment extensions, use of salvage regimens)	Composite strategy: occurrence of this intercurrent event is captured in the primary composite variable definition, a participant with this intercurrent event is considered unfavourable	Composite strategy: occurrence of this intercurrent event is captured in the primary composite variable definition, a participant with this intercurrent event is considered unfavourable
Failure to complete a protocol adherent course of allocated treatment	Treatment policy: same as above	Principal Stratum: population is the subgroup of participants who would adhere to treatment, whichever treatment they are allocated
Reinfection	Composite strategy: same as above	Composite strategy: same as above
Death	Composite strategy: same as above	Composite strategy: same as above

4 ANALYSIS POPULATIONS

4.1 Intention-to-treat (ITT)

All randomised patients will be included in the ITT analysis population.

4.2 Safety population

All randomised patients that have taken at least one dose of treatment will be included in the safety analysis population.

4.3 Modified intention-to-treat (mITT)

The mITT population is defined as all randomised patients that have a positive culture for *M. tuberculosis* at screening or randomisation, with the exception of patients with isolates taken before randomisation that are subsequently found to be susceptible to rifampicin, and patients with isolates taken before randomisation that are subsequently found to be

resistant to both fluoroquinolones and second-line injectables (i.e. XDR-TB) on phenotypic DST. Results from the central reference laboratory will take priority over any results from local laboratories where available. Genotypic DST from the central reference laboratory will be used when phenotypic DST is unavailable. Rifampicin susceptible phenotypic DST results will be confirmed by rpoB sequencing. If a participant has rifampicin DST results at screening and randomisation which are discordant they will be considered resistant and included in the mITT population. Participants randomised in error i.e. late screening failures, will be excluded from the mITT population.

4.4 Per protocol (PP)

The PP population will be the same as the mITT population with the exclusion of patients not completing a protocol-adherent course of treatment, other than for treatment failure, change of treatment for an adverse event or death. Treatment failure is defined as failure to attain and maintain culture negativity until the end of allocated treatment.

4.4.1 Definition of a protocol-adherent course of treatment

A patient will have completed a protocol-adherent course of treatment when they have taken 80% of doses within 120% of the duration in both the intensive phase and in the whole treatment period. For this purpose, a dose is defined as all the study medications at the correct dose for that particular day.

For Regimen B and Regimen C, **with or without** an extension of the intensive phase, a patient will have completed a protocol-adherent course of treatment if they have taken 80% of doses within 120% weeks. For example, in participants with no treatment extensions:

- 90 doses (80% of 16 weeks) within 134 days (120% of 16 weeks) in the intensive phase, and
- 224 doses (80% of 40 weeks) within 336 days (120% of 40 weeks) over the whole treatment period (i.e. the combined intensive and continuation phases),

or in participants with a four week extension:

- 112 doses (80% of 20 weeks) within 168 days (120% of 20 weeks) in the intensive phase, and
- 246 doses (80% of 44 weeks) within 370 days (120% of 44 weeks) over the whole treatment period (i.e. the combined intensive and continuation phases).

For Regimen D, **with or without** an extension of the intensive phase, a patient will have completed a protocol-adherent course of treatment if they have taken 80% of doses within 120% weeks. For example, in participants with no treatment extensions:

- 45 doses (80% of 8 weeks) within 67 days (120% of 8 weeks) in the intensive phase, and
- 157 doses (80% of 28 weeks) within 235 days (120% of 28 weeks) over the whole treatment period (i.e. the combined intensive and continuation phases) with no treatment extensions.

or in participants with a four week extension:

- 67 doses (80% of 12 weeks) within 101 days (120% of 12 weeks) in the intensive phase, and
- 179 doses (80% of 32 weeks) within 269 days (120% of 32 weeks) over the whole treatment period (i.e. the combined intensive and continuation phases).

The same algorithm will apply for Regimen A, the control regimen; the exact number of doses and days depends on the duration of the intensive and continuation phases of Regimen A as implemented locally.

4.5 Modified per protocol (mPP)

Sensitivity analyses for analyses of favourable outcome at Week 76 and Week 132 will use the definition of PP population as specified in version 10.0 of the STREAM protocol, denoted the modified per-protocol population (mPP). Specifically, the mPP population will be the same as the mITT population with the exclusion of patients not completing a protocol-adherent course of treatment, other than for treatment failure or death.

5 GENERAL ANALYSIS PRINCIPLES

Analysis of the primary outcome is described in Section 6. Otherwise, descriptive statistics will be reported overall and by randomised group, and percentages will be of non-missing values, with the number (%) of non-missing values given if data are not complete. Statistical tests will be 2-sided and estimates will be presented with a 2-sided 95% CI. Appropriate transformations for all variables will be applied after inspection of the data. Percentages will be reported to 0 decimal places, unless <0.5% when they will be given to one decimal place. P-values will be given to 2 significant figures.

5.1 Analysis populations

The analyses of the primary outcome will be based on both the mITT and the PP populations for determining non-inferiority. All analyses will be based on the mITT population for determining superiority, with the PP population as a sensitivity analysis. All patients included in the analysis will be analysed in the treatment group to which they were originally randomised.

5.2 Treatment and follow-up phase definitions

For the purpose of analysis, the treatment and follow-up phases for an individual patient will be defined as follows:

- **Screening phase**
  - Start: date of screening consent
  - End: day before randomisation
- **Treatment phase**
  - Start: date of randomisation.
  - End: date of last dose of any TB treatment defined as last dose of any TB treatment (including retreatment for relapse), plus 7 days.
- **Follow-up phase**
  - Start: the day after the end of the treatment phase.
  - End: date of the last patient contact (scheduled or unscheduled visit, or other contact e.g. phone call).

The treatment phase includes any extension of treatment or retreatment, and so the Allocated Treatment phase is defined as follows:

- **Allocated Treatment phase**
  - Start: date of randomisation.
  - End: date of last dose of trial treatment defined as last dose of allocated regimen or last dose before the addition of a new drug, whichever happens sooner, plus 7 days.

5.3 Visit window definitions

In Stage 2, patients will be assessed at screening, randomisation (Week 0), Week 1, Week 2, Week 3, Week 4, after which they will be seen 4-weekly until Week 52, after which they will

be seen 8-weekly until Week 84, after which they will be seen 12-weekly until Week 132 post randomisation.

Baseline is defined as the date of randomisation (Study day 1). For all variables the baseline measurement will be the one taken at the randomisation visit. If this does not exist the measurement from the screening visit will be used if available. For ECG measurements, the mean of all available measurements from the randomisation visit will be used.

For the purpose of analysis, each scheduled visit will have a window before and after the target date, calculated from date of randomisation. When referring to a visit hereon, this implies within the defined visit window as specified below in Table 2.

Table 1: Visit window definitions

Visit	Target study day	Analysis window (study days)
Screening / Baseline	1	Date of screening consent – 1
Week 4	29	2-42
Week 8	57	43-70
Week 12	85	71-98
Week 16	113	99-126
Week 20	141	127 – 154
Week 24	169	155 – 182
Week 28	197	183 – 210
Week 32	225	211 – 238
Week 36	253	239 – 266
Week 40	281	267 – 294
Week 44	309	295 – 322
Week 48	337	323 – 350
Week 52	365	351 – 392
Week 60	421	393 – 448
Week 68	477	449 – 490
Week 76	533	**491 - 574***
Week 84	589	575 – 630
Week 96	673	631 – 714†
Week 108	757	715 – 798†
Week 120	841	799 – 882†
Week 132	925	883 – no upper bound†

NB. Study day 1 = date of randomisation
* see section 3.1.1
† see section 5.3.1

Any visit, scheduled or unscheduled, that falls into the analysis window will be assigned to that visit for the purpose of analysis. If two visits fall within the same interval, the one closest to the target date will be used for analyses by visit, so that there is only one unique visit for each patient and analysis time-point.

There are additional study visits at Weeks 1, 2 and 3 only for ECG monitoring. For the analysis of ECG data only, there will be additional visit windows: Week 1 (2-11), Week 2 (12-18), Week 3 (19-25) and the Week 4 visit window will be modified to (26-42).

5.3.1 Extension of visit window due to COVID

The end dates of the Week 76 window will be extended for any Stage 2 patients whose Week 76 appointment is scheduled to occur during the COVID-19 pandemic and did not occur due to restrictions on movement, unacceptable risk of exposure to COVID-19 in connection with the scheduled visit, or any other reason related to the pandemic. For those patients, the Week 76 sputum samples must be taken in a window beginning six weeks prior to the scheduled visit date and ending within the Week 84 visit window.

A participant's scheduled last efficacy visit will be their latest scheduled trial visit on or before the projected Week 96 visit of the last patient randomised. The visit window for the scheduled last efficacy visit is defined as no more than six weeks prior to and up to six weeks after (or on 30 Nov 2021 if this is earlier) the scheduled last efficacy visit. The end dates of the window will be extended for any Stage 2 patients whose scheduled last efficacy visit was due during the COVID-19 pandemic and did not occur due to restrictions on movement, unacceptable risk of exposure to COVID-19 in connection with the scheduled visit, or any other reason related to the pandemic. For such patients, sputum samples must be taken in a window between six weeks prior to and twelve weeks after the last scheduled visit date .

5.4 Definition of a culture result

A culture result will be called positive for *M. tuberculosis* if the culture tests positive for the presence of microorganisms, at least one colony, and the microorganisms present are then identified as being *M. tuberculosis*. Identification will be based on tests performed at the central laboratory (ITM). If an identification test is not carried out for a particular culture, then for analysis purposes a culture will still be considered positive for *M. tuberculosis* if the culture tests positive for the presence of microorganisms and if that culture result is obtained fourteen days or more since the start date of sputum processing and incubation of the inoculated Ogawa or Löwenstein-Jensen (LJ) media. If the culture result is obtained less than fourteen days since the start date of sputum processing and incubation, the culture result will not be considered as positive for *M. tuberculosis* (if the identification test is not carried out), and the culture result will be considered missing in the analysis.

Culture results obtained using acidified Ogawa (Kudoh medium) will be used in analysis if available; results from LJ media will be used if the Ogawa result is missing. The only exception to this is within the baseline window, when Ogawa, LJ or MGIT culture results will be considered to determine if a participant is positive at baseline.

Any culture result that is missing because the patient is no longer able to produce sputum will be treated as a negative result, providing their last 2 available culture results (from sputum samples taken at separate visits) are negative.

If more than one culture result is available from sputum collected on the same day, this will be regarded as a single culture result for the purposes of all analyses with the following overall result:

i. **Positive**, if at least one of the culture results is positive
ii. **Negative**, if at least one of the culture results is negative and none of the culture results are positive

iii. **Contaminated** if at least one of the culture results is contaminated and none of the culture results are positive or negative.
iv. **Missing**, if no culture result is available

5.5 Definition of culture negative status

Sputum culture negative status is defined as two consecutive negative cultures from sputa collected on different days without an intervening positive. The date of the first of these negative cultures is the date at which negative culture status is obtained. Culture negative status continues until there are two positive culture results at different visits without an intervening negative culture, or until there is a single positive culture not followed by two negative cultures. Culture negative status can be achieved at any point during treatment or follow-up but before any re-treatment. Culture negative status can be re-established.

5.6 Definition of a smear result

A smear result will be called positive if it is graded as 'scanty' or 'rare AFB' or at least 1+. If more than one smear result is available from sputum collected on the same day, this will be regarded as a single smear result for the purposes of all analyses with the following overall result:

i. **Positive**, if at least one of the smear results is positive
ii. **Negative**, if at least one of the smear results is negative and none of the smear results are positive
iii. **Missing**, if no smear result is available.

5.7 Reference laboratory bacteriology

All positive isolates from Week 8 onwards will be sent from the STREAM sites to a reference laboratory at the Institute of Tropical Medicine (ITM) in Antwerp, Belgium. Drug sensitivity results from the reference laboratory will be used in all analyses in preference to those obtained from local site laboratories where available. Identification of MTB will be done at ITM; local culture results will be excluded from analysis if identified as non-MTB.

5.8 Adverse events

For all analyses of adverse events (AE), only treatment-emergent AEs will be included. An AE will be considered treatment-emergent in a particular phase if it is worse in severity than the corresponding baseline observation. If the baseline observation is missing, the AE is always considered as treatment-emergent. For ECG, laboratory events, weight changes and hearing loss, only the maximum grade within phase will be reported. For laboratory and weight AEs, a shift from 'abnormally low' at baseline to 'abnormally high' post baseline (or vice versa) is also treatment-emergent. Laboratory AEs starting on day 1 will be considered treatment emergent since it is impossible to know whether the blood samples were taken before or after administration of the first dose of trial treatment and this is the most conservative approach.

SAEs will be analysed as episodes, with all components of the same clinical SAE presented as one episode. Analyses of grade 3 or 4 adverse events will consider each component as separate events. Hearing loss reported in the left and right ear as two separate AEs but with overlapping start and stop dates will be reported in analyses as one AE of bilateral hearing loss at the maximum grade of the two unilateral events.

6 ANALYSIS OF PRIMARY OUTCOMES

6.1 Primary efficacy analyses

6.1.1 Modelling technique used in analysis

For the primary efficacy analysis the difference in proportions of favourable outcome between Regimen B (B_{mox} and B_{lev} combined) and Regimen C with corresponding 95% confidence intervals and p-values will be estimated using a stratified analysis of the risk difference from each stratum using Cochran-Mantel-Haenszel weights.[5] The analysis will be stratified by randomisation protocol and HIV status (as defined in Section 1.6).

6.1.2 Non-inferiority of Regimen C compared to Regimen B at Week 76

Non-inferiority will be shown if the upper bound of the 95% confidence interval of the difference in proportion of favourable outcomes at Week 76 between Regimens B and C is less than the 10% margin of non-inferiority; this must be shown in both the mITT and PP populations. An sensitivity analysis will be undertaken using the mPP population.

6.1.3 Superiority of Regimen C compared to Regimen B at Week 76

If non-inferiority is demonstrated, then superiority of Regimen C compared to Regimen B will be declared if the upper bound of the 95% confidence interval of the difference in proportion of favourable outcomes at Week 76 between Regimens B and C is less than zero,. For this analysis, the mITT population will be primary and the PP and mPP populations will be sensitivity analyses.

Secondary analyses will repeat the primary analysis but compare Regimen B_{lev} and Regimen C, Regimen B_{mox} and Regimen C, and Regimen B and Regimen D, restricting the population in each case to include only concurrent controls within site. For this analysis, the mITT population will be primary and the PP population will be sensitivity analyses. Comparisons of Regimen C and Regimen D were carried out at the time of the primary analysis but considered exploratory.

6.2 Tabulation of primary endpoint classification

Since the primary endpoint is a composite of various components, the actual reason (component) for outcome will also be tabulated by treatment arm. Patients will be classified by the first event that made the patient unfavourable (see section 3.2) and further sub-classified by whether their outcome is deemed to be bacteriologically related or not, based on a review of all data up to the time that the outcome occurred (see Section 6.2.1).

6.3 Subgroup analyses

This primary efficacy analysis will be repeated (in both mITT and PP analysis populations) in subgroups according to the following baseline characteristics:

- HIV infection status
- baseline drug resistance patterns (i.e. resistance to pyrazinamide and isoniazid)
- albumin levels (by severity grade)
- BMI (<16.0 kg/m^2, 16.0-18.4 kg/m^2, 18.5-24.9 kg/m^2, >25 kg/m^2)
- cavitation (presence, absence)
- study centre/country
- age (<25, 25-<45, ≥45)

- sex
- Smoking history (current, ex, never)
- Smear grade
- Weight band (<33kg, 33-50kg, >50kg)
- Race.

6.4 Sensitivity analyses

The primary efficacy analysis will be repeated:

A. Unadjusted for any covariates except randomisation protocol.
B. Adjusted for randomisation stratification factors HIV status and centre. Small strata with fewer than 10 patients will be combined within geographical regions within country.
C. Adjusted for randomisation stratification factors and any additional important covariate such as cavitation at baseline, baseline bacillary load or adherence.

D. Ignoring substitutions of levofloxacin for moxifloxicin and vice versa when counting the number of drug substitutions within the definition of unfavourable outcome.
E. Ignoring starting linezolid alone as a criterion for an unfavourable outcome.
F. Reclassifying any participants with reinfection (based on genotype or DST) as "non-assessable" rather than unfavourable.
G. Reclassifying any participants who died as a result of trauma (as adjudicated by independent review) as "non-assessable" rather than unfavourable.
H. Reclassifying any participants with missing Week 76 culture results because of COVID-19 as "non-assessable" rather than unfavourable.
I. Reclassifying any participants with missing Week 76 culture results because of COVID-19 as favourable if they meet the definition of favourable, with the latest of the 2 negative culture results being within the Week 68 window, unfavourable otherwise.
J. Requiring a 25 day delay between the last 2 negative culture results in participants classified as favourable. A participant's outcome will be unfavourable if previously classified as favourable but not meeting this definition. . .

K. In the ITT analysis population
L. In the safety analysis population

Analyses D to J will be repeated adjusted only for randomisation protocol.

6.5 Bayesian analysis of non-inferiority

An exploratory Bayesian analysis of non-inferiority provides an estimate of the probability that Regimen B has efficacy not much worse than Regimen C for different thresholds of what might be considered 'not much worse'.

Following methods described previously[6], we will use Bayesian binomial regression to estimate the distribution of the (unadjusted) difference in the proportion of favourable outcomes between regimens B and C. Gaussian Normal priors will be placed on the intercept term (mean = 0.0 and variance = 100) and on the difference in proportion between regimens (Flat: mean = 0.0 and variance = 100, Sceptical: mean = 0.02 and variance 0.05, and Expected: mean = -0.02 and variance 0.05). The Flat prior is an uninformative prior

with very large variance centred around zero representing weak prior information, the Sceptical prior is centred around an absolute 2% increase in proportion of favourable in Regimen B with a smaller variance, and the Expected prior represents the assumptions used in the sample size calculations representing an absolute 2% increase in the proportion of favourable outcomes in regimen C. Initial values for the Markov Chain Monte Carlo algorithm came from estimates from the frequentist binomial regression model. This analysis will be on the mITT population.

7 ANALYSIS OF SECONDARY OUTCOMES

7.1 Efficacy outcomes

Secondary efficacy analyses will be carried out on i) data censored at Week 76, and ii) data from all visits occurring on or before the projected Week 96 visit of the last patient randomised i.e. approximately end of November 2021 (long-term efficacy data). A participant's scheduled last efficacy visit will be their latest scheduled trial visit on or before this date. In addition to the comparisons detailed below, analyses of the long-term efficacy data will include comparisons of Regimen C with Regimen D. Comparisons of Regimen C and Regimen D using data up to Week 76 were carried out at the time of the primary analysis but considered exploratory.

Secondary efficacy outcomes will be analysed on the mITT analysis population with the exception of the analysis of favourable outcome at Week 132 to determine the non-inferiority of Regimen C to Regimen B which will be analysed on the mITT, PP and mPP populations.

7.1.1 Time to sputum smear and culture conversion

Time to sputum smear conversion is defined as the time from randomisation to the first of two consecutive negative sputum results, collected on separate days. All patients in the analysis population will be included in this analysis, except those with no positive smear result at screening and randomisation. Patients that never achieve smear conversion will be censored at the date of collection of sputum that yielded their last smear result.

Time to sputum culture conversion is defined as the time from randomisation to the first of two consecutive negative culture results, collected on separate days. Patients that never achieve culture conversion will be censored at the date of collection of sputum that yielded their last culture result.

Median time to sputum smear and culture conversion will be calculated for Regimen B, Regimen C, and Regimen D using the Kaplan-Meier estimator.

A hazard ratio with corresponding two-sided 95% confidence intervals and p-value will be estimated for the comparison of Regimen B with Regimen C, and Regimen B with Regimen D.A Cox Proportional Hazards model will be used, adjusted for randomisation protocol (as defined in Section 1.6) and the stratification factor HIV status with the three categories of: HIV negative; HIV positive with CD4 count ≥350 cells/mm^3; and HIV positive with CD4 count <350 cells/mm^3.

The equality of survivor functions for time to sputum conversion for Regimen B and Regimen C, and Regimen B and Regimen D, will each be tested using a Log rank test, stratified by randomisation protocol (as defined in Section 1 .6) and the randomisation

stratification factor HIV status with the three categories of: HIV negative; HIV positive with CD4 count ≥350 cells/mm^3; and HIV positive with CD4 count <350 cells/mm^3.

The assumption of proportional hazards between the two comparisons will be tested using the proportional hazards test based on the Schoenfeld residuals after fitting the Cox Proportional Hazards models. In the case where there is adequate evidence that the proportional hazard assumptions are violated at the 5% level (i.e. $p<0.05$), methods where proportional hazards is not a necessary assumption will be used, such as the Wilcoxon test of equality of survivor functions and restricted mean survival time.

These analyses of time to sputum smear conversion and time to sputum culture conversion will be repeated with the alternative definition as time from randomisation to the first negative culture or smear result respectively (without the need for a second negative culture or smear result to confirm). These analyses will only be carried out on data up to Week 76.

7.1.2 Time to unfavourable efficacy outcome

Time to unfavourable efficacy outcome is defined as the time from randomisation to the first event that results in the definition of an unfavourable efficacy outcome for that patient (as defined in Section 3.2). Patients that do not culture convert during the treatment and follow-up phases (i.e. fail to have 2 consecutive culture negative results), and have not otherwise been called unfavourable, will be called unfavourable at the date of the last visit when a culture positive result was obtained. Patients classified as favourable or not assessable will be censored in this analysis at the date of collection of sputum that yielded their last negative culture result.

Median time to unfavourable efficacy outcome will be calculated for Regimen B, Regimen C, and Regimen D using the Kaplan-Meier estimator.

A hazard ratio with corresponding two-sided 95% confidence intervals and p-value will be estimated for the comparison of Regimen B with Regimen C, and Regimen B with Regimen D. A Cox Proportional Hazards model will be used, adjusted for randomisation protocol (as defined in Section 1.6) and the stratification factor HIV status with the three categories of: HIV negative; HIV positive with CD4 count ≥350 cells/mm^3; and HIV positive with CD4 count <350 cells/mm^3.

The equality of survivor functions for time to unfavourable efficacy outcome for Regimen B and Regimen C, and Regimen B and Regimen D, will each be tested using a Log rank test, stratified by randomisation protocol (as defined in Section 1.6) and the randomisation stratification factor HIV status with the three categories of: HIV negative; HIV positive with CD4 count ≥350 cells/mm^3; and HIV positive with CD4 count <350 cells/mm^3.

The assumption of proportional hazards between the two comparisons will be tested using the proportional hazards test based on the Schoenfeld residuals after fitting the Cox Proportional Hazards models. In the case where there is adequate evidence that the proportional hazard assumptions are violated at the 5% level (i.e. $p<0.05$), methods where proportional hazards is not a necessary assumption will be used, such as the Wilcoxon test of equality of survivor functions and restricted mean survival time.

7.1.3 Time to failure or recurrence

Favourable outcome for each participant will be re-classified according to the likelihood that it was a *Failure or Recurrence (FoR) event* on a five-point Likert scale: Definite, Probable, Possible, Unlikely, and Highly Unlikely[7].

An event will be considered *Highly Unlikely* to be a FoR event only if there is evidence of durable cure; i.e. the primary outcome classification of favourable which required completion of follow-up with negative cultures. A *Definite* FoR event requires clear bacteriological evidence of failure or recurrence (excluding a proven reinfection with exogenous strain of *M. tuberculosis)*, a *Probable* FoR event requires some evidence for failure or recurrence (clinical, bacteriological, or radiological) in the absence of clear bacteriology. The remaining participants will be classified as possible or unlikely based on their data. Time of the FoR event/censoring is the same as defined in the time to unfavourable outcome analysis (see Section 7.1.2).The FoR classification will reviewed by an independent clinician, blinded to treatment.

In the time to FoR analyses, the main groups of interest are those classified as having a Definite or Probable FoR event, with censoring of Possible, Unlikely and Highly Unlikely events at the time of the censoring event which met criteria for Unfavourable or Not Assessable in the primary analysis.

Median time to FoR will be calculated for Regimen B, Regimen C, and Regimen D using the Kaplan-Meier estimator.

A hazard ratio with corresponding two-sided 95% confidence intervals and p-value will be estimated for the comparison of Regimen B with Regimen C, and Regimen B with Regimen D. A Cox Proportional Hazards model will be used, adjusted for randomisation protocol (as defined in Section 1.6) and the stratification factor HIV status with the three categories of: HIV negative; HIV positive with CD4 count ≥350 cells/mm^3; and HIV positive with CD4 count <350 cells/mm^3.

The equality of survivor functions for time to failure or recurrence for Regimen B and Regimen C, and Regimen B and Regimen D, will each be tested using a Log rank test, stratified by randomisation protocol (as defined in Section 1.6) and the randomisation stratification factor HIV status with the three categories of: HIV negative; HIV positive with CD4 count ≥350 cells/mm^3; and HIV positive with CD4 count <350 cells/mm^3.

The assumption of proportional hazards between the two comparisons will be tested using the proportional hazards test based on the Schoenfeld residuals after fitting the Cox Proportional Hazards models. In the case where there is adequate evidence that the proportional hazard assumptions are violated at the 5% level (i.e. $p<0.05$), methods where proportional hazards is not a necessary assumption will be used, such as the Wilcoxon test of equality of survivor functions and restricted mean survival time.

Sensitivity analyses will be conducted i) considering different dichotomies of the 5 point scale and ii) assessing the assumption of independent or non-informative censoring (i.e. that the likelihood of an FoR event at the time of censoring is assumed to be the same as for those in whom no censoring occurred) using inverse probability of censoring weights (IPCW)[8] and multiple imputation(MI)[9] respectively.

This failure or recurrence analysis will be repeated in subgroups according to baseline characteristics as described in Section 6.3.

7.1.4 Favourable outcome at Week 132

A patient's outcome will be classified as favourable if their last two sputum culture results are negative, unless they have previously been classified as unfavourable. These two sputum results must be from samples taken at separate visits; the latest must be no more than six weeks before their scheduled last efficacy visit (see Section 7.1).

A patient's outcome at their scheduled last efficacy visit will be classified as unfavourable if:
- They are unfavourable at Week 76
- They start any MDR-TB treatment between their Week 76 and scheduled last efficacy visits
- They die between their Week 76 and scheduled last efficacy visits
- At least one of their last two culture results, from specimens taken on separate occasions, is positive.

A patient who does not have a culture result within the window for their scheduled last efficacy visit, having not otherwise been classified as unfavourable, will be regarded as non-assessable if their last two cultures, from specimens taken on separate occasions, are negative. All other patients who do not have a culture result within the window for their scheduled last efficacy visit and do not meet this definition of non-assessable will be classified as unfavourable.

The secondary non-inferiority outcome at Week 132 for Regimen B compared to Regimen C will be analysed using time to unfavourable outcome, thereby using information on all participants. Data from participants whose scheduled last efficacy visit is before Week 132 will be censored at the time of their last visit, unless they have already become unfavourable. Non-assessable participants will be included in the time to event analysis and censored at the time of their last visit.

The survival curve for each combination of strata (randomised protocol and HIV status) will be calculated using a Cox model adjusting for stratification factors and randomised group. The average survival curve for each randomised group will be estimated as a weighted average of the corresponding stratum-specific survival curves, with weights proportional to the number of individuals in each stratum in the randomised group at baseline. The mean of these differences at week 132 is the point estimate for the difference in overall survival function between Regimen C and Regimen B. A two sided bias-corrected 95% CI for the difference in proportion (Regimen C – Regimen B) will be calculated with bootstrap standard errors. Bootstrap standard errors for the risk difference will be used so that the CI of the risk difference (estimated by the risk difference ±1.96se assuming the bootstrap distribution is approximately Normal) is bounded between -1 and 1. The bootstrapping will sample 1000 times and be stratified by stratification factors. The seed for the bootstrap simulations will be stored to enable replication of results.

Non-inferiority will be shown if the upper bound of the 95% confidence interval of the difference in proportion of favourable outcomes at Week 132 between Regimens B and C is less than the 10% margin of non-inferiority; this must be shown in both the mITT and PP populations. If non-inferiority is demonstrated, then superiority of Regimen C compared to Regimen B will be declared if the upper bound of the 95% confidence interval of the difference in proportion of favourable outcomes at Week 132 between Regimens B and C is less than zero. For this analysis, the mITT population will be primary and the PP and mPP populations will be sensitivity analyses.

A sensitivity analysis will estimate the proportion unfavourable at Week 132 on the subset of participants randomised at least 132 weeks on or before 30 November 2021 (see Section 6.1.1). For this analysis, non-assessable participants will be excluded. The mITT population will be primary and the PP and mPP populations will be sensitivity analyses.

The above analyses will be repeated for the following comparisons, restricting the population in each case to include only concurrent controls within sites: Regimen B_{lev} and Regimen C, Regimen B_{mox} and Regimen C, Regimen D and Regimen B, Regimen B and Regimen A, Regimen C and Regimen A, Regimen D and Regimen A. The analysis will also be repeated (in both mITT and PP analysis populations) in subgroups according to HIV status.

7.2 Safety outcomes

Safety outcomes will be analysed using the safety analysis population.

7.2.1 Placement of events by study phases

Adverse events are placed in study phases (see Section 5.2 for definitions) based on the start date.
In case of partial start dates, the following approach is used:

- **Missing day only**: The event is placed in the earliest phase that overlaps the given month and year for the event, excluding any phases that start after the end date of the AE (if specified).
- **Missing day and month only**: The event is placed in the earliest phase that overlaps the given year for the event, excluding any phases that start after the end date of the AE (if specified).
- **Missing start date:** The event is placed in the treatment phase, unless the end date of the AE is specified and is before randomisation, in which case the event is placed in the screening phase.

7.2.2 All-cause mortality during treatment or follow-up

All-cause mortality is defined as a patient who has died from any-cause (both TB- or non-TB-related) while in the trial either during treatment or during follow-up.

The number of patients who die during treatment and follow-up will be tabulated by treatment arm. Mortality rates, stratified by randomisation protocol and HIV status, will be calculated for the individual treatment arms and treatment group differences in mortality rates will be calculated together with a 95% confidence interval. For patients that do not die, time will be censored at their final visit.

The difference in proportion of patients dying during treatment and follow-up, between Regimen B and Regimen C, between Regimen B_{mox} and Regimen C, between Regimen B_{lev} and Regimen C, between Regimen B and Regimen D, and between Regimen B and Regimen A, with corresponding two-sided 95% confidence intervals and p-values will be estimated (see Section 6.1.1). The analysis will be stratified by randomisation protocol (as defined in Section 1.6) and HIV status with three strata: HIV negative, HIV positive with CD4 count less than 350 cells/mm^3, and HIV positive with CD4 count more than or equal to 350 cells/mm^3.

Survival analysis will be conducted for time to death. A hazard ratio with corresponding two-sided 95% confidence intervals and p-value will be estimated for the comparison of

Regimen B with Regimen C, and Regimen B with Regimen D, using a Cox Proportional Hazards model stratified by randomisation protocol (as defined in Section 1.6) and the randomisation stratification factor HIV status with the three categories of: HIV negative; HIV positive with CD4 count ≥350 cells/mm^3; and HIV positive with CD4 count <350 cells/mm^3.

The equality of survivor functions for time to death for Regimen B and Regimen C, and Regimen B and Regimen D, will each be tested using a Log rank test, stratified by randomisation protocol (as defined in Section 1.6) and the randomisation stratification factor HIV status with the three categories of: HIV negative; HIV positive with CD4 count ≥350 cells/mm^3; and HIV positive with CD4 count <350 cells/mm^3.

This analysis will be conducted on the whole study period, and separately for follow-up censored at Week 76. It will also be repeated comparing bedaquiline and non-bedaquiline regimens (Regimens A and B vs C and D) and comparing Regimen B_{mox} and Regimen C, and Regimen B_{lev} and Regimen C. A descriptive analysis will consider deaths by cause, as adjudicated by an independent review committee. In addition to the comparisons detailed above, analyses of the whole study period (up to Week 132) will include comparisons of Regimen C with Regimen D. Comparisons of Regimen C and Regimen D using data up to Week 76 were carried out at the time of the primary analysis but considered exploratory.

7.2.3 Proportion of patients experiencing a grade 3 or greater adverse event, as defined by the DAIDS criteria, during treatment and follow-up

The difference in proportion of patients experiencing a grade 3 or greater adverse event, as defined by the DAIDS criteria[10], during treatment and follow-up, between Regimen B and Regimen C, between Regimen B_{mox} and Regimen C, between Regimen B_{lev} and Regimen C, between Regimen B and Regimen D, and between Regimen B and Regimen A, with corresponding two-sided 95% confidence intervals and p-values will be estimated (see Section 6.1.1). The analysis will be stratified by randomisation protocol (as defined in Section 1.6) and HIV status with three strata: HIV negative, HIV positive with CD4 count less than 350 cells/mm^3, and HIV positive with CD4 count more than or equal to 350 cells/mm^3.

This analysis will be conducted on the whole study period, the study period censored at Week 76, and separately for each phase (Treatment censored at Week 76, and Allocated treatment).

Comparisons will be repeated in subgroups according to HIV infection status. In addition the proportion of participants with cardiac events, hearing loss and hepatotoxicity by Week 76 and Week 132 in Regimen C and Regimen D will be compared to Regimen B.

In addition to the comparisons detailed above, analyses of the whole study period (up to Week 132) will include comparisons of Regimen C with Regimen D. Comparisons of Regimen C and Regimen D using data up to Week 76 were carried out at the time of the primary analysis but considered exploratory.

7.2.4 Change of regimen for adverse events

A change of regimen for an adverse event is defined as when a patient's regimen is modified in any way (including stopping a drug, changing the dose of a drug or starting a new drug) with the reason being an adverse event.

The difference in proportion of patients who have a change of regimen for adverse events between Regimen B and Regimen C, between Regimen B_{mox} and Regimen C, between Regimen B_{lev} and Regimen C, and between Regimen B and Regimen D, will be calculated.

The MedDRA SOC/SMQ or preferred term of the adverse event and the type of treatment change, will be described by arm.

7.2.5 Pharmacokinetics/Pharmacodynamics

At a number of visits throughout the trial, a single blood sample is taken to determine the bedaquiline (BDQ) and M2 plasma concentrations. The BDQ plasma concentrations are used by a qualified vendor to estimate C_{trough}, C_{max} and AUC using empirical Bayes' estimation based on a population pharmacokinetic model. Individual ART plasma concentrations (LPV, RTV and NVP) and 4β-hydroxycholesterol/cholesterol ratio (as a measure of cytochrome P450 3A (CYP3A) activity) are also measured.

Participants randomised to Regimen C or Regimen D with PK data available, will be classified into 5 subgroups based on their HIV status and HIV drugs taken immediately before the PK blood draw:

- **HIV Neg:** HIV negative at baseline
- **HIV Pos, LPV/r :** HIV positive at baseline having received Liponavir/Ritonavir (but not Nevirapine) immediately before the PK blood draw.
- **HIV Pos, NVP:** HIV positive at baseline having received Nevirapine (but not Liponavir/Ritonavir) immediately before the PK blood draw.
- **HIV Pos, other ARV :** HIV positive at baseline having received ART (but not Liponavir/Ritonavir or Nevirapine) immediately before the PK blood draw.
- **HIV Pos, not taking ARV:** HIV positive at baseline but not receiving ART immediately before the PK blood draw.

Individual plasma concentrations will be listed and descriptive statistics generated at each sampling time, by treatment group and separately for the above defined subgroups.

Further details of the analyses of PK/PD will be described in the STREAM Stage 2 Extensive SAP.

7.3 Health Economics

In selected sites, acceptability of Regimens A, B, C and D to stakeholders will be analysed in terms of:

- Costs to the health system
- Household costs
- Patient treatment and support experiences
- Health worker experiences.

The analyses of health and household costs and patient and health worker experiences will be described in a separate document.

7.4 Testing multiple outcomes

We will make no adjustment to p-values or confidence intervals to allow for testing multiple secondary outcomes.

The STREAM primary outcome is a composite outcome of bacteriological failure, treatment failure for any reason (clinical, bacteriological, intolerance) and death. The primary analysis is non-inferiority of Regimen C versus Regimen B. If the 95% CI for the treatment effect

(difference in proportion unfavourable by 76 weeks (C-B)) lies below the non-inferiority margin, then we will also test for superiority; because this is a closed test procedure there is no issue of multiplicity.

Secondary outcomes are divided in the protocol into efficacy outcomes, safety outcomes and acceptability outcomes. For safety outcomes it is appropriate to test each independently since it is important to identify any risks associated with Regimen C. The secondary efficacy outcomes are mostly components of the primary outcome or very closely related to the primary outcome (bacteriological and clinical outcomes). We will not adjust for multiple testing for these since they are correlated with the primary outcome (so standard adjustments are conservative).

We will report significance tests for differences between treatment arms for acceptability outcomes, but if we have failed to demonstrate non-inferiority of Regimen C versus Regimen B for the primary outcome, we will not use significance tests for these acceptability outcomes to conclude superiority.

8 ADDITIONAL EXPLORATORY ANALYSES

The following analyses will all be completed after the last participant has reached Week 132 on the mITT population unless otherwise stated.

8.1 Time to cessation of clinical symptoms based on PI assessment

In stage 2 of STREAM, the presence of a productive cough, fever, and night sweats was only solicited at the randomisation visit. During follow-up these symptoms were recorded on the AE log if reported by the participant. There is concern that these events have been under reported. Therefore, the proportion of people reporting these symptoms in Regimen B will be compared to the proportion reported in stage 1 (when they were solicited at each visit). If this proportion is <80% of that seen in stage 1, the following analyses will not be conducted.

Time to cessation of clinical symptoms is defined as the time from randomisation to the first of two consecutive visits where cessation of **all three** of the TB symptoms: a productive cough, fever, and night sweats, as reported by the patient. Patients with none of the TB symptoms at screening and baseline with be excluded from this analysis. This definition matches the definition of time to culture conversion as the first of two consecutive symptom-free months. For patients who do not cease clinical symptoms, cessation of clinical symptoms will be censored at the patients' last visit up to and including week 76.

Median time to cessation of clinical symptoms will be calculated for Regimen B, Regimen C, and Regimen D.

A hazard ratio with corresponding two-sided 95% confidence intervals and p-value will be estimated for the comparison of Regimen B with Regimen C, and Regimen B with Regimen D. A Cox Proportional Hazards model will be used, adjusted for randomisation protocol (as defined in Section 1.6) and the stratification factor HIV status with the three categories of: HIV negative; HIV positive with CD4 count ≥350 cells/mm^3; and HIV positive with CD4 count <350 cells/mm^3.

The equality of survivor functions for time to cessation of clinical symptoms for Regimen B and Regimen C, and Regimen B and Regimen D, will each be compared using a (Wilcoxon)

Log rank test, stratified by randomisation protocol (as defined in Section 1.6) and the randomisation stratification factor HIV status with the three categories of: HIV negative; HIV positive with CD4 count ≥350 cells/mm^3; and HIV positive with CD4 count <350 cells/mm^3.

The assumption of proportional hazards between the two comparisons will be tested using the proportional hazards test based on the Schoenfeld residuals after fitting the Cox Proportional Hazards models. In the case where there is adequate evidence that the proportional hazard assumptions are violated at the 5% level (i.e. $p<0.05$), methods where proportional hazards is not a necessary assumption will be used, such as restricted mean survival time.

8.2 WHO classification of outcome

TB outcome at the end of treatment will be determined using the WHO DR-TB treatment outcomes definition[11]. Participants will be classified as cured, treatment completed, treatment failed, died, lost to follow-up or not evaluated, with the first two categories considered together as treatment success.

The modified WHO DR-TB treatment outcome definition will also be applied; this includes an additional category of relapse after treatment success, defined as bacteriological relapse after end of treatment cure[4,12].

The most recent WHO DR-TB treatment outcome definition[13] will also be applied; this moves away from bacteriological status and focuses more on the completion of treatment or reasons for incompletion.

Each classification will be tabulated by treatment arm and the unadjusted difference in treatment success comparing Regimen B with Regimen C, Regimen B with Regimen D and Regimen C with Regimen D calculated with 95% confidence intervals.

8.3 TB outcome irrespective of treatment changes at Week 132

This analysis will exclude all participants whose scheduled last efficacy visit is before Week 132.

TB outcome will be classified according to the patients' culture status at week 132 regardless of treatment changes or intermediate culture results – similar to a simplistic intention-to-treat analysis.

A participant is defined as "cured at Week 132" if the last two cultures were negative with the last no earlier than the Week 132 analysis window; other categories include death prior to Week 132, last culture positive at Week 132, last culture negative prior to Week 132, last culture positive prior to Week 132, or no cultures after baseline.

This classification will be tabulated by treatment arm, and the difference in proportion cured reported comparing Regimen B with Regimen C, Regimen B with Regimen D and Regimen C with Regimen D calculated with 95% confidence intervals, stratifying for randomisation protocol and HIV status.

8.4 Lung surgery

The proportion of participants having undergone lung surgery (resection or pneumonectomy) by Week 76 and Week 132 in Regimen C and Regimen D will be compared to Regimen B using Fisher's exact test on the safety population.

9 DATA SUMMARIES

9.1 Recruitment and baseline characteristics

9.1.1 Recruitment, screening, & eligibility

The number of patients screened, randomised, and treated will be tabulated by centre and treatment arm. The number of patients who failed screening, and the reasons for ineligibility will be presented by randomised group.

9.1.2 Exclusions from analysis

The number of patients excluded from the mITT and PP analysis populations will be tabulated by treatment arm and by reason for exclusion.

9.1.3 Baseline characteristics

All eligible patients randomised will be included in tables of baseline comparisons by treatment group. Characteristics will include sex, age, race, height, weight, BMI, albumin levels, and laboratory parameters such as, HIV status, CD4 count (if applicable), smear and culture status, and drug susceptibility status for a number of TB drug types. The baseline characteristics table will be repeated for each of the ITT, safety, PP and mITT populations.

9.2 Efficacy and adherence

Each analysis will be carried out using the mITT analysis population.

9.2.1 Sputum smear and culture

Sputum smear and culture results (positive or negative) will be tabulated by visit and treatment arm.

9.2.2 Adherence

Adherence will be summarised by treatment arm as the percentage of each of the intensive and continuation phase doses completed and overall across both phases.

9.2.3 Drug resistance

All drug resistance analyses will be based on phenotypic DST results unless unavailable when genotypic results will be used.

Drug resistance at screening or baseline will be tabulated by treatment arm.

Acquired resistance to any drugs will also be described and tabulated by treatment arm using i) the last available DST result for each drug for each patient, and ii) all available post-randomisation DST result (i.e. classifying as resistant if any result is resistant) for each drug for each patient ignoring isolated resistance unless it is the last result prior to the point of analysis. The proportion of participants developing resistance, especially resistance leading to the development of pre-XDR or XDR strains of TB in Regimen C and Regimen D will be compared to that in Regimen B for data up to Week 76, and for the Allocated treatment phase only (i.e. up to the point of starting a salvage regimen).

9.3 Safety

All the following analyses will be on the safety population.

9.3.1 ECG

Time to first QT/QTcF over 450ms and first QT/QTcF over 500ms and QTcF increase from baseline by 30ms and by 60ms analyses will be conducted. The proportion of participants experiencing of each of these events (ie whether a threshold was exceed or not) will be tabulated by treatment arm and country. Hazard ratios with corresponding two-sided 95% confidence intervals will be estimated for the comparison of Regimen B with Regimen C, Regimen B_{mox} with Regimen C, Regimen B_{lev} with Regimen C, and Regimen B with Regimen D. A Cox Proportional Hazards model will be used, adjusted for randomisation protocol (as defined in Section 1.6) and the stratification factor HIV status with the three categories of: HIV negative; HIV positive with CD4 count ≥350 cells/mm^3; and HIV positive with CD4 count <350 cells/mm^3.The outcomes will be censored at the patients' last visit.

The equality of survivor functions for time to QTcF over 450ms and over 500ms and QTcF increase from baseline by 30ms and by 60ms for Regimen B and Regimen C, and Regimen B and Regimen D, will each be tested using a Log rank test, stratified by randomisation protocol (as defined in Section 1.6) and the randomisation stratification factor HIV status with the three categories of: HIV negative; HIV positive with CD4 count ≥350 cells/mm^3; and HIV positive with CD4 count <350 cells/mm^3.

The assumption of proportional hazards will be tested using the proportional hazards test based on the Schoenfeld residuals after fitting the Cox Proportional Hazards models. In the case where there is adequate evidence that the proportional hazard assumptions are violated at the 5% level (i.e. $p<0.05$), methods where proportional hazards is not a necessary assumption will be used, such as using the Kaplan-Meier estimator restricted mean survival time.

QTcF will be estimated by visit and by treatment arm using a linear mixed model, adjusted for the stratification factors. Estimates with 95% confidence intervals will be plotted by visit and treatment arm. This will be repeated for change in QTcF from baseline.

It is likely that treatment and dose changes will impact on QTcF and so this analysis will be repeated ignoring any results after discontinuation or change of dose of any drug. Exploratory analyses will investigate the relationship between changes in QTcF and ART use.

9.3.2 Liver function

Bilirubin, ALT, and AST will be tabulated by DAIDS grade, visit and treatment arm. Mean (and SE) bilirubin, ALT, and AST will be presented by visit and treatment arm. The number of patients experiencing more than or equal to five times above the upper normal limit will be tabulated by arm.

9.3.3 Hearing impairment

The proportion of patients experiencing significant hearing loss (unilateral or bilateral) based on Brock's Criterion applied to Shoebox data, during treatment and follow-up will be tabulated by treatment arm. Significant hearing loss will be defined as hearing loss that persists such that it is still present at the last visit of the follow-up period being considered (Week 76 or last efficacy visit) and is grade 3 or higher.

9.3.4 Weight gain

Patient weight will be tabulated by treatment arm and visit in addition to change from baseline weight by visit and treatment arm.

9.3.5 SAE/NE/AE

SAEs and grade 3 or higher AEs will be tabulated by MedDRA System Organ Class (SOC) / Standardised MedDRA queries (SMQ) and Preferred Term and treatment arm. NEs will be tabulated by event type and treatment arm. The total number of events and the number of participants with at least one event will be given. SAEs will also be tabulated by event relatedness to study drugs and treatment arm. Pregnancy outcomes will be reported by treatment arm.

Grade 3 and 4 laboratory AEs will be tabulated by laboratory marker and treatment arm. This will be presented for results from the central laboratory alone.

These analyses will be conducted on the whole study period, and separately for each phase (Treatment censored at Week 76, and Allocated treatment).

9.3.6 HIV infection

HIV viral load (VL) and CD4 actual values and changes from baseline will be tabulated by visit and treatment arm. VL response, defined as <50 copies/mL, between 50 and less than 400 copies/mL, and ≥400 copies/mL, will be tabulated by visit and treatment arm, overall and by baseline VL. This analysis will be restricted to HIV positive patients and repeated by patient on/off ARV therapy at baseline.

10 REFERENCES

1. World Health Organization. Guidelines for the programmatic management of drug-resistant tuberculosis - 2011 update. Geneva: World Health Organization STOP TB Department, 2011.

2. Van Deun A, Maug AKJ, Salim MAH, et al. Short, Highly Effective, and Inexpensive Standardized Treatment of Multidrug-resistant Tuberculosis. *Am J Respir Crit Care Med* 2010; **182**(5): 684-92.

3. Aung KJ, Van Deun A, Declercq E, et al. Successful '9-month Bangladesh regimen' for multidrug-resistant tuberculosis among over 500 consecutive patients. *Int J Tuberc Lung Dis* 2014; **18**(10): 1180-87.

4. Ahmad N, Ahuja SD, Akkerman OW, et al. Treatment correlates of successful outcomes in pulmonary multidrug-resistant tuberculosis: an individual patient data meta-analysis. *The Lancet* 2018; **392**(10150): 821-34.

5. Mohamed K, Embleton A, Cuffe RL. Adjusting for covariates in non-inferiority studies with margins defined as risk differeneces. *Pharmaceutical statistics* 2011; **10**(5): 461-6

6. Nunn AJ, Phillips PPJ., Meredith SK. et. al. A Trial of a Shorter Regimen for Rifampin-Resistant Tuberculosis. *NEJM* 2019; **380**:1201-1213

7. Phillips PPJ, Van Deun A, Ahmed S et al. Investigation of the efficacy of the short regimen for rifampicin-resistant TB from the STREAM trial. *BMC Med* 2020; **18**, 314.

8. Dodd S, Williamson P, White IR. Adjustment for treatment changes in epilepsy trials: a comparison of causal methods for time-to-event outcomes. *Stat Methods Med Res* 2019; **28**(3):717–33.

9. Jackson D, White IR, Seaman S, et. al. Relaxing the independent censoring assumption in the Cox proportional hazards model using multiple imputation. *Stat Med*. 2014; **33**(27):4681–94.

10. DAIDS RSC. Division of AIDS table for grading the severity of adult and pediatric adverse events. Version 1.0, December 2004 (Clarification August 2009) ed. DAIDS Regulatory Support Centre; 2009.

11. World Health Organization. Definitions and reporting framework for tuberculosis. 2013 revision, updated December 2014. Geneva: World Health Organization; 2013.

12. World Health Organization. WHO consolidated guidelines on drug-resistant tuberculosis treatment. Geneva: World Health Organization; 2019.13. World Health Organization. Meeting report of the WHO expert consultation on drug-resistant tuberculosis treatment outcome definitions,17–19 November 2020. Geneva: World Health Organization; 2021.