Implications of missing data in tuberculosis non-inferiority clinical trials

Sunita Rehal

Submitted for the degree of
Doctor of Philosophy
at University College London

October 2018

Declaration

I, Sunita Rehal, confirm that the work presented in this thesis is my own.

Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

I was the lead author of the paper which forms the basis of the work presented in Chapter 2. In addition, Tim Morris (a colleague at the MRC CTU at UCL) and my supervisors were co-authors.

Abstract

Non-inferiority designs have been increasingly used in randomised clinical trials in recent years. However, there remain several key issues with this design that can have important implications for the primary analysis and its interpretation. Specifically, choosing the population for inclusion in the primary analysis and how to deal with missing values, remains unclear.

This thesis tackles three related methodological issues in tuberculosis (TB) clinical trials: (i) a lack of clear guidance on design and reporting; (ii) the need for a valid approach to missing data and (iii) how to perform sensitivity analysis.

First, widely available guidance documents on non-inferiority trials are critiqued, highlighting differences in recommendations between them on fundamental issues. These differences are reflected in inconsistent reporting from a systematic review we conducted, and make suggestions for improvements.

Second, using data from two recent TB non-inferiority trials, we compare and contrast (i) different imputation approaches, (ii) inverse probability weighting with marginal models, and (iii) multi-state Markov models, for handling missing outcome data under the missing at random assumption. We find a form of multiple imputation is the best practical approach.

Third, we explore sensitivity analysis to the missing at random assumption, and show how a "reference based" method provides an accessible, practical approach.

In conclusion, more appropriate guidelines and analyses for non-inferiority trials in TB are needed, and some proposals are made to this end. Based on these findings, it is proposed that missing data in TB non-inferiority trials should be handled using the "two-fold" multiple imputation algorithm for imputing the missing data. By imputing the data in this way uses all the information available and allows for the trials defined primary outcome to be determined for each patient. Following this, reference based sensitivity analysis should be utilised.

Contents

List of Figures

List of Tables

Acknowledgements

Firstly, I would like to thank my primary supervisor Dr. Patrick Phillips. I sincerely appreciate the insightful advice and encouragement he has given me during my PhD. There are not enough words to express my gratitude to Professor James Carpenter. He has always been on hand to provide invaluable expertise, time and great enthusiasm. Qualities that are echoed by all his colleagues. Thanks also goes to Dr. Katherine Fielding for all her help and enthusiasm she has given me throughout my PhD.

I would like to thank the MRC CTU at UCL for funding my research and for providing a friendly environment to work in. I thank all the other PhD students and colleagues, both past and present, at the MRC CTU at UCL for all the support and fun times.

Lastly, I thank my family and friends for continuing to motivate me throughout my time as a PhD student. I specifically thank my parents who have always been at the forefront of my education.

Chapter 1

Introduction

1.1 Non-inferiority

Non-inferiority trials are designed to show that when compared to an active control an alternative treatment is not much worse by a pre-specified, acceptable, margin. These trials are only appropriate if the alternative treatment has some other benefit that a standard treatment (or care) does not, for example a less intensive treatment[1]. This design differs from the more familiar superiority trials, where the aim is to discover a treatment which performs better compared with placebo or active control.

Over the last 10 years, there has been a vast increase in the number of trials that use a non-inferiority design. In PubMed using the search term "non-inferiority trial" or "noninferiority trial" in titles and abstracts and filtering the publication date from 1st January 2007 to 31st December 2007, yields a record of 46 articles compared with 222 records from 1st January 2017 to 31st December 2017 (search done on 12th June 2018). The increasing use of non-inferiority trials highlights the importance for non-inferiority trials being well designed and appropriately analysed.

The focus of this thesis is on tuberculosis non-inferiority trials. However non-inferiority trials in general pose design challenges since there are more statistical issues to consider than for superiority trials. This introductory chapter focuses on some critical elements when designing non-inferiority trials in a general setting. The second chapter investigates how the research community are reporting these types of

trials. After drawing out some recommendations, we then introduce tuberculosis trials; these studies are our motivating examples of non-inferiority studies.

In this chapter, currently available guidelines for non-inferiority trials are reviewed and critiqued. Particular issues within non-inferiority trials that could influence the results and conclusions are discussed, and we highlight some key differences in the advice given within these guidelines. Finally, the objectives of this thesis are set out.

1.1.1 Existing guidelines on non-inferiority

The guidelines critiqued here include those from the International Council for Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH), which are related to all clinical trials. The European Medicines Agency (EMA) provides guidance for the evaluation and safety of medicines in the European Union[2] and the Food and Drug Administration provide guidance for the United States of America (U.S. FDA). In the U.S. the FDA oversee medicinal products and tobacco, foods, global regulatory operations and policy[3]. The Consolidated Standards of Reporting Trials (CONSORT) involves different groups of researchers, depending on their expertise, working to improve inadequate reporting of randomised controlled trials[4]. The Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) provides guidance for trial protocols with the aim of improving how clinical trials are conducted, from the design stage onwards[5].

The following, currently available guidelines contain information on the design and analysis of non-inferiority trials to assist researchers that may choose this design (in order of publication date):

ICH Harmonised Tripartite Guideline Statistical Principles for Clinical Trials E9 (ICH E9 1998)[6]

Explanations and recommendations have been given within this guideline for all trials. That is for: superiority, equivalence and non-inferiority trials. Guidance is given surrounding the population to include in analyses, type of outcomes to consider (primary/secondary), and ways to avoid bias. Methods for analysts to consider are also explained within this document. This is the earliest formal guideline available

which considers the statistical and design aspects of clinical trials and which provides some information about non-inferiority.

ICH Harmonised Tripartite Guideline Choice of Control Group and Related Issues in Clinical Trials (ICH E10 2001)[7]
The ICH E10 guidelines focus on choosing the control group for different trial designs. This document also contains some important guidance on how to determine the non-inferiority margin based on historical evidence of test drugs.

Committee for Proprietary medicinal products (CPMP) Points to Consider on Switching Between Superiority and Non-inferiority (EMA 2000)[8]
The focus of this document is how to extend and interpret a non-inferiority trial as a superiority trial, and vice versa. This document also contains some information with regards to the design and interpretation of non-inferiority trials.

Committee for Proprietary medicinal products (CPMP) Guideline on the choice of the non-inferiority margin (EMA 2006)[9]
Details of the non-inferiority margin and points to consider when determining the margin are included here along with interpretations of non-inferiority.

Reporting of Noninferiority and Equivalence Randomised trials. An Extension of the CONSORT statement. (CONSORT 2006)[1]
The CONSORT 2006 statement describes specific items to be reported in medical journals for non-inferiority trials. Justifications for each and examples are also provided.

Reporting of Noninferiority and Equivalence Randomised trials. Extension of the CONSORT 2010 statement. (CONSORT 2012)[10]
The CONSORT 2012 statement provides some additional clarification on the original CONSORT 2006 document.

CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. (CONSORT 2010)[11]
This document provides guidance for superiority trials but has been included for

review as it is linked with the CONSORT 2006 and CONSORT 2012 statements. The CONSORT 2010 is favoured over the original CONSORT statement[12] for superiority studies as this is the most recent document.

Non-inferiority clinical trials to establish effectiveness. Guidance for industry (November 2016). (U.S. FDA 2016)[13]

This document focuses completely on trials that have a non-inferiority design. All aspects and issues of non-inferiority designs are discussed and considered, such as the choice and calculation of the non-inferiority margin, sample size, hypotheses for non-inferiority trials, rationale for choosing a non-inferiority design and potential biases in non-inferiority studies are addressed. As this document was finalised in November 2016, the guidance has been updated in this chapter accordingly. There are however some references to the draft U.S. FDA guidance written in 2010[14] since this was referred to by researchers until 2016. This document is also relevant for the research reported in this thesis.

SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials (SPIRIT 2013)[15]

This document is written for protocols for all study designs, including non-inferiority. This guidance has been included in the review as these are the most recent guidelines which contain up-to-date methods, some of which are relevant to non-inferiority studies.

1.1.2 A brief note on non-inferiority and equivalence

Non-inferiority and equivalence are often used interchangeably. Although there are some similarities, they actually relate to two different designs with some subtle differences. Equivalence trials are designed to show that a new intervention performs not much worse and not much better than a standard intervention. The ICH E9[6], ICH E10[7] and SPIRIT[15] guidelines encompass superiority, equivalence and non-inferiority study designs, while the CONSORT 2006 and 2012[1,10], EMA 2000 and 2006[8,9] and FDA 2016[13] guidelines consider equivalence and non-inferiority study designs. In all guidelines, methods for equivalence are discussed alongside non-inferiority for ease of relating one to the other as the two concepts are similar. Equivalence and

non-inferiority studies are very similar in design in terms of margins and bounds of confidence intervals, but they answer different questions. Discussing two very similar methods interchangeably may be the cause of confusion between the two. In terms of interpretation of the analysis, the most important difference is that equivalence considers two margins and therefore conclusions are based on both sides of the confidence interval, whereas one margin and one bound of the confidence interval are used for conclusions of non-inferiority. This thesis will only be considering non-inferiority studies.

1.2 Issues to consider surrounding non-inferiority studies

There are several statistical issues to consider when designing any trial, but there are more challenges for non-inferiority studies which are more susceptible to bias. Poor trial quality can bias trial results towards achieving no difference between treatments[14]. This creates more challenges in non-inferiority trials than for superiority trials as this bias can produce false positive results for non-inferiority.

1.2.1 Non-inferiority margin

For superiority trials, when looking at differences in treatment effects, the aim is to reject the null hypothesis that a treatment is equal to an active control/placebo. For non-inferiority trials, the aim is to reject the null hypothesis that a new treatment is some pre-specified amount worse than the standard treatment. This "pre-specified amount" is the non-inferiority margin denoted by Δ or δ.

All guidance documents state that the non-inferiority margin should be specified, reported and justified on clinical grounds. ICH E10 2000[7], EMA 2006[9] and the U.S. FDA 2016[13] are the only guidelines which explicitly state that statistical justifications should also be considered alongside clinical justifications. The U.S. FDA (2016) guideline gives two detailed approaches to how the non-inferiority margin can be calculated. The first FDA recommendation is the fixed margin approach. Here the margin is pre-determined based on historical evidence comparing the standard of care treatment with placebo, and using clinical judgement. The second is the synthesis approach[13]. The synthesis approach entails a combination of the estimate of the

treatment effect relative to the standard of care from an ongoing non-inferiority trial with the estimate of the standard of care from a meta-analysis of historical trials[13]. The confidence interval of the two estimates after combining them is then used to test the non-inferiority hypothesis; pre-specification of an acceptable fraction of the control therapy's effect that should be retained by the new treatment being tested is judged on a clinical basis[13]. This approach is not often used because the margin is calculated during research and cannot be pre-specified[13]. The CONSORT (2006 & 2012) statements[1,10] imply through examples that clinical and/or statistical justifications for the margin need to be considered. Justification on how the margin was determined is not explicitly requested in other guidelines, and so it can be hard to judge in studies whether or not a margin has been arbitrarily chosen[16].

1.2.2 Population included in analyses

Most guidelines[1,6,8,10,13] agree that an intention-to-treat (ITT) analysis is defined as all patients who were randomised into a study, where patients are analysed based on what treatment was allocated at the time of randomisation regardless of what may occur afterwards. Most guidelines agree that the per-protocol (PP) analysis contains one or more exclusions. There are some other definitions of who should be included in analyses. Of note, the draft U.S. FDA 2010[14] guidelines suggested an "as-treated" analysis but failed to provide a definition. Other literature suggests that an as-treated analysis means analysing treatment differences based on what patients had actually received in treatment rather than according to their randomised treatment[17–19]. This terminology has since been removed in the final U.S. FDA (2016) guidelines so that no particular analysis is formally recommended.

The ICH E9 1998 guidelines[6] propose another type of analysis; a full analysis set. This is similar to the ITT definition, but can exclude patients provided that the reasons for exclusion are not treatment dependent. This seems to be similar to how the newly emerging modified intention-to-treat (mITT) analysis is defined, although consistent definitions are lacking[15].

The explanation of "full analysis set" provided by ICH E9 (1998) is confusing. What ICH E9 (1998) consider to be acceptable exclusions while maintaining the ITT definition (i.e. a "full analysis set") intertwines with their list of PP exclusions:

eligibility violations could be regarded as major protocol violations and so could fit into the PP definition; failure to take at least one dose of trial medication could fit into a PP population where "exposure to treatment" is considered; missing outcome data, for example, could fit into the ICH E9 (1998) definition of "full analysis set" as this is a "lack of post randomisation data", but equally could fit into the PP definition of "the availability of measurements".

The very first CONSORT statement, published in 2001[12], defines ITT as all patients randomised with an outcome. It is later acknowledged in the CONSORT 2010 guidelines, that an ITT population where all patients have an outcome is rarely achievable due to missing data. These guidelines "favour a clear description of exactly who was included in each analysis"[11]. The difference between the two definitions has led to SPIRIT classing all randomised patients that have an outcome, where any missing outcome data are resolved, as a "classic" ITT and ITT as all patients randomised regardless of adherence.

It is unclear in guidelines which analysis is most appropriate for non-inferiority trials. ICH E9 1998[6] state: "preservation of the initial randomisation is important in preventing bias and in providing a secure foundation for statistical tests"; whereas CONSORT 2006 & 2012[1,10] state: "non-ITT analyses might be desirable as a protection from ITTs increase in the Type I error". For non-inferiority trials, both the ITT and PP analysis have their biases. The ITT analysis can bias towards the null treatment effect, which may lead to false claims of non-inferiority[20]. The PP analysis, which excludes patients, fails to preserve a balance of patient numbers between treatment arms (i.e. randomisation) that ITT analysis does, and so may cause bias in either direction, depending on who the analysis excludes[21]. It is, therefore, often recommended for both a PP analysis and an ITT analysis to be analysed for primary analyses in non-inferiority trials with any disagreements investigated.

Given that both the ITT and PP analysis are recommended, it is unclear whether one analysis ought to take precedence over the other. The EMA 2000 guidelines[8] state that analyses on both an ITT analysis and PP analysis are equally important and the same is implied in the SPIRIT 2013 guidelines[15]. However, CONSORT 2006[1] appear to give

the researcher a choice into whether an ITT analysis, PP analysis or both should take precedence in the primary analysis to determine non-inferiority, but emphasise that whatever is chosen should be clearly stated.

Although the U.S. FDA (2010) guidelines no longer recommend a particular analysis[14], issues relative to the widely used PP analysis such as loss to follow-up or treatment switching are highlighted. The advice in the final U.S. FDA (2016) guidelines[13] is to minimise these potential problems while planning a study without really providing a useful solution. They do suggest imputing missing data for patients whose outcome data are missing, and this will be explored throughout this thesis.

1.2.3 Confidence intervals

In superiority studies, conclusions are determined based on where the treatment effect and its confidence interval lie relative to the null. For non-inferiority, only one bound of the confidence interval is required for inference and conclusions are determined based on where the confidence interval for the treatment effect lies relative to the pre-defined non-inferiority margin[10].

All guidance documents[1,6–10,13,15] clearly state that inferences for non-inferiority should be made on one bound of the confidence interval. The ICH E9 1998 and E10 2000[6,7] documents are the only guidelines that do not inform the reader at what level the confidence interval should be set. All other guidelines recommend two-sided 95% confidence intervals, where conclusions of non-inferiority are based upon the upper (or lower) bound of the two-sided 95% confidence interval. The CONSORT 2006 guidelines mention that 90% confidence intervals may be appropriate in certain situations[1].

There is some ambiguity around whether 90% confidence intervals or 95% confidence intervals should be reported for non-inferiority trials. The emphasis on reporting 95% two-sided confidence intervals seems to have originated from the EMA guidelines 2000[8], where "if a one-sided confidence interval is used then the 97.5% should be used". This is further explained in the EMA 2006 guidelines[9] "statistical significance is generally assessed using the 0.05 level of significance (or one-sided 0.025). An

alternative way of stating this requirement is that the lower bound of the two-sided 95% confidence interval (or one-sided 97.5% confidence interval) for the difference between active and [control] should be above [the non-inferiority margin]" for a positive outcome. This is supported in the CONSORT 2006 statement[1] and endorsed in all other subsequent guidelines.

The EMA (2000) explain that the two-sided confidence interval "should lie entirely on the right side of delta". This is reiterated again in the later 2006 EMA guidelines where inferences should be made on the lower bound of the confidence interval (for a control minus treatment comparison). On the other hand, U.S. FDA (2016) suggest the upper bound of the two-sided confidence interval is used for inference relative to the margin: "the upper bound of the 95% confidence interval is typically used to judge the effectiveness of the test drug in the non-inferiority study" and "the 95% confidence interval upper bound for control minus treatment, is used to provide a reasonably high level of assurance that the test drug does, in fact, have an effect greater than zero"[13].

There is some inconsistency between the EMA (2000 & 2006) and the FDA (2016) guidelines, and therefore perhaps some lack of clarity about which side of the confidence interval to use (to the left or right of delta) to make inferences on. The answer to this depends on the question a researcher is asking. Assume a treatment minus control comparison. In a superiority study, if the way an outcome is defined is a success (e.g. better quality of life or decrease in hospital admissions) a result less than 0 would indicate that the treatment is unfavourable. If the outcome is defined as failure (e.g. worse quality of life or increase in hospital admissions) then a result greater than 0 would indicate that the treatment is unfavourable. The same rationale can be applied to non-inferiority studies. Taking increase in hospital admissions as an example, the non-inferiority margin would be a value greater than 0 as this value would indicate how much of an increase in hospital admissions is "acceptably worse". Therefore, conclusions would be made on the upper bound of the confidence interval, relative to the margin. An outcome which is defined as a success (e.g. better quality of life or decrease in hospital admissions) the pre-defined non-inferiority margin would be less than 0. Therefore, conclusions would be made on the lower bound of the

confidence interval, relative to the margin. Recently, JAMA have introduced a policy of presenting the lower bound of the confidence interval with the upper bound tending towards infinity[22]. Their rationale being that the results are distinguished from a superiority or equivalence study and since only one bound of the confidence interval is used for inference, the other bound of the confidence interval does not matter for making conclusions about non-inferiority. This policy has been put into practice in recent non-inferiority trials[23–26].

1.2.4 Missing data

If one or more values were not recorded on collected data but were intended to be, then this is classed as missing data. Missing data results in loss of information: high amounts of missing data results in the conclusions of a study losing validity.

The discussion of the issues raised by missing data that addresses non-inferiority trials is sparse. The ICH E9 1998 guideline[6], suggests single imputation techniques, such as last observation carried forward (LOCF) to handle missing data. The guideline also suggests "complex mathematical models" but does not suggest what those could be. The U.S. FDA 2016 guidelines[13] acknowledge that "Conducting any poor quality studies should always be avoided, but with non-inferiority studies, sloppiness in study design/conduct is particularly problematic, because it introduces bias towards the alternative hypothesis of non-inferiority". The guideline suggests imputation of missing data without suggesting a particular method to counteract any bias due to missing observations. It fails to highlight that imputation methods also carry assumptions that ought to be reported.

Last observation carried forward assumes that the average of the unobserved outcomes in each randomised group do not change over time[27]. This is a quite a strong and definitive assumption to make on the unobserved data of patients who are lost to follow up and so needs to be carefully thought of before being used. Over time, post-1998, the LOCF method is no longer recommended as a favourable analysis and instead has been cautioned against due to the underlying assumptions made. A simulation study performed by Cook et al[28] assessed the bias, empirical standard errors, and type I and type II error rates and coverage probabilities including

investigation of LOCF. The authors found that there was severe bias in the estimates and inflation of the type I error rate. A contributing factor to these biases was due to observed response and withdrawal: "the most influential [factor] appear to be whether there is a trend in the responses over time, and whether the conditional probability of drop-out is different for those in the treatment and control groups"[28]. Molenberghs et al[29] used three anti-depressant clinical trials as case studies to assess the impact simple imputation methods have on the overall conclusions. The authors found that these imputation methods altered the conclusions of the case studies used, and conclude: "there is little justification for analysing incomplete data from longitudinal clinical trials by means of such simple methods such as LOCF and complete case"[29].

Other simplistic analyses that are considered in clinical trials to address the issue surrounding missing data include the best case/worst case scenario (§3.4.3). The best case method replaces missing values in the reference arm with the worst value and treatment arm(s) with the best value. The worst case method replaces missing values in the reference arm with the best value and the treatment arm(s) with the worst value. However, imputing the data in this way replaces the missing values with certainty and can therefore bias the standard error downwards if uncertainty of the imputed value is ignored[30]. Unnebrink et al[31] performed a simulation study based on a two-arm osteoporosis trial to assess what impact the best case and worst case analysis has on the significance and power between the treatment regimen and control. The authors found when a worst case scenario was used the power decreased with increasing rate of withdrawal and the type I error rate also decreased. This means that using the worst case scenario biases towards the null hypothesis, favouring the control treatment arm. Assuming a best case scenario showed that with increasing rates of withdrawal, the type I error rate and power increased, therefore biasing towards the alternative hypothesis favouring the treatment regimen. The authors note that using a worst case and best case scenario are "too extreme"[31].

Inverse probability weighting (IPW), a conceptually simple method can also be used to account for missing data based on the availability of patient observations. Two models are required for IPW; a model relating the outcome to the explanatory

variables and a model for the probability of missing observations[32]. IPW then calculates a weighted average, allocating more weight to patients with a lower chance of being observed. However, IPW is restricted for monotone missing data patterns (i.e. a patient who is missing data over follow-up is never observed in the future)[33]. Chapter 4 investigates this approach in more detail.

A more advanced approach to deal with the issue of missing data is multiple imputation[34]. For multiple imputation, a model for the distribution of the missing data given the observed data is specified. Missing values are replaced several times with random values from this model to create several imputed data sets. By replacing missing values several times and creating multiple imputed datasets accounts for the uncertainty of the imputed value. This methodology is investigated in detail in Chapter 3.

There are several methods that make more reasonable assumptions than those considered within the guidelines reviewed here and therefore may be preferable. We investigate some of these methods within this thesis. Regardless of what method is used each carry assumptions which should be stated, investigated and reported in publications.

1.2.5 Sensitivity analyses

In the context of missing data, sensitivity analyses explore how robust conclusions from the primary analysis are to different assumptions about the missing data. If conclusions of the sensitivity analysis are similar to that of the primary analysis, then this enhances confidence in the results of a trial as they are robust to a range of plausible assumptions about the missing data.

ICH E9 1998[6] state for all trials: "an investigation should be made concerning the sensitivity of the results of analysis to the method of handling missing values, especially if the number of missing values is substantial". The SPIRIT 2013 guidelines[15] support this: "sensitivity analyses are highly recommended to assess the robustness of trial results under different methods of handling missing data". EMA (2000) follow the ICH E9 (1998) guideline where, to show sensitivity of a study, it is

suggested that both a full analysis set and analysis on the PP population should be done to enhance the integrity of results. CONSORT 2012[10] do not suggest sensitivity analyses but imply that if an ITT or PP analysis is chosen, the other counts as a sensitivity analysis. CONSORT (2012) guidelines for non-inferiority suggest through an example that if one analysis is on either an ITT or PP population, then an analysis of the other would be a sensitivity analysis: "Study endpoints were analysed primarily for the per protocol population and repeated, for sensitivity reasons, for the intention-to-treat (ITT) population"[6–9]

The ITT and PP analysis actually ask different questions about the behaviour of patients in the analysis and do not test the assumptions made about the missing data. As pointed out by Carpenter and Kenward[35], "focus[ing] on comparing results of certain methods, which can make similar assumptions about the missingness mechanism, rather than comparing the sensitivity of the conclusions to varying the assumptions about the missingness mechanism misses the point of using sensitivity analysis and can led to misleading conclusions".

1.2.6 Other considerations

Other important criteria to consider for non-inferiority studies include blinding, randomisation, assay sensitivity and biocreep.

Blinding and randomisation

Blinding is a method that can protect against bias since the knowledge of what treatment has been allocated to participants can create bias[36]. Randomisation also protects against potential bias as the method ensures random allocation of treatment to patients[37].

Assay sensitivity

Assay sensitivity is the ability to detect an effective treatment from a less effective treatment[7]. The impact of missing data can lead to false conclusions of non-inferiority[38]. This impact depends on the missingness mechanism (see §3.3). In cases where the data are MCAR or MAR, the impact is marginal as the missingness reason is known[39]. Therefore in cases where patients are lost to follow-up for reasons

due to the administered treatment, collecting follow-up data for these patients can maintain assay sensitivity[40]. In non-inferiority studies, it is often assumed that an active control is superior compared to placebo[41].

Biocreep

If a treatment is found to be "non-inferior", then the treatment is acceptably worse. This slightly inferior treatment may become an active control for the next generation of non-inferiority trials and so on until the active control is no better than a placebo[20].

1.2.7 Summary

The statistical guidelines reviewed show that the suggestions made to deal with some of the issues that arise in non-inferiority trials are inconsistent, particularly surrounding the choice of the primary analysis. This is concerning given the increase of non-inferiority trials being performed. Following this review of the guidelines, a systematic review of selected journals was conducted described in Chapter 2. The rationale being that the inconsistency highlighted within this Chapter between the guidelines are highly likely to have some impact on the design and reporting of published non-inferiority trials.

1.3 Thesis objectives

The overarching goal of this thesis is to find better methods for analysing the primary outcome of non-inferiority clinical trials. The PP analysis is often preferred since the analysis emphasises patients who adhered to the protocol. As a consequence, patients who deviate from the protocol or who miss a follow-up visit may be excluded from the primary analysis. The recommended "conservative" methods to handle missing outcome data for all trials proposed by regulators (and used by trialists), such as the best case/worst case scenario[30] and used by trialists are simplistic and require strong assumptions about the nature of the missing data[31]. Building on the review in Chapter 2, the specific objective of this thesis is therefore to find alternative methods to include these missing observations within the primary analysis of non-inferiority trials, thereby providing a valid, more powerful analysis.

The methods we will investigate make the untestable missing at random assumption (defined in §3.3) about the missing data, and so a further aim is to look at sensitivity analyses to investigate departures from this assumption made.

For the last 5 years, tuberculosis (TB) has been classed as the leading cause of death from an infectious disease by WHO[42]. Due to the intensity of the treatment over 6 months, some patients struggle to take the full course of treatment. In some cases, patients feel much better very quickly after receiving treatment and therefore do not feel the need to take the full course of treatment. In both cases, this leads to them contracting TB again. There is a real need to shorten these treatment regimens. The two datasets used in this thesis, the REMoxTB and RIFAQUIN trials, aimed to do this but failed to show non-inferiority. It is possible that the exclusion of patients who seemed to be disease-free but were missing their last follow-up visits at the end of a study could have impacted on the trial results, since the proportion of patients lost to follow-up (around 10-15%) is larger than the 6% non-inferiority margin chosen for these studies. This thesis will investigate this. Additionally, the definition of the primary outcome for TB trials requires a confirmatory result to indicate a patient is cleared of TB over 18 months of follow-up. Unobserved results add an extra complexity to determining the outcome of a patient and so any missing results are usually ignored in the analysis. This means only a small proportion of patients will have completed data. TB trials could benefit substantially by including information from patients with missing observations within the primary analysis, in a statistically valid way.

Although the focus of this thesis is on tuberculosis trials, there are very few phase III non-inferiority trials that exist within that disease area. Therefore in Chapter 2 the systematic review was performed for non-inferiority trials across multiple disease areas. Chapters 3-6 then take two phase III tuberculosis studies which are used as motivating examples to assess the implications of missing data in these non-inferiority studies. Chapter 3 investigates different patterns of the missing data within the exemplar datasets and uses different multiple imputation techniques to include missing observations, resulting in "completed" datasets. Chapter 4 applies Generalised Estimating Equations and uses inverse probability weighting to account

for the missing data within these models. Chapter 5 introduces multi-state Markov models, focussing on hidden Markov models, and applies them to the two TB datasets. The results are compared to those produced from multiple imputation and from the original analysis. Reference-based sensitivity analysis via multiple imputation is introduced in Chapter 6 and we extended the methodology for use of binary outcome data. These methods are then illustrated in the REMoxTB and RIFAQUIN datasets. We finish with a summary and discussion in Chapter 7.

Chapter 2

Systematic review

2.1 Introduction

The inconsistency between statistical guidelines discussed in Chapter 1 and summarised in Table 2.1, led to the hypothesis that poor reporting would be associated with demonstrating non-inferiority. Given these inconsistencies between the guidelines, we explore what guidance is taken on board by researchers using a non-inferiority design for clinical trials that have been conducted, what is being ignored and what can be improved on. A systematic review is appropriate to summarise the methods that are currently adopted by researchers[43]. This review investigates the quality of conduct and reporting in a selection of high impact journals over a 5 year period for non-inferiority trials. This work was jointly done with Tim Morris, Katherine Fielding, James Carpenter and Patrick Phillips. At the time this review was performed, the U.S. FDA guidelines for non-inferiority[13] were not finalised and so in this chapter we refer to the draft U.S. FDA (2010) guidelines for non-inferiority clinical trials[14]. The results of this review have been published in BMJ Open[44].

2.2 Methods

Medical journals (general and internal medicine) with an impact factor greater than 10 according to the ISI web of knowledge[45] were included in the review (correct at time of search on 31th May 2015), the rationale being that articles published in these journals are likely to have the highest influence on clinical practice and be the most

rigorously conducted and reported due to the thorough editorial process. The search terms "noninferior", "non-inferior", "noninferiority" and "non-inferiority" were used in Ovid (Medline) in titles and abstracts between 1st January 2010 and 31st May 2015 in New England Journal of Medicine with an impact factor of 54.4; Lancet with an impact factor of 39.2; JAMA with an impact factor of 30.4; British Medical Journal with an impact factor of 16.4; Annals of Internal Medicine with an impact factor of 16.1; PLOS medicine with an impact factor of 14.0 and Archives of Internal Medicine with an impact factor of 13.2. From 2013, Archives of Internal Medicine was renamed JAMA Internal Medicine, and therefore both journals have been included in this review. This search was cross checked with the PubMed database using the same search terms in titles/abstracts between 1st January 2010 and 31st May 2015.

2.2.1 Inclusion/exclusion criteria

Tim Morris and I independently assessed the eligibility of articles by reviewing the abstracts of articles. Articles included were non-inferiority randomised controlled clinical trials. Articles were excluded if the primary analysis was not for non-inferiority. Systematic reviews, meta-analyses and commentaries were also excluded. A few trials were designed and analysed using Bayesian methods, and were therefore excluded for consistent comparability between trials that analysed according to frequentist methods, which were the vast majority.

Table 2.1: Summary of non-inferiority guidelines.

Guideline	Justification of margin	Who is included in analysis	Confidence interval	Missing data	Sensitivity analysis
CONSORT 2006 [1]	"Margin should be specified and preferably justified on clinical grounds"	"Non-ITT analyses might be desirable as a protection from ITTs increase in type I error. There is greater confidence in results when the conclusions are consistent." Intent-to-treat: "Analysing all patients within their randomized groups, regardless of whether they completed allocated treatment is recommended" Per-protocol: "Alternative analyses that exclude patients not taking allocated treatment or otherwise not protocol -adherent could bias the trial in either direction. The terms on-treatment or per-protocol analysis are often used but may be inadequately defined."	"Many noninferiority trials based their interpretation on the upper limit of a 1-sided 97.5% CI, which is the same as the upper limit of a 2-sided 95% CI." "Although both 1-sided and 2-sided CIs allow for inferences about noninferiority, we suggest that 2-sided CIs are appropriate in most noninferiority trials. If a 1-sided 5% significance level is deemed acceptable for the noninferiority hypothesis test (a decision open to question), a 90% 2 -sided CI could then be used."		
CONSORT 2012 [10]		"Should be indicated if conclusions are related to PP analysis, ITT analysis or both and if the conclusions are stable between them."	"The two-sided CI provides additional information, in particular for the situation in which the new treatment is superior to the reference treatment"		Sensitivity analysis is discussed through an example: "Study endpoints were analysed primarily for the per protocol population and repeated, for sensitivity reasons, for the intention-to-treat (ITT) population."
Draft U.S. FDA 2010 [14]	"Whether M1 (the effect of the active control arm relative to placebo) is based on a single study or multiple studies, the observed (if there were multiple studies) or anticipated (if there is only one study) statistical variation of the treatment effect size should contribute to the ultimate choice of M1, as should any concerns about constancy. The selection of M2 (the largest clinically acceptable difference of the test treatment compared to the active control) is then based on clinical judgement regarding how much of the M1 active comparator treatment effect can be lost. The exercise of clinical judgement for the determination of M2 should be applied after the determination of M1 has been made based on the historical data and subsequent analysis"	"It is therefore important to conduct both ITT and 'as-treated' analyses in non-inferiority studies." Intent-to-treat: "preserve the principle that all patients are analyzed according to the treatment to which they have been randomized even if they do not receive it"	"Typically, the one-sided Type I error is set at 0.025, by asking that the upper bound of the 95% CI for control-treat be less than the NI margin. If multiple studies provide very homogeneous results for one or more important endpoints it may be possible to use the 90% lower bound rather than the 95% lower bound of the CI to determine the active control effect size"	"Poor quality can reduce the drug's effect size and undermine the assumption of the effect size of the control agent, giving the study a 'bias towards the null".	

ICH E9 [6]	"This margin is the largest difference that can be judged as being clinically acceptable"	"In confirmatory trials it is usually appropriate to plan to conduct both an analysis of the full analysis set and a per protocol analysis In an equivalence or non-inferiority trial use of the full analysis set is generally not conservative and its role should be considered very carefully." Intent-to-treat: "subjects allocated to a treatment group should be followed up, assessed and analysed as members of that group irrespective of their compliance to the planned course of treatment" Full analysis set: "The set of subjects that is as close as possible to the ideal implied by the intention-to-treat principle. It is derived from the set of all randomised subjects by minimal and justified elimination of subjects." Per-protocol: "The set of data generated by the subset of subjects who complied with the protocol sufficiently to ensure that these data would be likely to exhibit the effects of treatment, according to the underlying scientific model. Compliance covers such considerations as exposure to treatment, availability of measurements and absence of major protocol violations."	"For non-inferiority trials a one-sided interval should be used. The choice of type I error should be a consideration separate from the use of a one-sided or two-sided procedure."	"Imputation techniques, ranging from LOCF to the use of complex mathematical models may be used to compensate for missing data"	"An investigation should be made concerning the sensitivity of the results of analysis to the method of handling missing values, especially if the number of missing values is substantial."
ICH E10 [7]	"The determination of the margin in a non-inferiority trial is based on both statistical reasoning and clinical judgement"	Use an example where "non-inferiority would be claimed if both ITT and PP analysis show conclusions of NI." Intent-to-treat: "In order to preserve the unique benefit of randomisation as a mechanism to avoid selection bias, an "as randomised" analysis retains participants in the group to which they were originally allocated. To prevent attrition bias, out-come data obtained from all participants are included in the data analysis, regardless of protocol adherence." Per-protocol and modified intention-to-treat: "Some trialists use other types of data analyses (commonly labelled as modified intention to treat or "per protocol") that exclude data from certain participantssuch as those who are found to be ineligible after randomisation or who deviate from the intervention or follow-up protocols. This exclusion of data from protocol non-adherers can introduce bias, particularly if the frequency of and the reasons for non-adherence vary between the study groups.			
SPIRIT [15]				"Multiple imputation can be used to handle missing data although relies on untestable assumptions"	"Sensitivity analyses are highly recommended to assess the robustness of trial results under different methods of handling missing data"

EMA 2006 [9]	"The choice of delta must always be justified on both clinical and statistical grounds"		"A two-sided 95% CI (or one-sided 97.5% CI) is constructed. The interval should lie entirely on the positive side of the margin. Statistical significance is generally assessed using the two-sided 0.05 level of significance (or one-sided 0.025)"	
EMA 2000 [8]		"ITT and PP analyses have equal importance and their use should lead to similar conclusions for robust interpretation"	"A two-sided confidence interval should lie entirely to the right of delta. If one-sided confidence is used then 97.5% should be used"	"It will be necessary to pay particular attention to demonstrating the sensitivity of the trial by showing similar results for the full analysis set and PP analysis set"

2.2.2 Data extraction

Before performing the review, a data extraction form (see Appendix A) was developed to extract information from articles. The form was tested by two reviewers (Tim Morris and I) on articles, included in this review, until agreement was achieved between both reviewers. The form was standardised to collect information on year of publication, non-inferiority margin (and how the margin was justified), randomisation, type of intervention, disease area, sample size, analysis performed (how this was defined and what was classed as primary/secondary), primary outcome, p-values (and whether this was for a superiority hypothesis), significance level of confidence intervals (and whether both bounds were reported), imputation techniques for missing data, sensitivity analyses, conclusions of non-inferiority and whether a test for superiority was pre-specified.

Non-inferiority margin

The size of the margin chosen was recorded and justification of the margin was noted according to what the authors reported. Any attempt the authors made at justifying the margin was recorded as a justification. These were subsequently classed according to what basis the margin was justified by two reviewers (Patrick Phillips and I), due to the subjectivity of classifications.

Analysis chosen by authors

Participants to be included in analyses were classed according to what was stated within articles along with how the analysis for the population was defined by authors within the main text (if available). Definitions of the analysis was summarised according to authors definition. If the type of analysis was not explicitly defined within articles, the definition was recorded according to what was shown in the CONSORT flow chart or otherwise implicitly. ITT, PP, mITT and as-treated were considered to be the most well-known. If the analyses was defined by authors but did not include a classification then they were categorised in the following way:

- All patients randomised into the study were analysed was classed as an intention-to-treat analysis;

- Patients who were excluded after administration of treatment (e.g. withdrawals, loss to follow up, compliance) was classed as a per-protocol analysis;
- Patients who were excluded after administration of treatment, but the exclusion was not treatment related (e.g. patients who did not have the disease of interest) was classed as a modified intention-to-treat analysis; and
- Analysis based on what treatment patients actually received as opposed to the treatment that was allocated at the time of randomisation was classed as an as-treated analysis.

Information on whether the analysis was considered as a primary analysis or secondary analysis (for the same primary outcome) was collected. The analysis was assumed primary if only one analysis was reported. If more than one analysis was performed but it was not clearly described which was to be taken as the primary and/or secondary analysis, the primary analysis was assumed to be whatever was presented in the results section of the abstract and secondary if not presented in the abstract but stated elsewhere within the article. If all results were presented for all populations in the abstract, then both were assumed as primary unless non-inferiority was concluded on only one analysis. Analysis was assumed secondary if the patient population was stated but not defined or if the results of the analysis were not presented in the article.

Sample size

Power and significance levels were recorded according to the planned sample size calculation from the methods section of articles, and whether the significance level was calculated for a one-sided or two-side test.

Confidence intervals

The significance level of the confidence interval was recorded and whether the judgement about non-inferiority was made on the upper or lower bound of a confidence interval. We also recorded whether the interval was presented for both bounds or for one bound.

Missing data

The amount of missing data in the primary outcome was recorded, either from the text, tables or the CONSORT flow chart. A range of missing data, or a minimum/maximum was taken if the exact amount of missing data could not be determined. Whether or not any imputation techniques to handle missing data were considered for the primary outcome was recorded along with the technique used. If multiple imputation was used, then whether the number of repetitions was stated and whether authors stated assumptions made about missing data according to Rubin's Rules[34] were recorded.

Sensitivity analysis

We recorded whether sensitivity analyses were considered for the primary outcome, what they were and if they were performed to test assumptions made about the missing data for primary analyses.

Hypothesis

A quality grading system was developed based on whether the margin was justified (yes vs. no/poor), how many analyses were performed on the primary outcome (<2 vs. ≥ 2) and whether the type I error rate was consistent with the significance level of the confidence interval (yes vs. no/unclear). Articles were classed as "excellent" if all these criteria were fulfilled and were classed as "poor" if none were fulfilled. Articles which satisfied one criterion were classed as "fair" and articles that provided two of the three criteria were classed as "good". The results of this grading were compared to inferences on non-inferiority to assess if the quality of reporting was associated with concluding non-inferiority.

Subgroup of trials with published protocols

Additional supplementary content was only accessed if it specifically referred to the information we were extracting within articles. As a sub-study, all statistical methods, outcomes and sample sizes from protocols and/or supplementary content were reviewed from New England Journal of Medicine as the journal is known to specifically request and publish protocols and statistical analysis plans alongside accepted publications.

Assessments

If more than two arms were compared, information was taken from the first comparison that was presented in the abstract, unless the comparison was for superiority (e.g. placebo vs. standard treatment vs. experimental arm). For articles that reported more than one trial, information was captured from the trial whose results were presented first in the abstract. If more than one primary outcome was assessed, information was recorded based on the primary outcome that was presented first in the results of the abstract.

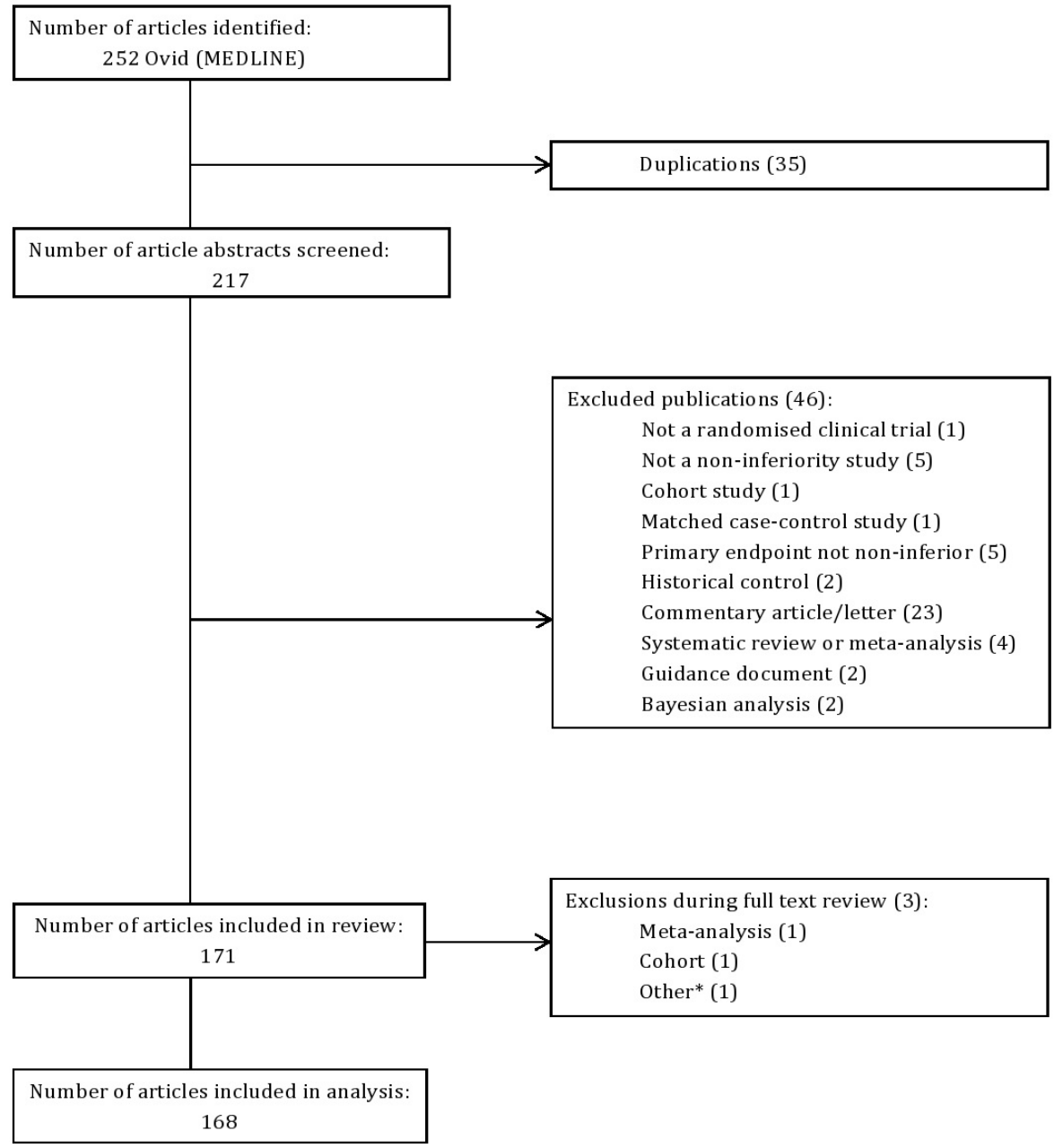
I performed all the assessments and a random selection of 5% of articles were independently reviewed by Patrick Phillips. Any assessments that required a second opinion were independently reviewed by Tim Morris. Any discrepancies were resolved by discussion between the reviewers. All analyses were conducted using Stata version 14.

2.3 Results

The search found 252 articles. After 35 duplicate publications were removed, 217 were screened for eligibility using their titles and abstracts. A total of 46 articles were excluded (Figure 2.1) leaving 171 articles to be reviewed. Three articles were excluded during the full-text review: one was a meta-analysis; one presented results of follow up data from a study which had already been included in the review and one was a cohort study. Therefore, a total of 168 articles were reviewed.

General characteristics are shown in Table 2.2. Most articles (164; 98%) specified the threshold of the margin to determine non-inferiority. Almost half of the articles defined a composite primary outcome 78 (46%). Non-inferiority trials were most common for those investigating heart disease 30 (18%), followed by HIV 18 (11%), cancer 16 (10%), bleeding 14 (8%) and diabetes 11 (7%). Statistical power used for sample size calculations ranged between 71 and 99%, with 80% or 90% power being more common: 61 (36%) and 65 (39%) trials respectively.

Figure 2.1: Flow chart of eligibility of articles.



* Secondary analyses. Primary analyses for the same study was included in the review

Table 2.2: General characteristics.

	All articles n=168	**Including NEJM protocols (n=61)**
Characteristics	n (%)	n(%)
Journal		
NEJM	61 (36%)	61
Lancet	64 (38%)	
JAMA	19 (11%)	
BMJ	8 (5%)	
Annals of Internal Medicine	5 (2%)	
PLOS Medicine	7 (4%)	
Archives of Internal Medicine	2 (1%)	
JAMA of Internal Medicine	2 (1%)	
Year of publication		
2010	26 (15%)	9 (15%)
2011	27 (16%)	9 (15%)
2012	29 (17%)	8 (13%)
2013	39 (23%)	19 (31%)
2014	27 (16%)	10 (16%)
2015	20 (12%)	6 (10%)
Type of intervention		
Drug	112 (67%)	44 (72%)
Surgery	22 (13%)	7 (11%)
Other	34 (20%)	10 (16%)
Randomisation		
Patient	163 (97%)	59 (97%)
Cluster	5 (3%)	2 (3%)
Margin specified		
Yes	164 (98%)	58 (95%)
No	4 (2%)	3 (5%)
Power		
80%	61 (36%)	19 (31%)

85%	11 (7%)	5 (8%)
90%	65 (39%)	26 (43%)
71 to 99% (Excluding the above)	21 (12%)	11 (18%)
Not reported/unclear	10 (6%)	0
Composite outcome		
Yes	78 (46%)	37 (61%)
No	90 (54%)	24 (39%)
Disease		
Heart disease	30 (18%)	13 (21%)
Blood disorder	19(11%)	6 (10%)
Cancer	16 (10%)	8 (13%)
Diabetes	11 (7%)	2 (3%)
Thromboembolism	6 (4%)	6 (10%)
Skin infection	3 (2%)	2 (3%)
Urinary tract infection	3 (2%)	0
Arthritis	3 (2%)	1 (2%)
Ophthalmology	3 (2%)	1 (2%)
Pneumonia	3 (2%)	1 (2%)
Complications in pregnancy	3 (2%)	0
Stroke	3 (2%)	2 (3%)
Testing method	3 (2%)	1 (2%)
Appendicitis	2 (1%)	1 (2%)
Depression	2 (1%)	0
Other Non-infectious disease	18 (11%)	7 (11%)
HIV	18 (11%)	2 (3%)
Tuberculosis	6 (4%)	4 (7%)
Malaria	4 (2%)	1 (2%)
Skin infection	2 (1%)	0
Hepatitis C	2 (1%)	2 (3%)
Other infectious disease	8 (5%)	1 (2%)

Table 2.3: Justification of choice of margin, total number of patient populations considered for analyses and patient population included in analysis.

	All articles (n=168)	**Including NEJM protocols (n=61)**
	n (%)	n(%)
Justification of NI margin		
Made no attempt for justification	90 (54%)	22 (36%)
Clinical basis. No evidence for consultation with external expert group, and no reference to previous trials of the control arm	32 (19%)	11 (18%)
Preservation of treatment effect based on estimates of control arm effect from previous trials	13 (8%)	14 (23%)
Expert group external to the authors. No reference to previous trials of the control arm	6 (4%)	3 (5%)
The same margin as was used in other similar trials	5 (3%)	2 (3%)
10-12% recommended by disease specific U.S. FDA guidelines	4 (2%)	1 (2%)
General comment that margin was decided according to U.S. FDA/regulatory guidance	4 (2%)	0
Clinical basis and based on previous similar trial. No evidence for consultation with external expert group, and no reference to previous trials of the control arm	3 (2%)	0
Based on registry/development program	0	2 (3%)
Based on previous trial. No evidence for consultation with external expert group, and no reference to previous trial of the control arm	1 (1%)	1 (2%)
Based on unpublished data. No evidence for consultation with external expert group, and no reference to previous trials of the control arm	1 (1%)	0

Clinical basis and based on previous trials and guidelines. No evidence for consultation with external expert group, and no reference to previous trials of the control arm	1 (1%)	0
Clinical basis. Attempted to justify based on preservation of treatment effect, but were unable to do so due to paucity of previous trials	1 (1%)	0
Expert group external to the authors and previous trial. No reference to previous trials of the control arm	1 (1%)	0
Justified based on treatment effect of control, but margin actually bigger than control arm treatment effect	1 (1%)	1 (2%)
Placebo controlled study. Clinical basis, previous trials and literature review	1 (1%)	0
Preservation of treatment effect. Reference to separate paper justifying margin	1 (1%)	1 (2%)
Regulatory guidelines (WHO), but recommendation is for superiority. No evidence for consultation with external expert group, and no reference to previous trials of the control arm	1 (1%)	0 (0%)
Synthesis approach	1 (1%)	0 (0%)
Unclear	1 (1%)	0 (0%)
General comment that margin was decided according to U.S. FDA request	0	1 (2%)
Preservation of treatment effect based on estimates of control arm effect from previous trials and clinical basis	0	1 (2%)
Preservation of treatment effect based on estimates of control arm effect from previous trials, clinical basis and according to U.S. FDA guidelines	0	1 (2%)
Number of analyses		
One	65 (39%)	15 (25%)
Two	91 (54%)	38 (62%)
Three	10 (6%)	7 (11%)

Not defined	2 (1%)	1 (2%)
Analysis		
ITT	129 (77%)	44 (72%)
PP	90 (54%)	35 (57%)
mITT	34 (20%)	17 (28%)
As-treated	4 (2%)	6 (10%)
Other	20 (12%)	10 (16%)
Unclear	2 (1%)	2 (3%)

Justification of the non-inferiority margin

The non-inferiority margin was justified in less than half of articles 76 (45%). The most common justification was on a clinical basis (32 (19%)) which was often worded ambiguously and with little detail. A total of 13 (8%) used previous findings from past trials or statistical reviews to justify the choice of the margin (Table 2.3).

Analyses performed

Over a third of articles 65 (39%) performed only one analysis (Table 2.3) rather than presenting both an ITT and PP analyses as recommended in most guidelines. A total of 129 (77%) articles presented at least one ITT analysis, of which 68/129 (53%) defined ITT analysis as "all patients randomised into the study" and 21/129 (16%) defined the ITT population as "all patients randomised who took at least one dose of treatment/intervention" (whichever was appropriate; Table 2.4). The number of studies that performed only an ITT analysis or both ITT and PP analyses was about the same: 54 (32%) and 56 (33%) respectively (Table 2.5). PP analyses were performed in 90 (54%) trials (Table 2.3) of which 11 (12%) did not define what was meant by "per-protocol" (Table 2.4). Definitions of the PP population contained various exclusions, mostly regarding errors in randomised treatment or treatment received. Some exclusions were with respect to protocol adherence or deviations, other exclusions were due to missing outcome data, ineligibility, withdrawals or errors in randomisation. There were a variety of other definitions 20 (12%), including a modified per-protocol analysis which excluded all but one of the protocol deviations; the per-protocol analysis excluded all protocol deviations. The majority of studies

classed ITT analysis as primary and PP analyses as secondary (Figure 2.2). There were 10 (6%) studies that chose to perform three or more analyses on different patient populations.

Type I error rate

Consistency between the type I error rate and confidence intervals reported was moderate at 95 (57%): 11 (42%) in 2010; 15 (56%) in 2011; 15 (52%) in 2012; 24 (62%) in 2013; 19 (70%) in 2014 and 11 (55%) by May 2015 (Table 2.6). Most articles, 69 (41%), used a one-sided 2.5% or (numerically equivalent) two-sided 5% significance level (Table 2.7) and some used a one-sided 5% significance level or (numerically equivalent) two-sided 10% significance level, 48 (28%). The majority of articles presented two-sided confidence intervals (147; 88%) and 19 (11%) articles presented one-sided confidence intervals. Most two-sided confidence intervals were at the 95% significance level: 125 (74%).

Missing data

Imputation techniques to account for missing primary outcome data were carried out for 57 (34%) trials. A total of 99 (59%) trials did not report whether or not any imputation was done and only 12 (7%) explicitly declared that no imputation was used. Assuming a worst-case scenario or multiple imputation were the most common methods used, 19/57 (33%) and 11/57 (19%) respectively (Table 2.8). The number of imputations used for multiple imputation was specified in 8/11 articles and 4/11 stated at least one of the assumptions from Rubin's rules[34].

Table 2.4: Definition of analysis

Analysis	Definition	n (%)
ITT		**129**
	All patients randomised	68 (53%)
	All patients randomised who received at least one dose of treatment/intervention	21 (16%)
	All patients randomised excluding missing data	7 (5%)
	All patients randomised excluding errors in randomisation	3 (2%)
	All patients randomised who received at least one dose of treatment/intervention, excluding missing data	1 (1%)
	All patients randomised with exclusions from one centre which was removed due to misconduct	1 (1%)

	Other	17 (13%)
	Unclear	1 (1%)
	Not defined	10 (8%)
PP		**90**
	Patients who received allocated treatment/intervention	8 (9%)
	Excluding patients with major protocol violations	5 (6%)
	Patients who completed allocated treatment/intervention as intended	4 (4%)
	Patients who adhered to treatment	2 (2%)
	Excluding patients with protocol deviations	2 (2%)
	Patients with no exclusion criteria and who received specific amount of treatment/intervention	2 (2%)
	Patients who received allocated treatment/intervention, no major protocol violations with outcome	2 (2%)
	Excluding patients who switched treatment	1 (1%)
	Patients who received at least one dose of treatment/intervention	1 (1%)
	Patients who adhered to the protocol	1 (1%)
	Patients who completed the assigned study regimen or adhered to treatment before an event	1 (1%)
	Patients who received correctly allocated treatment/intervention excluding withdrawals	1 (1%)
	Patients who received specific amount of treatment/intervention and adhered to protocol	1 (1%)
	Patients who received allocated treatment/intervention, excluding non-adherence	1 (1%)
	Patients who adhered to protocol excluding withdrawals	1 (1%)
	Excluded patients with protocol deviations in addition to mITT definition	1 (1%)
	Excluded patients that received rescue medication and protocol violations	1 (1%)
	Patients who received at least one dose of drug/intervention and received allocated treatment/intervention excluding missing outcome data	1 (1%)
	All patients who received at least one dose of treatment/intervention and did not have major protocol violations and were followed for event while receiving drug	1 (1%)
	All patients who received at least one dose of treatment/intervention and did not have major protocol violations	1 (1%)
	Excluding patients who were ineligible, excluding patients who were administered the incorrect dose of medication and excluding patients who were allocated the incorrect treatment	1 (1%)
	All patients randomised who received at least one dose of treatment/intervention with an outcome, completed the study and complied with protocol	1 (1%)
	Non-adherence, patients who declined follow up, errors in randomisation, recurrent atrial fibrillation before randomisation were excluded	1 (1%)

	The per-protocol population (which consisted of the modified intention-to-treat population with the exclusion of patients with major protocol deviations and a compliance rate of <80%) was of primary interest, since a noninferiority analysis that is based on the modified intention-to-treat population is deemed to be not conservative	1 (1%)
	Patients were not eligible for per-protocol analysis for the following reasons: no follow-up visit; systemic treatment with other antimicrobial drugs up to day 28 (visit three); or missing more than one dose of the study drug during the first week of treatment or more than two doses during the whole treatment period	1 (1%)
	Excluded missing inclusion criteria; incorrect dosing; received prohibited medication; missing assessments	1 (1%)
	Per-protocol analyses excluded participants who had missing data at 1 month or who had major protocol violations (e.g., death, pregnancy, withdrawal from the study, loss to follow-up, or noncompliance). NB: Two results were presented for PP where compliance was included and excluded.	1 (1%)
	Per-protocol prespecified analyses included children with complete follow-up or a confirmed treatment failure, and excluded those treated for malaria without confirmatory microscopy, those for whom the alternative Plasmodium species was detected, and those who defaulted from follow-up despite repeated attempts at contact Flow chart includes: "and followed protocol"	1 (1%)
	Patients who, during the intended treatment period, had a venogram adjudicated as assessable, who developed confirmed deep vein thrombosis or pulmonary embolism, or who died from any cause); patients who had important protocol violations were excluded from the per-protocol analysis.	1 (1%)
	The per-protocol population was defined as all patients included in the ITT analysis, excluding those who did not receive the regimen as prescribed. These were patients who received less than 6 weeks of treatment (42 days of daily treatment or 36 days of 6-days-a-week treatment) or more than 9 weeks of treatment (63 days of daily treatment or 54 days of 6-days-a-week treatment) in the intensive phase and those who received less than 42 doses (ie, 4 weeks of missed treatment) or more than 60 doses (ie, 2 weeks of extra treatment) in the continuation phase (the protocol requirement is that patients receive 18 weeks of 3- times-weekly treatment, ie, 54 doses). Also excluded were patients whose treatment was modified for reasons other than bacteriological failure or relapse (including patients changing treatment for adverse drug reactions, following return after default, or attributable to concomitant HIV infection).	1 (1%)
	Per-protocol snapshot analysis, which included all participants who were enrolled, received at least one dose of study drug, and did not meet any of the following pre-specified criteria: discontinuation of study drug before week 48 or HIV RNA data missing in week 48 analysis window (accounting for 80% of excluded patients), and adherence in the bottom 2.5th percentile (accounting for 20% of the excluded patients)	1 (1%)
	The perprotocol group consisted of all patients who were enrolled, had no major protocol deviation, received the full treatment, and were assessed at day 15 or 31, day 45, and 6 months (-2 to $+6$ weeks).	1 (1%)

	Criteria to exclude patients from this set were violation of major in- or exclusion criteria, change of treatment arm, early treatment discontinuation or relevant dose deviations of chemo- or radiotherapy unless caused by death or progression, radiotherapy without PET panel recommendation or omission of radiotherapy against recommendation, PET panel decision to take the patient off protocol treatment, or missing documentation of treatment	1 (1%)
	The per-protocol analysis set additionally excludes patients with change of treatment arm, early treatment discontinuation or relevant dose deviations of chemo- or radiotherapy unless caused by death or progression, or missing documentation of treatment	1 (1%)
	The perprotocol analysis was based on all participants who received 3 doses of vaccine according to 1 of the studys vaccine dosing schedules, were seronegative to the relevant HPV type at baseline, and had a valid serology result after the third dose of the HPV vaccine	1 (1%)
	Not defined. Taken from flow chart: Patients not meeting the definition of having received adequate treatment provided they have not already had an unfavourable response to treatment. Other exclusions done as well, but are not defined in flow chart	1 (1%)
	All patients who underwent randomization, completed a full treatment course or had early treatment failure before treatment was completed, had outcome data for the primary efficacy end point on day 28, and complied with the protocol to the extent that would allow efficacy evaluation	1 (1%)
	We also conducted a perprotocol analysis, which included those who completed the 2-month visit while receiving treatment (108 oral, 113 intratympanic) because intention-to-treat analyses may bias toward noninferiority. Flow chart also shows patients who withdrew before the 2m follow up, those who discontinued treatment but completed follow up and those who completed treatment but missed 2m follow up were excluded.	1 (1%)
	Which consisted of participants who received all three doses of vaccine within 1 year, did not have the HPV type being analyzed (i.e., were seronegative on day 1 and PCR-negative from day 1 through month 7), and had no protocol violations	1 (1%)
	A total of 12 (10%) patients in each group did not undergo PEG for anatomical reasons. Between the PEG procedure and the follow-up visit, five patients died, one patient pulled out the PEG catheter without ensuing complications, three patients were lost to follow-up, and one patient who was randomised to cefuroxime received co-trimoxazole instead.	1 (1%)
	Will include all subjects in the MITT population grouped by randomized treatment assignment regardless of treatment received with the exception of the following additional exclusions 1. Subjects not meeting the definition of having received an adequate amount of their allocated study regimen (see below for definition), provided they have not already been classified as having an unfavourable outcome 2. Subjects lost to follow-up or withdrawn before the Month 6 visit, unless they have already been classified as having an unfavourable outcome.	1 (1%)

	3. Subjects whose treatment was modified or extended for reasons (e.g. an adverse drug reaction or pregnancy) other than an unfavourable therapeutic response to treatment, unless they have already been classified as having an unfavourable outcome 4. Subjects who are classified as "major protocol violations" (see section 6.5), unless they have already been classified as having an unfavourable outcome on the basis of data obtained prior to the protocol violation	
	The per-protocol analysis excluding the 6 patients who were lost to follow-up and the 3 patients who received postoperative corticosteroids (including the 4 patients who experienced primary bleeding events)	1 (1%)
	Excluded patients who received a platelet transfusion for reasons not recommended in the protocol	1 (1%)
	We also did a per-protocol analysis of the medical outcomes, excluding outpatients discharged more than 24 h after randomisation and inpatients discharged 24 h or less after randomisation.	1 (1%)
	The perprotocol population was defined as intention-to-treat patients with (1) successful procedure outcome, (2) treatment solely with the zotarolimus-eluting stent, (3) dual antiplatelet therapy according to randomization, and (4) complete clinical follow-up information.	1 (1%)
	Not defined. Flow chart shows the following exclusions: had another histology or malignancy; withdrew informed consent; had an allergic reaction on first rituximab infusion and consecutively other treatment; only had radiotherapy; received incorrectly allocated treatment; did not meet inclusion or exclusion criteria; no therapy; death before therapy	1 (1%)
	Not defined. Flow chart suggests patients were excluded if they did not receive the protocol and withdrawals	1 (1%)
	Censoring of events if any component of the initial randomised trial treatment was stopped	1 (1%)
	Not defined. Flow chart shows inclusion/exclusion criteria violated, non-adherence, prohibited medication and missing results were excluded	1 (1%)
	Participants who did not follow protocol and/or were seropositive or polymerase chain reaction-positive for HPV-16, HPV- 18, HPV-6, or HPV-11 at enrolment were excluded from the per-protocol population analysis but retained for the intention-to-treat population analysis. Participants were eligible to continue with the 18- and 36-month follow-up if they had all of their doses of vaccine and a 7-month blood sample collected. If participants were excluded from the per-protocol population analysis at 7 months, they remained excluded for the remainder of the study but were retained for intention- to-treat analysis.	1 (1%)
	The per-protocol population included all patients who completed the study (1 year), and for whom the second reading of a CT-scan confirmed the diagnosis of uncomplicated appendicitis.	1 (1%)

	For analyses based on the per-protocol population, patients were analysed according to their randomly assigned treatment group. To be included in the perprotocol population, a patient was required to meet the following criteria: Had a mean baseline hemoglobin $\geq$ 8.0 and <11.0 g/dl; Completed the study through at least week 36, and at least 5 hemoglobin values were obtained during the evaluation period; Had no missing administrations of study medication between weeks 21 and 35, inclusive; Had not received any RBC or whole blood transfusions within the 12 weeks prior to randomization; Had not received any RBC or whole blood transfusions for reasons other than lack of effect of study medication (lack of effect of study medication was documented as "Anemia of CRF" on the case report form) between weeks 21 and 35, inclusive; Had not received any ESA other than the assigned study treatment between weeks 21 and 35, inclusive; Had adequate iron status at baseline and during the evaluation period (defined as serum ferritin $\geq$ 100 ng/ml and TSAT $\geq$ 20% during weeks 24, 28, and 32)	1 (1%)
	Not defined. Flow chart shows exclusions: caesarean section or forceps; short umbilical cord or nuchal cord; need for resuscitation; team became unavailable; weight scale malfunctioned; parent withdrew consent	1 (1%)
	Completers (observed cases; included patients in the full analysis set who did not have important protocol violations, completed at least 684 days of treatment, and had HbA1c measured at week 104)	1 (1%)
	For analyses based on the per-protocol population, patients were analyzed according to their randomly assigned treatment group. To be included in the per-protocol population, a patient was required to meet the following criteria: Had a mean baseline hemoglobin $\geq$ 10.0 and $\leq$ 12.0 g/dl; Completed the study through at least week 36, and at least six haemoglobin values were obtained during the evaluation period.; Received 75% of total prescribed (i.e., expected) doses of study medication between weeks 25 and 35, inclusive (detailed algorithms for this determination were specified in the Statistical Analysis Plan).; Had not received any RBC transfusions within the 12 weeks prior to randomization.; Had not received any RBC transfusions for reasons other than lack of effect of study medication (lack of effect of study medication was documented as "Anemia of CRF" on the case report form) between weeks 25 and 36, inclusive.; Had not received any ESA other than the assigned study treatment between weeks 25 and 35, inclusive.; Had adequate iron status at baseline and at week 36 (defined as serum ferritin $\geq$ 100 ng/ml and TSAT $\geq$ 20%).	1 (1%)
	This population included all patients who underwent randomisation and who completed the study procedures to month 6.	1 (1%)
	We also performed a per-protocol analysis, which notably excluded patients in the antibiotic group who had been switched from amoxicillin plus clavulanic acid to another antibiotic.	1 (1%)
	We did a per-protocol snapshot analysis, which included all participants who were randomly assigned treatment, received at least one dose of study drug, and did not meet any of the following prespecified criteria: discontinuation of study drug before week 48 or HIV RNA results missing in the week 48 analysis window, and adherence in the bottom 2.5th percentile.	1 (1%)

	Patients were included in the per-protocol population if they met the criteria for inclusion in the modified intention-to-treat population, underwent an adequate assessment of venous thromboembolism not later than 2 days after administration of the last dose of study drug, and had no major protocol violations.	1 (1%)
	The perprotocol population comprised patients in the modified intention-to-treat group who received treatment for at least 3 days (in the case of patients with treatment failure) or at least 8 days (in the case of patients with clinical cure), had documented adherence to the protocol, and underwent an end-of-therapy evaluation.	1 (1%)
	The per-protocol analysis set consisted of participants with exposure to treatment for at least 12 weeks who did not have any major protocol violations that could affect the primary endpoint and had a valid glycated haemoglobin (HbA1c) assessment at baseline and at (or after) 12 weeks.	1 (1%)
	Not defined	11 (12%)
mITT		**34**
	All patients randomised who received at least one dose of treatment/intervention	10 (29%)
	All patients randomised who received at least one dose of treatment/intervention, excluding missing data	6 (18%)
	All patients randomised with at least one dose of treatment/intervention excluding patients/site with violations of GCP	2 (6%)
	All randomised patients who received at least one dose of treatment/intervention excluding patients without disease or excluding patients resistant to one of the drug combinations. Excluding patients whose death was not related to the disease or had reinfection after being cured or patients who were classed as unassessable at the endpoint	1 (3%)
	Patients were excluded if they were resistant to two of the treatment combinations and patients who were unassessable and had not reached endpoint	1 (3%)
	On-treatment which included events that occurred within 30 days after the last dose of study medication was administered	1 (3%)
	Patients were excluded if they had missing/contaminated outcome data or could not produce an assessment or were lost to follow up or had death not related to disease or had confirmed reinfection	1 (3%)
	Excluded if consent withdrawn, non-compliance, moved and other (other not defined)	1 (3%)
	Other	11 (32%)
As-treated		**4**
	All patients randomised who received intervention	1 (25%)
	Not defined	3 (75%)
Other		**20**
	Full analysis set	4 (20%)
	On treatment analysis	3 (15%)
	Complete follow up data	1 (5%)
	ITT efficacy	1 (5%)

	PP and modified PP	1 (5%)
	Should be classed as PP. All patients who completed study with no major protocol deviations	1 (5%)
	Should be classed as mITT	2 (10%)
	Should be classed as mITT (ITT with no exclusion criteria)	1 (5%)
	Should be as treated (treatment received)	1 (5%)
	Other	5 (25%)
Unclear		**2**

Table 2.5: Type of primary analysis chosen.

	All articles (n=168)	**Including NEJM protocols (n=61)**
Analysis	n (%)	n(%)
ITT only	54 (32%)	12 (20%)
PP only	3 (2%)	0
mITT only	8 (5%)	3 (5%)
ITT & PP	56 (33%)	17 (28%)
ITT & mITT	3 (2%)	2 (3%)
ITT & as-treated	4 (2%)	4 (7%)
ITT & other definition	6 (4%)	2 (3%)
PP & mITT	17 (10%)	9 (15%)
PP & other definition	4 (2%)	2 (3%)
PP & other definition	4 (2%)	2 (3%)
mITT & as-treated	0	1 (2%)
mITT & other definition	1 (1%)	1 (2%))
ITT, PP & mITT	1 (1%)	1 (2%)
ITT, PP & as-treated	0	1 (2%)
ITT, PP & other definition	5 (3%)	5 (8%)
ITT,PP & other definition	4 (2%)	0
Unclear	2 (1%)	1 (2%)

NB: One study performed ITT and PP analyses but it was unclear which of the two was taken as primary and secondary

Figure 2.2: Chosen analysis by primary or secondary analysis.

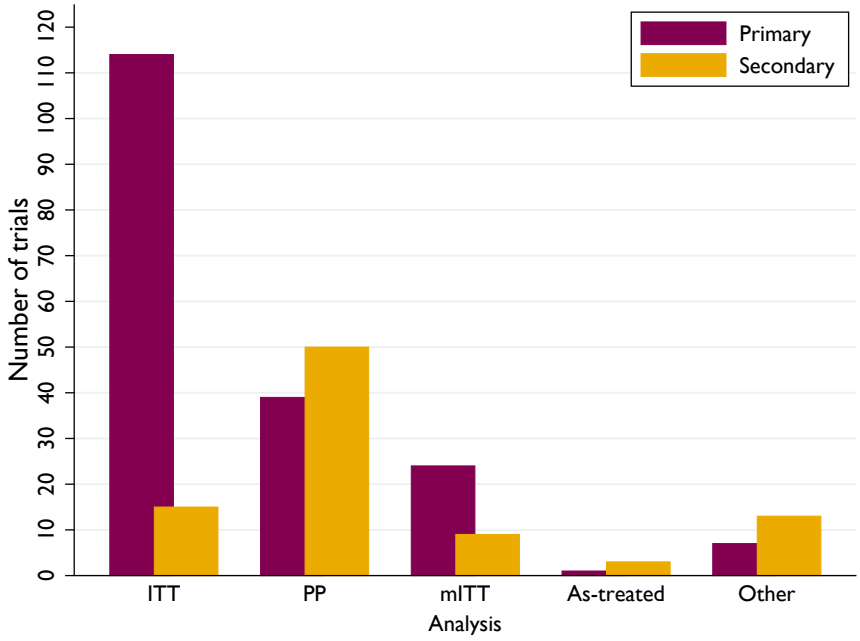


Table 2.6: Consistency of type I error rate with significance levels of confidence intervals over year.

	2010	2011	2012	2013	2014	2015	Total
All articles N=168							
Yes	11 (42%)	15 (56%)	15 (52%)	24 (62%)	19 (70%)	11 (55%)	**95**
No	5 (19%)	4 (15%)	4 (14%)	5 (13%)	5 (19%)	3 (15%)	**26**
Not reported	10 (38%)	8 (30%)	10 (34%)	10 (26%)	3 (11%)	6 (30%)	**47**
NEJM subgroup N=61							
Yes	7 (78%)	6 (67%)	5 (63%)	14 (74%)	8 (80%)	4 (67%)	**44**
No	1 (11%)	2 (22%)	2 (25%)	3 (16%)	2 (20%)	1 (17%)	**11**
Not reported	1 (11%)	1 (11%)	1 (13%)	2 (11%)	0	1 (17%)	**6**

Table 2.7: Significance level of a) type I error rate and b) confidence intervals for all articles.

a) Type I error rate			
	One sided	Two sided	Not reported
0.8	0	1 (1%)	0
1.25	3 (2%)	0	0
2.45	1 (1%)	0	0
2.5	40 (24%)	2 (1%)	2 (1%)
5	46 (27%)	29 (17%)	15 (9%)
10	1 (1%)	2 (1%)	0
Not reported	3 (2%)	0	23 (14%)
b) Significance level of confidence interval			
	One sided	Two sided	Not reported
90	1 (1%)	14 (8%)	1 (1%)
95	14 (8%)	125 (74%)	0
97.5	4 (2%)	7 (4%)	0
Other	0	1 (1%)	0
Not reported	0	0	1 (1%)

Table 2.8: Reporting of a)missing data and b)sensitivity analyses.

	n (%)
a) Imputation performed	
Yes	**56 (33%)**
Worst case scenario	19 (34%)
Multiple imputation	11 (20%)
Last observation carried forward	8 (14%)
Complete case analysis	6 (11%)
Best case scenario	2 (4%)
Last observation carried forward and worst case scenario	2 (4%)
Best case/worst case scenario	3 (5%)
Mean imputation	1 (2%)
Complete case analysis, multiple imputation using propensity scores and multiple imputation using regression modelling	1 (2%)
Other and worst case scenario	1 (2%)
Other	1 (2%)
No	12 (7%)
Not reported	99 (59%)
Unclear	1 (1%)
Including NEJM protocols (N=61)	
Yes	22 (36%)
No	7 (11%)
Not reported	31 (51%)
Unclear	1 (2%)
b) Sensitivity analyses performed	
Yes	**64 (38%)**
Patient population	13 (20%)
Competing risks	2 (3%)
Statistical modelling	2 (3%)
Adjusted for baseline variables	1 (2%)

Excluded protocol violations	1 (2%)
On-treatment	1 (2%)
Patient population/other	1 (2%)
Unclear	2 (3%)
Other	15 (23%)
Missing data	**27 (42%)**
Best case/worst case scenario	5
Complete case analysis	3
Imputation of missing values	3
Multiple imputation	3
Worst case scenario	3
Baseline observation carried forward	1
Baseline observation carried forward and complete case analysis	1
Complete case analysis, multiple imputation using propensity scores and multiple imputation using regression modelling	1
Complete case analysis and missing not at random	1
Complete case analysis and best case scenario	1
Different methods	1
Last observation carried forward	1
Modelling	1
Observed-failure	1
Worst case scenario and last observation carried forward	1
No	**103 (61%)**
Unclear	**1 (1%)**
Including NEJM protocols	
Yes	38 (62%)
No	23 (38%)

Sensitivity analyses

A total of 64 (38%) trials reported using sensitivity analyses to test robustness of conclusions of the primary outcome; of these 27/64 (42%) were related to assumptions about the missing data (Table 2.8). Performing additional analyses on a patient population was considered as a sensitivity analysis in 13/64 (20%).

Study conclusions

There were 7 (4%) articles that could not make definitive conclusions (noted as "other", Table 2.9). For example, if all analyses conducted had to demonstrate non-inferiority to conclude a treatment was non-inferior, and only one of the analyses did, then non-inferiority could not be concluded and could not be rejected. Non-inferiority was declared in 132 (79%) articles. 10 of these had some association with equivalence. Of the articles that were designed as non-inferiority trials, two articles stated the trial was non-inferiority, but had drawn equivalence graphs with two thresholds denoting margins; one article stated the trial was for non-inferiority but then calculated the sample size to determine equivalence; one article concluded that their study did not show equivalence; one concluded equivalence; one article stated that the margin was an equivalence margin; one stated that they would test for equivalence; one concluded non-inferiority as the confidence interval was within $\pm$margin; one concluded equivalence in the abstract but non-inferiority in the main paper and one stated that "results were consistent with showing non-inferiority (i.e. equivalence)".

Superiority analyses were performed in 37 (22%) trials after declaring non-inferiority, of which 27/37 (73%) had explicitly pre-planned for superiority analyses. P-values were reported in 98 (58%) articles, of which 29/98 (30%) were actually testing a superiority hypothesis.

Subgroup of trials with published protocols

Additional information from protocols published by NEJM was extracted for 57 of 61 articles. Including this additional information provided by NEJM improved the reporting of results across all criteria: 39 (64%) articles justified the choice of the non-inferiority margin compared to 19 (31%); most planned two or more analyses 45 (74%) compared to 37 (61%) (there were a couple of cases where two analyses were planned in the protocol but only one was stated in the published article); consistency between type I error rates and confidence intervals was 44 (72%) compared with 36 (59%); imputation techniques were considered in 29 (48%) compared with 17 (28%) articles and sensitivity analyses were considered in 38 (62%) articles compared with 25 (41%). The majority of articles concluded non-inferiority with 8 (13%) not determining

non-inferiority. A total of 14 (23%) articles concluded superiority, of which most were pre-planned; 9/14 (64%). Few articles 8/40 (20%) presented superiority p-values.

Association between quality of reporting and conclusions

Trials that were classed as having some "other" conclusion about non-inferiority were excluded from the analysis. Overall, there was a suggestive difference between the quality of reporting and concluding non-inferiority: $\chi^2_1 = 3.76$; $p = 0.05^*$ (Cochran-Armitage test; Table 2.9). Trials classed as "Excellent" or "Good" (66 articles) were just as likely to conclude non-inferiority than those classed as "Fair" or "Poor" (66 articles). However, those classed as "Excellent" concluded non-inferiority more often (11; 73% compared to 2; 13% that did not) than those classed as "Poor" (18; 62% compared to 10; 34%). The numbers are however too small to make definitive conclusions.

Table 2.9: Quality of reporting of trials associated with conclusions of non-inferiority.

	Yes (n=132)	No (n=29)	Other (n=7)	Total (n=168)
Grade	n (%)	n (%)	n (%)	n (%)
Excellent[1]	11 (73%)	2 (13%)	2 (13%)	15
Good[2]	55 (86%)	9 (14%)	0 (0%)	64
Fair[3]	48 (80%)	8 (13%)	4 (7%)	60
Poor[4]	18 (62%)	10 (34%)	1 (3%)	29

χ^2_1=3.76; p=0.05* (Cochran-Armitage test), excluding trials that concluded "other".

1. Excellent if margin justified, ≥ 2 analyses on patient population performed, type I error rate consistent with significance level of confidence interval.
2. Good if fulfilled two of the following: margin justified, ≥ 2 analyses on patient population performed, type I error rate consistent with significance level of confidence interval.
3. Fair if fulfilled one of the following: margin justified, ≥ 2 analyses on patient population performed, type I error rate consistent with significance level of confidence interval.
4. Poor if margin not justified, < 2 analyses on patient population performed, type I error rate not consistent with significance level of confidence interval.

2.4 Discussion

Reporting of non-inferiority trials is poor and is partly due to disagreement between guidelines on vital issues. There are some aspects that guidelines agree on, such as a requirement for the non-inferiority margin to be justified, but the results showed that this recommendation is completely neglected by most authors. It is remarkable that several authors performed only one analysis for the primary outcome and the lack of consistency between the significance level chosen in sample size calculations and the confidence interval reported further highlights confusion of non-inferiority trials. Not knowing how to deal with missing data nor appropriate sensitivity analyses, also adds to the confusion. The combination of these recent findings taken from high impact journals and the inconsistency in guidelines indicate:

1. The non-inferiority design is not well understood by those using the design, and
2. Methods for non-inferiority designs are yet to be optimised.

We anticipated that poor reporting of articles would bias towards concluding non-inferiority, however, the poorly reported trials were less likely to demonstrate non-inferiority. This is somewhat reassuring. Nevertheless, it is essential to ensure that what is reported at the end of a trial was pre-specified before the start of a trial: scientific credibility and regulatory acceptability of a non-inferiority trial rely on the trial being well-designed and conducted according to the design[46].

Almost 80% of studies concluded non-inferiority, although it is unclear whether this is due to how these studies have been reported or publication bias. It appears that positive results (i.e. alternative hypotheses) are published more often, regardless of trial design, as this number is consistent with other studies that found that more than 70% of published superiority trials demonstrated superiority[47].

More than half of articles reported p-values, of which approximately a third reported p-values for a two-sided test for superiority. P-values, if reported, should be calculated for one-sided tests corresponding to the non-inferiority hypothesis; that is, with H_0: δ = margin. P-values for superiority should not be presented unless following demonstration of non-inferiority, where a pre-planned superiority hypothesis is tested[48].

2.4.1 Comparison with other studies

The value of the non-inferiority margin was almost always reported but more than half of articles made no attempt to explain how the choice was justified. While justification of the margin is low, this is actually an improvement from Schiller et al who reported 23% articles made a justification[49], although this difference could be because only high impact journals were included in this review. This result is consistent with a more recent review performed by Althunian et al, published after our systematic review, which found that reporting the choice of the non-inferiority margin had not improved over time[50]. The authors included articles from 1966 to 2015 and only investigated double-blinded randomised controlled trials. There were equally as many articles that planned and reported an ITT analysis compared with articles that performed ITT and PP analyses. This is surprising given that CONSORT 2006 state that an ITT analysis can bias non-inferiority trials towards showing non-inferiority[1]. These results were lower than found by Wangge et al[51] who reported 55% used either an ITT or PP and 42% used both ITT and PP. Most articles presented two-sided 95% confidence intervals which is consistent with results from Le Henanff et al[52].

2.4.2 Non-inferiority margin

This review showed that the value of the non-inferiority margin was almost always reported, but surprisingly, less than half of articles justified the margin. Those that did and stated that the choice of the margin was based on clinical considerations[1,6,7,9,10,14] often had poor justifications, such as "deemed appropriate" or "consensus among a group of clinical experts" without any detail on how consensus was achieved or how clinical experts were selected. In one article, included in the review, the authors justified the choice of the margin on unpublished data, but had not provided any additional information with regards to the nature of the data. The authors could have included details of the unpublished data within the supplementary content to provide more clarity of the choice of the margin as well as providing additional information for other studies. Non-inferiority is only meaningful if it has strong justification in the clinical context and so should be reported. Guidelines recommend that the choice of

margin should be justified primarily on clinical grounds, however, previous trials and historical data should also be considered if available. For example, Gallagher et al[53] justified the choice of the margin providing as much information as possible by including references to all published reports and providing data from the institution where the senior author is based.

A statement often used in articles reviewed was "the choice of the margin was clinically acceptable". This statement does not contain enough information to justify the choice of the non-inferiority margin. If the choice of the margin is based on a group of clinical experts, authors should provide information on how many experts were involved and how many considered the choice of the margin being acceptable: a consensus among a group of 3 clinicians from one institution is different from a consensus of 20 clinicians representing several institutions. Radford et al[54] justify the choice of the non-inferiority margin after performing a delegate survey at a symposium. This method may be a way forward for researchers to obtain clinical assessment from a large group of clinicians. Even better would be to obtain formal assessments, using for example the Delphi method[55] which has been used in the COMET initiative[56], after presenting the proposed research at a conference or symposium for clinicians to really engage with the question at hand.

There were very few articles that referred to preserving the treatment effect based on estimates of the standard of care arm from previous trials. It is vital that this is acknowledged when reporting non-inferiority studies to ensure the standard of care is effective. If the control were to have no effect at all in the study then finding a small difference between the standard of care and new intervention would be meaningless[14].

2.4.3 Analyses

Definitions provided by authors were inconsistent under what they classed as ITT, PP, mITT and as-treated, for example "all patients randomised who received at least one dose of treatment" was defined at least once in each classification. According to the guidelines, the PP definition excludes patients from the analysis but it is unclear what

those exclusions are. The ambiguity of how per-protocol is defined was evident in this review as definitions provided by authors could not be succinctly categorised. Some defined per-protocol as "patients who received the allocated treatment" while others stated "patients who received the allocated treatment and no major protocol violations". The SPIRIT guidelines encourage authors to avoid the "ambiguous use of labels" and favour for definitions to be clearly defined within the trial protocol[15]. This would avoid other ambiguous classifications found such as "on treatment analysis", "treatment received analysis" and in one case a "modified per-protocol analysis".

Many articles performed only one analysis, despite most guidelines recommending at least two analyses[1,6,8,10,14]. Unfortunately, guidelines differ in their advice on which of the two analyses should be chosen to base conclusions on. This regrettable state of affairs was clearly reflected in our review where some chose ITT to be the primary analysis, others chose PP as the primary analysis.

Both the ITT and PP analyses have their biases and so neither can be taken as a "gold standard" for non-inferiority trials. The analysis of the primary outcome is the most important result for any clinical trial, and so the per-protocol analysis, only including patients in the analysis who take treatment as they were supposed to, is what is of interest in non-inferiority trials. It should be pre-defined in the protocol what patients should adhere to and should be considered at the design stage what can be done to maximise adherence. It should be made clear exactly who is included in analyses given the variety of definitions provided by various authors, particularly for per-protocol analyses where definitions are subjective. Most authors included treatment related exclusions such as "received treatment", "completed treatment" or "received the correct treatment". Such differences in definitions may be may be superficially small but could in fact make critical differences to the results of a trial.

2.4.4 Significance level

Poor reporting of whether the hypothesis test was one-sided or two-sided or absence of the type I error rate in the sample size calculation meant over a quarter of articles were not clearly consistent with regards to the type I error rate and corresponding

confidence interval. Most guidelines advise presenting two-sided 95% confidence intervals and this is what most articles presented. However, this recommendation may cause some confusion between equivalence and non-inferiority trials. A 5% significance level is maintained using 95% confidence intervals in equivalence trials for two-sided hypotheses whereas non-inferiority takes a one-sided hypothesis and so a two-sided 90% confidence interval should be calculated. If a one-sided type I error rate of 2.5% is used in the sample size calculation then this corresponds to the stricter two-sided 95% confidence intervals, not a one-sided 95% confidence interval[57].

The power and type I error rate should be clearly reported within sample size calculations and whether α is for a one-sided or two-sided test. For example, the CAP-START trial used a one-sided significance test of 0.05 with two-sided 90% confidence intervals and the authors provide exact details of the sample size calculation in the supplementary appendix[58].

2.4.5 Missing data

Most trials reported ITT analyses but had not considered imputation techniques to test missing data assumptions. Most trials that used multiple imputation stated the number of imputations used but few discussed the assumptions made, which are particularly critical in this context. Some missing data are inevitable, but naïve assumptions and/or analysis threaten trial validity for both ITT and per-protocol analyses[15], particularly in the non-inferiority context where more missing data can bias towards demonstrating non-inferiority[59].

Trials should report whether imputation was or was not performed. If imputation was used it should be clearly stated what method was used along with any assumptions made, following the guidelines of Sterne et al[60].

2.4.6 Sensitivity analyses

Only about a third of articles reviewed reported using sensitivity analyses. There was some confusion between sensitivity analyses for missing data, and secondary analyses. Sensitivity analyses for missing data should keep the primary analysis model, but vary

the assumptions about the distribution of the missing data, to establish the robustness of inference for the primary analysis to the inevitably untestable assumptions about the missing data. By contrast, secondary analysis with regards to excluding patients for the primary outcome tries to answer a separate, secondary question[61]. Thus, while EMA 2000[8] and CONSORT 2012[10] describe this as sensitivity analysis (and many papers we reviewed followed this), in general this will not be the case, and conflating the two inevitably leads to further confusion.

2.4.7 Subgroup of trials with published protocols

The mandatory publication of protocols taken from NEJM publications improved results for all criteria assessed. This reiterates the findings from Vale et al[62] who evaluated the risk of bias assessments in systematic reviews assessed from published reports, but had also accessed protocols directly from the trial investigators and found that deficiencies in the medical journal reports of trials does not necessarily reflect deficiencies in trial quality. Given this, it is clear that a major improvement in the reporting of non-inferiority trials would result if all journals followed the NEJM practice. Since publication of e-supplements is very cheap, there appears to be no reason not to do this.

Supplementary content should also be taken advantage of and explicitly referred to within articles as this provides the opportunity for authors to provide and describe the details of important methods and rationale for criteria which cannot be included in the main publication due to word limits in journals.

2.4.8 Strengths and limitations

This research demonstrates the inconsistency in the recommendations for non-inferiority trials provided by the available guidelines, which was also reflected within this review. We have highlighted the importance of missing data and using sensitivity analyses specific to non-inferiority trials. There are also some limitations in this review. Firstly, a justification of the choice of the margin was recorded as such if any attempt was made to do so, and so one could argue that inadequate attempts were counted as a "justification", however there was good agreement between reviewers when independently assessed. Secondly, only one reviewer extracted

information from all articles and therefore assessments may be subjective. However, there was good agreement when a random 5% of papers were independently assessed, and the categorisation of the justification of the non-inferiority margin was also independently assessed in all papers where a justification was given. Thirdly, an update of the CONSORT statement for non-inferiority trials was published during the period of the search in 2012[10], which could improve the reporting of non-inferiority trials over the next few years. However, the first CONSORT statement for non-inferiority trials published in 2006[1] was released well before the studies included in our search and we have found that reporting of non-inferiority trials remains poor.

2.5 Conclusion

This review shows a lack of clear reporting that address some key statistical issues for non-inferiority trials. Although it is unclear whether poor reporting largely impacts on the conclusions made about non-inferiority, trials that fail to clearly report the items discussed above should be interpreted cautiously. It is essential that justification of the choice of the non-inferiority margin becomes standard practice, providing the information early on when planning a study including as much detail as possible. If the choice of the non-inferiority margin changes following approval from an ethics committee, justification for the change and changes to the original sample size calculation should be explicit. If journals enforced a policy where authors must justify the choice of the non-inferiority margin prior to accepting publication, this would encourage authors to provide robust justifications for something so critical given that clinical practice may be expected to change if the margin of non-inferiority is met.

Sample size calculations include consideration of the type I error rate, which should be consistent with the confidence intervals as these provide inferences made for non-inferiority when compared against the margin. Inconsistency between the two may distort inferences made, and stricter confidence intervals may lack power to detect true differences for the original sample size calculation. If any imputation was performed then this should be detailed along with any underlying assumptions, supplemented with sensitivity analyses under different assumptions about the missing data.

Information that is partially pre-specified before the conduct of a trial may inadvertently provide opportunities to modify decisions that were not pre-specified at the time of reporting without providing any justification. A compulsory requirement from journals to publish protocols as e-supplements and even statistical analysis plans along with the main article would avoid this ambiguity.

2.6 Summary

The findings from this review presented in this chapter suggest clear violations of available guidelines. This includes the CONSORT 2006 statement (published four years before the first paper in this review) which concentrates on improving how non-inferiority trials are reported and is widely endorsed across medical journals, particularly surrounding analyses conducted.

2.7 Overview of thesis

As a result of the systematic review in this chapter it was clear that the primary analysis chosen from non-inferiority trials and definitions for it varied between published articles, particularly for the PP analysis. For non-inferiority trials, the interest is to obtain estimates from patients who behaved the way they were supposed to in the trial and the PP analysis is key to answering this. This is due to the ITT analysis potentially being anti-conservative; including patients who fail to reach the end of follow-up and therefore do not complete the full course of treatment including all patients in the analysis as defined by ITT can make the treatment look similar to the standard of care arm (or vice versa) therefore biasing towards demonstrating non-inferiority[20]. It was apparent from the review that there is a real need to find a way to deal with missing data in the PP analysis rather than excluding these observations from patients who failed to reach the end of the study.

Many articles reviewed failed to acknowledge and discuss the implications of missing outcome data in relation to the primary outcome and articles rarely performed sensitivity analyses to investigate the assumptions made about the missing data.

Often, patients were excluded from analyses due to protocol deviations but some of these patients excluded could have had some information contributing to the primary outcome prior to deviating and this was rarely addressed in the articles reviewed. Although there was some improvement when protocols from articles published in NEJM were also reviewed in our subgroup analysis, reporting still remained poor. This further supports the need to find a more appropriate analysis in non-inferiority trials that includes missing observations for analyses and adequately tests for the assumptions made about the missing data. The remainder of this thesis aims to do this using data sets as examples from TB trials.

From Chapter 3 onwards, we focus on two non-inferiority trials that aimed to shorten treatment regimens for patients diagnosed with tuberculosis. These datasets are used to help find better analyses for non-inferiority trials, that is, to recover patient information within the primary analysis, in a statistically valid way. In tuberculosis trials, patients are excluded due to loss of follow-up from the modified intention-to-treat analysis and per-protocol analysis depending on completion of treatment. Given that patients can be excluded from the modified intention-to-treat analysis and per-protocol analysis due to loss of follow up, the general aim is to include all patients in these analysis without imposing extreme assumptions about the missing data under an intent-to-treat type of analysis. Chapter 3 investigates single imputation methods and multiple imputation to impute trial participants' outcome data that are missing and these are then applied to our two example datasets. Patterns of missing data are explored in our datasets splitting observations into separate observation times. The partitioned visits are kept to investigate a simpler approach to multiple imputation using inverse probability weighting with Generalised Estimating Equations in Chapter 4. We then explore using multi-state Markov models in Chapter 5 as a possibly new and alternative analysis to re-capture missing observations using two datasets from tuberculosis trials. Chapter 6 introduces reference-based sensitivity analyses, extending the methods for applicability to binary data and applying these methods to the two exemplar datasets. We end with an overview of the results presented in this thesis in Chapter 7 and discuss some ideas for future research.

Chapter 3

Missing data

The results from the systematic review in Chapter 2 demonstrated some confusion surrounding the choice of the primary analysis and how it should be defined for non-inferiority trials. Missing outcome data were rarely considered in the articles reviewed, and when considered that missing data could be problematic for a trial, patients were often excluded from the PP analysis to deal with the issue. This demonstrates that there is some uncertainty of how to deal with missing data within non-inferiority trials.

In this chapter, we investigate tuberculosis (TB) non-inferiority trials. These trials are particularly interesting since patient outcomes are collected at multiple visits and determination of the primary outcome requires a confirmatory result. A consequence of patients who are lost to follow-up leads to uncertainty of whether they are free of the disease at the end of follow-up. Such patients are commonly dealt with in primary analyses by total exclusion if perceived to be disease free prior to being lost to follow-up, irrespective of the history of their outcomes before being lost to follow-up. The aim is to recover this information within the primary analysis.

We proceed by first defining the algorithm commonly used to calculate the primary outcome in TB trials with the aim of shortening treatment regimens. We then introduce the REMoxTB and RIFAQUIN studies which will be used as examples for subsequent analyses. We then investigate single imputation methods and more complex imputation methods used to handle missing observations, and apply some of

these methods to our two example datasets. Doing so, will result in a complete dataset that can be used to determine the overall outcome for each patient randomised into the study.

3.1 Definition of the primary outcome for tuberculosis studies

To assess whether a patient is cured from TB or not, sputum samples are taken from patients to determine a patient's culture result and repeated over several weeks, until study completion. These samples are sent to laboratories and inserted into machines to determine whether patients have a positive culture result (i.e. presence of TB) or a negative result (i.e. absence of TB)[63]. Sometimes a result may be contaminated. Typically a contaminated result arises from within the sample itself due to non-TB commensal bacteria. Other reasons can include clerical errors, contamination of clinical equipment or laboratory cross-contamination. In terms of analyses, these contaminated results are considered missing since a clear-cut result cannot be determined from a contaminated result.

Generally in TB trials, the intensive phase can be anywhere between 2 to 8 months and the continuation phase can be between 4 to 18 months. In the TB trials that are used in this thesis the intensive phase is 4 or 6 months (depending on allocated treatment) and the continuation phase from 4 or 6 months to 18 months of treatment. The intensive phase administers several drugs for treatment to kill most of the TB bacteria living within the patients lungs. However, some of these bacilli may still remain and so the second phase of treatment aims to kill any remaining TB bacteria still present within a patient using fewer drugs.

The primary outcome for TB trials is a binary composite outcome of treatment failure and relapse. In order for a patient to be classed as a success i.e. cured from TB, they must achieve two consecutive negative cultures at two different visits thereby reaching stable negative culture conversion. Patients who achieved two consecutive negative (-) cultures at separate visits but then have two consecutive positive (+) cultures at different visits after treatment are classed as relapses (i.e. $-,-,+,+$) and therefore are considered as "failures" unless able to achieve two consecutive negative cultures at separate visits again (i.e. $-,-,+,+,-,-$). Isolated positive culture results

are not indicative of relapse and are therefore assumed to be classed as "negative" if patients were successful before having a single positive result provided they have a negative culture after the single positive result. Contaminated results which are re-classed as "missing" and missing culture results due to loss of follow up across visits are ignored and therefore do not influence whether a patient is classed overall as a success or failure at the end of follow up.

3.2 Datasets

Non-inferiority trials are appropriate in TB to shorten treatment duration[64] where the standard care involves patients taking several drugs for at least 6 months. In 2014, three major phase III non-inferiority trials were published that aimed to shorten treatment duration from 6 months to 4 months, however all failed to show that a 4 month treatment regimen was non-inferior[65–67]. We use two of these studies as example datasets. A summary of the REMoxTB and RIFAQUIN studies follow and different imputation methods that can deal with missing data are explored and applied. Finally, we investigate patterns of missing data within visit windows. This is to see whether patients mostly have negative cultures towards the end of the study (and are therefore "successful") are similar for different missing data patterns. Data from the REMoxTB and RIFAQUIN studies were accessed from TB-PACTS[68], a publicly available data repository, following an application that I applied for and was granted.

3.2.1 The REMoxTB study

The REMoxTB study[65] was a double-blind, placebo controlled non-inferiority trial that aimed to shorten treatment regimens in patients with newly diagnosed mycobacterium tuberculosis. A total of 1931 patients were randomised to receive one of: a combination of rifampicin (R), isoniazid (H), pyrazinamide (Z) and ethambutol (E) for 8 weeks followed by 18 weeks of HR (control group; 2EHRZ/4HR); a combination of R, H, Z and moxifloxacin (M) for 17 weeks followed by 9 weeks of placebo (isoniazid group; 2MHRZ/2MHR) or a combination of R, M, Z and E for 17 weeks followed by 9 weeks of placebo (ethambutol group); 2EMRZ/2MR. The

primary endpoint was to find a difference in proportions for patients with an unfavourable outcome. An unfavourable outcome was defined as treatment failure or relapse, within 18 months of treatment. The results were compared to a 6% non-inferiority margin. A per-protocol analysis and modified intention-to-treat analysis were performed and the results of the per-protocol analysis were considered primary. Patients who were lost to follow up before the 6 month visit were treated as unfavourable outcomes in the modified intention-to-treat analysis. Additionally, patients who were lost to follow up after the 6 month visit until the end of the study (at 18 months) were excluded, unless already classed as unfavourable. For the per-protocol analysis, any patients lost to follow up were excluded. Patients who were positive when last seen were considered to be treatment failures (unless they were a relapse). Sensitivity analyses to the missing data were performed for the modified intention-to-treat analysis to assess the impact of missing data at the final visit (i.e. patients lost to follow up), first assuming complete case analysis where these patients were excluded from analyses, second assuming a worst case scenario where patients were classed as having an unfavourable outcome and third assuming a best case scenario where patients were classed as having a favourable outcome. For both the modified intention-to-treat and per-protocol populations, analyses were adjusted for weight and centre.

3.2.2 The RIFAQUIN study

The RIFAQUIN study[66] was an open-label non-inferiority trial that investigated two new treatment regimens replacing isoniazid used in the control regimen (a combination of rifampicin, isoniazid, pyrazinamide and ethambutol for 2 months followed by 4 months of isoniazid and rifampicin) with moxifloxacin for 2 months of treatment followed by different doses of rifapentine either at 2 months (i.e. total treatment duration of 4 months) or 4 months (a total treatment duration of 6 months) in patients with newly diagnosed smear positive drug sensitive tuberculosis. The primary composite endpoint was to find a difference in proportions for patients with an unfavourable outcome, defined as treatment failure or relapse, comparing the results to a 6% non-inferiority margin. Per-protocol and modified intention-to-treat analyses were performed; non-inferiority was concluded if both analyses demonstrated non-inferiority. Patients who were lost to follow up were excluded from

both analyses if they reached stable negative culture conversion when last seen. Sensitivity analyses to the missing data were performed by classing those excluded from analyses due to reinfection and separately those who died during the study as unfavourable for both PP and mITT analyses. A worst case scenario for the mITT analysis was performed for all patients except those who were classed as a screening failure. For both the modified intention-to-treat and per-protocol populations, analyses were adjusted for centre only. Unlike the REMoxTB study, weight was not adjusted for.

In both studies, patients could have been seen outside of the scheduled follow up during the study. Some of these unscheduled visits occurred after the final 18 month scheduled visit in these studies. Therefore, for all analyses explored in this thesis, visits up to the final scheduled visit observed in the study will be included. Any unscheduled visits that occur after the final 18 month scheduled follow-up visit will be ignored.

Next, we describe single imputation and multiple imputation methods (§3.3 to §3.5.4), used to result in a "completed" dataset, before describing the pattern of missing data within the REMoxTB and RIFAQUIN datasets (§3.6 to §3.5.4).

3.3 Methods used for imputation of missing data

We now describe different imputation methods to impute the positive or negative sputum culture results that are missing. Using a statistically valid method that allows us to use all the information within a dataset, will inevitably result in a "completed" dataset that will then allow us to determine each patient's outcome of treatment failure (§3.1). These methods (described in §3.4 to §3.5.4) are then applied to the REMoxTB and RIFAQUIN datasets. Imputation of missing data involves methods to replace data that were meant to be collected but for some reason were not. One of the more common reasons for this in clinical trials is loss to follow up or withdrawals. Several imputation methods exist to handle missing observations in trial data, but in order to apply these methods, it is important to think about how the data collected are missing to ensure assumptions made concerning the data are valid. There are three missing data mechanisms to consider:

Missing completely at random (MCAR)

The probability of missing data does not depend on any unobserved or observed variable relevant to the analysis. For example, if sputum was collected from patients to determine whether they had TB but the machine to detect tuberculosis was broken then the outcome is MCAR as the probability of the machine breaking is equal for all patients.

Missing at random (MAR)

The probability of data being missing is independent of the unobserved data, conditional on the observed data. For example, if culture results are taken from patients and males are less likely to have a culture result but this does not depend on the culture result itself, after accounting for gender. In other words, culture results are MAR dependent on gender.

Missing not at random (MNAR)

The probability of missing data depends on an unobserved variable, after accounting for observed variables. For example, if culture results are taken from patients and it is more likely for younger patients to have an outcome, but age is not recorded, then culture results are MNAR as the reason for patients missing culture results depends on data that has not been observed (age).

In this thesis, the analyses explored between Chapters 3-5 assume MAR which is an untestable assumption. We therefore also look at departures from this assumption under MNAR in Chapter 6 using sensitivity analyses.

For TB studies, missing data can occur due to loss of follow-up, missed visit, a contaminated result which is re-classed as missing or because the patient is unable to produce a sputum sample. For a patient who is unable to produce a sputum sample at a visit is considered to have a negative culture result. Given that persistent coughing is a major symptom of TB, if following treatment a patient is no longer able to cough and is therefore unable to produce a sputum sample then they are clinically considered as not having TB bacteria at that visit. As a result, if a patient is missing a

result at a visit but it is known that the patient was unable to produce sputum, the missing result is considered to be negative. If a culture result is missing because the sputum sample is contaminated, it is reasonable to assume MAR because the reason for the result being missing depends on the sputum sample that was produced by the patient. If a patient happens to be missing a visit but is observed again in the future, MAR is assumed since it is likely that the missing visit occurred by chance. For instance, this might occur if a patient is seen a few days later outside of the scheduled visit window. If a patient is lost to follow-up and the visit is related to what occurred previously (e.g. relapse) or if the missing visit is related to other variables collected within the data (such as an adverse event or a withdrawal reason), then the missing data are classed as MAR. This is a reasonable assumption since the reason for a patient being lost to follow-up is observed within the data. In instances where it is unknown why a patient is lost to follow-up or if they are not observed for many months, the reason for missing a visit might not be so clear. This means that the missing data are MNAR. This aspect of missing data is explored in Chapter 6 of this thesis.

Having defined various assumptions that can be made about the distribution of the missing data, we now describe single imputation methods and multiple imputation methods that can be used to impute the missing data. These methods are then used for the REMoxTB study and RIFAQUIN study.

3.4 Single imputation methods

The use of single imputation methods imputes missing observations with one value. The values imputed are considered as if they were known with certainty. We consider the complete case analysis, last observation carried forward and best case/worst case scenarios discussing the assumptions made for each.

3.4.1 Complete case analysis

One of the simplest methods to handle missing data is a *complete case analysis*. Here, patients with missing data are completely excluded from analyses. This can be a valid analysis provided that the reason for having a missing outcome does not depend on

observed or unobserved data, i.e. the missing mechanism is MCAR. If a baseline covariate contains missing data, then a complete case analysis can be valid under the MAR assumption as long as the probability of outcome data being missing is independent of the observed data, conditional on the baseline covariate[69].

Under MAR, a complete case analysis can be problematic for longitudinal data, where patients are followed up several times over a study for outcomes. The long sequence of data collected means patients are more likely to have at least one observation missing over time. A complete case analysis would exclude those patients, wasting all other information collected about a patient and severely decreasing the sample size (and therefore power) in a study.

3.4.2 Last observation carried forward

For longitudinal data, last observation carried forward (LOCF) assumes that the last observation seen is unchanged over time for future observations, which were expected but unseen, during a study. The method works by taking the result of the last observation recorded for a patient and replaces missing observations in the future using that observation. LOCF can be unbiased if the average observed and unobserved outcomes in each randomised group do not change over time[27].

3.4.3 Best case/worst case scenario

Using a best case or worst case scenario, often produces extreme results[70]. A best case scenario replaces missing observations with the best value observed in the treatment arms and the worst values in the control arm.

A worst case scenario is analogous to this, replacing missing values with the worst value observed in the treatment arms and the best value in the control arm.

The above definitions are a more accurate reflection of a best case and worst case scenario to take into account the behaviour within treatment regimens rather than using blanket statements imputing all patients with the best or worst value regardless of treatment arm. These methods assigning all patients to a best or worst value

regardless of treatment administered are commonly recommended in statistical guidelines, but do not accurately reflect a best case or a worst case scenario.

The methods described so far from §3.4.1 to §3.4.3 assume decisive decisions which are rarely, if ever, plausible. Another method to consider when accounting for missing observations for longitudinal data are mixed effects models.

Mixed effects models

For longitudinal data, a mixed effects regression model can be valid under MAR to predict missing outcomes from a model based on observed data. This model can be used to model observed outcomes within each patient while allowing for correlation between patient outcomes observed. For a mixed effects model, using interim follow up data of an outcome for patients who withdraw before the final follow up visit can can make best use of the MAR assumption as the interim data informs the measurement at the final time point[71].

For TB trials, the outcome is determined based on an accumulation of the whole data collected rather than at a singular time point (i.e. at the last visit), and therefore a mixed effects regression analysis would not be able to determine whether a patient should be classed as a success or a failure. Instead a statistically valid method that imputes all missing observations so that we have a complete dataset of negative and positive culture results is necessary to then implement the algorithm for the primary outcome that determines whether patients were or were not cured of TB.

We now consider multiple imputation methods which take into account the uncertainty of the imputed value.

3.5 Multiple imputation methods

The single imputation approaches described above make strong assumptions about the missing data mechanism and this often leads to extreme results. Methods that are based on multiple imputation, assuming the distribution of the missing data is MAR, impute the data accounting for the uncertainty of the result imputed under a range of

possible values. These values are then combined for each imputed set using Rubin's rules[34], outlined in §3.5.1.

3.5.1 Multiple imputation

Multiple imputation was introduced by Donald Rubin[34]. The concept of multiple imputation is that missing data are imputed more than once based on the distribution of the observed data, and includes randomness to reflect the uncertainty about the missing values[27], replacing the missing value with an estimated value. Multiple imputation assumes data are MAR. That is, data are missing conditional on the observed data. For one imputation, the imputations are drawn from the joint posterior predictive distribution $f(Y_{mis}|Y_{obs})$ for missing observations for variable Y (Y_{mis}) conditional on those observed (Y_{obs}) of a Bayesian model. Common practice is to create more than one imputation to give I completed datasets. The datasets are analysed individually but identically to give I estimates and I estimates of the variance of the model. These estimates are then combined to get an overall estimate and variance using Rubin's rules[34].

Rubin's rules are applied after creating an imputed dataset[34]. Let $Comp$ equal the imputed complete-data estimates that range from 1 to I imputations. Then the mean is:

$$\overline{Comp}_I = \frac{1}{I}\sum_{v=1}^{I} Comp_v \tag{3.1}$$

Let W equal the within imputed complete-data variances that range from 1 to I imputations. The within variance is:

$$\overline{W}_I = \frac{1}{I}\sum_{v=1}^{I} W_v \tag{3.2}$$

Let $Betw$ equal the between variance among the "I" complete-data estimates, then:

$$Betw_I = \frac{1}{I-1}\sum_{v=1}^{I}(Comp_v - \overline{Comp}_I)^2 \tag{3.3}$$

The total variance combines 3.2 and 3.3 such that:

$$Var(Comp) = \overline{W}_I + (1+\frac{1}{I})Betw_I \tag{3.4}$$

For TB trials (see §3.2) where the sputum test result is binary, a logistic regression model is used such that:

$$\text{logit}(\rho_k) = log(\tfrac{\rho_k}{1-\rho_k}) = \beta_0 + \beta_1 X_{1,1} + ... + \beta_v X_{k,v}, \tag{3.5}$$

where ρ represents the probability of having a "yes" or "no" outcome for patient $k = 1, 2, ..., N$. β are a vector of logistic regression parameters:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_v \end{bmatrix} \tag{3.6}$$

and X are a vector of observed covariates:

$$X = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,v} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,v} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{k,1} & X_{k,2} & \cdots & X_{k,v} \end{bmatrix} \tag{3.7}$$

where v is the v^{th} covariate for the k^{th} patient. From this model, the posterior mean ($\widehat{\beta}$) and the variance-covariance matrix ($V(\widehat{\beta})$) of the estimate (β) is approximated. The posterior distribution of β_k follows a multivariate normal distribution.

To impute one dataset to obtain the primary outcome of interest, (for the TB data this would be the sputum test culture results, described in §3.3), assuming covariates have no missing data[34]:

1. Draw a random parameter β_k from $\sim N(\widehat{\beta}, V(\widehat{\beta}))$.
2. Impute from the inverse logit function using the values drawn from β_k for each missing observation $\tilde{z}$:

$$logit^{-1}(\beta_k X_{k,v}) = \frac{exp(\beta_k X_{k,v})}{1 + exp(\beta_k X_{k,v})} \tag{3.8}$$

3. Independently draw random numbers, $Rand_k$, from a uniform probability distribution, U(0,1).
   (a) If $Rand_k > logit^{-1}(\beta_k X_{k,v})$, 0 is imputed for the missing observation, otherwise 1 is.

The process is iterated for I imputations and combined using Rubin's Rules (3.1 to 3.4), assuming normal approximation as is common practice[34].

While multiple imputation can be easily implemented in statistical software in Stata[72], SAS[73] and R[74], difficulties may still occur for binary longitudinal data, particularly where patients are followed up for several visits. In TB data for example, there are often long sequences where the probabilities tend to 0 or 1 and so this can create numerical problems for calculations. Different approaches must also be considered to analyse these types of data. First we consider a non-parametric approach using hot deck imputation, then we describe a simple version of the fully conditional specification (FCS) method before using an extension of this method the two-fold FCS multiple imputation. The final consideration is ordinal imputation, stepping back from classing culture results as "positive" or "negative".

3.5.2 Hot Deck Imputation

Hot-deck imputation is a non-parametric imputation procedure which fills in missing patient observations with observations from patients who are retained by matching on variables that are observed in both types of patients, such as matching covariates. Hot deck imputation avoids distributional assumptions, instead the imputed values will have the same distribution as the observed data[34]. Hot deck imputation is quite often used to handle binary outcome data as it avoids computational issues since a missing value will always be imputed assuming MCAR or MAR[75]. However, hot deck imputation underestimates variability as the method treats the imputed values as if they were known with certainty. Another disadvantage of hot deck imputation is that it is inappropriate for data that are longitudinal as the procedure does not condition on the rest of the data. This method will therefore not be considered any further since the primary outcome for TB depends on repeated observations from patients.

3.5.3 Two-fold fully conditional specification multiple imputation

The two-fold fully conditional specification (FCS) method may be a useful alternative for imputing longitudinal missing observations. Briefly, the method works by iteratively taking each visit and imputing missing observations for that visit based on

observations either side of that visit, at a specified length, for all follow up visits in a study.

The two-fold FCS approach first introduced by Nevalainen, Kenward and Virtanen constructs the joint distribution for missing observations conditional on those observed for the outcome on a flexible selection of univariate imputation distributions without needing to formally specify the joint multivariate density[76]. Before describing this method in §3.5.3, we first describe the fully conditional specification approach which is then extended to the two-fold FCS multiple imputation.

Fully conditional specification

In TB studies, where the sputum culture result (Y) is imputed by treatment arm (X_{trt}) which has no missing values, the multiple imputation model can be denoted by $f(Y_{mis,I}, Y_{obs}, X_{trt})_I$ for I imputed data sets where Y_{mis} and Y_{obs} corresponds to missing and observed positive or negative sputum test results and X_{trt} corresponds to the fully observed treatment covariate that we include for imputation. To impute missing observations, using $\tilde{z}$ to represent the imputed values, for one imputed dataset as follows:

1. For a vector of unknown parameters, say $\hat{\phi}$, calculate the posterior distribution $p(\hat{\phi}|Y_{obs}, X_{trt})$.

2. Draw a random parameter, $\hat{\phi}^*$, from the multi-variate normal distribution: $\hat{\phi}^* \sim N(\hat{\phi}, Var(\hat{\phi}))$.

3. Set $\hat{\phi} = \hat{\phi}^*$ and draw a value ,$\tilde{z}$, from the conditional posterior distribution of $p(Y_{mis}|Y_{obs,X_{trt}}, \hat{\phi} = \hat{\phi}^*)$.

The FCS algorithm proposed by van Bureen et al samples iteratively to account for any dependence on the estimated model parameters[77]. For multivariate Y, which is incomplete, using FCS imputes the missing data one variable at a time. This is an iterative process cycling through all variables, possibly with different conditional specifications, several times[78]. For one cycle, the joint model multiple imputation is approximated by regressing the observed part of the sputum test result (Y_{obs}) on all other remaining variables which are of interest. Following this first cycle, the initial

starting values are replaced by imputed values[79]. A number of cycles are run and the imputations are taken from one final cycle through univariate models[76].

The use of FCS only means that the missing values of the variable itself are not substituted for use at the next point of imputation and so the regression model is estimated using only those patients whose outcome is observed at that time point. There are several disadvantages to using a simple FCS method. Imputing separately for each visit does not take into account correlations between observations per patient and for several follow up visits the model may be over-fitted[76]. Therefore, this method has been proposed using a "two-fold" approach by Nevalainen et al[76] and validated by Welch et al[78] to account for the correlation of observations between patients using a two-fold approach, described below.

Two-fold fully conditional specification

Two-fold FCS assumes data are MAR (see 3.3). It is doubly iterative, taking imputed values from the imputation model to only condition on previous $(t - \nu)$ and subsequent $(t + v)$ time points, t where $v \in 1, 2, ..., V$. The length, v, is specified at the user's discretion, thus allowing for imputation within smaller visit windows of the data. This approach is useful in situations where, using standard multiple imputation (see §3.5.1) may be computationally impossible. For instance, the underlying distribution to the observed data cannot be fitted across all time points at once because there is insufficient data (often at later follow-up times due to withdrawal) over several follow-up visits. Therefore, using the two-fold methodology to impute the data may be an attractive approach. For the TB datasets we use, the previous $(t - 1)$ and next $(t + 1)$ scheduled follow-up visit at each time t is used. Then to impute Y[78]:

1. Having cycled around the imputation model using the standard FCS method (§3.5.3), *within-time* (b_W) iterations are calculated at (i.e. within) each time point t, using logistic regressions, to create an imputation set ($\tilde{z}$) for each patient k in $Y_{k,t}$ by treatment arm (X_{trt}) until the last visit, regress:

$$\mathrm{Y}_{k,t,obs} \text{ on } Y_{k,t-1}, Y_{k,t+1}, X_{k,trt}, \tag{3.9}$$

   The missing values of $Y_{k,t}$ is imputed conditional on values of the sputum test

results within the specified window. This is:

$$Y_{k,t,mis}|Y_{k,t-1,obs}, Y_{k,t+1,obs}, Y_{k,t,obs}, X_{k,trt}. \tag{3.10}$$

2. The *between-time* (b_{Bm}) iteration is calculated applying the *within-time* iteration forwards over time for values at the $t+1$ visit until the final follow-up visit. This algorithm is then repeated for b_{Bm} iterations[76] to form one imputed dataset.

This process is then repeated for I imputations, propagating the observed and imputed values forwards at time t from the previous $t-1$ visit. The imputations are combined using Rubin's rules[34] for $\tilde{z}$ imputations resulting in a completed data set. The first visit collected at the time of randomisation only has one immediate visit afterwards and the final 18 month visit only has one immediate visit preceding it. Therefore, missing observations in the first and last visit can only be imputed based on the next visit and prior visit respectively.

3.5.4 Ordinal Imputation

In TB studies, patients are classed depending on counts of bacteria seen on a sputum sample. If no bacteria are seen on the sputum sample produced by a patient then they are considered to have a negative culture result. If bacteria are seen, each sample is graded 1, 2 or 3 depending on the amount of bacteria seen. Any bacteria means the patient has TB, and patients are therefore recorded as having a "positive" culture result and these outcomes are taken for statistical analyses. However, since the smear results are graded, ordinal multiple imputation and ordinal two-fold FCS can be considered where the gradings of each culture result follow a natural order to potentially enrich the results as an alternative to imputing the determined positive or negative culture results.

The remainder of this chapter is as follows. First, we describe patients who will remain in our analyses in order to impute missing observations resulting in a complete dataset to determine each patient's primary outcome. We then describe single imputation and multiple imputation methods and use some of these methods for the REMoxTB data. The results are compared to the original results presented in the REMoxTB study to investigate whether these imputation methods have a large impact on the conclusions

of the study. We then look at patterns of missing data, splitting visits into windows to see how the proportion of negative culture results differ over time. We then repeat these analyses for the RIFAQUIN study.

3.6 Application to the REMoxTB study

In tuberculosis phase III trials, the aim is to shorten treatment and the overall outcome is determined through an algorithm of longitudinal outcomes collected over many weeks using risk differences (see §3.1 for details). Given the intensity of treatment, patients may withdraw or change treatment and are either completely excluded from primary analyses or treated as failures (unless classed as a failure beforehand).

To investigate the impact of missing data, methods described above in §3.5.1 that deal with missing data are explored on data from the REMoxTB study. Patients who had contaminated results were re-classed as having missing results, as defined in the trial protocol.

3.6.1 Patients included in REMoxTB analyses

In TB trials, there are certain situations where patients need to be excluded from analyses, for example, resistance to treatment drug or patients who never had TB but had very similar symptoms of the disease. For the analyses presented in this thesis, patients are excluded for reasons not related to treatment. For our intention-to-treat analyses, we exclude patients if they were resistant to rifampicin and/or fluroquinolones as the interventions used in the study were not intended for this group of patients (Table 3.1). Other standard drugs exist to cure these patients[80]. Patients who had no confirmed positive culture results within 2 weeks after randomisation were also excluded from analyses as it would be unclear whether or not patients truly had TB at all at the start of the study. Protocol violations at enrolment were also excluded as patients were withdrawn from the study prior to taking any medication after enrolment. Patients who were enrolled at Pinetown and Mexico were also excluded from the analysis due to inaccurate testing of culture results at those centres. These patients were withdrawn prior to completing treatment and follow up. Therefore a total of 146 patients were excluded from the imputation analyses.

Table 3.1: Tabulation of patients to be excluded from future analyses, by treatment arm for REMoxTB.

	Control (N=640)	Isoniazid N= (655)	Ethambutol (N=636)	Total (N=1931)
Pinetown/Mexico	10 (2%)	8 (1%)	18 (3%)	36
Resistant to rifampicin /fluroquinolone	23 (4%)	20 (3%)	18 (3%)	61
Protocol violations at enrolment	5 (1%)	7 (1%)	8 (1%)	20
No positive cultures $\leq$2 weeks from randomisation	12 (2%)	11 (2%)	6 (1%)	29
Total	50 (8%)	46 (7%)	50 (8%)	146
Total not to be excluded	590	609	586	1,785

3.6.2 Imputation analyses for the REMoxTB study

For reasons described in §3.4.2, LOCF will not be investigated. Before applying multiple imputation methods we consider a complete case analysis and best case/worst case scenarios.

Brief note on using a mixed effects regression model for tuberculosis trials

Another model to assess binary outcomes with longitudinal data is a logistic mixed effects regression model which would account for missing data. However, this model is inappropriate for tuberculosis trials as a mixed effects regression model does not readily take into account the necessary requirement of patients achieving two consecutive negative culture results at separate visits (i.e. cure) and will not be considered any further.

Complete case analysis

For the complete case analysis, patients were included only if culture results were reported for all visits. Patients missing any one or more culture results post-randomisation were excluded from this analysis. After excluding patients who

had at least one observation missing during their scheduled follow up and an extra 8 patients according to the exclusion criteria in Table 3.1, a total of 259 patients were included in the analyses (Table 3.2). The risk difference model did not converge when adjusting for centre and weight, therefore 100 iterations were used for this adjusted model. The results from this analysis are consistent with that of the primary analysis and non-inferiority is not shown; the upper bound of the 97.5% confidence interval crosses the 6% margin and therefore non-inferiority cannot be concluded.

Best case scenario

For the best case scenario, patients were only excluded if they were randomised at Pinetown/Mexico, were resistant to rifampicin, had protocol violations or did not have any positive cultures within 2 weeks of randomisation (Table 3.1). Missing culture results were imputed as positive for patients randomised to the control arm and negative otherwise. We then proceeded to determining whether or not patients reached stable negative culture conversion. A total of 1785 patients were included in this analysis. The results in Table 3.2 demonstrate how extreme these results are, showing a strong case for non-inferiority in the ethambutol regimen (upper bound of the 97.5% CI: -23.8%) and an even stronger case for non-inferiority in the isoniazid regimen (upper bound of the 97.5% CI: -26.8%).

Worst case scenario

Missing observations for patients randomised to the control regimen were imputed as negative and imputed positive for patients randomised to the treatment regimens. A total of 1785 patients were included in this analysis after applying our exclusion criteria from Table 3.1. The results from this analysis (Table 3.2) show extreme results, where the upper bound of the 97.5% confidence interval fails to demonstrate non-inferiority in both treatment regimens (upper bound of the 97.5% CI: 49.3% for the isoniazid regimen and 49.7% for the ethambutol regimen).

Multiple imputation

Scheduled visits were used for multiple imputation. Following imputation, unscheduled visits with an observed sputum test result that may have occurred from randomisation until the final 18 month follow-up visit are then included to determine

the primary outcome (§3.1). Unscheduled visits were not included for imputation for two reasons. Firstly unscheduled visits were specific to patients and so including them in the imputation model would unnecessarily impute results for patients who did not need to have extra visits. Secondly, because unscheduled visits could occur at any time between the 17 scheduled visits over the 78 weeks of scheduled follow up patients would not necessarily be seen at the same time. This would result in an even longer sequence of data to be imputed increasing the risk of the imputation model failing to converge.

Imputing positive and negative culture results across all visits using multiple imputation failed due to perfect prediction. That is, the imputation failed to include a random element to take into account the uncertainty about the missing values. Even after accounting for perfect prediction in the model, imputing for all 17 scheduled visits was computationally impossible. Instead, we proceeded to the two-fold fully conditional specification multiple imputation method (see §3.5.3) to impute scheduled visits for 1785 patients (see Table 3.1).

Two-fold fully conditional specification multiple imputation

A total of 50 imputations were used. Data were imputed at each visit based on results on either side of that visit. For example, imputations at week 2 would rely on observed data from week 1 and week 3. Patients who were lost to follow up had their missing data imputed at expected visits based on patients who had outcome data at that visit and at either side of the missing visit. Imputations were performed separately within each of the three treatment groups and the resulting analyses were compared to the standard of care arm. A total of 1785 patients were included in this analysis after applying our exclusion criteria from Table 3.1. The results from this analysis were consistent with the primary ITT and PP analyses failing to demonstrate non-inferiority since the upper bound of the 97.5% CI was 10.03% for the isoniazid regimen (5.25%; 97.5% CI: 0.48% to 10.03%) and 12.3% for the ethambutol regimen (7.07%; 97.5% CI: 1.84% to 12.3%).

Table 3.2: Difference in proportions of unfavourable outcome using different imputation methods for the REMoxTB study

Analysis	**Isoniazid**	**Ethambutol**
	Risk difference (97.5% CI)	**Risk difference (97.5% CI)**
Primary analysis (PP) from REMoxTB (n=1548)*		
Unadjusted results	6.74% (2.25% to 11.24%)	11.61% (6.81% to 16.40%)
Adjusted results[1]	6.09% (1.71% to 10.47%)	11.36% (6.70% to 16.10%)
Primary analysis (mITT) from REMoxTB (n=1674)*		
Unadjusted results	7.56% (2.30% to 12.83%)	8.28% (2.94% to 13.63%)
Adjusted results[1]	7.80% (2.70% to 13.0%)	9.01% (3.80% to 14.20%)
Complete case analysis (n=259)*		
Unadjusted results	7.46% (-3.66% to 18.57%)	12.32% (0.37% to 24.27%)
Adjusted results[1,2]	12.49% (1.29% to 23.68%)	14.97% (2.41% to 27.53%)
Best case scenario (n=1785)*		
Unadjusted results	-32.97% (-38.49% to -27.44%)	-30.06% (-35.78% to -24.34%)
Adjusted results[1]	-32.30% (-37.85% to -26.76%)	-29.59% (-35.32% to -23.85%)
Worst case scenario (n=1785)*		
Unadjusted results	44.29% (39.21% to 49.38%)	44.76% (39.59% to 49.92%)
Adjusted results[1,2]	44.39% (39.46% to 49.33%)	44.69% (39.67% to 49.72%)
Two-fold FCS MI[3] **(n=1785)***		

Unadjusted results	5.44% (0.58% to 10.30%)	7.62% (2.44% to 12.80%)
Adjusted results[1]	5.25% (0.48% to 10.03%)	7.07% (1.84% to 12.30%)

*Number of patients included in the analysis.

[1]Adjusted for weight and centre;

[2]Model did not converge, therefore 100 iterations were used;

[3]Fully Conditional Specification (FCS) Multiple imputation (MI).

Ordinal multiple imputation

We proceeded to ordinal imputation, where the bacteria seen on samples were graded 1, 2 or 3. Bacteria that are not seen will always result in a negative culture result and culture results graded 1, 2 or 3 will always be classed as positive. Some patients had more than one sample collected at the same visit and so a condition is required within the imputation model to account for these patients. Grades were averaged for more than one sample at the same follow-up visit and the result rounded up to 1, 2 or 3 to enable imputation across all scheduled visits, otherwise there would be very few patients with different proportions of grading, making imputation difficult to perform. Again, the imputation failed due to perfect prediction and therefore we proceeded to the two-fold FCS multiple imputation where the data were ordered as 0, 1, 2 or 3 according to the number of bacteria present in each sample and imputed missing observations using observed data either side of that visit as above (§3.6) using 50 imputations.

It was computationally impossible to impute the REMoxTB data for all 17 scheduled visits with an ordinal regression model, even after accounting for issues relating to perfect prediction. The two-fold FCS multiple imputation algorithm also failed where the outcome was ordinal and therefore no results have been presented for this nor the preceding ordinal imputation analyses.

3.7 Discussion

So far, we have investigated single imputation methods and multiple imputation methods for the REMoxTB study. The results from the complete case analysis were consistent with that of the primary analysis. However the exclusion of over 1000

patients resulted in a huge loss of information which was reflected in the wider confidence intervals and consequently created greater uncertainty surrounding the estimates. The inflation of the estimates and the upper bounds of the confidence interval indicates potential bias towards favouring the control regimen. If the primary analysis marginally demonstrated or failed to demonstrate non-inferiority the complete case analysis would not assist with the conclusions of the study, due to the increase in uncertainty of the estimates as a consequence of deleting such a large amount of data. Therefore, using a complete case analysis is inefficient as patients have been excluded from the analysis in spite of having data that, clearly, contributes to the primary outcome.

The results from the best case scenario and worst case scenario produced inflated results in either direction. These scenarios differ from those performed in the original study since we imputed according to treatment arm. That is, for a best case scenario patients randomised to the control arm were imputed as "treatment failures" and those randomised the treatment arms were classed as reaching stable negative culture conversion. The worst case scenario is analogous to this. The best case scenario showed a strong case for non-inferiority in both treatment regimens and the worst case scenario failed to demonstrate non-inferiority in the treatment regimens, where the estimates of the upper bound of the 97.5% confidence intervals were over 3 times wider than those shown for the primary analysis of the study. These analyses demonstrate the effect such strong assumptions about the missing data can have, and clearly alternative methods to explore the nature of missing data without inferring such extreme assumptions are necessary.

The two-fold fully conditional specification multiple imputation method seemed to work well for the REMoxTB data, producing consistent results with that of the main study. However, the results from this method were not as extreme as those from the original PP and mITT analyses, as demonstrated by the upper bound of the 97.5% CI.

Regressing back to bacterial culture results and using ordinal imputation failed when imputing for all scheduled visits and when using the two-fold FCS imputation method. At the beginning of TB treatment, most patients are expected to have grades

of 1, 2 or 3; a positive culture result which indicates that they have TB and towards the end of treatment most patients are expected to have negative culture results indicating no presence of TB or "cure". A small number of patients have negative culture results in the first few weeks of the study, which then diminishes over time. This means that patients are mostly positive in the first few weeks of follow up and are mostly negative towards the end, and so the fitted probabilities from using a logistic regression model for multiple imputation are very close to either 1 or 0 resulting in perfect prediction[81]. Therefore, by splitting patients' positive results into grades of 1, 2 or 3 has dichotomised an already sparse group of patients into even smaller groups by the end of follow-up. Ordinal imputation is not an approach that will work for TB studies and will not be considered any further.

Next we summarise the proportion of missing data by treatment arm to see whether there are any major differences between the missing data patterns by treatment. We then investigate patterns of missing data for the REMoxTB study to see how the proportion of negative culture results differ over time.

3.8 Investigating patterns of missing data for the REMoxTB study

In tuberculosis (TB) trials, the interest is in whether a patient is cured of TB (i.e. are successful) or not. To be classed "successful" patients need to produce a confirmatory negative culture result, immediately following a negative result at separate visits. Patients are followed up at several visits over the course of 18 months over two phases; 1) the treatment phase and 2) the follow up phase. The combination of collecting results, repeatedly, over the 18 months and the requirement of a confirmatory negative culture result provide plenty of opportunity for patients to miss one or more visits (and are therefore missing outcome results) at any point during a study. This can be problematic for patients missing one or more visits as it can create uncertainty around whether a patient maintains stable negative culture conversion, particularly where there are sporadic results missing or more than one consecutive missing result.

Per-protocol analyses often exclude patients with missing data, for example, patients who fail treatment, patients who are not assessable at the last follow up visit or patients who never enter into the continuation phase. The aim is to include such patients in analyses therefore providing an alternative analysis to the customary per-protocol analysis, often ambiguously defined, used in non-inferiority trials while making best use of the data collected.

A natural intuition is to use multiple imputation to "fill in" outcomes of patients who are unobserved at visits who should have been observed. Multiple imputation assumes the data are MAR (§3.3). Using data from the REMoxTB study as an example, different patterns of missing data are investigated to support the assumption that the data are MAR. As previously discussed, using multiple imputation across many different visits for binary data causes severe issues within the data, especially perfect prediction. As the REMoxTB study has a total of 17 visits, the visits are now grouped into four clinically meaningful visit windows to reduce the vast number of different patterns that may occur, before performing any further analyses. This will provide an overview of how the culture results for patients behave within these trials.

3.8.1 Summary of culture results for the REMoxTB study

A total of 1785 patients were included having applied the exclusion criteria (see Table 3.1) in the REMoxTB study. Table 3.3 shows the total number of culture results collected at each scheduled visit week. The total number of patients that died at each week are cumulative over time. Patients were followed weekly up to week 8 from baseline (week 0). The remaining visits for the treatment phase were at week 12, 17, 22 and 26 and additionally at weeks 39, 52, 65 and 78 during the follow up phase.

Table 3.3: Summary of culture results for 1785 patients who are included after applying the exclusion criteria for REMoxTB.

Week	Culture result	Control (N=590)	Isoniazid (N=609)	Ethambutol (N=586)
	Negative	37 (6.27%)	27 (4.43%)	35 (5.97%)

0	Positive	529 (89.66%)	558 (91.63%)	534 (91.13%)
	Missing	23 (3.90%)	24 (3.94%)	17 (2.90%)
	Died	1 (0.17%)	0	0
1	Negative	74 (12.54%)	46 (7.55%)	59 (10.07%)
	Positive	407 (68.98%)	492 (80.79%)	445 (75.94%)
	Missing	108 (18.31%)	71 (11.66%)	82 (13.99%)
	Died	1 (0.17%)	0	0
2	Negative	89 (15.08%)	79 (12.97%)	87 (14.85%)
	Positive	388 (65.76%)	429 (70.44%)	421 (71.84%)
	Missing	111 (18.81%)	99 (16.26%)	78 (13.31%)
	Died	2 (0.34%)	2 (0.33%)	0
3	Negative	115 (19.49%)	136 (22.33%)	119 (20.31%)
	Positive	339 (57.46%)	369 (60.59%)	370 (63.14%)
	Missing	134 (22.71%)	102 (16.75%)	97 (16.55%)
	Died	2 (0.34%)	2 (0.33%)	0
4	Negative	167 (28.31%)	185 (30.38%)	181 (30.89%)
	Positive	284 (48.14%)	328 (53.86%)	298 (50.85%)
	Missing	136 (23.05%)	94 (15.44%)	106 (18.09%)
	Died	3 (0.51%)	2 (0.33%)	1 (0.17%)
5	Negative	217 (36.78%)	247 (40.56%)	226 (38.57%)
	Positive	235 (39.83%)	234 (38.42%)	239 (40.78%)
	Missing	135 (22.88%)	125 (20.53%)	119 (20.31%)
	Died	3 (0.51%)	3 (0.49%)	2 (0.34%)
6	Negative	273 (46.27%)	305 (50.08%)	289 (49.32%)
	Positive	176 (29.83%)	173 (28.41%)	183 (31.23%)
	Missing	137 (23.22%)	128 (21.02%)	111 (18.94%)
	Died	4 (0.68%)	3 (0.49%)	3 (0.51%)
7	Negative	334 (56.61%)	359 (58.95%)	367 (62.63%)
	Positive	114 (19.32%)	110 (18.06%)	104 (17.75%)
	Missing	138 (23.39%)	135 (22.17%)	112 (19.11%)
	Died	4 (0.68%)	5 (0.82%)	3 (0.51%)
8	Negative	362 (61.36%)	405 (66.50%)	411 (70.14%)
	Positive	82 (13.90%)	75 (12.32%)	61 (10.41%)

	Missing	142 (24.07%)	124 (20.36%)	111 (18.94%)
	Died	4 (0.68%)	5 (0.82%)	3 (0.51%)
12	Negative	404 (68.47%)	471 (77.34%)	467 (79.69%)
	Positive	26 (4.41%)	13 (2.13%)	21 (3.58%)
	Missing	156 (26.44%)	120 (19.70%)	93 (15.87%)
	Died	4 (0.68%)	5 (0.82%)	5 (0.85%)
17	Negative	436 (73.90%)	452 (74.22%)	452 (77.13%)
	Positive	17 (2.88%)	13 (2.13%)	21 (3.58%)
	Missing	133 (22.54%)	137 (22.50%)	108 (18.43%)
	Died	4 (0.68%)	7 (1.15%)	5 (0.85%)
22	Negative	435 (73.73%)	426 (69.95%)	404 (68.94%)
	Positive	11 (1.86%)	16 (2.63%)	29 (4.95%)
	Missing	139 (23.56%)	159 (26.11%)	146 (24.91%)
	Died	5 (0.85%)	8 (1.31%)	7 (1.19%)
26	Negative	435 (73.73%)	403 (66.17%)	408 (69.62%)
	Positive	10 (1.69%)	34 (5.58%)	49 (8.36%)
	Missing	140 (23.73%)	163 (26.77%)	121 (20.65%)
	Died	5 (0.85%)	9 (1.48%)	8 (1.37%)
39	Negative	438 (74.24%)	416 (68.31%)	394 (67.24%)
	Positive	24 (4.07%)	50 (8.21%)	61 (10.41%)
	Missing	119 (20.17%)	133 (21.84%)	123 (20.99%)
	Died	9 (1.53%)	10 (1.64%)	8 (1.37%)
52	Negative	424 (71.86%)	414 (67.98%)	423 (72.18%)
	Positive	19 (3.22%)	28 (4.60%)	33 (5.63%)
	Missing	133 (22.54%)	155 (25.45%)	122 (20.82%)
	Died	14 (2.37%)	12 (1.97%)	8 (1.37%)
65	Negative	426 (72.20%)	407 (66.83%)	411 (70.14%)
	Positive	14 (2.37%)	29 (4.76%)	24 (4.10%)
	Missing	136 (23.05%)	161 (26.44%)	143 (24.40%)
	Died	14 (2.37%)	12 (1.97%)	8 (1.37%)
78	Negative	443 (75.08%)	428 (70.28%)	429 (73.21%)
	Positive	11 (1.86%)	22 (3.61%)	18 (3.07%)
	Missing	121 (20.51%)	146 (23.97%)	130 (22.18%)

	Died	15 (2.54%)	13 (2.13%)	9 (1.54%)

NB: Missing includes contaminated results re-classed as "missing".

Table 3.3 shows that there are higher proportions of patients missing culture results early on in the study on the control arm, however this imbalance levels out after week 22. The number of patients missing culture results is generally consistent during the follow up phase (weeks 39 to 78) at around 20-25%. At 6 weeks, the number of negative culture results exceeds the number of positive culture results. It is at this time point where most patients begin to culture convert.

3.8.2 Patterns of missing data for the REMoxTB study

We now investigate principal patterns of missing data to get a sense of how the proportion of negative culture results varies between the different patterns of missing data, if at all, and to see how the probability of negative culture results differs across time between patients. As there were a vast amount of scheduled visits in the REMoxTB study, the 17 scheduled visits were grouped into clinically meaningful visit windows, smoothing the data as follows:

- Weeks 0 to 4 includes weeks 0, 1, 2, 3, 4;
- Weeks 5 to 8 includes weeks 5, 6, 7, 8;
- Weeks 12 to 26 includes weeks 12, 17, 22 and 26;
- Weeks 39 to 78 includes weeks 39, 52, 65, 78.

To get a sense of the missingness pattern for the REMoxTB study and to group more patients into similar missing data patterns, the culture results within each visit window for patients were grouped into the following pattern types: "completers" (indicated by "O"), "intermittent" (indicated by "Δ") or "missing" (indicated by "."). Patients were allocated one of these patterns depending on how frequently a patient was observed within that visit window and on how many consecutive results were observed. If within a visit window, patients only missed one of their scheduled visits, they were considered to be "completers" since most of the results were complete in that window. Similarly, if a patient was only observed at one of their scheduled visits

within a visit window, they were considered to be "missing" since the majority of the patient's results was missing. For weeks 0 to 4, which has 5 visits within that visit window, a patient with three observations was considered to be "intermittent" if the patient was missing results between any of those three observations. If the patient had three consecutive observations between 0 to 4 weeks, the pattern was classed as "completers", since there no missing observation occurred between observed results. Similarly, a patient with three missing results was classed as "intermittent" if a result was observed between those missing results at any visit and "missing" if a patient was missing at three consecutive visits between weeks 0 to 4. This is done to reflect that the patient was not re-observed within that visit window for a long period.

Formally, for for weeks 0 to 4:

- "Completers" if a result was observed at all visits, or if one visit was missing at any one of the scheduled visits within the visit window. Patients were also classed as "completed" if patients were observed at any three consecutive visits within the visit window;
- "Intermittent" if there were two or three results missing intermittently between the 5 scheduled visits;
- "Missing" if all results were missing, or if only one result was observed between weeks 0 to 4. Patients were also classed as "missing" if patients were missing results at any three consecutive visits within the visit window.

For the remaining visit windows (weeks 5 to 8; 12 to 26 and 39 to 78), patients were classed as follows:

- "Completers" if a result was observed at all visits, or if one visit was missing at any one of the scheduled visits within the grouped visit;
- "Intermittent" if two results were missing;
- "Missing" if all results were missing or if only one result was observed within the visit window.

Table 3.4 shows the total number of patients within each pattern (in descending order) for all 1785 patients. The table also shows the total number of negative culture results in each pattern and the final column shows how many patients were classed overall as "successful" at the end of the study (by week 78).

Table 3.4 summarises patients whose overall missingness pattern includes patients who have most of their culture results observed across the visit windows (i.e. "completers" or a combination of "completers" and "intermittent"), or patients who are "missing" most of their results or those who died (indicated with a "D"). Table 3.5 describes missing data patterns for patients whose missingness pattern is mostly "intermittent" across visit windows. Table 3.6 summarises patients who have various combinations of "completers" pattern, "intermittent" pattern, "missing" or death across visit windows.

The proportion of patients with negative culture results at each visit window was also calculated for each treatment arm (Appendix B) according to each pattern as follows:

1. The proportion of negative cultures in each pattern at each week per patient was calculated by taking the total number of negative culture results and dividing over the total number of culture results observed;

2. The proportion of patients being negative in each pattern at each week was averaged over visit windows;

3. Those classed as "missing" had very little data and therefore were assumed as having no observations within that visit window.

Table 3.4: Number of negative culture results and proportion of patients who achieved negative culture conversion for patients with most culture results observed (i.e. completers') over visit windows for REMoxTB[1].

Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78	Total (N=1785)	Number of negative culture results Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78	Total number of negative culture results	Treatment success n/no. patient per pattern
O	O	O	O	879 (49.24%)	799/4188 = 19.08%	2261/3307 = 68.37%	3142/3260 = 96.38%	3147/3336 = 94.33%	9349	790 (89.87%)
O	O	Δ	O	106 (5.94%)	79/483 = 16.36%	238/379 = 62.80%	198/212 = 93.40%	353/381 = 92.65%	868	92 (86.79%)
O	O	O	.	86 (4.82%)	70/417 = 16.79%	214/318 = 67.30%	272/309 = 88.03%	0/60 = 0%	556	15 (17.44%)
O	O	O	Δ	79 (4.43%)	69/377 = 18.30%	192/296 = 64.86%	272/284 = 95.77%	148/158 = 93.67%	681	68 (86.08%)
O	Δ	O	O	67 (3.75%)	52/302 = 17.22%	94/134 = 70.15%	218/231 = 94.37%	229/247 = 92.71%	593	57 (85.07%)
Δ	O	O	O	50 (2.80%)	22/141 = 15.60%	134/179 = 74.86%	170/178 = 95.51%	179/188 = 95.21%	505	46 (92%)
O	O	.	.	35 (1.96%)	40/166 = 24.10%	91/130 = 70%	0/13 = 0%	0/8 = 0%	131	0
.	.	.	.	35 (1.96%)	0/44 = 0%	0	0/1 = 0%	0/1 = 0%	0	0
.	.	O	O	31 (1.74%)	0/30 = 0%	0/10 = 0%	108/110 = 98.18%	108/113 = 95.58%	216	26 (83.87%)
O	.	O	O	28 (1.57%)	29/117 = 24.79%	0/22 = 0%	101/104 = 97.12%	96/107 = 89.72%	226	22 (78.57%)
O	O	.	O	27 (1.51%)	23/127 = 18.11%	61/96 = 63.54%	0/24 = 0%	92/94 = 97.87%	176	21 (77.78%)
O	.	.	.	22 (1.23%)	23/82 = 28.05%	0/5 = 0%	0/1 = 0%	0/1 = 0%	23	0
.	O	O	O	15 (0.84%)	0/22 = 0%	48/53 = 90.57%	56/59 = 94.92%	57/58 = 98.28%	161	13 (86.67%)
O	D	D	D	7 (0.39%)	10/32 = 31.25%	3/4 = 75%	0	0/0 = 0	13	0
O	.	.	O	6 (0.34%)	6/27 = 22.22%	0/5 = 0%	0/4 = 0%	16/22 = 72.73%	22	3 (50%)
O	O	O	D	5 (0.28%)	2/24 = 8.33%	5/19 = 26.32%	16/19 = 84.21%	2/4 = 50%	25	0
.	.	.	O	5 (0.28%)	0/7 = 0%	0/1 = 0%	0/2 = 0%	19/19 = 100%	19	1 (20%)
O	.	O	.	4 (0.22%)	4/14 = 28.57%	0/1 = 0%	11/14 = 78.57%	0/1 = 0%	15	1 (25%)
.	.	O	.	4 (0.22%)	0/5 = 0%	0/1 = 0%	11/12 = 91.67%	0/2 = 0%	11	0
D	D	D	D	4 (0.22%)	2/5 = 40%	0	0	0	2	0
.	O	O	.	2 (0.11%)	0/3 = 0%	5/7 = 71.43%	6/6 = 100%	0/1 = 0%	11	0
O	O	D	D	2 (0.11%)	0/9 = 0%	7/8 = 87.50%	2/2 = 100%	0	9	0
.	O	.	D	1 (0.06%)	0/2 = 0%	4/4 = 100%	0/1 = 0%	0	4	0
.	.	O	D	1 (0.06%)	0/1 = 0%	0/1 = 0%	3/3 = 100%	0	3	0
.	O	.	.	1 (0.06%)	0/2 = 0%	4/4 = 100%	0	0/1 = 0%	4	1 (100%)
O	O	.	D	1 (0.06%)	1/4 = 25%	2/3 = 66.67%	0/1 = 0%	0	3	0
O	.	.	D	1 (0.06%)	0/3 = 0%	0	0/1 = 0%	0	0	0
.	.	D	D	1 (0.06%)	0/2 = 0%	0	0	0	0	0

[1]Where O=Most results observed within a window; Δ=Intermittent results observed within a window; D=Death or .=Missing.

Table 3.5: Number of negative culture results and proportion of patients who achieved negative culture conversion for patients with most culture results intermittently observed over visit windows for REMoxTB[1].

Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78	Total (N=1785)	Number of negative culture results Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78	Total number of negative culture results	Treatment success n/no. patient per pattern
Δ	Δ	O	O	24 (1.34%)	21/64 = 32.81%	37/48 = 77.08%	81/84 = 96.43%	85/89 = 95.51%	224	22 (91.67%)
O	O	Δ	Δ	22 (1.23%)	12/108 = 11.11%	55/78 = 70.51%	42/44 = 95.45%	41/44 = 93.18%	150	19 (86.36%)
Δ	O	Δ	O	13 (0.73%)	6/38 = 15.79%	23/41 = 56.10%	26/26 = 100%	49/49 = 100%	104	13 (100%)
O	Δ	Δ	O	12 (0.67%)	3/52 = 5.77%	14/24 = 58.33%	22/24 = 91.67%	41/42 = 97.62%	80	10 (83.33%)
O	Δ	O	Δ	9 (0.50%)	7/40 = 17.50%	17/18 = 94.44%	28/29 = 96.55%	16/18 = 88.89%	68	8 (88.89%)
Δ	.	.	.	8 (0.45%)	3/22 = 13.64%	0/3 = 0%	0/1 = 0%	0	3	0
Δ	Δ	Δ	O	6 (0.34%)	3/17 = 17.65%	9/12 = 75%	12/12 = 100%	23/23 = 100%	47	5 (83.33%)
Δ	Δ	Δ	.	3 (0.17%)	0/8 = 0%	0/6 = 0%	6/6 = 100%	0/2 = 0%	6	1 (33.33%)
O	Δ	Δ	Δ	3 (0.17%)	3/14 = 21.43%	4/6 = 66.67%	5/6 = 83.33%	6/6 = 100%	18	3 (100%)
Δ	.	Δ	Δ	2 (0.11%)	0/6 = 0%	0/1 = 0%	4/4 = 100%	4/4 = 100%	8	2 (100%)
Δ	.	Δ	.	2 (0.11%)	0/4 = 0%	0/1 = 0%	2/4 = 50%	0	2	0
Δ	O	O	Δ	2 (0.11%)	0/5 = 0%	5/7 = 71.43%	7/7 = 100%	4/4 = 100%	16	2 (100%)
.	.	Δ	Δ	2 (0.11%)	0/4 = 0%	0/1 = 0%	4/4 = 100%	3/4 = 75%	7	1 (50%)
Δ	Δ	.	.	2 (0.11%)	2/4 = 50%	3/4 = 75%	0/2 = 0%	0/1 = 0%	5	1 (50%)
Δ	Δ	O	Δ	1 (0.06%)	1/3 = 33.33%	1/2 = 50%	3/3 = 100%	2/2 = 100%	7	1 (100%)
Δ	Δ	Δ	Δ	1 (0.06%)	1/3 = 33.33%	2/2 = 100%	2/2 = 100%	2/2 = 100%	7	1 (100%)
Δ	Δ	.	Δ	1 (0.06%)	2/3 = 66.67%	0/2 = 0%	0	1/2 = 50%	3	0
Δ	.	.	Δ	1 (0.06%)	1/3 = 33.33%	0	0	2/2 = 100%	3	0

[1]Where O=Most results observed within a window; Δ=Intermittent results observed within a window or .=Missing.

Table 3.6: Number of negative culture results and proportion of patients who achieved negative culture conversion for patients with a mixture of observed, intermittent and missing culture results within visit windows for REMoxTB[1].

Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78	Total (N=1785)	Number of negative culture results				Total number of negative culture results	Treatment success n/no. patient per pattern
					Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78		
Mixture of completed	culture results,	hit and miss	culture results	and missing results						
O	O	Δ	.	23 (1.29%)	6/102 = 5.88%	35/83 = 42.17%	43/46 = 93.48%	0/12 = 0%	84	2 (8.70%)
O	.	Δ	O	18 (1.01%)	15/77 = 19.48%	0/9 = 0%	34/36 = 94.44%	55/63 = 87.30%	104	14 (77.78%)
O	Δ	.	.	10 (0.56%)	4/45 = 8.89%	11/20 = 55%	0/3 = 0%	0/1 = 0%	15	0
.	Δ	O	O	8 (0.45%)	0/6 = 0%	16/16 = 100%	29/29 = 100%	30/30 = 100%	75	8 (100%)
O	Δ	.	O	8 (0.45%)	5/34 = 14.71%	9/16 = 56.25%	0/8 = 0%	27/29 = 93.10%	41	4 (50%)
O	Δ	O	.	8 (0.45%)	7/36 = 19.44%	10/16 = 62.50%	24/28 = 85.71%	0/5 = 0%	41	2 (25%)
O	O	Δ	D	7 (0.39%)	6/32 = 18.75%	20/25 = 80%	14/14 = 100%	2/2 = 100%	42	0
Δ	O	O	.	7 (0.39%)	8/20 = 40%	20/24 = 83.33%	22/26 = 84.62%	0/2 = 0%	50	0
O	O	.	Δ	7 (0.39%)	5/33 = 15.15%	21/26 = 80.77%	0/5 = 0%	14/14 = 100%	40	6 (85.71%)
.	.	Δ	O	6 (0.34%)	0/8 = 0%	0/1 = 0%	12/12 = 100%	22/22 = 100%	34	5 (83.33%)
O	.	Δ	.	6 (0.34%)	2/25 = 8%	0/5 = 0%	12/12 = 100%	0/4 = 0%	14	0
Δ	.	O	O	6 (0.34%)	6/16 = 37.50%	0/5 = 0%	21/21 = 100%	23/23 = 100%	50	6 (100%)
Δ	Δ	O	.	6 (0.34%)	4/16 = 25%	8/12 = 66.67%	16/20 = 80%	0/5 = 0%	28	0
O	.	O	Δ	5 (0.28%)	3/20 = 15%	0/4 = 0%	18/18 = 100%	10/10 = 100%	31	5 (100%)
O	Δ	Δ	.	4 (0.22%)	5/18 = 27.78%	5/8 = 62.50%	7/8 = 87.50%	0/1 = 0%	17	0
O	.	.	Δ	4 (0.22%)	2/17 = 11.76%	0/2 = 0%	0/1 = 0%	7/8 = 87.50%	9	0
Δ	Δ	.	O	4 (0.22%)	3/10 = 30%	5/8 = 62.50%	0/4 = 0%	12/13 = 92.31%	20	3 (75%)
Δ	O	Δ	.	3 (0.17%)	3/9 = 33.33%	8/10 = 80%	5/6 = 83.33%	0/1 = 0%	16	1 (33.33%)
Δ	O	.	O	3 (0.17%)	1/9 = 11.11%	8/11 = 72.73%	0/2 = 0%	9/10 = 90%	18	3 (100%)
Δ	O	.	.	3 (0.17%)	5/8 = 62.50%	10/11 = 90.91%	0/1 = 0%	0	15	0
.	Δ	O	.	3 (0.17%)	0/5 = 0%	5/6 = 83.33%	8/11 = 72.73%	0/3 = 0%	13	1 (33.33%)
Δ	.	D	D	2 (0.11%)	1/5 = 20%	0/2 = 0%	0	0	1	0
.	.	O	Δ	2 (0.11%)	0/2 = 0%	0/1 = 0%	4/7 = 57.14%	4/4 = 100%	8	1 (50%)
O	.	Δ	Δ	2 (0.11%)	7/9 = 77.78%	0	4/4 = 100%	4/4 = 100%	15	2 (100%)
Δ	O	.	Δ	2 (0.11%)	1/6 = 16.67%	5/6 = 83.33%	0/2 = 0%	3/4 = 75%	9	1 (50%)
O	Δ	.	Δ	2 (0.11%)	6/9 = 66.67%	4/4 = 100%	0/2 = 0%	4/4 = 100%	14	2 (100%)
Δ	.	Δ	O	2 (0.11%)	2/5 = 40%	0/1 = 0%	4/4 = 100%	6/6 = 100%	12	2 (100%)
Δ	.	.	O	2 (0.11%)	0/4 = 0%	0	0/2 = 0%	7/7 = 100%	7	0
Δ	O	O	D	2 (0.11%)	4/5 = 80%	5/6 = 83.33%	6/8 = 75%	1/1 = 100%	16	0
Δ	.	O	Δ	1 (0.06%)	0/2 = 0%	0	4/4 = 100%	2/2 = 100%	6	1 (100%)

[1]Where O=Most results observed within a window; Δ=Intermittent results observed within a window or .=Missing.

Figure 3.1: Proportion of negative culture results for completers' in REMoxTB.

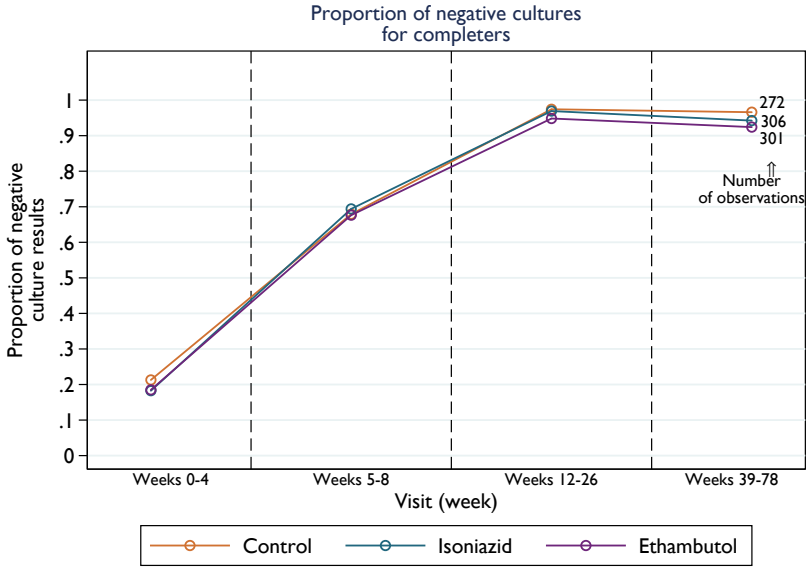


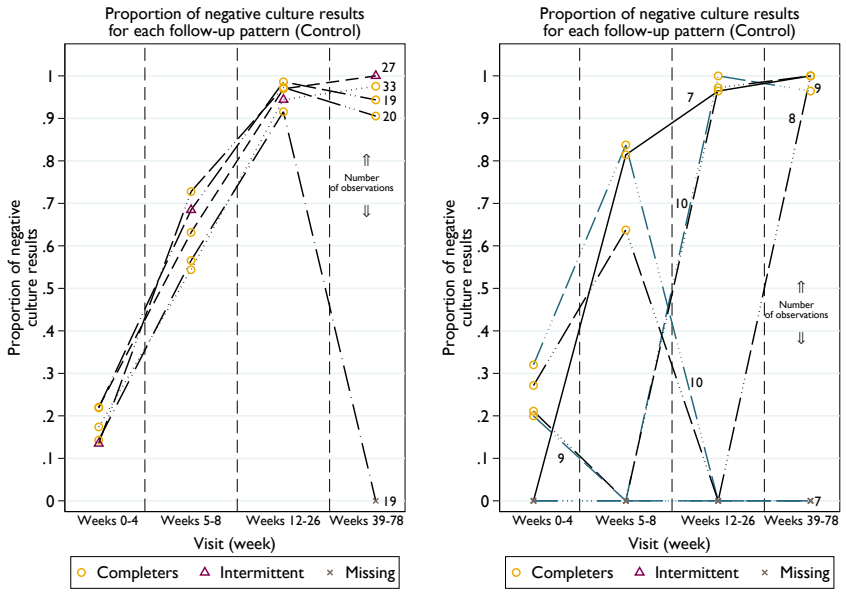
Table 3.4 shows that almost half of patients (49%) had complete results, most of whom were successful (Tables B1 to B7 in Appendix B show these patterns for each treatment arm). Patients who are missing culture results towards the end of the study (i.e. in the follow-up phase) are not so successful in achieving two negative culture results. This is due to the definition of the primary outcome for REMoxTB where these patients who are unobserved are classed as "unassessable" if culture results are not observed before the final 78 week visit.

Figures 3.1 to 3.10 show the most common missing data patterns by treatment arm for patients who have mostly completed/missing results across all four visit windows, intermittent/missing results across all four visit windows or a mixture of results across all four visit windows. Completers are define as patients who only had one result within the visit window that was not a negative culture result. Figure 3.1 shows that the proportion of negative culture results for patients who have completed results at all four visit windows (from weeks 0 to 78) are very similar between treatment regimens. The proportion of negative culture results levels at around 95% over weeks 12 to 26 and drops slightly a year later, over weeks 39 to 78.

Figures 3.2, 3.3 and 3.4 show the remaining patients whose missingness pattern have the majority of their culture results observed within each visit window and also those who are "missing" between week 0 to 78. The lines in blue indicate which patterns were mostly missing culture results across all visits during the study. The different lines in these graphs correspond to the same missing data pattern in different treatment groups. This is for ease of comparing each different missingness pattern between treatment arms. Numbers noted within the graphs relate to the number of patients that follow a particular pattern. Figures 3.5 to 3.7 is a graphical representation of Table 3.5 by treatment arm, showing the most common patterns for patients whose culture results are mostly intermittent. Figures 3.8 to 3.10 is a graphical representation of patients in Table 3.6 by treatment arm whose pattern has a mixture of "completers", "intermittent", "missing" or death across visit windows.

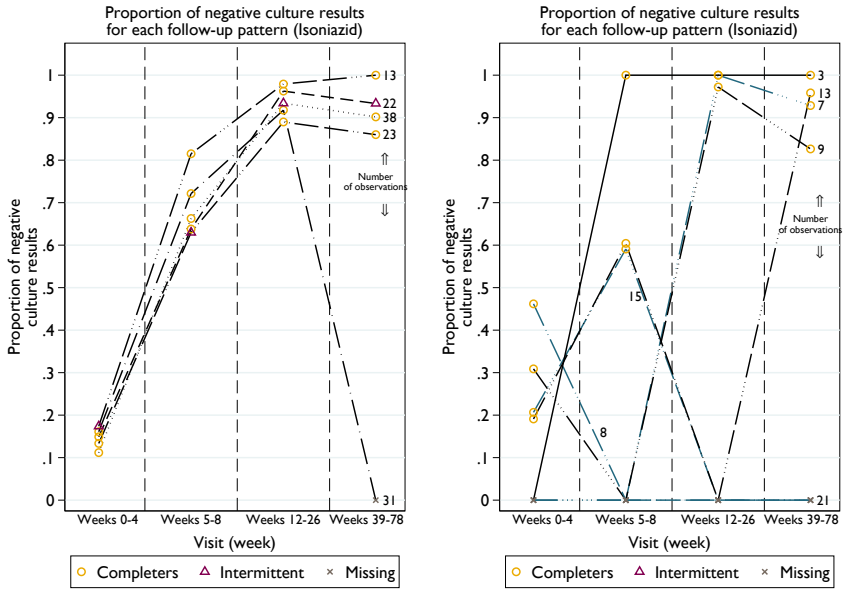
The left hand panel in Figure 3.2 shows that before withdrawal (i.e. the 19 patients who are missing results between weeks 39 to 78) have a very similar pattern to the 99 patients in the other patterns who do reach the end of the study. The right hand panel in Figure 3.2 show similar proportions of negative culture results between different patterns during follow up at weeks 39 to 78, irrespective of whether or not patients are observed beforehand. Figures 3.3 to 3.10 show similar trends. The ethambutol arm in Figure 3.4 shows the 36 patients who are mostly observed within each visit window up to week 26 and are missing results between weeks 39 to 78 have slightly lower proportions of negative cultures between weeks 12 to 26 in comparison to the remaining patterns for patients who had completed results between weeks 39 to 78. Patients who were randomised to the control arm that have intermittent patterns of missing data from weeks 0 to 8 and have completed results from weeks 12 to 78 (Figure 3.5) have lower proportions of negative culture results compared to the treatment arms (Figures 3.6 and 3.7).

Figure 3.2: Proportion of negative culture results for completers' pattern in REMoxTB.



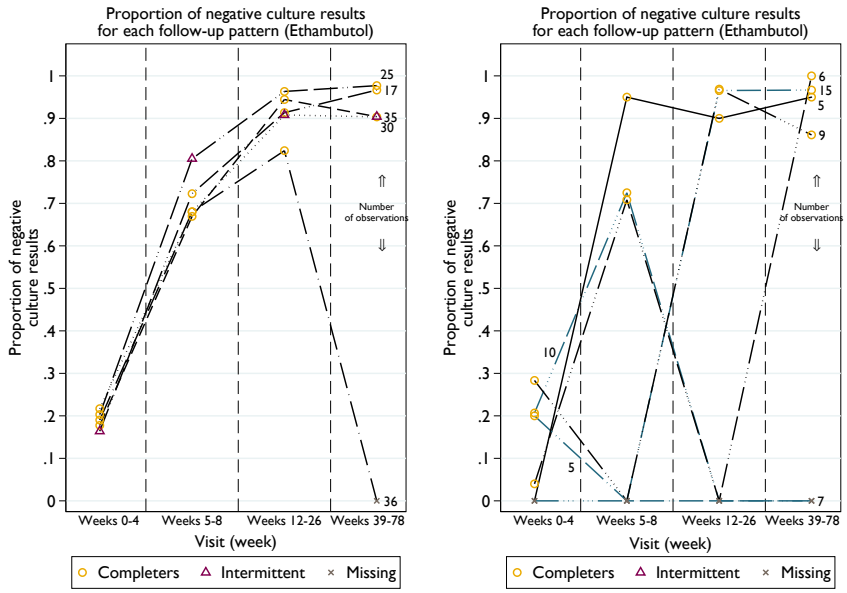
NB: Due to the vast number of patterns, graphs are split into two for ease of viewing each pattern.

Figure 3.3: Proportion of negative culture results for completers' pattern in REMoxTB.



NB: Due to the vast number of patterns, graphs are split into two for ease of viewing each pattern.

Figure 3.4: Proportion of negative culture results for completers' pattern in REMoxTB.



NB: Due to the vast number of patterns, graphs are split into two for ease of viewing each pattern.

Figure 3.5: Proportion of negative culture results for intermittent pattern in REMoxTB.

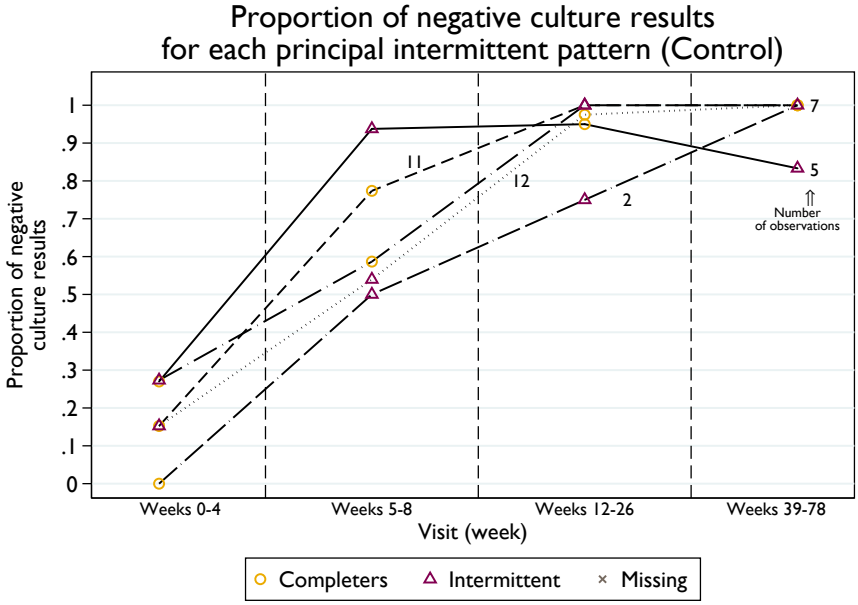


Figure 3.6: Proportion of negative culture results for intermittent missing results in REMoxTB.

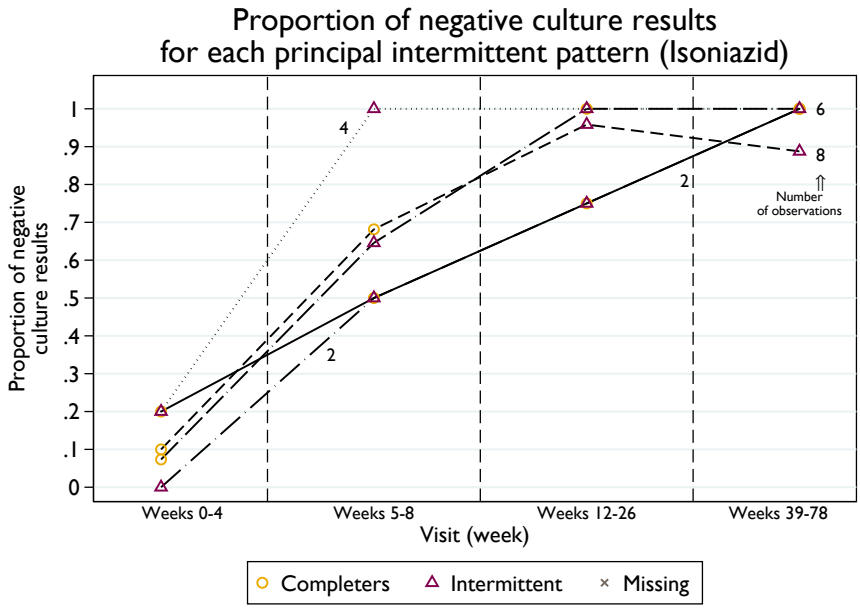


Figure 3.7: Proportion of negative culture results for intermittent missing results in REMoxTB.

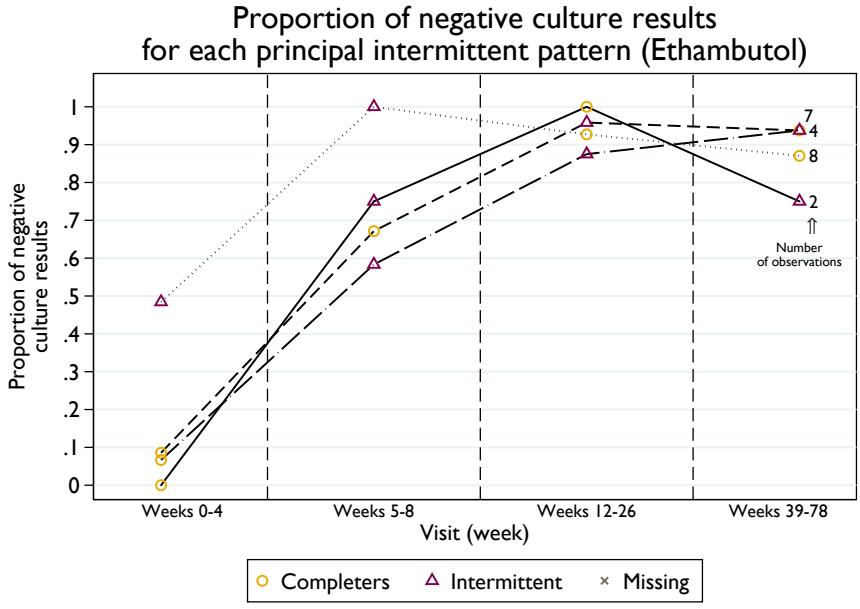


Figure 3.8: Proportion of negative culture results for a mixture of results in REMoxTB.

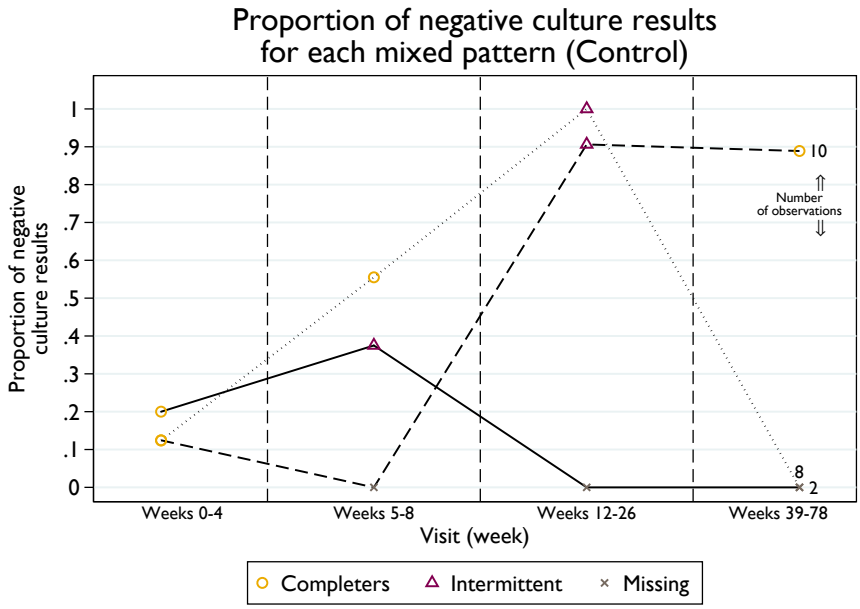


Figure 3.9: Proportion of negative culture results for a mixture of results in REMoxTB.

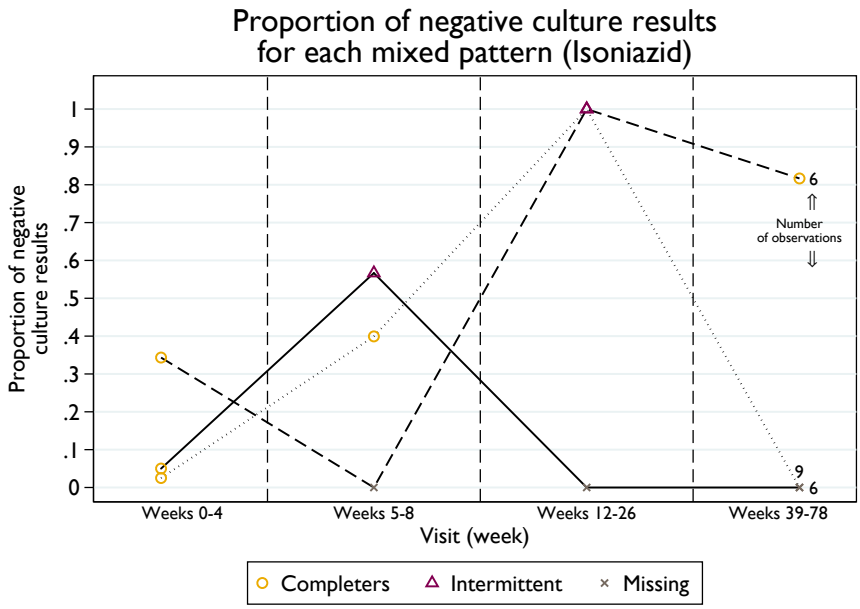
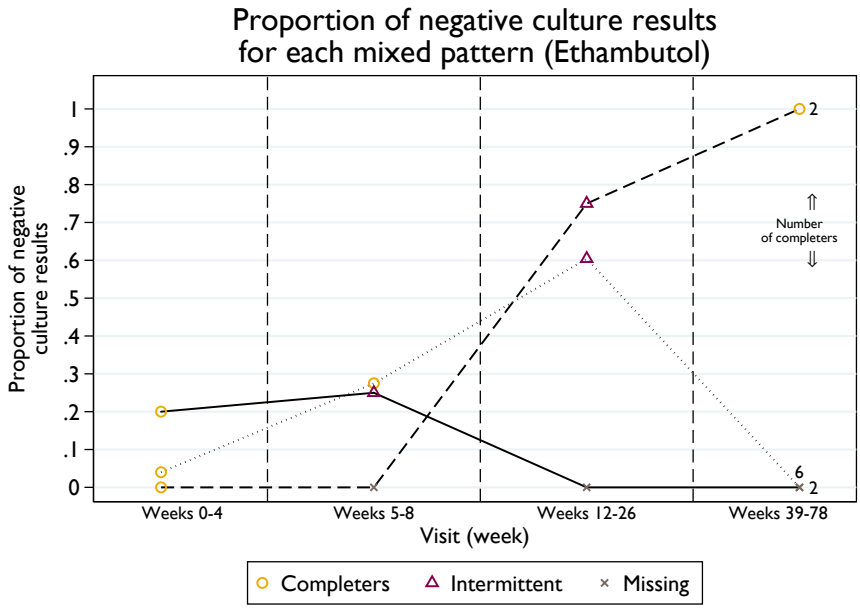


Figure 3.10: Proportion of negative culture results for a mixture of results in REMoxTB.



3.8.3 Discussion

In this section we investigated the proportion of missing data between treatment arm and then explored patterns of missing culture results by treatment arm. This was done splitting the data into visit windows to reduce the amount of missing data patterns while looking at the trend of the missing data over time. We found that the proportion of patients having negative culture results is similar between patients irrespective of the pattern and is also similar between treatment arms. Patients whose data are mostly missing generally follow similar patterns to those that have data observed within the same visit window. Even patients who are missing culture results before the follow-up phase (weeks 39 to 78) have similar proportions towards the end of follow up as those that have completed data. This suggests that the reason for patients being missing does not depend on treatment received. The methods applied here thus far for the REMoxTB study are now applied to the RIFAQUIN study.

3.9 Application to the RIFAQUIN study

Analyses from the REMoxTB study (§3.6.1 to §3.8.2) are applied similarly to the RIFAQUIN study (see §3.2.2). Even though both the PP and mITT analysis were used to determine non-inferiority for the RIFAQUIN study, we focus on the PP analysis. We first define patients who will be included in analyses for this study, impute patients' sputum culture results that are missing and then focusing on these patients we describe patients who were missing outcome observations at each scheduled follow-up visit. We then investigate analyses to impute the missing observations of patient outcomes using single imputation methods (complete case analysis, best case/worst case scenarios) and then we explore multiple imputation and the two-fold fully conditional specification multiple imputation method. As the RIFAQUIN data is much smaller than that for the REMoxTB data, we do not explore ordinal multiple imputation. We then investigate different patterns of negative culture results splitting the scheduled follow-up visits into windows.

3.9.1 Patients included in analyses for the RIFAQUIN study

Table 3.7 shows patients who are excluded from subsequent analyses. As for REMoxTB, we aim for an ITT analysis where patients who were resistant to the drugs used in the study and those whose diagnosis of TB were not confirmed within the first two weeks of randomisation were excluded. Patients who might have been screened for TB, but were subsequently found as not having TB much later were also excluded. This can sometimes occur due to the lag in time attaining confirmation of TB after sending sputum samples to laboratories. A total of 730 patients will be included in our analyses suggesting that more information is to be gained from the 216 patients who were excluded from the primary analysis due to the PP criteria.

Table 3.7: Tabulation of patients to be excluded from analyses, by treatment arm for RIFAQUIN.

	Control (N=275)	4 month regimen (N=275)	6 month regimen (N=277)	Total (N=827)
Late screening failure, previous TB resistance	1 (0.4%)	0	0	1

Resistant to rifapentine/ isoniazid/moxifloxacin	12 (4%)	16 (6%)	13 (5%)	41
No positive cultures $\leq$2 weeks from randomisation	22 (8%)	20 (7%)	13 (5%)	55
Total	35 (13%)	36 (13%)	26 (9%)	97
Total not to be excluded	240	239	251	730

Table 3.8 shows the proportion of patients that will be included in our analyses who have a positive, negative or missing culture result or who died at each scheduled follow up visit. Patients who were recorded to be "contaminated" were assumed to be missing. The proportion of patients with negative culture results was consistently lower for patients who were randomised to the 4 month regimen in comparison to the control and for the 6 month regimen from month 3 onwards. The proportion of patients who are missing culture results is slightly higher on the treatment regimens at 4 months but then levels out by month 7. The proportion of patients missing is slightly higher on the 6 month regimen at month 11 and 12, but again levels out towards the end of the study.

Table 3.8: Summary of culture results for 730 patients who are included after applying the exclusion criteria for RIFAQUIN.

Visit	Culture result	Control (N=240)	4 month regimen (N=239)	6 month regimen (N=251)
Month 0	Positive	240 (100%)	239 (100%)	251 (100%)
Month 2	Negative	188 (78.33%)	196 (82.01%)	203 (80.88%)
	Positive	31 (12.92%)	17 (7.11%)	21 (8.37%)
	Missing	20 (8.33%)	26 (10.88%)	26 (10.36%)
	Died	1 (0.42%)	0	1 (0.40%)
Month 3	Negative	208 (86.67%)	197 (82.43%)	211 (84.06%)
	Positive	3 (1.25%)	8 (3.35%)	7 (2.79%)
	Missing	28 (11.67%)	34 (14.23%)	32 (12.75%)
	Died	1 (0.42%)	0	1 (0.40%)
	Negative	214 (89.17%)	204 (85.36%)	213 (84.86%)

Month 4	Positive	4 (1.67%)	5 (2.09%)	3 (1.20%)
	Missing	21 (8.75%)	30 (12.55%)	34 (13.55%)
	Died	1 (0.42%)	0	1 (0.40%)
Month 5	Negative	203 (84.58%)	193 (80.75%)	207 (82.47%)
	Positive	5 (2.08%)	3 (1.26%)	2 (0.80%)
	Missing	30 (12.50%)	42 (17.57%)	41 (16.33%)
	Died	2 (0.83%)	1 (0.42%)	1 (0.40%)
Month 6	Negative	194 (80.83%)	174 (72.80%)	204 (81.27%)
	Positive	6 (2.50%)	12 (5.02%)	1 (0.40%)
	Missing	38 (15.83%)	51 (21.34%)	45 (17.93%)
	Died	2 (0.83%)	2 (0.84%)	1 (0.40%)
Month 7	Negative	183 (76.25%)	168 (70.29%)	196 (78.09%)
	Positive	3 (1.25%)	17 (7.11%)	2 (0.80%)
	Missing	51 (21.25%)	51 (21.34%)	52 (20.72%)
	Died	3 (1.25%)	3 (1.26%)	1 (0.40%)
Month 8	Negative	178 (74.17%)	167 (69.87%)	192 (76.49%)
	Positive	2 (0.83%)	20 (8.37%)	2 (0.80%)
	Missing	57 (23.75%)	49 (20.50%)	56 (22.31%)
	Died	3 (1.25%)	3 (1.26%)	1 (0.40%)
Month 9	Negative	184 (76.67%)	163 (68.20%)	195 (77.69%)
	Positive	3 (1.25%)	13 (5.44%)	4 (1.59%)
	Missing	50 (20.83%)	59 (24.69%)	50 (19.92%)
	Died	3 (1.25%)	4 (1.67%)	2 (0.80%)
Month 10	Negative	174 (72.50%)	169 (70.71%)	182 (72.51%)
	Positive	4 (1.67%)	5 (2.09%)	4 (1.59%)
	Missing	58 (24.17%)	60 (25.10%)	62 (24.70%)
	Died	4 (1.67%)	5 (2.09%)	3 (1.20%)
Month 11	Negative	164 (68.33%)	156 (65.27%)	182 (72.51%)
	Positive	4 (1.67%)	4 (1.67%)	4 (1.59%)
	Missing	68 (28.33%)	72 (30.13%)	62 (24.70%)
	Died	4 (1.67%)	7 (2.93%)	3 (1.20%)
Month 12	Negative	169 (70.42%)	166 (69.46%)	192 (76.49%)
	Positive	5 (2.08%)	2 (0.84%)	2 (0.80%)

	Missing	61 (25.42%)	64 (26.78%)	54 (21.51%)
	Died	5 (2.08%)	7 (2.93%)	3 (1.20%)
Month 15	Negative	153 (63.75%)	147 (61.51%)	167 (66.53%)
	Positive	4 (1.67%)	5 (2.09%)	2 (0.80%)
	Missing	78 (32.50%)	79 (33.05%)	78 (31.08%)
	Died	5 (2.08%)	8 (3.35%)	4 (1.59%)
Month 18	Negative	138 (57.50%)	129 (53.97%)	152 (60.56%)
	Positive	4 (1.67%)	6 (2.51%)	2 (0.80%)
	Missing	93 (38.75%)	95 (39.75%)	90 (35.86%)
	Died	5 (2.08%)	9 (3.77%)	7 (2.79%)

NB: Missing includes contaminated results re-classed as "missing".

3.10 Analysis using imputation methods for the RIFAQUIN study

As for the REMoxTB analyses, we analyse the RIFAQUIN study by using a complete case analysis and best case/worst case scenarios (§3.3) followed by multiple imputation (§3.5.1) over 18 months of treatment and then apply the two-fold FCS multiple imputation method (§3.5.3). Table 3.9 shows the results of these analyses, along with the results for the PP and mITT analysis of the original study.

Table 3.9: Difference in proportions of unfavourable outcome using different imputation methods for the RIFAQUIN study.

Analysis	**4 month regimen**	**6 month regimen**
	Risk difference (95% CI)	**Risk difference (95% CI)**
Primary analysis (PP) from RIFAQUIN (n=514)*		
Unadjusted results	13.27% (6.52% to 20.03%)	-1.68% (-5.86% to 2.49%)
Adjusted results[1]	13.60% (7.0% to 20.20%)	-1.80% (-6.90% to 3.30%)

Primary analysis (mITT) from RIFAQUIN (n=593)*		
Unadjusted results	12.58% (4.56% to 20.60%)	-0.68% (-7.50% to 6.14%)
Adjusted results[1]	13.10% (5.60% to 20.60%)	-0.40% (-5.70% to 6.60%)
Complete case analysis (n=182)*		
Unadjusted results	3.22% (-7.26% to 13.70%)	-2.85% (-11.25% to 5.54%)
Adjusted results[1,2]	7.47% (-5.42% to 20.36%)	2.16% (-10.33% to 14.64%)
Best case scenario (n=730)*		
Unadjusted results	-43.70% (-51.22% to -36.18%)	-52.25% (-58.94% to -45.56%)
Adjusted results[1]	-43.63% (-51.11% to -36.14%)	-52.40% (-59.0% to -45.80%)
Worst case scenario (n=730)*		
Unadjusted results	60.69% (54.10% to 67.28%)	49.20% (42.49% to 55.91%)
Adjusted results[1,2]	63.09% (56.80% to 69.39%)	51.54% (45.09% to 57.98%)
Two-fold FCS MI[3] **(n=730)***		
Unadjusted results	9.80% (2.36% to 17.24%)	-3.30% (-8.60% to 1.99%)
Adjusted results[1]	10.26% (2.71% to 17.81%)	-2.69% (-7.98% to 2.60%)

*Number of patients included in the analysis.

[1]Adjusted for centre.

[2]Model did not converge, therefore 100 iterations were used.

[3]Fully Conditional Specification (FCS), Multiple imputation (MI).

The complete case analysis, patients were included only if culture results were reported for all visits. Patients missing any one or more culture results post-randomisation were excluded from this analysis. A total of 182 patients are included in the analysis, excluding an extra patient who had no positive culture results within 2 weeks of randomisation, a vast reduction from the 514 patients

included in the PP analysis. The adjusted results from the complete case analysis and worst case scenario failed to converge, due to there being few patients within each centre, and so 100 iterations were used. The results from the adjusted analysis for the four month treatment regimen (Table 3.9) are consistent with that of the primary outcome even though there are far fewer patients in the complete case analysis; non-inferiority cannot be concluded on the 4 month regimen as the upper bound of the 95% confidence interval (around 20% for the PP, mITT analyses and complete case adjusted analyses) exceeds that of the pre-defined 6% non-inferiority margin. Non-inferiority can be concluded on the 6 month regimen for the unadjusted analysis since the upper bound of the 95% confidence interval lies below the 6% non-inferiority margin; 3.3% for the primary PP adjusted analysis and 5.54% for the complete case analysis.

A total of 730 patients were included in the best case scenario and worst case scenario analyses after applying the exclusion criteria in Table 3.7. The best case scenario shows the 4 month treatment regimen and 6 month treatment regimen are non-inferior since the upper bound of the 95% confidence interval lies far below the 6% non-inferiority margin (upper bound of the 95% CI adjusted analysis: -36.14% for the 4 month regimen and 45.8% on the 6 month regimen). The worst case scenario fails to demonstrate non-inferiority in the 4 month regimen (95% CI: 69.39%) and in the 6 month treatment regimen (95% CI: 57.98%) for the adjusted analyses.

Multiple imputation was performed for all patients with an outcome and excluded 97 patients described in Table 3.7. As for the REMoxTB study, performing multiple imputation for across all 14 scheduled visits was computationally infeasible even after accounting for issues with perfect prediction. We therefore proceeded with the two-fold FCS multiple imputation, taking each scheduled visit in turn and imputing missing observations based on outcomes observed adjacent either side of that visit.

The results from the two-fold FCS multiple imputation (Table 3.9) were consistent with that of the primary analysis; non-inferiority cannot be concluded on the 4 month regimen but can be concluded on the 6 month regimen since the upper bound of the 95% CI is 2.6% on the adjusted analyses which lies below 6%. The result from the

two-fold FCS multiple imputation show that the control regimen performs slightly worse in comparison to the two treatment regimens as the estimates are less than that of the primary outcome; 10.26% on the 4 month regimen and -2.69% on the 6 month regimen after imputation compared with 14.30% on the 4 month regimen and -1.07% on the 6 month regimen from the primary analysis for adjusted analyses.

3.11 Missing data patterns for the RIFAQUIN study

To investigate trends of negative culture results for patients in the RIFAQUIN study, scheduled visits were split into windows for an overview of the missing data pattern and to reduce the number of possible sequences for patient profiles. Even though the timings of the treatment phase and follow-up phase differed by treatment regimen, patients were split into the same four visit windows:

- Months 0-3: month 0, month 2 and month 3;
- Months 4-6: month 4, month 5 and month 6;
- Months 7-10: month 7, month 8, month 9 and month 10;
- Months 11-18: month 11, month 12, month 15 and month 18.

The data were split in this way to reduce the number of missingness patterns across the whole dataset, allowing us to obtain a general overview of the missingness patterns while ensuring the windowing is clinically meaningful. Each visit window has a maximum of 4 visits in a visit window for ease of classing observed results within each window. As for the REMoxTB study (§3.8.2), culture results were grouped as "completers" (indicated by "O"), "intermittent" (indicated by "Δ") or "missing" (indicated by ".") within each visit window. If within a visit window, patients only missed one of their scheduled visits, they were considered to be "completers" since the majority of the results were complete in that window. Similarly, if a patient was only observed at one of their scheduled visits within a visit window, they were considered to be "missing" since most of the patient's results are missing within that visit window.

Patients were classed as as follows:

- "Completed" if a result was observed at all visits, or if one visit was missing at any one of the scheduled visits within the grouped visit;
- "Intermittent" if two results missing;
- "Missing" if all results were missing or if only one result was observed within the visit window.

Tables 3.10 to 3.12 shows the proportion of patients with negative culture results across different patterns of missing data in each visit window, and also shows the proportion of patients who achieved culture negative status at the end of the study in each missing data pattern. Table 3.10 describes patients who have most of their culture results observed across all visit windows (i.e. "completers" or a combination of "completers" and "intermittent"), or patients who are "missing" most of their results or those who died (indicated with a "D"). Table 3.11 summarises patients whose missing data pattern is mostly "intermittent" across visit windows. Table 3.12 summarises patients whose overall missingness pattern contains a mixture of "completers", "intermittent", "missing" or death across the four visit windows. A total of 384 (53%) patients were "completers" across all four visit windows of which 353 (92%) had achieved culture conversion by the end of the study.

Table 3.10: Number of negative culture results and proportion of all patients who achieved negative culture conversion for patients with most culture results observed (i.e completers' over visit windows for RIFAQUIN[1].

Months 0-3	Months 4-6	Months 7-10	Months 11-18	Total (N=730)	Number of negative culture results Months 0-3	Months 4-6	Months 7-10	Months 11-18	Total number of negative culture results	Treatment success n/no. patient per pattern
O	O	O	O	384 (52.60%)	695/1124 = 61.83%	1088/1112 = 97.84%	1415/1458 = 97.05%	1383/1415 = 97.74%	4581	353 (91.93%)
O	O	O	Δ	60 (8.22%)	108/176 = 61.36%	173/177 = 97.74%	215/225 = 95.56%	115/120 = 95.83%	611	56 (93.33%)
O	O	O	.	42 (5.75%)	76/122 = 62.30%	119/120 = 99.17%	139/146 = 95.21%	29/29 = 100.00%	363	38 (90.48%)
O	O	.	.	40 (5.48%)	73/118 = 61.86%	99/106 = 93.40%	9/11 = 81.82%	9/9 = 100.00%	190	17 (42.50%)
O	.	.	.	31 (4.25%)	45/84 = 53.57%	15/15 = 100.00%	1/1 = 100.00%	4/5 = 80.00%	65	3 (9.68%)
.	.	.	.	26 (3.56%)	0/26 = 0.00%	0/0 = .%	0/0 = .%	2/3 = 66.67%	2	0 (0.00%)
O	Δ	O	O	22 (3.01%)	40/65 = 61.54%	42/44 = 95.45%	83/84 = 98.81%	82/83 = 98.80%	247	21 (95.45%)
Δ	O	O	O	19 (2.60%)	17/38 = 44.74%	53/54 = 98.15%	68/74 = 91.89%	67/69 = 97.10%	205	16 (84.21%)
O	O	Δ	O	13 (1.78%)	22/37 = 59.46%	34/35 = 97.14%	23/26 = 88.46%	46/47 = 97.87%	125	12 (92.31%)
O	.	O	O	8 (1.10%)	14/23 = 60.87%	6/6 = 100.00%	27/27 = 100.00%	28/29 = 96.55%	75	6 (75.00%)
.	O	O	O	4 (0.55%)	0/4 = 0.00%	10/10 = 100.00%	15/15 = 100.00%	14/14 = 100.00%	39	4 (100.00%)
O	.	.	O	3 (0.41%)	4/8 = 50.00%	2/2 = 100.00%	2/2 = 100.00%	9/9 = 100.00%	17	2 (66.67%)
O	O	O	D	3 (0.41%)	6/9 = 66.67%	8/8 = 100.00%	10/10 = 100.00%	0/0 = .%	24	3 (100.00%)
O	O	D	D	2 (0.27%)	4/6 = 66.67%	4/4 = 100.00%	0/0 = .%	0/0 = .%	8	0 (0.00%)
O	D	D	D	2 (0.27%)	4/6 = 66.67%	2/2 = 100.00%	0/0 = .%	0/0 = .%	6	0 (0.00%)
.	.	.	O	2 (0.27%)	0/2 = 0.00%	0/0 = .%	1/1 = 100.00%	8/8 = 100.00%	9	0 (0.00%)
O	O	.	D	2 (0.27%)	3/5 = 60.00%	6/6 = 100.00%	2/2 = 100.00%	0/0 = .%	11	1 (50.00%)
O	O	.	O	2 (0.27%)	4/6 = 66.67%	6/6 = 100.00%	2/2 = 100.00%	6/6 = 100.00%	18	2 (100.00%)
.	.	O	O	2 (0.27%)	0/2 = 0.00%	2/2 = 100.00%	7/7 = 100.00%	7/7 = 100.00%	16	0 (0.00%)
D	D	D	D	2 (0.27%)	0/2 = 0.00%	0/0 = .%	0/0 = .%	0/0 = .%	0	0 (0.00%)
.	.	.	D	1 (0.14%)	0/1 = 0.00%	0/0 = .%	0/0 = .%	0/0 = .%	0	0 (0.00%)
.	O	O	.	1 (0.14%)	0/1 = 0.00%	3/3 = 100.00%	3/3 = 100.00%	1/1 = 100.00%	7	0 (0.00%)

[1]Where O=Most results observed within a window; Δ=Intermittent results observed within a window; D=Death or .=missing.

Table 3.11: Number of negative culture results and proportion of patients who achieved negative culture conversion for patients with most culture results intermittently observed over visit windows for RIFAQUIN[1].

O	O	Δ	Δ	9 (1.23%)	16/26 = 61.54%	26/26 = 100.00%	16/18 = 88.89%	18/18 = 100.00%	76	8 (88.89%)
Δ	O	O	Δ	8 (1.10%)	6/16 = 37.50%	21/22 = 95.45%	28/29 = 96.55%	16/16 = 100.00%	71	7 (87.50%)
O	Δ	O	Δ	3 (0.41%)	6/9 = 66.67%	6/6 = 100.00%	12/12 = 100.00%	6/6 = 100.00%	30	3 (100.00%)
Δ	O	Δ	O	1 (0.14%)	1/2 = 50.00%	3/3 = 100.00%	2/2 = 100.00%	3/3 = 100.00%	9	1 (100.00%)
Δ	Δ	O	O	1 (0.14%)	1/2 = 50.00%	2/2 = 100.00%	4/4 = 100.00%	4/4 = 100.00%	11	1 (100.00%)
Δ	Δ	O	Δ	1 (0.14%)	1/2 = 50.00%	2/2 = 100.00%	4/4 = 100.00%	2/2 = 100.00%	9	1 (100.00%)
.	.	Δ	Δ	1 (0.14%)	0/1 = 0.00%	1/1 = 100.00%	2/2 = 100.00%	2/2 = 100.00%	5	0 (0.00%)
O	Δ	Δ	O	1 (0.14%)	1/3 = 33.33%	2/2 = 100.00%	2/2 = 100.00%	3/3 = 100.00%	8	1 (100.00%)
Δ	Δ	Δ	O	1 (0.14%)	1/2 = 50.00%	2/2 = 100.00%	2/2 = 100.00%	4/4 = 100.00%	9	1 (100.00%)

[1]Where O=Most results observed within a window; Δ=Intermittent results observed within a window or .=missing.

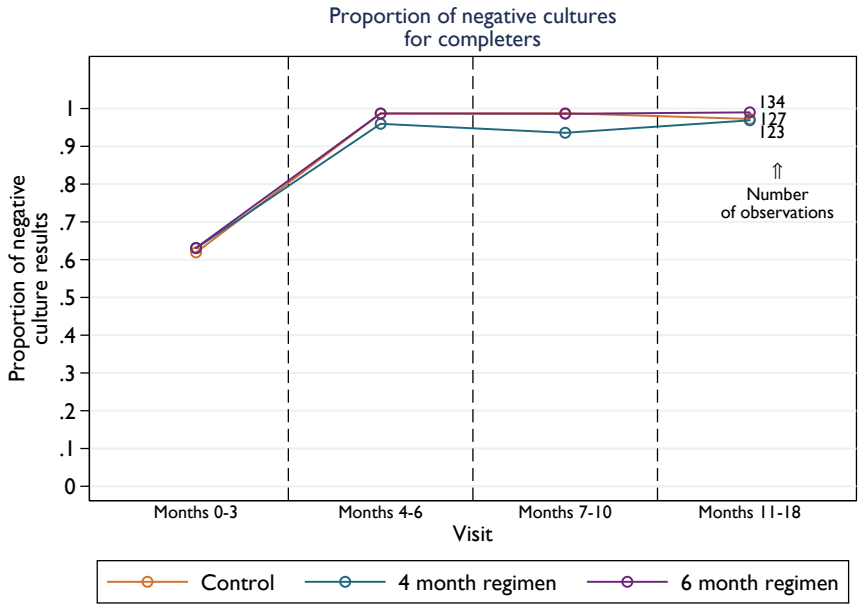
Table 3.12: Number of negative culture results and proportion of patients who achieved negative culture conversion for patients with a mixture of observed, intermittent and missing culture results over visit windows for RIFAQUIN[1].

O	O	Δ	.	12 (1.64%)	23/36 = 63.89%	33/33 = 100.00%	24/24 = 100.00%	4/4 = 100.00%	84	11 (91.67%)
O	.	Δ	O	5 (0.68%)	8/15 = 53.33%	4/4 = 100.00%	9/10 = 90.00%	19/19 = 100.00%	40	4 (80.00%)
O	O	.	Δ	3 (0.41%)	6/9 = 66.67%	9/9 = 100.00%	2/2 = 100.00%	6/6 = 100.00%	23	3 (100.00%)
O	.	.	Δ	2 (0.27%)	3/5 = 60.00%	2/2 = 100.00%	1/1 = 100.00%	4/4 = 100.00%	10	2 (100.00%)
O	.	O	Δ	1 (0.14%)	1/3 = 33.33%	1/1 = 100.00%	3/3 = 100.00%	2/2 = 100.00%	7	1 (100.00%)
O	Δ	O	D	1 (0.14%)	2/3 = 66.67%	2/2 = 100.00%	3/3 = 100.00%	0/0 = .%	7	1 (100.00%)
Δ	.	O	O	1 (0.14%)	1/2 = 50.00%	1/1 = 100.00%	3/3 = 100.00%	3/3 = 100.00%	8	1 (100.00%)
O	Δ	O	.	1 (0.14%)	2/3 = 66.67%	2/2 = 100.00%	4/4 = 100.00%	1/1 = 100.00%	9	1 (100.00%)
.	O	O	Δ	1 (0.14%)	0/1 = 0.00%	3/3 = 100.00%	1/4 = 25.00%	2/2 = 100.00%	6	0 (0.00%)
O	O	Δ	D	1 (0.14%)	1/2 = 50.00%	3/3 = 100.00%	2/2 = 100.00%	0/0 = .%	6	1 (100.00%)
Δ	Δ	O	.	1 (0.14%)	1/2 = 50.00%	2/2 = 100.00%	4/4 = 100.00%	1/1 = 100.00%	8	1 (100.00%)
Δ	.	.	O	1 (0.14%)	1/2 = 50.00%	1/1 = 100.00%	1/1 = 100.00%	3/3 = 100.00%	6	1 (100.00%)
O	Δ	.	Δ	1 (0.14%)	2/3 = 66.67%	2/2 = 100.00%	0/0 = .%	2/2 = 100.00%	6	1 (100.00%)
O	Δ	D	D	1 (0.14%)	2/3 = 66.67%	2/2 = 100.00%	0/0 = .%	0/0 = .%	4	0 (0.00%)
O	Δ	.	.	1 (0.14%)	2/3 = 66.67%	2/2 = 100.00%	0/0 = .%	0/0 = .%	4	0 (0.00%)

[1]Where O=Most results observed within a window; Δ=Intermittent results observed within a window; D=Death or .=missing.

Figures 3.11 to 3.20 describe the proportion of patients who had negative culture results over different missing data patterns over each visit window between treatment arms. The numbers in each of the figures show how many patients followed a type of missingness pattern.

Figure 3.11: Proportion of negative culture results for completers in RIFAQUIN.



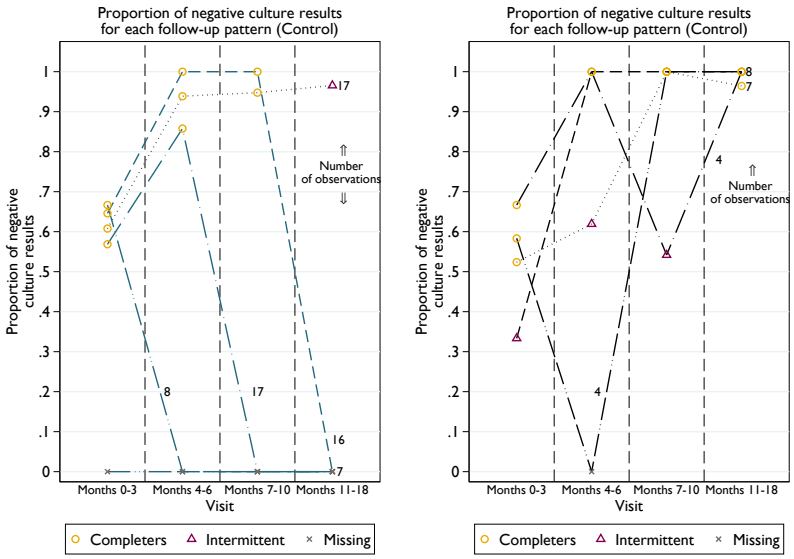
The proportion of negative culture results for completers, i.e. those with one or fewer missing culture results in a visit window, (Figure 3.11) shows that the 4 month regimen starts off as performing slightly better than the 6 month and control regimens, but worsens over the next two periods.

Figures 3.12 to 3.20 show the most common missing data patterns for patients who have the majority of their culture results observed/missing results, intermittent/missing culture results or a mixture of "completers", "intermittent" or "missing" results across visit windows. Figures 3.12 to 3.14 is a graphical representation of Table 3.10 by treatment arm, showing the most common patterns for patients who are nearly always observed. Figures 3.15 to 3.17 show the most common patterns from Table 3.11 by treatment arm for patients whose culture results are mostly intermittent. Figures 3.18 to 3.20 is a graphical representation of patients in

Table 3.12 by treatment arm whose pattern has a mixture of "completers", "intermittent", "missing" or death across visit windows. Numbers noted within the graphs relate to the number of patients that follow a particular pattern.

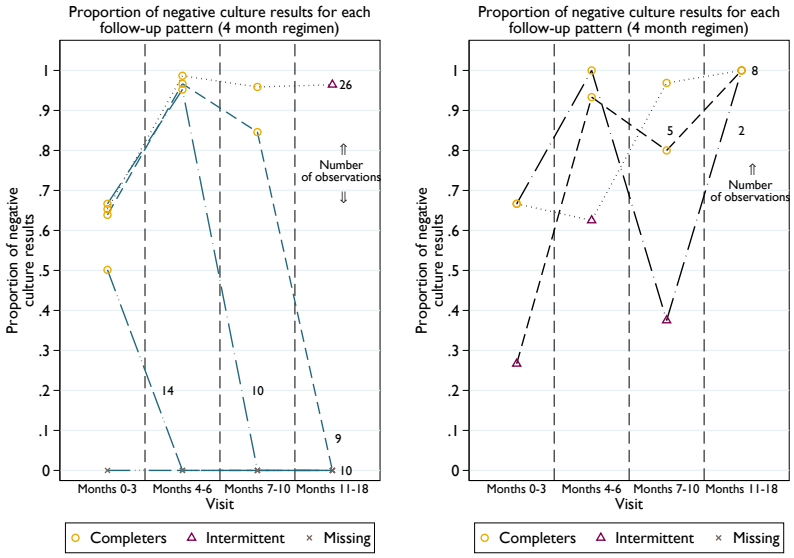
Figures 3.12 to 3.20 are consistent with what we found in the REMoxTB study, where patients whose pattern includes observations that are either missing or intermittently missing within one or more visit windows generally have similar proportions of negative culture results with the other treatment arms when observed. This finding suggests that using the worst case analysis, as recommended by regulators, may be an overly conservative analysis for these data. These figures also suggest that the MAR assumption is reasonable. That is, to assume that a patient's observed culture result towards the end of the study given what they were at the beginning is different from other patients within treatment arms.

Figure 3.12: Proportion of negative culture results for completers' pattern in RIFAQUIN.



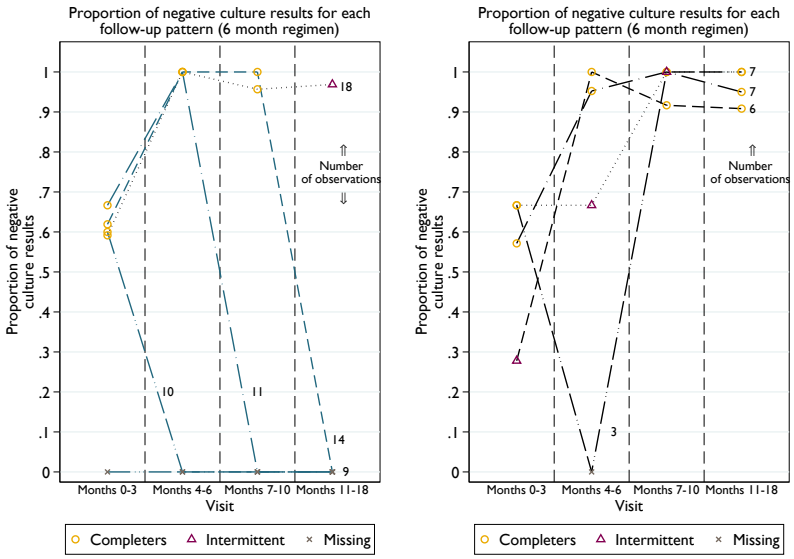
NB: Due to the vast number of patterns, graphs are split into two for ease of viewing each pattern.

Figure 3.13: Proportion of negative culture results for completers' pattern in RIFAQUIN.



NB: Due to the vast number of patterns, graphs are split into two for ease of viewing each pattern.

Figure 3.14: Proportion of negative culture results for completers' pattern in RIFAQUIN.



NB: Due to the vast number of patterns, graphs are split into two for ease of viewing each pattern.

Figure 3.15: Proportion of negative culture results for intermittent missing pattern.

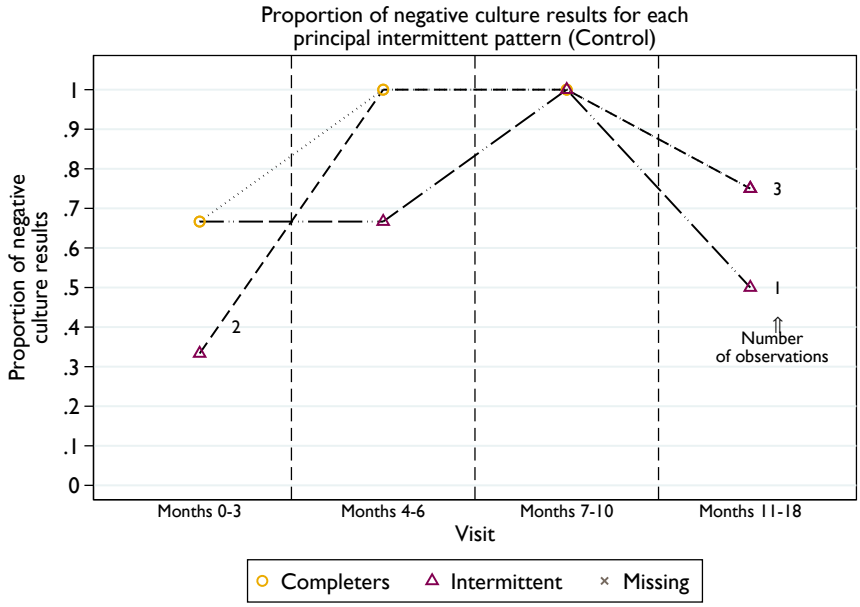


Figure 3.16: Proportion of negative culture results for intermittent missing pattern in RIFAQUIN.

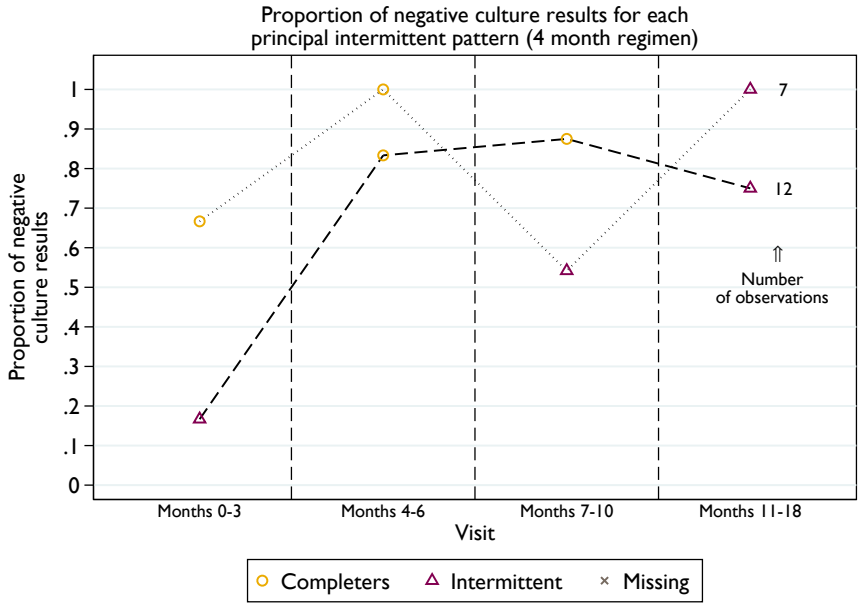


Figure 3.17: Proportion of negative culture results for intermittent missing pattern in RIFAQUIN.

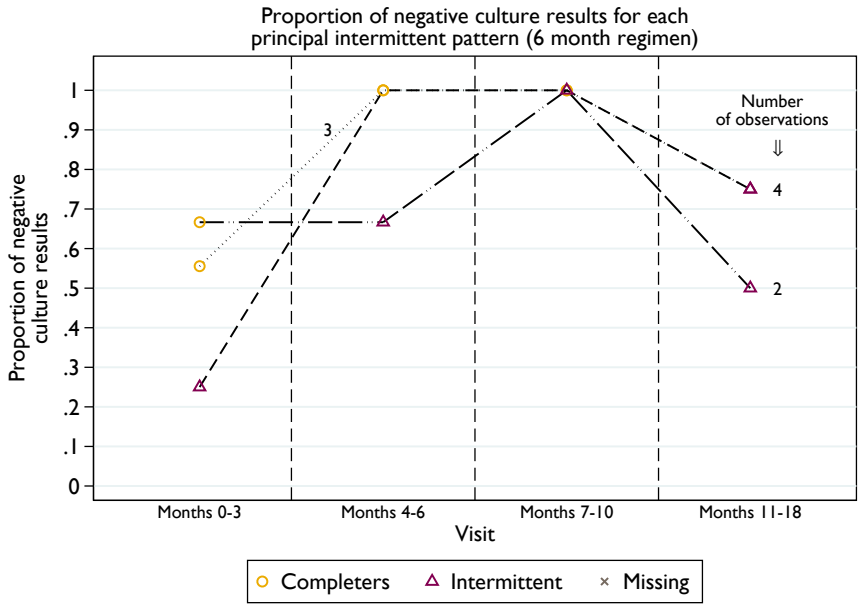


Figure 3.18: Proportion of negative culture results for a mixture of results in RIFAQUIN.

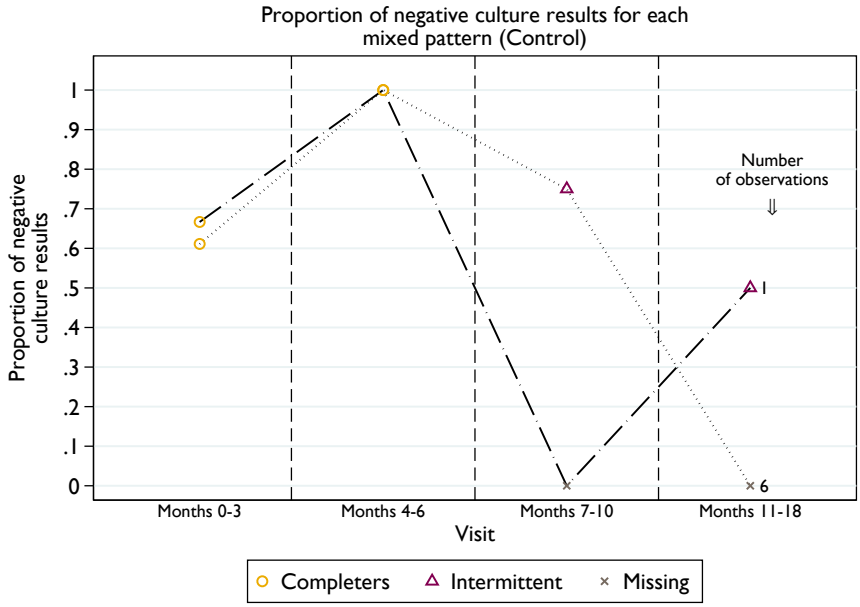


Figure 3.19: Proportion of negative culture results for a mixture of results in RIFAQUIN.

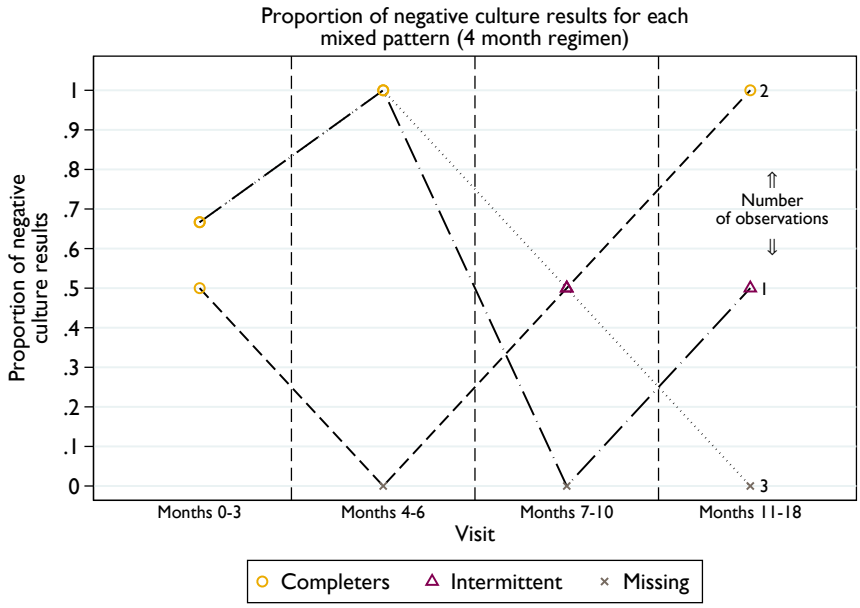
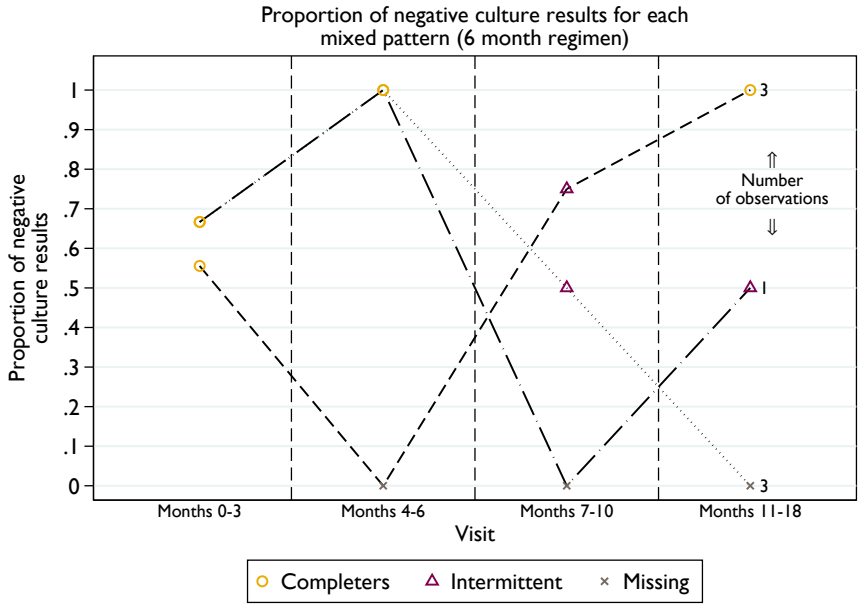


Figure 3.20: Proportion of negative culture results for a mixture of results in RIFAQUIN.



3.11.1 Discussion

In this section we applied different imputation methods to the RIFAQUIN dataset resulting in a "completed" dataset for this study. This enabled us to impute each patient's culture result. We began with single imputation methods using a complete case analysis and best case/worst case scenarios. We then used multiple imputation and two-fold fully conditional specification multiple imputation. Following this, we investigated the proportion of missing data for the RIFAQUIN study by treatment arm. We then explored different patterns of missing culture results by treatment arm, grouping visits into visit windows.

Although the results from the complete case analysis were consistent with that of the primary analysis, the exclusion of 78% of patients who were randomised clearly creates greater uncertainty surrounding the estimates and therefore greater uncertainty as to whether non-inferiority was met. The unadjusted results from the complete case analysis lean towards failing to demonstrate non-inferiority, suggesting the analysis is biased towards favouring the standard of care regimen. The results from the best case and worst case scenario show extreme estimates and confidence intervals demonstrating and failing to demonstrate non-inferiority respectively. While the worst case scenario bears more weight according to regulatory guidelines, and is consistent with the primary results of the RIFAQUIN study, the results may be overly extreme. This is supported by the figures showing the proportion of negative culture results for different patterns of missingness, since patterns which show patients who are missing most of their results within a visit window (and comparing these to results observed for other patterns) suggest that it is most likely they would have had a high proportion of having negative culture results if they were actually observed. Following two-fold FCS multiple imputation, including all patients within the analysis was consistent with the primary results found by the study. For both treatment regimens, the analyses from the two-fold FCS suggest the results are not as extreme as those found from the primary analyses, although there is not a huge gain in information when using this method.

3.12 Summary

In this chapter, we applied different imputation methods to the REMoxTB and RIFAQUIN datasets. Doing so resulted in a "completed" dataset for each study, including more patient information. This enabled us to impute each patient's sputum culture result as a treatment failure or as reaching stable negative culture conversion. It is clear from both these studies, that the complete case analysis creates extreme confidence intervals. This is due to excluding patients, and therefore information that clearly contributes to the primary outcome, who are missing a culture result from at least one of their follow-up visits from the analysis. Even though the complete case failed to demonstrate non-inferiority, which is broadly consistent with the primary analysis, the resulting confidence intervals are misleading are they are more extreme compared to the primary analysis.

The best case and worst case scenario analyses provide a threshold for the very best and worst circumstances, however, the results are extreme. These analyses demonstrate a need for better methods to deal with missing data rather than using implausible assumptions about the missing observations.

Using standard multiple imputation on both datasets to impute missing observations at all follow-up visits was problematic due to issues with perfect prediction. This is perhaps not surprising for TB trials, where patients are unlikely to be cured in the first few weeks of treatment and therefore most (if not all) patients have a positive culture result. This causes problems when trying to impute the data computationally as the perfect prediction leads to infinite regression parameters[81]. Therefore, we proceeded with an alternative imputation method; two-fold fully conditional specification multiple imputation. This method takes each follow-up visit in turn, imputing results based on observations either side of that visit and propagating that information forwards for future imputations. The results from the two-fold FCS multiple imputation were similar to results from the primary analysis in both studies, but showed a lack of benefit using this method. By using a smaller window of observed visits that surround the visit we wish to impute culture results for may have inflated the confidence intervals slightly. More investigation is required to ensure whether the

two-fold fully conditional specification multiple imputation model used here is satisfactory to impute missing observations, and consequently determine the primary analysis.

The missing data patterns explored for the REMoxTB study and RIFAQUIN study suggest that even if patients were missing culture results over the duration of follow-up, the probability of having a negative culture is likely to be much higher than 0 if a patient was actually observed. For sensitivity analyses, regulators recommend using a worst case scenario where patients who are lost to follow-up and therefore missing observations are considered "unfavourable". This assumption means it is assumed that the probability of missing patients having a negative culture result is 0. However, the figures presented in this chapter that show different missing data patterns by treatment arm suggest that this recommended sensitivity analysis is overly conservative. It is most likely that the probability for having a negative result if missing is similar to other patients who were actually observed at that missing visit and is therefore greater than 0.

The figures plotted in this chapter for the REMoxTB and RIFAQUIN studies showed that missing data patterns across visits were similar between the treatment arms. The figures generally showed that within each visit window the proportion of negative culture results were similar for each pattern. Therefore, the MAR assumption is a reasonable primary assumption for the REMoxTB and RIFAQUIN studies since there appears to be no particular reason why the probability that patients' observed culture results towards the end of the study given what they were at the beginning of the study is different from other patients within treatment arms.

Next we explore using inverse probability weighting (IPW) as an alternative analysis to handle missing data in REMoxTB. The data are kept within visit windows as the long sequence of data leads to issues with perfect prediction. That is that the fitted probabilities are very close to 1 at the start of follow up as most patients have a positive result or 0 towards the end of follow up where most patients become disease free and have a negative culture result. IPW is arguably a simpler method than multiple imputation which up-weights those with a higher probability of being

missing relative to those who have a lower probability of being missing. IPW is conceptually simpler as the method is marginal and so does not condition on previous observations in a dataset like multiple imputation does[82]. IPW assumes data follow a monotone missing pattern where patients who withdraw from a study are never observed again at future visits. Given that data from the REMoxTB study does not follow this pattern, a monotone pattern is imposed. Next, multiple imputation is investigated where the missingness pattern is non-monotone. Generalised estimating equations (GEEs) are used with IPW, providing an alternative to maximum likelihood based methods applied and discussed here in this chapter.

Since the visits are grouped into blocks of visits, another analysis is to count the number of negative culture results within each visit window. This is investigated using a multilevel mixed-effects Poisson regression model to analyse the number of negative culture results within each of the four visit windows. These analyses are then repeated for the RIFAQUIN study.

Chapter 4

Inverse probability weighting

In this Chapter, we explore a different analysis using inverse probability weighting (IPW), assuming MAR, to try and correct for the missing data. In the context for longitudinal data, this means that the probability of a missing response is independent of its current and future responses conditional on the observed past responses and covariates[83]. This is a reasonable assumption as described in §3.8.3. Each observation is weighted by the conditional probability of that observation being observed:

$$\text{P(seen at T)} = \prod_{t=1}^{T} \text{P}(\text{r}_{k,t} = 1 | X_{k,t-1}, Y_k)^{-1} \tag{4.1}$$

where $r_{k,t}$ = 1 if the outcome, $y_{k,t}$ is observed and 0 otherwise for each patient k at time t,...,T. This is the probability of being observed at time t given a patient was observed at the previous time point, $t-1$.

IPW can be used provided that data follow a *monotone* missing pattern[84]. In longitudinal analyses, a monotone pattern describes the pattern of missingness for patients within a study as those who are missing an observation at a visit and at all future visits and/or patients who complete the study with all visits observed. An example of this would be patients who are lost to follow up where patients would never be re-observed at future visits.

For the studies investigated here, the data are kept within visit windows. Given the issues of perfect prediction in Chapter 3 when trying to multiply impute the missing observations across all visit windows, the data are kept within our clinically

meaningful visit windows to ensure that the weights are more stable within each window. Doing so will allow for a simpler approach to analyse these studies.

Before applying IPW to the REMoxTB study, we investigate any possible predictors of failure on the primary outcome and from patient withdrawal. Any predictors will then be included within the weights for the IPW model. We then explore Generalised Estimating Equation (GEE) models with and without weighting, imposing a monotone missingness pattern to our data before exploring a Poisson regression model as an alternative analysis.

4.1 Predictions of outcome failure and withdrawals for the REMoxTB study

First, to better represent the population sample within the REMoxTB study, we investigate different predictors for outcome failure in addition to withdrawal to account for the fact that outcome failure is dependent on being observed. Variables that are predictive of *both* outcome failure and withdrawal will be included within the weighting. If weights are included from variables which are only predictive of withdrawal and not related to the effect of the treatment (i.e. the chance of outcome failure), then using inverse probability weighting will be ineffective because patient probability weights that increase or decrease are independent of whether or not a patient is going to fail treatment or not. However, if important predictors of both withdrawal and outcome failure are found, then using inverse probability weighting and including those variables could improve the weighting and impact the results compared to when no weighting is used. Therefore including predictors from both outcome failure and withdrawal will account for the fact that the proportion of patients who withdraw from a study may influence the overall outcome by the end of the study.

By the definition of the primary outcome, patients who are missing culture results towards the end of the study are classed as "unfavourable". Each patient is classed as a "success" or "failure" within each visit window. Outcome failure was defined as follows:

- Patients who were never "successful" (i.e. never achieved two consecutive negative culture results at separate visits);
- Patients who had more than two positive culture results and did not culture convert back to a stable negative status during the treatment phase (weeks 0 to 26).

Patients who had a positive culture result followed by a negative culture result and preceded by at least two consecutive negative culture results when last seen were classed as treatment success. This definition is amended slightly to that of the primary outcome where patients whose last positive result was not followed by at least two negative results when last seen were considered as "unfavourable". For analyses in this chapter, we only look at results collected at the time of scheduled follow-up. All unscheduled results are ignored since not all patients have an unscheduled result and the number of unscheduled results vary from patient to patient. This enables the results to be more comparable within visit windows.

Reasons for withdrawing permanently from follow up included:

- Patients who did not reach the end of the treatment phase (week 26);
- Patients who died (either due to non-violent death or due to TB/respiratory distress);
- Patients who withdrew from the study due to pregnancy;
- Patients who moved away or withdrew consent.

Logistic regression models were used to identify variables that predicted outcome failure and withdrawals. Baseline covariates included in the model were time to positivity, weight band ($\leq$40kg, 40-45kg, $\geq$45-55kg or $\geq$55kg-$\geq$75kg), age (years), x-ray cavitation (yes or no), smoking status (never, past or present), race (Asian, black, mixed race or other), HIV (positive or negative), sex (male or female) and grouped centre (Stellenbosch, Cape Town, Other South African centre, India, Kenya/Zambia/Tanzania or Other (East Asia)). Other covariates included were production of sputum samples taken from patients (yes or no) and time to not producing sputum. Patients were classed as not being able to produce sputum if they

had at least one occurrence of not being able to produce a sputum sample at any time over their follow-up. Time to not producing sputum was taken from the time of randomisation until the first occurrence of not being able to produce sputum. For patients who were able to produce sputum, without any occurrence of not being able to produce sputum, time was taken until they were last observed in the study. Patients who had not produced sputum and were lost to follow-up were assumed to have not produced sputum, and their time to not producing sputum was taken up to the point they were last observed. The Nelson-Aalen estimate of time to not producing sputum was taken. The Nelson-Aalen estimate is a cumulative hazard function ($\hat{H}(t)$) which is used to predict how censored patients evolve over the remaining duration of a study. This estimate looks at time from being able to produce sputum to no longer being able to produce sputum. The estimate is denoted by:

$$\hat{H}(t) = \sum_{t_k<t} \frac{g_k}{n_k}, \tag{4.2}$$

where g_k is the number of events that occur and n_k is the total number of patients included in the analysis for time t_k.

Body Mass Index (BMI) was also collected at baseline but was not included within the model due to strong correlation with weight. Covariates were removed from the regression model using the backward stepwise technique, using a 5% level of significance.

A total of 212 (12%) patients were classed as failures and a total of 98 (5%) patients withdrew from the study. Table 4.1 and Table 4.2 show estimates from our final model of covariates that could be included in our weights model for IPW. These models show predictions of outcome failure and predictions of withdrawals. Unadjusted and adjusted results from all covariates included in the models are shown in Appendix D.

Table 4.1: Final model showing adjusted odds ratios (OR) and confidence intervals (CI) for predicting outcome failure for REMoxTB.

Covariate	Adjusted OR[1]	95% CI	P-value
Treatment			
Isoniazid	1.800	(1.279, 2.532)	0.001
Ethambutol	1.970	(1.403, 2.766)	0.000
Centre[2]			
Cape Town	0.853	(0.521, 1.398)	0.529
Other South Africa	0.394	(0.811, 2.433)	0.226
India	3.134	(1.982, 4.957)	0.000
Kenya/Zambia/Tanzania	1.445	(0.911, 2.293)	0.118
Other East Asia	1.099	(0.642,1.882)	0.732
Sex			
Female	0.674	(0.480, 0.947)	0.023
Smoking status[3]			
Past	1.948	(1.355, 2.801)	<0.001
Current	1.912	(1.287, 2.840)	0.001
Sputum produced			
No	0.231	(0.150, 0.358)	<0.001
Time to not producing sputum (weeks)	0.082	(0.037, 0.181)	<0.001

[1]Adjusted for all other covariates in the model.

[2]Likelihood-ratio test for centre P< 0.001.

[3]Likelihood-ratio test for smoking status P$= 0.001$.

Table 4.2: Final model showing adjusted odds ratios (OR) and confidence intervals (CI) for variables predictive of withdrawals for REMoxTB.

Covariate	Adjusted OR	95% CI	P-value
Sputum produced			
No	0.018	(0.007, 0.042)	<0.001
Time to not producing sputum (years)	0.0004	(0.0001, 0.002)	<0.001

Predictions of failure include treatment, centre, sex, smoking status, inability to produce sputum and time to not producing sputum and for withdrawals inability to produce sputum and time to not producing sputum were predictors (Table 4.2). We can see that treatment, smoking status and most centres all increase the chance of outcome failure. Interestingly, being female was associated with a reduction of around 33% in outcome failure (OR: 0.67; 95% CI: 0.48 to 0.95) which is similar to what Phillips et al found where being male was predictive of an unfavourable outcome[85]. Patients randomised in Cape Town had less chance of failing treatment compared to patients randomised in Stellenbosch. Not being able to produce sputum reduces the chance of outcome failure by around 98% (OR: 0.018; 95% CI: 0.007 to 0.042). A shorter time to not producing sputum is associated with a decrease in the chance of withdrawing by 99.96% (OR:0.0004; 95% CI: 0.01 to 0.04).

4.2 Discussion

In this section we investigated predictors of outcome failure and loss to follow-up. Predictions of outcome failure were treatment, centre, sex, smoking status, inability to produce sputum and time to not producing sputum. For withdrawal, not producing sputum and time to not producing sputum were predictors. Having investigated important predictions for outcome failure and withdrawals, we need to include covariates that predict both outcome failure and withdrawal since outcome failure is dependent on withdrawal. Therefore, inability to produce sputum and the Nelson-Aalen estimate for time to not producing sputum will be included as weights in our IPW model. We now investigate using IPW for the REMoxTB data. First, we

impose a monotone missing data pattern and then investigate Generalised Estimating Equations (GEEs) without weights and then with IPW. We then use multiple imputation, without imposing a monotone pattern to the data.

4.3 Application of inverse probability weighting to the REMoxTB study

For the REMoxTB study we use IPW (see Chapter 4) to calculate the probability of a patient being observed within each visit window, and then weight each observation in the analysis by the inverse of that conditional probability. This method could potentially improve the efficiency of the estimates since the method allows us to include the missing observations in the analysis. The REMoxTB study actually follows a *non-monotone* missing data pattern where patients are missing visits but are observed at future visits. To explore IPW, we first need to impose a monotone missing pattern. Patients were classed as a success (i.e. achieving two consecutive negative culture results) or a failure within each visit window (see §3.8.2), thus creating binary observations within each visit window. Patients were assumed to be missing future visits from the first point of withdrawal, even if patients were subsequently re-observed at future visits. In instances where there were two consecutive negative culture results in different visit windows, patients were classed as a success at the point of achieving culture negative status. For example, patients who have their first negative result at week 4 and a negative result at week 5 would be classed "successful" between weeks 0 to 4.

Table 4.3 shows the proportion of patients in each treatment arm in each visit window. At weeks 5 to 8, the proportion of patients who are successful in the ethambutol arm is 75.4% and is slightly lower in the control arm (402 (68.1%)). A total of 92 (15.6%) patients are missing culture results in the control arm. A total of 82 (13.5%) and 73 (12.5%) patients are missing in the isoniazid and ethambutol treatment arms respectively between 5 to 8 weeks. These differences settle towards the end of treatment by week 78 with the control and ethambutol regimens achieving around a 69% proportion of success. The proportion of success is slightly lower for the isoniazid regimen at around 66%. The proportion of patients who are missing between weeks 39

to 78 are similar between treatment arms, ranging between 26% to 29%, but the proportion of failure is over twice as high in the two treatment regimens than in the control regimen (over 4% vs. 1.86%). Although, it is important to note that there are a small proportion of patients within these groups.

Table 4.3: Proportion of patients considered to be a "success", "failure" or "missing" imposing a monotone missingness pattern in REMoxTB.

Visit	Outcome	Control	Isoniazid	Ethambutol
Week 0 to 4	Success	171 (28.98%)	191 (31.36%)	185 (31.57%)
	Fail	381 (64.58%)	375 (61.58%)	364 (62.12%)
	Missing	38 (6.44%)	43 (7.06%)	37 (6.31%)
Week 5 to 8	Success	402 (68.14%)	452 (74.22%)	442 (75.43%)
	Fail	96 (16.27%)	75 (12.32%)	71 (12.12%)
	Missing	92 (15.59%)	82 (13.46%)	73 (12.46%)
Week 12 to 26	Success	456 (77.29%)	466 (76.52%)	462 (78.84%)
	Fail	7 (1.19%)	17 (2.79%)	22 (3.75%)
	Missing	127 (21.53%)	126 (20.69%)	102 (17.41%)
Week 39 to 78	Success	410 (69.49%)	404 (66.34%)	407 (69.45%)
	Fail	11 (1.86%)	24 (3.94%)	25 (4.27%)
	Missing	169 (28.64%)	181 (29.72%)	154 (26.28%)

Table 4.4 shows the proportion of patients who have negative culture results over time within each treatment according to their monotone pattern. Patients who are observed are denoted as "O" within the table. The majority of patients are observed across all visit windows (over 70%) for all treatment arms.

Table 4.4: Monotone missing data pattern for patients with negative results in REMoxTB, by treatment arm.

Treatment	Week 0 to 4	Week 5 to 8	Week 12 to 26	Week 39 to 78	Number of patients[1] n(%)
Control (N = 590)	O	O	O	O	421 (71.36%)
	O	O	O	.	42 (7.12%)
	O	O	.	.	35 (5.93%)
	O	.	.	.	54 (9.15%)
	.	.	.	.	38 (6.44%)
Isoniazid (N = 609)	O	O	O	O	428 (70.28%)
	O	O	O	.	55 (9.03%)
	O	O	.	.	44 (7.22%)
	O	.	.	.	39 (6.40%)
	.	.	.	.	43 (7.06%)
Ethambutol (N = 586)	O	O	O	O	432 (73.72%)
	O	O	O	.	52 (8.87%)
	O	O	.	.	29 (4.95%)
	O	.	.	.	36 (6.14%)
	.	.	.	.	37 (6.31%)

[1] at the end of the study.

Figures 4.1 to 4.3 show similar trends of negative culture results within each monotone missing pattern for the control and ethambutol regimens, although the proportion of negative culture results are slightly lower in the ethambutol regimen for those observed between weeks 12 to 26. In the isoniazid regimen, the proportion of patients with negative culture results for patients that are missing at week 12 to 26 and week 39 to 78 is slightly lower at around 62% compared with the control and ethambutol regimen which has around 70% negative culture result, and is slightly higher for patients who are fully observed or are fully observed until the final visit window.

Figure 4.1: Proportion of negative culture results in control regimen imposing a monotone missing pattern for REMoxTB.

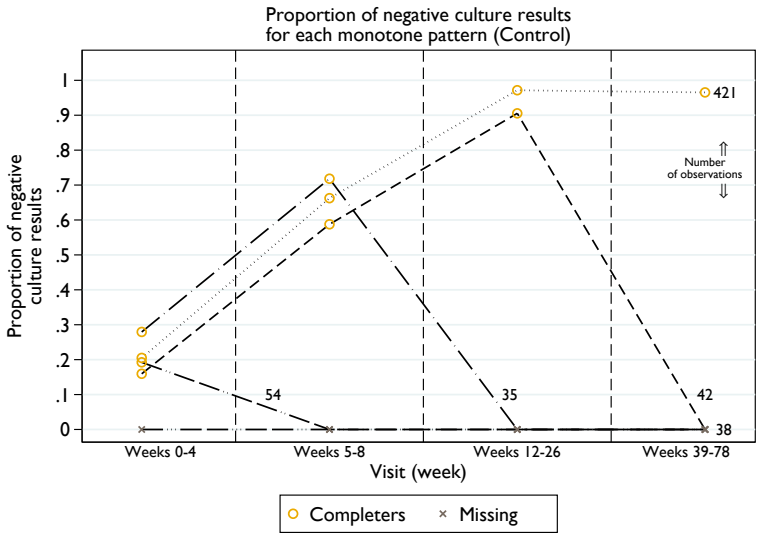


Figure 4.2: Proportion of negative culture results in isoniazid regimen imposing a monotone missing pattern for REMoxTB.

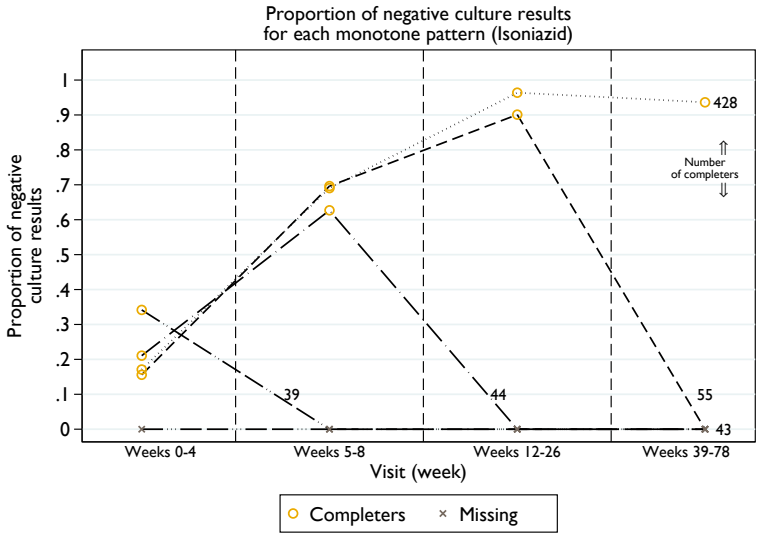
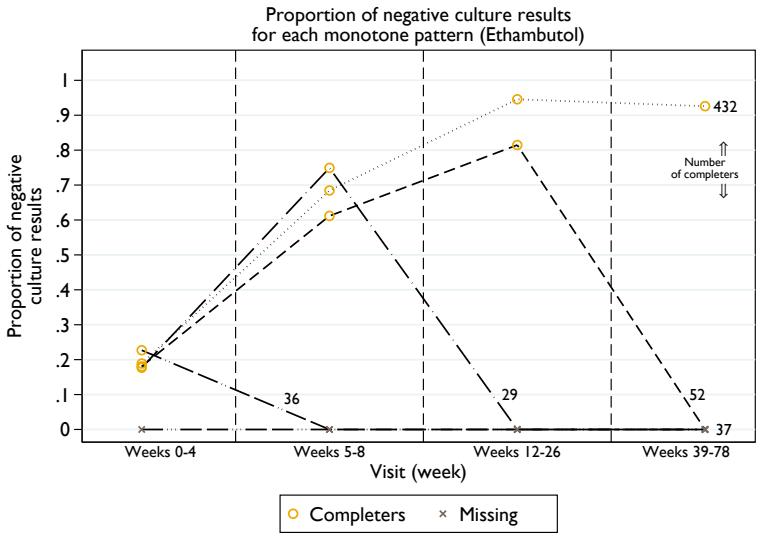


Figure 4.3: Proportion of negative culture results in ethambutol regimen imposing a monotone missing pattern for REMoxTB.



4.3.1 Discussion

Having imposed a monotone pattern on the REMoxTB data, failure decreases over time but the amount of missing data increases. The proportion of negative culture results within each monotone missing pattern are similar between treatment arms.

Next, generalised estimating equations (GEEs) are used to provide a flexible approach for modelling the average population using the observed data from the REMoxTB study, keeping patients classed as "success" or "failure" within each visit window. First, patients are only included if observed at all four visit windows and then all patients with an observed outcome over visit windows are included in the analysis. Secondly, inverse probability weighting is investigated within the GEE models including baseline covariates that were predictive of both outcome failure and withdrawal while imposing a monotone pattern to the data. In this model, we estimate the probability of patients being observed or withdrawing at each visit window given that they were observed at the previous visit window, and this probability will be incorporated into the IPW GEE analysis. We then use multiple

imputation without imposing a monotone pattern to the data. Doing so reflects the true structure of the data and so we investigate whether imposing a monotone pattern to the data has major implications in the results of these analyses.

4.4 Generalised Estimating Equations

Generalised estimating equations (GEEs) provide an alternative to maximum likelihood based methods, which model the population-average (or marginal) effect of covariates. GEE models are more flexible as they do not assume a particular type of distribution for repeated outcomes observed over time. Instead, the method links each marginal mean to a linear predictor and provides a working assumption about the correlation for the variance-covariance structure of the repeated outcomes observed over time. The sandwich estimator of the variance can be used so that even if the working assumption is misspecified, then the standard errors are still reasonably estimated provided there is enough data.

In relation to our two example trial datasets, the GEE model extends the logistic regression model (see equation 3.5) and is denoted by:

$$\log\left(\frac{\pi_{k,t}}{1-\pi_{k,t}}\right) = \log\left(\frac{p(fail_{k,t}=1)}{1-p(fail_{k,t}=1)}\right) = \beta_0 + \beta_1 trt_k + \beta_2 time_{k,t} + \beta_3 trt_k time_{k,t} \quad (4.3)$$

for patient k at time t. The correlation structure needs to be specified in the model to account for the correlated observations for repeated outcome measurements. We specify independent and unstructured working correlation matrices which makes different assumptions about the relationship between the repeated observations. See Appendix E for other correlation matrices.

1. Independence which assumes repeated observations are independent.

$$R_{k,t} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

2. Unstructured which assumes the correlation structure cannot be appropriately modelled and each correlation must be estimated.

$$R_{k,t}(\rho) = \begin{pmatrix} 1 & \rho_{21} & \rho_{31} & \rho_{41} \\ \rho_{21} & 1 & \rho_{32} & \rho_{42} \\ \rho_{31} & \rho_{32} & 1 & \rho_{43} \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 \end{pmatrix}$$

If the missing data are MCAR (§3.3), and if the working correlation is incorrectly specified the resulting estimates remain consistent provided the marginal model is correct, although there will be some loss in efficiency reflected in the larger standard errors. In this case independence is often assumed and the sandwich estimator of the variance is assumed. Otherwise, we could assume a simple structure such as first order autoregression (see Appendix E) and use the sandwich estimator of the variance. However, if there is missing data then the choice of the variance-covariance matrix is no longer a nuisance parameter and is therefore no longer something where if incorrect the correct inferences will be drawn from the data asymptotically. When patients withdraw from a study, the way information on those patients from their initial visits flows through to the treatment estimates is critically dependent on the variance-covariance matrix. The variance-covariance matrix becomes important and is no longer a nuisance parameter. If the sandwich estimator is incorrect, then the wrong inferences will be drawn from the data even asymptotically due to patient withdrawal. We therefore try to use IPW in addition to the GEEs to take into account the missing data[86].

4.4.1 Calculation for risk differences

For comparison to the original results reported in the studies we explore here we convert the odds ratios estimated by the GEE models (4.3) and calculate them as risk differences. To convert the odds ratios into risks, the inverse of the log odds ratio is calculated as follows:

$$RD = \left[\frac{e^{\beta_0 + \beta_1 trt}}{1 + e^{\beta_0 + \beta_1 trt}} - \frac{e^{\beta_0}}{1 + e^{\beta_0}} \right] \tag{4.4}$$

where β_0 represents the log odds of treatment failure for patients randomised to the control regimen and β_1 is the difference in the log odds of treatment failure comparing treatment trt (where trt=isoniazid or ethambutol for REMoxTB and trt=4m regimen or 6m regimen for RIFAQUIN) to the control regimen.

4.4.2 Application to REMoxTB

For the REMoxTB study, visits were kept in visit windows (see §3.8.2) and analyses were performed within each visit window to investigate the difference between the treatment regimens and standard of care over time to focus on the trend of treatment failure over 78 weeks of follow-up, and this was done for all subsequent analyses. Patients are excluded for reasons not related to treatment (Table 3.1). Treatment effects are estimated only from patients who had, by definition (§3.8.2), completed results from randomisation to week 78. GEE models are then used without including weights using all the information from all patients from randomisation to week 78. An unstructured variance-covariance matrix is used as it assumes no two pairs of observations are equally correlated and also allows for different variance terms along the diagonal of the matrix[87]. Finally IPW is used to include observed weights within the GEE model. GEE models were fitted using a binomial distribution and a logit link and were back-transformed into risk differences (see §4.4.1) to compare these results with the results from the REMoxTB study. Results from all analyses were compared to a 6% non-inferiority margin to assess how the effect of treatment changes over time.

Results

For all GEE analyses performed, a total of 118/1785 (6.6%) patients had no observations at any of the four visit windows and were therefore excluded from all analyses.

The GEE model that included patients who were observed at all visit windows (i.e. completers) is shown in Table 4.5. The risk differences are presented for treatment failure and relapse by treatment arm at each visit window from the GEE model assuming an unstructured variance-covariance matrix. This analysis includes 1281/1667 (77%) patients. Non-inferiority is met between weeks 5 to 8 and weeks 12 to 26 and can be concluded for patients randomised to the ethambutol arm between weeks 0 to 4 since the upper bound of the 97.5% CI is 4.8%. In the final visit window non-inferiority cannot be concluded since the upper bound of the 97.5% CI just lies above the 6% pre-determined non-inferiority margin.

Table 4.5: GEE model for a difference in proportions of treatment failure including "completers" assuming an unstructured variance-covariance matrix, by treatment arm for REMoxTB.

	Risk difference	SE	97.5% CI
Week 0 to 4			
Isoniazid	-0.007	0.032	(-0.078, 0.065)
Ethambutol	-0.024	0.032	(-0.096, 0.048)
Week 5 to 8			
Isoniazid	-0.036	0.025	(-0.091, 0.020)
Ethambutol	-0.044	0.025	(-0.099, 0.011)
Week 12 to 26			
Isoniazid	0.011	0.009	(-0.009, 0.032)
Ethambutol	0.018	0.010	(-0.004, 0.040)
Week 39 to 78			
Isoniazid	0.030	0.014	(-0.0004, 0.060)
Ethambutol	0.032	0.014	(0.001, 0.062)

Table 4.6 presents the results from using a GEE model that includes all patients who had at least one outcome during follow up. Table 4.6 shows the risk difference of treatment effects from the control regimen for failure, that is patients who did not reach culture negative status (§4.1), at each visit window from the GEE model assuming an unstructured variance-covariance matrix. The results from this analysis draw similar conclusions to the GEE analysis which included patients observed across all four visit windows (Table 4.5). That is, at the 6% margin, non-inferiority could be concluded between weeks 0 to 4 on the ethambutol regimen, weeks 5 to 8 for both treatments and non-inferiority failed to be demonstrated in the final visit window (weeks 39 to 78). The association between outcome failure and treatment arm when all patients are included in the analysis, decreases between weeks 5 to 8 and then increases until the end of the study. Between weeks 12 to 26, the upper bound of the 97.5% confidence intervals tend more towards failing to demonstrate non-inferiority (4.3% for isoniazid and 5.5% for ethambutol) compared to 3.2% for isoniazid and 4.0% for ethambutol when patients are fully observed in all four visit windows (Table 4.5).

Table 4.6: GEE model for a difference in proportions of treatment failure including all patients in the analysis assuming unstructured variance-covariance matrix, by treatment arm for REMoxTB.

	Risk difference	Robust SE's	97.5% CI
Week 0 to 4			
Isoniazid	-0.028	0.028	(-0.090, 0.035)
Ethambutol	-0.027	0.028	(-0.090, 0.036)
Week 5 to 8			
Isoniazid	-0.050	0.023	(-0.103, 0.018)
Ethambutol	-0.054	0.023	(-0.107, -0.002)
Week 12 to 26			
Isoniazid	0.020	0.010	(-0.003, 0.043)
Ethambutol	0.030	0.011	(0.006, 0.055)
Week 39 to 78			
Isoniazid	0.030	0.014	(-0.0005, 0.060)
Ethambutol	0.032	0.014	(0.001, 0.062)

Having looked at the differences between treatment arms within each window, we are able to see whether demonstrating or failing to demonstrate non-inferiority changes over the course of follow-up. The results show that at earlier follow up visits, patients randomised to either the isoniazid or ethambutol treatment arms are associated with fewer failures in comparison to the standard of care arm. This changes, after week 8, where patients who were randomised to either the isoniazid or ethambutol treatment arm were associated with an increase in failure over time. The results up to week 26 demonstrate non-inferiority when compared to a 6% non-inferiority margin, however there is a large decrease in the upper bound of the resulting CI between week 12 - 26 from the previous visit window (week 5 to 8) on both treatment regimens. By week 39 to 78, patients randomised to the isoniazid regimen are associated with higher failures (3%: 97.5% CI; -0.05% to 6%) and similarly for patients randomised to the ethambutol regimen (3.2%: 97.5% CI; 0.1% to 6.2%). Non-inferiority cannot be concluded as the upper bounds of the two-sided 97.5% CIs are only just greater than the 6% non-inferiority margin between weeks 39 to 78.

4.4.3 Discussion

The results from the GEE model that includes patients who were observed in all four visit windows (i.e. "completers") and the results where all patients were included irrespective of completeness were very similar. However, when concluding non-inferiority the results from the GEE model only including patients who were observed at all visits fluctuated between concluding and failing to conclude non-inferiority at the 6% margin at that time point. Assuming an unstructured variance-covariance matrix can create more noise surrounding the parameter estimates or may not even be able to provide sensible estimates for our data, however the results are consistent with the conclusions made from the study.

Next, we add in weights to the GEE model to better represent the patients included in the REMoxTB study.

4.5 Weighted Generalised Estimating Equations

Weights are included in the GEE model using inverse probability weights. As explained in §4.3, IPW weights complete records by the inverse probability of observing the data. We assign patients different weights at each visit rather than at patient level across all visits, and the probabilities are calculated for the weights based on predictions of failure and loss to follow up found in §4.1. These covariates are included in the model so that patients with similar characteristics are weighted according to this. Including the covariates assumes the observations and therefore the correlation between any repeated measurements at each time point are independent of each other. Given this assumption of independence when weights from covariates are included, an independent structure sandwich estimator of the variance also needs to be specified to reflect this. Consider the following marginal mean model for the REMoxTB study:

$$\log\left(\frac{\pi_{k,t}}{1-\pi_{k,t}}\right) = \beta_0 + \beta_1 trt_k + \beta_2 week_{k,t} + \beta_3 trt_k week_{k,t}$$

for patient k. Visit windows 0 to 4 weeks, 5 to 8 weeks, 12 to 26 weeks and 39 to 78 weeks are represented as 1-4 for each window, t. To obtain weights for this model

consider the following logistic regression model for patients who are missing culture results:

$$\text{logit}(y_{k,t-1}) = \gamma + \gamma_1 r_{k,t-1} + \gamma_2 treat_{k,iso} + \gamma_3 trt_{k,eth} + \gamma_4 nosputum_{k,t-1} + \gamma_5 timenosputum_{k,t-1}$$

where $y_{k,t-1}$ represents the observed response (i.e. success or failure) at the previous visit window *t-1* for each patient *k*; where if patients are observed on the control arm then $treat_{k,iso}$=0 and $treat_{k,eth}$=0, if patients are observed on the isoniazid arm then $treat_{k,iso}$=1 and $treat_{k,eth}$=0 and if patients are observed on the ethambutol arm then $treat_{k,iso}$=0 and $treat_{k,eth}$=1 for each patient *k*; $nosputum_{k,t-1}$ represents the observed response of not being able to produce sputum at the previous visit window *t-1* for each patient *k* and $timenosputum_{k,t-1}$ is the Nelson Aalen estimate for time to not producing a sputum result at the previous visit window *t-1* for each patient *k*.

We only include covariates that were significant from both loss to follow up and outcome failure to take into account patients who fail treatment. That is not producing sputum and time to not producing sputum. Even though treatment was not a significant predictor for patients who withdrew (§4.1), it is an important covariate and we therefore keep this in the model. Predictions were obtained from the marginal model at weeks 5 to 8 (pr8), 12 to 26 (pr26) and 39 to 78 (pr78). Observation weights were calculated as follows:

1. Weeks 0 to 4: all patients are observed and therefore the weight equals to 1 for all patients included in the analysis,

2. Weeks 5 to 8: 1/pr8 where pr8 is the probability of being observed between weeks 5 to 8 given patients were observed between weeks 0 to 4,

3. Weeks 12 to 26: 1/(pr8*pr26) where pr8*pr26 is the probability of being observed between weeks 12 to 26 given patients were observed between weeks 0 to 4 and 5 to 8,

4. Weeks 39 to 78: 1/(pr8*pr26*pr78) where pr8*pr26*pr78 is the probability of being observed between weeks 39 to 78 given patients were observed between weeks 0 to 4, 5 to 8 and 12 to 26.

An independent variance-covariance matrix is used since the weights created for the observations within the dataset assume independence. Figures 4.4 to 4.6 summarises

the probability weights calculated, by treatment arm and by outcome (success/failure) at each week. The probability weights calculated for patients who are classed as success are consistent across the visit windows at around 0.9. However, for patients classed as failures, the probability weights between week 39 to 78 are much lower at around 0.7 in comparison to failures in other weeks which have weights at around 0.9. This is expected since there are mostly successes towards the end of the study between weeks 39 to 78.

Figure 4.4: Histogram of probability weights between weeks 5 to 8 given weeks 0 to 4 for REMoxTB.

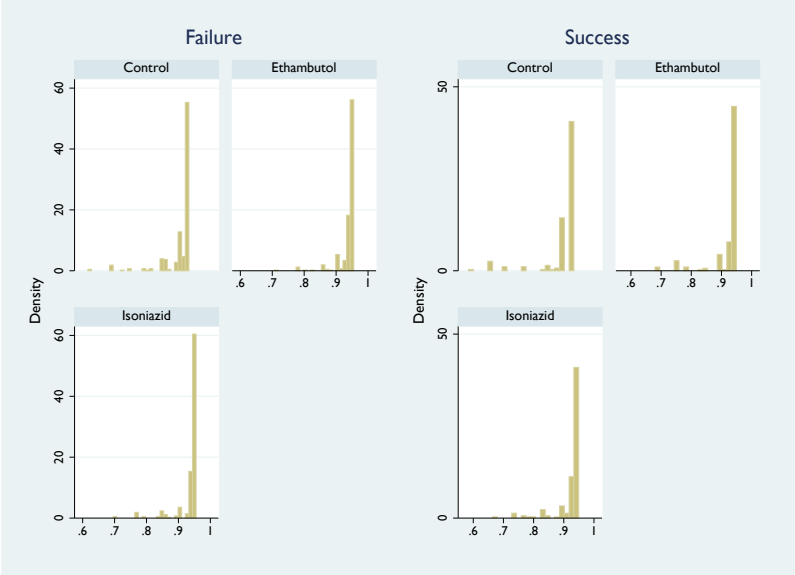


Figure 4.5: Histogram of probability weights between weeks 12 to 26 given weeks 0 to 4 and weeks 5 to 8 for REMoxTB.

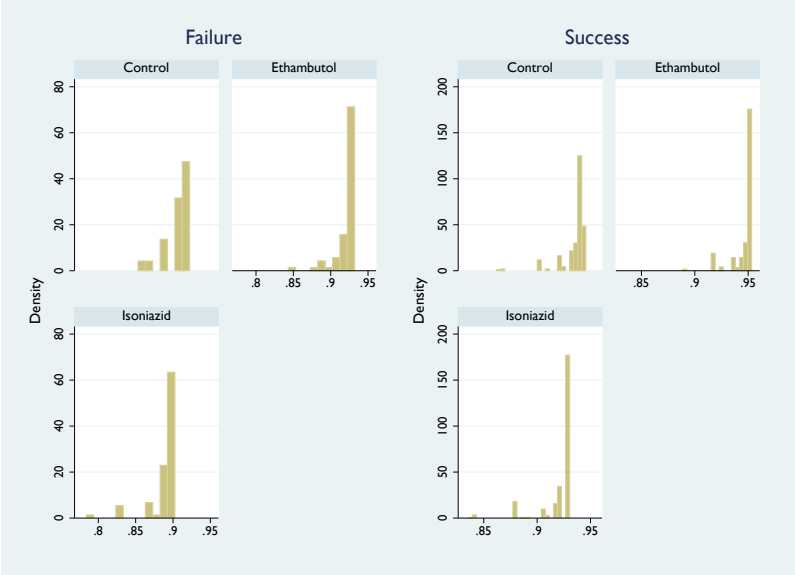


Figure 4.6: Histogram of probability weights at week 39 to 78 given week 0 to 4, week 5 to 8 and week 12 to 26 for REMoxTB.

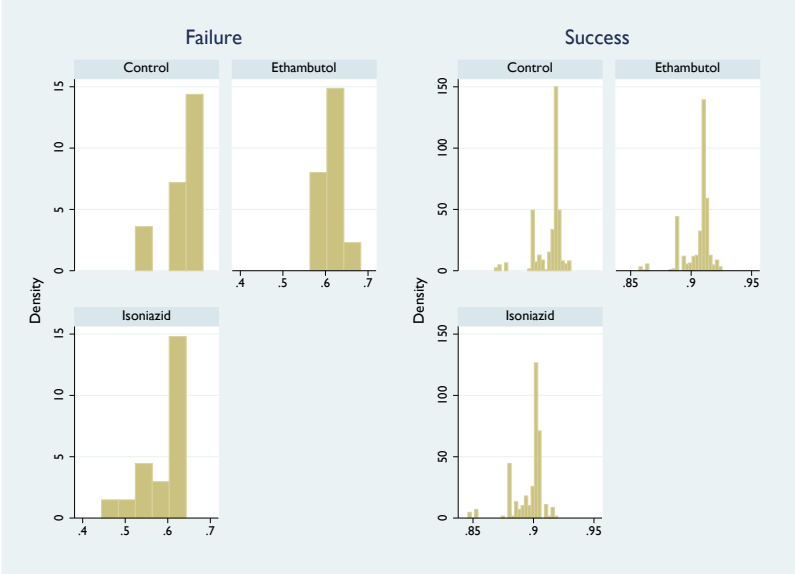


Table 4.7 shows the results from the GEE model including observed weights calculated by using predicted probabilities from the logistic regression model. The results from this model and conclusions made about non-inferiority are similar to the GEE model when including all patients in the analysis without weights; non-inferiority can be concluded at the 6% margin up to week 26 but non-inferiority can no longer be concluded between week 39 to 78. Looking at the upper bound of the 97.5% confidence intervals, results from weeks 39 to 78 show that we are less likely to conclude non-inferiority (6.6% for isoniazid and 6.8% for ethambutol) compared to the GEE model including all patients without including weights (6.0% for isoniazid and 6.2% for ethambutol; Table 4.6).

Table 4.7: GEE model for a difference in proportions using estimated weights from data observed assuming an independent variance-covariance matrix, by treatment arm for REMoxTB.

	Risk difference	SE	97.5% CI
Week 0 to 4			
Isoniazid	-0.028	0.028	(-0.090, 0.035)
Ethambutol	-0.027	0.028	(-0.090, 0.036)
Week 5 to 8			
Isoniazid	-0.050	0.023	(-0.103, 0.002)
Ethambutol	-0.055	0.023	(-0.107, -0.003)
Week 12 to 26			
Isoniazid	0.020	0.010	(-0.002, 0.043)
Ethambutol	0.031	0.011	(0.006, 0.056)
Week 39 to 78			
Isoniazid	0.033	0.015	(-0.0001, 0.066)
Ethambutol	0.035	0.015	(0.002, 0.068)

4.5.1 Discussion

Having investigated using GEE models we have demonstrated that inclusion of patients who have complete observations, that is patients who are observed at every

single visit, rather than including patients who are observed at least once over the follow-up of the study can lead to different conclusions. This is perhaps unsurprising since excluding patients who have information contributing to an analysis will always result in loss of information reflected in the wider confidence intervals. As a consequence of the greater uncertainty, the ability to conclude non-inferiority decreases since the wider confidence intervals are more likely to include the value of the pre-determined margin. The observed weights that included weighting on sputum production and the Nelson-Aalen estimate for time to not producing sputum which predicted outcome failure and withdrawals were added into the GEE model along with treatment. The results from this model are consistent with results from the GEE model including all patients in the analysis without weights and are consistent with conclusions made with regards to non-inferiority with a 6% margin.

Next multiple imputation is investigated keeping the data monotone to check the results with the GEE models performed in this section. Data from the REMoxTB study follow a non-monotone pattern and so multiple imputation imposing a non-monotone missingness pattern will also be performed.

4.6 Multiple Imputation for monotonic and non-monotonic missing patterns

Multiple imputation is used to confirm the results produced from the GEE models used above in §4.4 to §4.5. To do this we keep the data in a monotone pattern. We then revert back to the original structure of the data and investigate whether there is any gain of information by using multiple imputation for a non-monotone pattern where follow-up visits are kept within windows.

4.6.1 Monotone pattern

To check on the validity of the results from the GEE model, data were imputed by treatment arm keeping the pattern of missing data monotone within visit windows using multiple imputation. Data were imputed using logistic regression and 100 imputations were used. As presented for the original study, risk differences for treatment failure between the treatment and control regimens were calculated using a

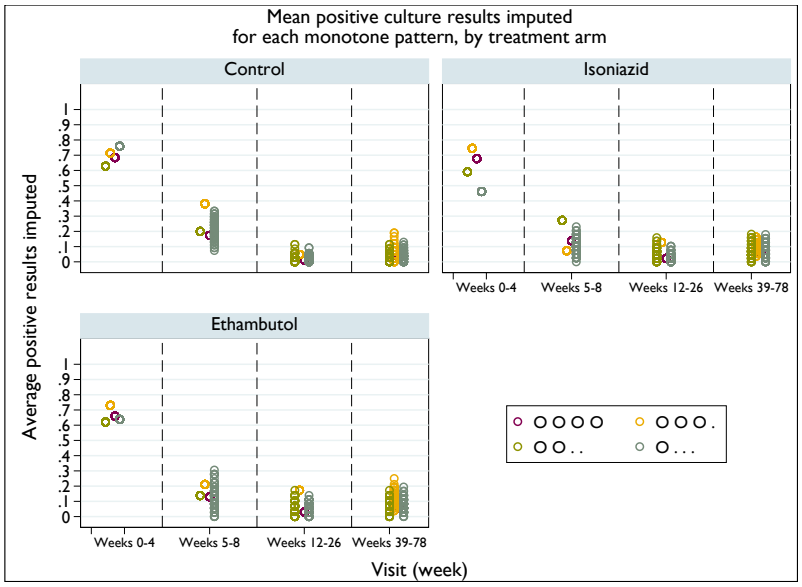
generalised linear model with an identity link for treatment failure. The results following imputation are shown in Table 4.8 and are similar to the GEE model that includes all patients who had at least one observed outcome with and without weighting (Table 4.6 and 4.7).

Table 4.8: Difference in proportions for treatment failure following multiple imputation, by treatment arm for REMoxTB.

	Risk difference	SE	97.5% CI
Week 0 to 4			
Isoniazid	-0.028	0.028	(-0.090, 0.035)
Ethambutol	-0.027	0.028	(-0.090, 0.036)
Week 5 to 8			
Isoniazid	-0.055	0.023	(-0.106, -0.003)
Ethambutol	-0.056	0.023	(-0.109, -0.004)
Week 12 to 26			
Isoniazid	0.020	0.010	(-0.003, 0.043)
Ethambutol	0.030	0.011	(0.005, 0.055)
Week 39 to 78			
Isoniazid	0.031	0.015	(-0.002, 0.064)
Ethambutol	0.035	0.015	(0.002, 0.069)

Figure 4.7 shows imputed results of positive culture results for each monotone missing pattern from the imputation model. The average number of positive culture results imputed were calculated by treatment arm and missingness pattern for each of the 100 imputed datasets created. Figure 4.7 shows the imputation model has imputed sensible results in each treatment arm for each missingness pattern since the imputed values are closely fitted to the observed values, and there are no outliers.

Figure 4.7: Imputed results of mean positive cultures where a monotone pattern is imposed for REMoxTB.



Data used for these models had a monotone missingness pattern imposed whereas the data from REMoxTB actually follow a non-monotone missingness pattern. To see whether there is any more information to be gained, multiple imputation where a non-monotone missingness pattern is imposed.

4.6.2 Non-monotone pattern

The REMoxTB study follows a non-monotone missing pattern, where patients who are observed and are missing follow-up visits are then observed again at later visits as described in §4.3. Table 4.9 shows the proportion of patients in each treatment arm in each visit window where a non-monotone missing pattern is imposed on the data.

Table 4.9: Proportion of patients with a non-monotone missingness pattern imposed for REMoxTB.

Visit	Outcome	Control	Isoniazid	Ethambutol
Week 0 to 4	Success	171 (28.98%)	191 (31.36%)	185 (31.57%)
	Fail	381 (64.58%)	375 (61.58%)	364 (62.12%)

	Missing	38 (6.44%)	43 (7.06%)	37 (6.31%)
Week 5 to 8	Success	414 (70.17%)	460 (75.53%)	452 (77.13%)
	Fail	96 (16.27%)	75 (12.32%)	71 (12.12%)
	Missing	80 (13.56%)	74 (12.15%)	63 (10.75%)
Week 12 to 26	Success	508 (86.10%)	509 (83.58%)	511 (87.20%)
	Fail	11 (1.86%)	18 (2.96%)	23 (3.92%)
	Missing	71 (12.03%)	82 (13.46%)	52 (8.87%)
Week 39 to 78	Success	485 (82.20%)	460 (75.53%)	464 (79.18%)
	Fail	16 (2.71%)	29 (4.76%)	27 (4.61%)
	Missing	89 (15.08%)	120 (19.70%)	95 (16.21%)

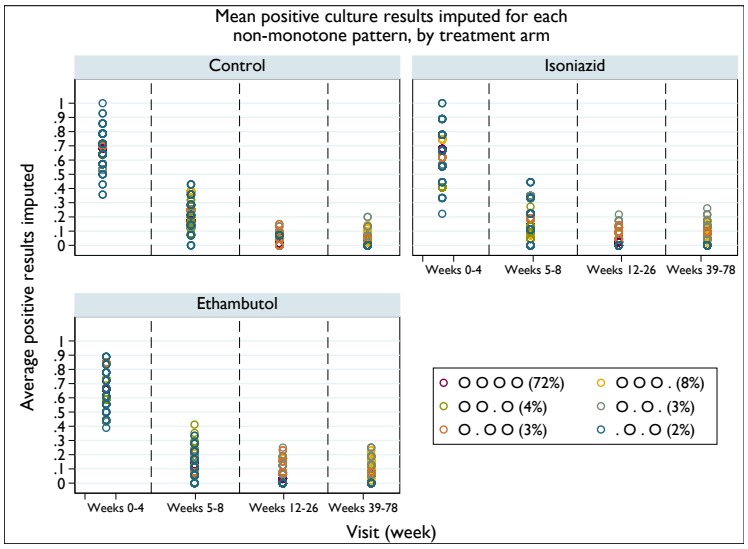
The proportion of patients who are successes, failures or missing are similar to the proportion of patients who are successes, failures or missing following a monotone pattern up to week 8 (Table 4.3). Between week 12 to 26 and week 39 to 78, there are more patients who are successful across all treatment regimens when a monotone missing pattern is not imposed compared to when a monotone missing pattern is imposed in Table 4.3. Given there are in fact 10% more patients who achieve culture negative status in the study, the monotone pattern may be producing biased treatment effects at later visits given that this data has been thrown away, wasting resources.

Table 4.10: Difference in proportions for treatment failure following multiple imputation where the pattern is non-monotone, by treatment arm for REMoxTB.

	Risk difference	SE	97.5% CI
Week 0 to 4			
Isoniazid	-0.026	0.028	(-0.089, 0.036)
Ethambutol	-0.026	0.028	(-0.090, 0.037)
Week 5 to 8			
Isoniazid	-0.054	0.020	(-0.105, -0.002)
Ethambutol	-0.057	0.015	(-0.109, -0.005)
Week 12 to 26			
Isoniazid	0.013	0.010	(-0.011, 0.037)
Ethambutol	0.021	0.011	(-0.004, 0.046)

Week 39 to 78			
Isoniazid	0.030	0.014	(-0.001, 0.061)
Ethambutol	0.027	0.014	(-0.004, 0.058)

Figure 4.8: Imputed results for the mean rate of positive culture results following principal patterns of non-monotone data for REMoxTB.



Data were imputed by treatment arm using logistic regression and with 100 imputations. Table 4.10 shows the difference in proportions for treatment failure between treatment and control regimens from imputing the data following a non-monotone missing pattern. The difference in proportions were calculated using a general linear regression model with an identity link function in each visit window. The results show a similar trend to when a monotone missing pattern is imposed, where the upper bound of the 97.5% confidence intervals are slightly lower in comparison. Figure 4.8 shows the imputed results of positive culture results for each monotone missing pattern from the imputation model. The average number positive culture results imputed were calculated by treatment arm and each non-monotone missingness pattern for each of the 100 imputed datasets created. We can see that the imputation model has imputed sensible results in each treatment arm for each missingness pattern since the imputed values are closely fitted to the observed values, and there are no outliers.

4.7 Discussion

The results from keeping a non-monotone missing pattern are broadly similar to the imputation results when a monotone missing pattern is imposed; the ethambutol regimen is associated with fewer failures in comparison to the control arm whereas the isoniazid regimen is associated with slightly higher failures between weeks 0 to 4. Between weeks 5 to 8 the treatment regimens have a lower associated risk of failure in comparison to the control arm when not imposing a monotone missing pattern to the data. However, this effect is reversed by weeks 12 to 26 and further worsens by weeks 39 to 78. Non-inferiority is demonstrated during the treatment phase between weeks 0 to 4, 5 to 8 and 12 to 26 but not during the follow up phase (weeks 39 to 78) which is consistent with the primary analysis originally performed for REMoxTB. The confidence intervals from the non-monotone imputation are not as extreme as those when a monotone missingness pattern is imposed, and seem to work better. A plausible reason for this is because a non-monotone pattern better reflects the data than imposing a monotone pattern which removes information. Therefore, the results from the non-monotone imputation show that there is some useful information when a non-monotone pattern is followed which needs to be included within the analysis.

We now apply the methods used here for the REMoxTB study to our second dataset, the RIFAQUIN study. We begin by exploring predictions of outcome failures and withdrawals before proceeding with our GEE models with and without weights. Next, we check the results from the weighted GEE model using multiple imputation. We then use multiple imputation where the data are non-monotone, reflecting the true nature of the data for the RIFAQUIN study.

4.8 Application to the RIFAQUIN study

The methods used for the REMoxTB study (from §4.1 to 4.6.2) are applied to the RIFAQUIN study and visits are kept within visit windows for analysis. Results from all analyses were compared to a 6% non-inferiority margin to assess how the effect of treatment changes over time. Covariates that predict outcome failure and withdrawals are included as weights for inverse probability weighting. Marginal GEE models are

explored as they have more flexibility in their assumptions, imposing a monotone missingness pattern to the data. Weighted GEE models are also used including IPW, so that the population sample is more accurately represented of the whole study. We then confirm the results found in the GEE models using multiple imputation keeping the pattern of missing data monotone before investigating any gains in information with a non-monotone structure.

4.9 Predictions of outcome failure and withdrawals

The following tables show predictions for outcome failure and withdrawals. Unadjusted results are in Appendix F. A backwards stepwise procedure was used and covariates were excluded at the 5% significance level. Covariates included in the model were randomised treatment, x-ray cavitations (0, 1 or 2), sex (male or female), ethnicity (black or other), age (years), HIV (positive or negative), baseline time to positivity, centre (Johannesburg, Cape Town or other), weight band ($\leq$40kg, 40-45kg, $\geq$45-55kg or $\geq$55kg), smoking status (never, past or present), production of sputum sample taken from patients (yes or no) and the Nelson Aalen estimate of time to not producing sputum. Patients were censored at the first instance of not being able to produce a sputum sample.

Treatment, sex, not producing sputum and the Nelson-Aalen estimate of time to not producing sputum were significant predictors of outcome failure (Table 4.11). Patients randomised to the 6 month regimen were about 35% less likely to fail treatment (OR: 0.649; 95% CI: 0.362 to 1.162) in comparison to those randomised to the 4 month regimen who were 58% more likely to fail (Table 4.11). Being female seemed to have a protective effect against outcome failure (i.e. treatment failure or relapse); female patients are half as likely to fail treatment and a similar result was also found for the REMoxTB study. The odds of patients who did not produce sputum at any time during the study decreased by 76% for outcome failure (OR: 0.024; 95% CI: 0.006 to 0.094). Patients who are unable to produce sputum quicker than those who did not led to a reduction of association with outcome failure.

Table 4.11: Final model showing adjusted odds ratios (OR) and confidence intervals (CI) for predicting outcome failure for RIFAQUIN.

Covariate	Adjusted OR	95% CI	P-value
Treatment[1]			
4 month regimen	1.581	(0.938, 2.663)	0.085
6 month regimen	0.649	(0.362, 1.162)	0.146
Sex			
Female	0.490	(0.299, 0.804)	0.005
Sputum produced			
No	0.024	(0.006, 0.094)	<0.001
Time to not producing sputum (years)	7.61×10^{-24}	$(1.26\times10^{-29}, 4.61\times10^{-18})$	<0.001

[1]Likelihood-ratio test for treatment P<0.006.

Table 4.12 shows predictors of withdrawing are centre, X-ray cavities, inability to produce sputum and the Nelson Aalen estimate of time to not producing sputum. Patients who were randomised to other sites (Harare, Marondera, Zambia and Botswana) have a 76% reduction in odds of withdrawals (OR: 0.24; 95% CI 0.096 to 0.61) compared to those from Johannesburg. Patients randomised at Cape Town had a 22% reduction in odds of withdrawing from the study (OR: 0.78; 95% CI: 0.356 to 1.208) in comparison to those from Johannesburg (Table 4.12). Those unable to produce sputum (OR: 0.018) and the less time it took for patients to not be able to produce sputum, the less chance of them withdrawing from the study.

Table 4.12: Final model showing adjusted odds ratios (OR) and confidence intervals (CI) predicting withdrawals for RIFAQUIN.

Covariate	Adjusted OR	95% CI	P-value
Centre[1]			
Cape Town	0.780	(0.356, 1.708)	0.534
Other	0.242	(0.096, 0.609)	0.003

X-ray cavities			
1	0.917	(0.405, 2.073)	0.835
2	2.171	(0.892, 5.289)	0.088
Sputum produced			
No	0.018	(0.003, 0.129)	<0.001
Time to not producing sputum (years)[2]	9.216×10^{-20}	$(5.52\times10^{-28}, 2.100\times10^{-13})$	<0.001

[1]Likelihood-ratio test for centre P< 0.002.

[2]NB: years presented as estimates are small.

Therefore, not being able to produce sputum at any point during the study and the Nelson-Aalen estimate of time to the first occurrence of not being able to produce sputum are both important covariates in the RIFAQUIN study and will be included in our weights for the IPW model. We also include treatment in our weights model as this is an important covariate. Now, we impose a monotone pattern on the RIFAQUIN study before exploring GEE models, with and without weights. We then compare the results from the IPW GEE model to data that does not impose a monotone pattern using multiple imputation.

4.10 Inverse probability weighting for the RIFAQUIN study

Table 4.13 shows the proportion of patients who were classed as a success, failure or missing within each visit window across all treatment arms when a monotone missing pattern is imposed for the RIFAQUIN study. There are slightly more patients missing in the treatment regimens between months 0-3 and months 4-6 which balances out by the final visit window.

Table 4.13: Proportion of patients imposing a monotone missingness pattern in RIFAQUIN.

	Outcome	Control (N=240)	4 month regimen (N=239)	6 month regimen (N=251)
	Success	219 (91.25%)	214 (89.54%)	223 (88.84%)

Months 0-3	Fail	11 (4.58%)	15 (6.28%)	14 (5.58%)
	Missing	10 (4.17%)	16 (6.69%)	14 (5.58%)
Months 4-6	Success	211 (87.92%)	200 (83.68%)	217 (86.45%)
	Fail	4 (1.67%)	4 (1.67%)	0
	Missing	25 (10.42%)	35 (14.64%)	34 (13.55%)
Months 7-10	Success	194 (80.83%)	178 (74.48%)	199 (79.28%)
	Fail	1 (0.42%)	10 (4.18%)	2 (0.80%)
	Missing	45 (18.75%)	51 (21.34%)	50 (19.92%)
Months 11-18	Success	167 (69.58%)	167 (69.87%)	181 (72.11%)
	Fail	4 (1.67%)	3 (1.26%)	1 (0.40%)
	Missing	69 (28.75%)	69 (28.87%)	69 (27.49%)

Patients are assumed to be missing after the first occurrence of a missing observation in a visit window. Table 4.14 shows the proportion of patients with negative culture results within each monotone missing pattern by treatment arm by the end of the study. There are fewer patients who have completed results on the 4 month regimen in comparison to the control and 6 month regimen.

Table 4.14: Monotone missing pattern for patients with negative results in RIFAQUIN, by treatment arm.

Treatment	Months 0-3	Months 4-6	Months 7-10	Months 11-18	Number of patients[1] (n(%))
Control (N=240)	O	O	O	O	171 (71.25%)
	O	O	O	.	24 (10.00%)
	O	O	.	.	20 (8.33%)
	O	.	.	.	15 (6.25%)
	.	.	.	.	10 (4.17%)
4 month regimen (N=239)	O	O	O	O	170 (71.13%)
	O	O	O	.	18 (7.53%)
	O	O	.	.	16 (6.69%)
	O	.	.	.	19 (7.95%)

	.	.	.	.	16 (6.69%)
6 month regimen (N=251)	O	O	O	O	182 (72.51%)
	O	O	O	.	19 (7.57%)
	O	O	.	.	16 (6.37%)
	O	.	.	.	20 (7.97%)
	.	.	.	.	14 (5.58%)

[1]at the end of the study.

Figures 4.9 to 4.11 show the proportion of patients with negative culture results in each visit window for each monotone missing pattern across treatment arm. The proportion of patients who are observed have around 90% of negative culture results after 3-4 months of treatment. Patients on the control regimen who are observed 0 to 3 months and 4 to 6 months and then are subsequently missing have slightly fewer negative culture results in comparison to others observed within that visit window. The proportion of negative results are broadly similar across all visit windows for patients who are observed on treatment arms.

Figure 4.9: Proportion of negative culture results in control regimen imposing a monotone missing pattern for RIFAQUIN.

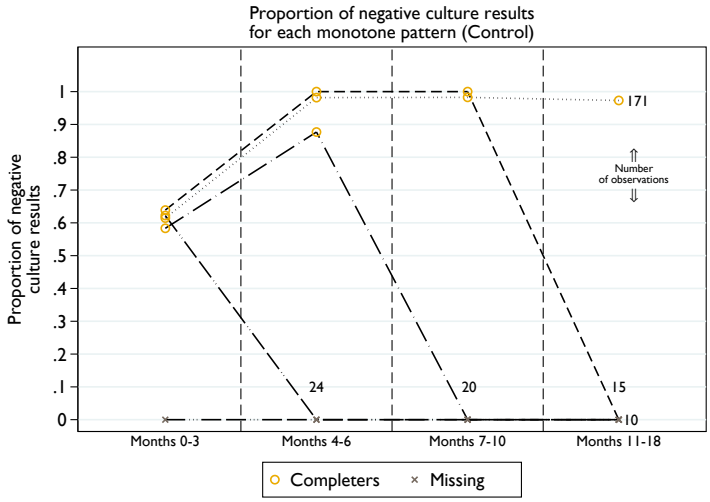


Figure 4.10: Proportion of negative culture results in the 4 month regimen imposing a monotone missing pattern for RIFAQUIN.

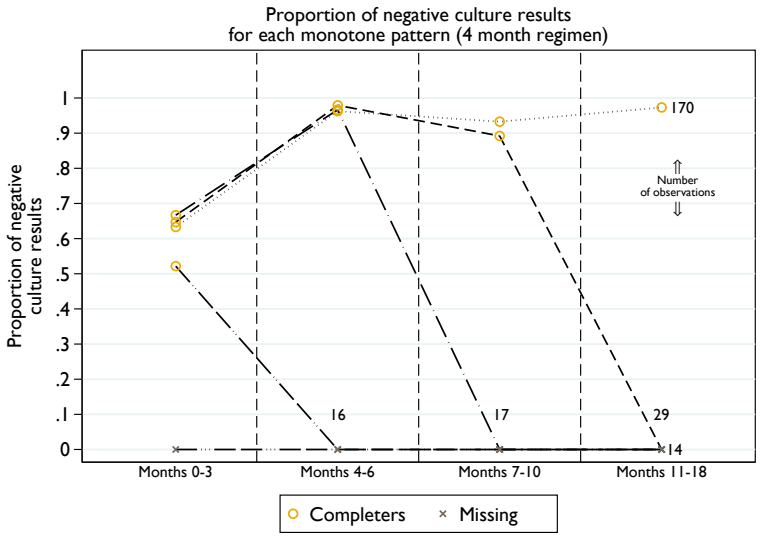
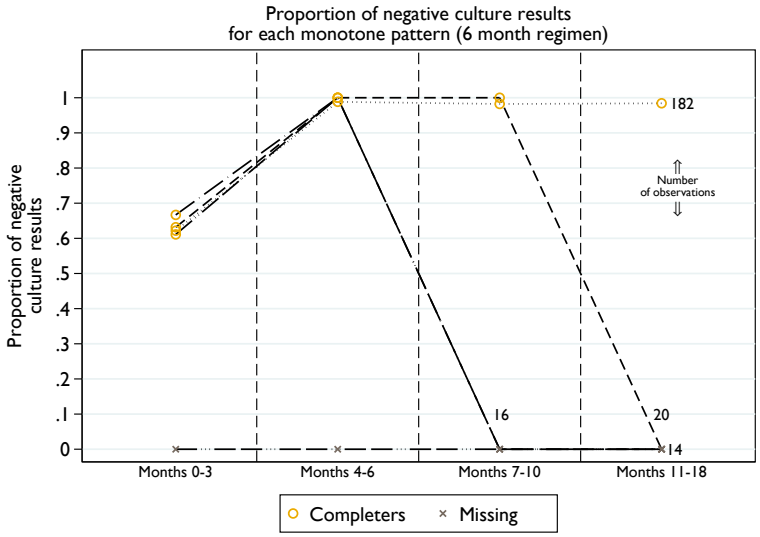


Figure 4.11: Proportion of negative culture results in the 6 month regimen imposing a monotone missing pattern for RIFAQUIN.



Next we explore GEE models, without weights before using IPW to impute the missing observations for patients in the RIFAQUIN study.

4.11 Generalised Estimating Equations applied to the RIFAQUIN study

The results from the Generalised Estimating Equations (GEEs) for the RIFAQUIN study are presented. We aim for an intention-to-treat type analysis, excluding patients for reasons not related to treatment (Table 3.7). First, a GEE model for patients observed within each visit window (i.e. "completers") was explored, comparing the upper bound of each confidence interval to a 6% margin. Then we investigate a GEE model including patients with at least one observation in a visit window. GEEs are used as an alternative method to multiple imputation and are used to estimate the population-average effect of randomised treatment to the study.

Table 4.15 shows the results from the GEE model for patients who had an outcome in each visit window (i.e. "completers"). A difference in proportions between treatment (6m regimen or 4m regimen) and control for treatment failure (i.e. patients who did not reach negative culture status) are presented from the GEE model for each visit window and an unstructured variance-covariance matrix was assumed. Since a monotone missingness pattern was imposed on the data, patients missing an observation during the first visit window were subsequently missing at future follow up visits and therefore were excluded from the analysis. This analysis includes 523/690 patients (76%) who had observations across all visit windows. Non-inferiority is demonstrated on the 4 month regimens between months 4-6 (1.2%: 95% CI; -1.6% to 4.0%) and all patients on the 6 month regimen were successful (hence no estimates are available). In the third visit window, non-inferiority is shown on the 6 month regimen between months 7-10, since the upper bound of the 95% confidence interval is below the 6% non-inferiority margin (5%; 95% CI: -1.4% to 2.4%), but borderline on the 4 month regimen; 3.5% (95% CI: 0.3% to 6.7%). In the final visit window, non-inferiority is demonstrated for both regimens (upper bound of the 95% CI: 2.4% for the 4 month regimen and 0.7% for the 6 month regimen). It is likely that patients who were failing would have been withdrawn from the study. As a result, patients who remain in follow-up are likely to be performing well on treatment and this is what is seen in the final visit window.

Table 4.15: GEE model for a difference in proportions of treatment failure for "completers" assuming an unstructured variance-covariance matrix, by treatment arm for RIFAQUIN.

	Risk difference	SE	95% CI
Months 0-3			
4 month regimen	0.018	0.022	(-0.026, 0.061)
6 month regimen	0.025	0.023	(-0.019, 0.070)
Months 4-6			
4 month regimen	0.012	0.014	(-0.016, 0.040)
6 month regimen[1]	-	-	-
Months 7-10			
4 month regimen	0.035	0.016	(0.003, 0.067)
6 month regimen	0.005	0.010	(-0.014, 0.024)
Months 11-18			
4 month regimen	-0.006	0.015	(-0.036, 0.024)
6 month regimen	-0.018	0.013	(-0.043, 0.007)

[1] All patients were "successful".

Table 4.16 shows the risk difference of the 4 month and 6 month regimens from the control arm for failure at each visit window from the GEE model assuming an unstructured variance-covariance structure for all 690 patients. The results from this analysis are broadly consistent with results from the complete case analysis. However, for the 6 month regimen non-inferiority can be concluded during the first visit window between 0-3 months, since the upper bound of the 95% confidence interval lies below the 6% non-inferiority margin (1.1%; 95% CI: -3.0% to 5.2%). In the third visit window between 7 and 10 months, the 4 month regimen shows non-inferiority is less likely to be demonstrated (upper bound of 95% CI: 8.2%) compared to patients who are observed across all visits are included (Table 4.15; upper bound of 95% CI: 6.7%). This suggests that by excluding patients who are not observed at all visits, additional information in relation to the primary outcome is also lost.

Table 4.16: Difference in proportions of treatment failure including all patients in the analysis assuming unstructured variance-covariance matrix, by treatment arm for RIFAQUIN.

	Risk difference	SE	95% CI
Months 0-3			
4 month regimen	0.019	0.022	(-0.024, 0.062)
6 month regimen	0.011	0.021	(-0.030, 0.052)
Months 4-6			
4 month regimen	0.001	0.013	(-0.025, 0.027)
6 month regimen[1]	-	-	-
Months 7-10			
4 month regimen	0.048	0.017	(0.014, 0.082)
6 month regimen	0.005	0.009	(-0.012, 0.022)
Months 11-18			
4 month regimen	-0.006	0.015	(-0.036, 0.024)
6 month regimen	-0.018	0.013	(-0.043, 0.007)

[1] All patients were "successful".

4.12 Discussion

In this section, we investigated using GEE models imposing a monotone missingness pattern to the RIFAQUIN data without including weights in the model. The results from the GEE model which includes patients observed across all visit windows and the results from the GEE model that includes patients with at least one observation in a visit window during follow-up were similar. The results from the analysis which included patients with at least one observed outcome across visit windows showed stronger suggestions of not being able to conclude non-inferiority. This was shown by the upper bounds of the 95% confidence intervals which were larger compared to those from the completers' analysis.

Next, we look at including inverse probability weights in the GEE model to account for the missing data.

4.13 Weighted Generalised Estimating Equations applied to the RIFAQUIN study

Inverse probability weights are included in our GEE model to better represent the patients randomised to the RIFAQUIN study. Again, patients are excluded for reasons not related to treatment: late screening failure, drug resistance and no positive culture results in the first 2 weeks of randomisation (Table 3.7). To obtain weights, treatment, inability to produce a sputum culture result and the Nelson-Aalen estimate of time to not producing sputum (see §4.9) were included.

Figures 4.12 to 4.14 summarises the probability weights calculated by treatment arm and by outcome (success/failure) for each visit window. The probability weights for patients who are classed as successes are consistent at around 0.9. The probability weights for patients who are classed as failures are weighted lower at around 0.6 since there are more patients who are successful in each visit window. For patients randomised to the 6 month treatment regimen who were included in this analysis were successful, and therefore no weights are calculated at months 7 to 10 given months 0 to 3 and months 4 to 6.

Figure 4.12: Histogram of probability weights at months 4 to 6 given months 0 to 3 for RIFAQUIN.

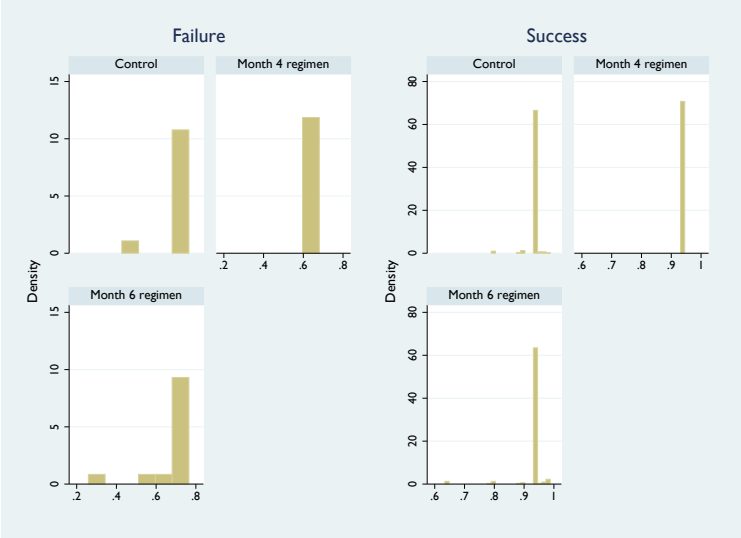


Figure 4.13: Histogram of probability weights at months 7 to 10 given months 0 to 3 and months 4 to 6 for RIFAQUIN.

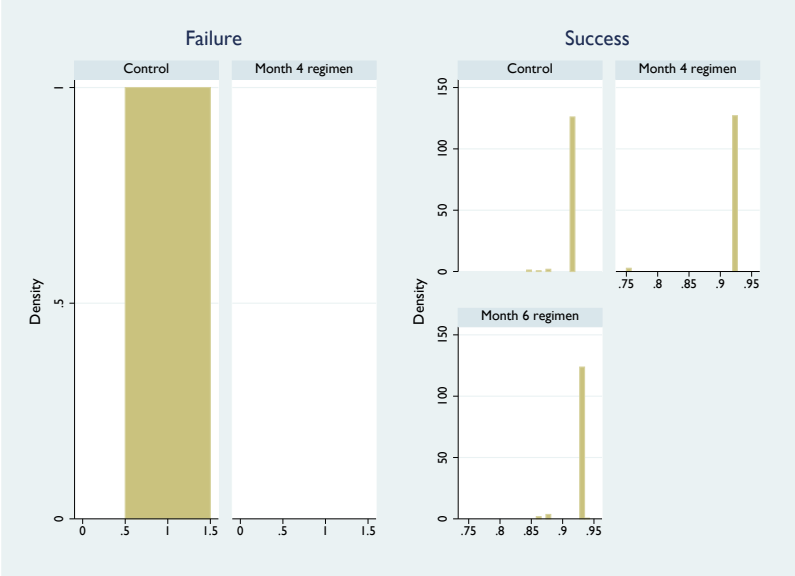


Figure 4.14: Histogram of probability weights at completion (months 11 to 18) given months 0 to 3, months 4 to 6 and months 7 to 10 for RIFAQUIN.

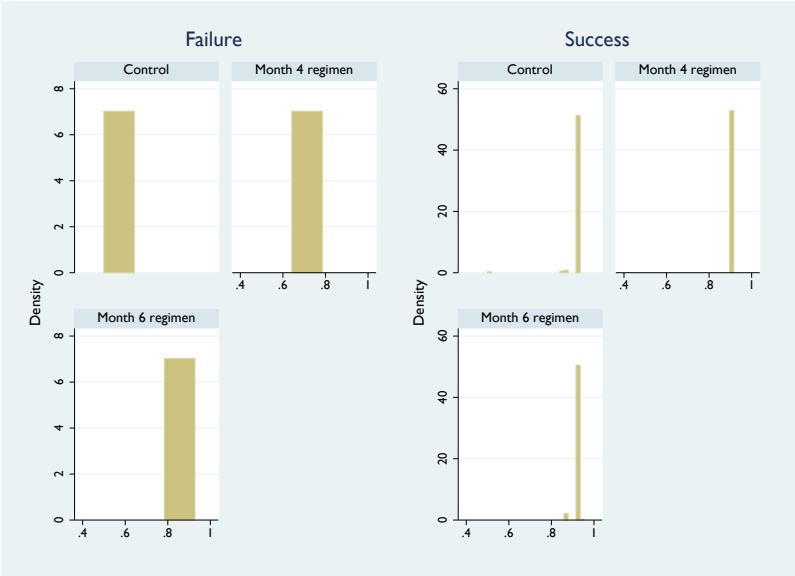


Table 4.17: GEE model for a difference in proportions of treatment failure using estimated weights from data observed assuming an independent variance-covariance matrix, by treatment arm for RIFAQUIN.

	Risk difference	SE	95% CI
Months 0-3			
4 month regimen	0.019	0.022	(-0.024, 0.062)
6 month regimen	0.011	0.021	(-0.030, 0.052)
Months 4-6			
4 month regimen	0.001	0.016	(-0.029, 0.032)
6 month regimen[1]	-	-	-
Months 7-10			
4 month regimen	0.046	0.020	(0.006, 0.086)
6 month regimen	0.0004	0.011	(-0.022, 0.023)
Months 11-18			
4 month regimen	-0.025	0.023	(-0.070, 0.021)
6 month regimen	-0.033	0.022	(-0.076, 0.010)

[1] All patients were "successful".

Table 4.17 shows the results from the GEE model including observed weights that were calculated using predicted probabilities from the logistic regression model. As for the REMoxTB study (see §4.5), an independent structure sandwich estimator of variance is included as it is assumed observations at each time point were measured independently. The results from this model are similar to the GEE model that did not include any weights (Table 4.16) and are consistent with the GEE model that only included patients who were observed at all visit windows (Table 4.15).

4.14 Discussion

Having investigated using GEE models for the RIFAQUIN study, there appears to be little difference between results when we restrict the analysis to include patients who are observed at all visits compared to including all patients who are observed at least

once throughout the study. However, by only including patients who are observed at each visit window throughout the study has the potential to lead to different conclusions as the upper bounds of the 95% confidence intervals were smaller than when all patients were included in the analysis. Therefore, excluding patients from the GEE analysis is not recommended. Sputum production and the Nelson-Aalen estimate for time to not producing sputum were important predictors for both outcome failure and withdrawals. These covariates were used to calculate observed weights separately for each treatment to include in the weighted GEE model. The results from this model were broadly consistent with the GEE models that did not include weights when considering whether or not non-inferiority could be concluded in each visit window. In all three analyses, no results were available between months 4 to 6 for the 6 month regimen since all patients in this visit window were successful.

We now use multiple imputation where a monotone missingness pattern is imposed for the RIFAQUIN study. The impact of this assumption is investigated, reversing this restriction, by investigating a non-monotone missingness pattern.

4.15 Multiple imputation for monotonic and non-monotonic missing patterns

To verify the results from the GEE models used for the RIFAQUIN study, multiple imputation is used. First, the data is kept in a monotone pattern to directly compare the results to the GEE models and then we return to the true non-monotone pattern of the RIFAQUIN study to explore whether there is any gain in information. For both missing data patterns, visits are kept in visit windows.

4.15.1 Monotonic missing data pattern

Multiple imputation was used by restricting the data to follow a monotone missingness pattern to check on the accuracy of results produced from the weighted GEE model.

Missing data were imputed by treatment arm using a logistic regression model, using 100 imputations. Given that all patients randomised to the 6 month regimen were

successful between months 4 to 6, patients who were missing in this group were assumed to be successful prior to proceeding to the imputation. To obtain a difference in proportions for treatment failure between the trial treatment regimens and the control, a generalised linear model was used to estimate the treatment effects assuming an identity link function.

Figure 4.15: Imputed results of mean positive cultures where a monotone pattern is imposed for RIFAQUIN.

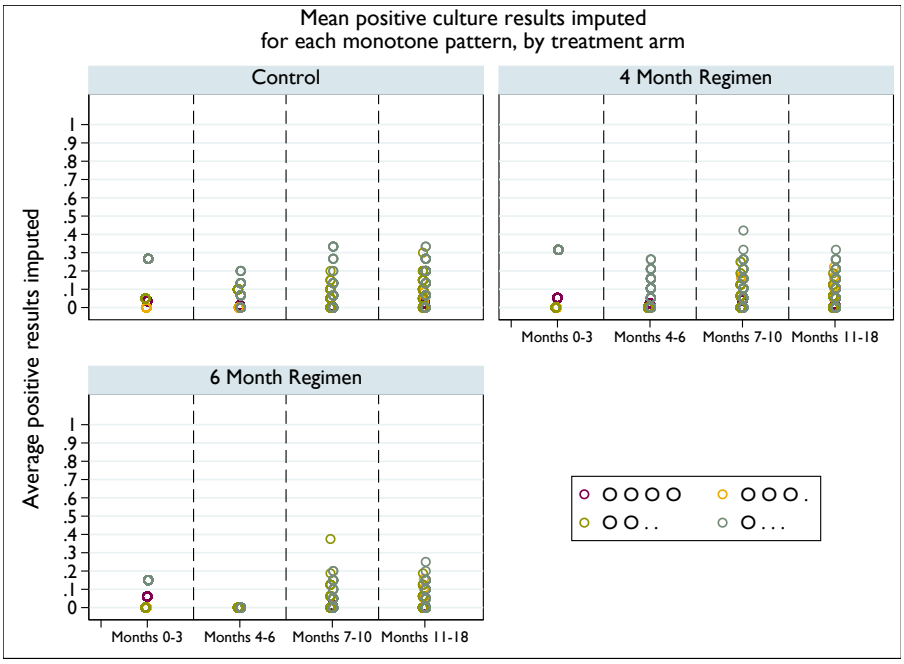


Figure 4.15 shows imputed results of patients who failed for each monotone missingness pattern. The average number positive culture results imputed were calculated by treatment arm and each non-monotone missingness pattern for each of the 100 imputed datasets created. The imputed values are close to values that were observed and shows the imputation has been performed accurately as the imputations show there are no outliers. Table 4.18 shows that the results following imputation are identical to the GEE models with and without weighting (see §4.11 to §4.13). The model explored here confirms the results of the weighted GEE model.

Table 4.18: Difference in proportions for treatment failure following multiple imputation, by treatment arm for RIFAQUIN.

	Risk differences	SE	95% CI
Months 0-3			
4 month regimen	0.019	0.022	(-0.023, 0.062)
6 month regimen	0.011	0.021	(-0.030, 0.052)
Months 4-6			
4 month regimen	0.004	0.016	(-0.027, 0.035)
6 month regimen[1]	-	-	-
Months 7-10			
4 month regimen	0.045	0.021	(0.004, 0.086)
6 month regimen	0.0001	0.014	(-0.028, 0.028)
Months 11-18			
4 month regimen	-0.011	0.020	(-0.049, 0.028)
6 month regimen	-0.025	0.017	(-0.059, 0.009)

[1] All patients were "successful".

4.15.2 Non-monotonic missing data pattern

A non-monotone missingness pattern is used to see whether there is any additional information to be gained from when a monotone missing pattern is imposed. Table 4.19 shows the proportion of patients in each treatment arm who were classed as successful, failures and who were missing across all visit windows. The proportion of patients classed as a success, failure or missing are similar to when a monotone missingness pattern is imposed between months 0 to 3 and months 4 to 6, but there are fewer patients missing between months 7 to 10 and months 11 to 18 suggesting there may be some information lost when data are restricted to a monotone missingness pattern though this may not impact on conclusions made.

Table 4.19: Proportion of patients with a non-monotone missingness pattern imposed for RIFAQUIN.

	Outcome	Control (N=240)	4 Month regimen (N=239)	6 Month regimen (N=251)
Months 0-3	Success	219 (91.25%)	208 (87.03%)	223 (88.84%)
	Fail	11 (4.58%)	15 (6.28%)	14 (5.58%)
	Missing	10 (4.17%)	16 (6.69%)	14 (5.58%)
Months 4-6	Success	212 (88.33%)	202 (84.52%)	220 (87.65%)
	Fail	4 (1.67%)	4 (1.67%)	0
	Missing	24 (10.00%)	33 (13.81%)	31 (12.35%)
Months 7-10	Success	200 (83.33%)	186 (77.82%)	209 (83.27%)
	Fail	1 (0.42%)	10 (4.18%)	2 (0.80%)
	Missing	39 (16.25%)	43 (17.99%)	40 (15.94%)
Months 11-18	Success	177 (73.75%)	177 (74.06%)	198 (78.88%)
	Fail	4 (1.67%)	3 (1.26%)	1 (0.40%)
	Missing	59 (24.58%)	59 (24.69%)	52 (20.72%)

For data that follow a non-monotone missingness pattern, missing data were imputed by treatment arm using a logistic regression model using 100 imputations. Again, patients who received the 6 month regimen were all successful between months 4 to 6 and therefore anyone missing within this visit window who received the 6 month regimen were assumed to be successful. Figure 4.16 shows the imputed results of positive culture results for each monotone missing pattern from the imputation model. The average number positive culture results imputed were calculated by treatment arm and each non-monotone missingness pattern for each of the 100 imputed datasets created. Figure 4.16 shows that the imputation model has imputed sensible results in each treatment arm for each missingness pattern since the imputed values are closely fitted to the observed values, and there are no outliers.

Risk differences were calculated using a logistic regression model assuming an identity link function. The results from the imputation (Table 4.20) are nearly identical to the results where a monotone missingness pattern is imposed (Table 4.18).

Figure 4.16: Imputed results for the mean rate of positive culture results following principal patterns of non-monotone data for RIFAQUIN.

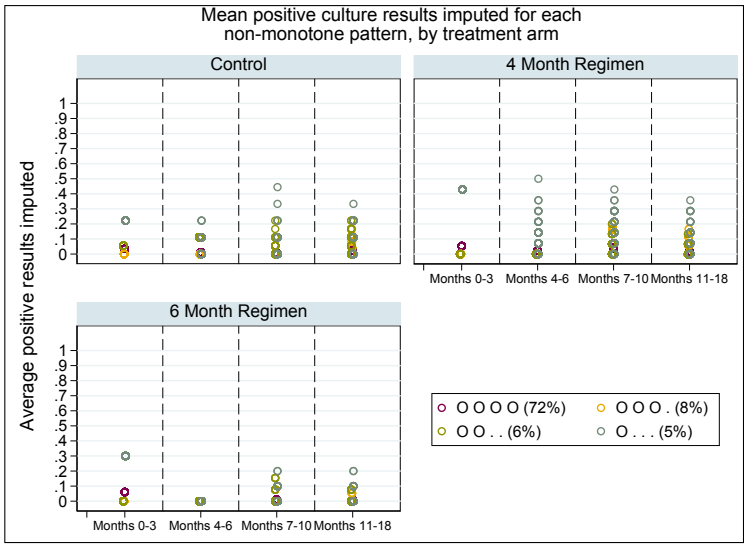


Table 4.20: Difference in proportions for treatment failure following multiple imputation where the pattern is non-monotone, by treatment arm for RIFAQUIN.

	Risk differences	SE	95% CI
Months 0-3			
4 month regimen	0.019	0.022	(-0.024, 0.062)
6 month regimen	0.010	0.021	(-0.030, 0.051)
Months 4-6			
4 month regimen	0.005	0.015	(-0.025, 0.035)
6 month regimen[1]	-	-	-
Months 7-10			
4 month regimen	0.048	0.018	(0.012, 0.084)
6 month regimen	0.002	0.011	(-0.019, 0.024)
Months 11-18			
4 month regimen	-0.006	0.018	(-0.040, 0.029)
6 month regimen	-0.022	0.014	(-0.050, 0.005)

[1] All patients were "successful".

4.16 Discussion

In this section, we investigated using GEE models to impute the missing observations of trial participants' outcome data imposing a monotone missingness pattern to the RIFAQUIN data.

The results from the different analyses performed here were broadly similar to each other. This could be due to having less frequent visits over the study period since patients were followed up monthly with two 3 month visits towards the end of follow up. As a consequence the frequency for patients to switch into different states of having negative and positive culture results is reduced.

All analyses showed that the treatment arms performed worse than the control between months 4 to 6 and months 7 to 10 and performed better than the control in the final visit window between months 11 to 18. This final visit window consists of visits 1, 2, 5 and 8 months after the last visit in the third visit window (10 month after randomisation) and therefore the final window consists of a long time where patients are not seen as the final two follow up visits were 3 months apart after the last (15 and 18 months). During this final visit window, patients who were failing or were not getting better on treatment would have been withdrawn from the study and would have switched treatment or have had their treatment modified. Due to this, it is expected for both treatment regimens to perform better than the control by the end of the study within the final visit window since patients who are culture converting to negative status are more likely to stay enrolled within the study, and this was reflected in our analyses. Non-inferiority could be concluded on the 4 month regimen between months 4 to 6 and all patients were successful on the 6 month regimen whereas between months 7 to 10, non-inferiority could not be concluded on the 4 month regimen but could be concluded on the 6 month regimen. These observations are consistent with those reported in the RIFAQUIN study, where non-inferiority could not be concluded on the 4 month regimen (13.60%, 95% CI: 7.0% to 20.20%) and non-inferiority was concluded on the 6 month regimen (1.80%, 95% CI: -6.90% to 3.30%) for the PP analysis.

The results from using multiple imputation, imposing and not imposing a monotone missingness pattern to the data produced similar results. Although this is reassuring, a non-monotone missingness pattern is preferred thus reflecting the true nature of the data collected.

So far in this chapter, a binary outcome was imposed within visit windows. Instead of making assumptions about the nature of the data within these visit windows, the number of negative culture results within each visit window could be counted and a Poisson regression model could be used to assess treatment differences of outcome failure.

Next, a Poisson regression model is explored for the REMoxTB and RIFAQUIN studies to assess the rate of negative culture results over time.

4.17 Poisson regression

Poisson regression models are useful for analysing the number of times an event occurs over time. Estimating the rate of negative culture results over each of the four visit windows is now considered.

In the REMoxTB and RIFAQUIN studies, patients were not always observed within each visit window, for example some patients may have been observed at all scheduled visits within a visit window whereas others may have been observed only once. It is therefore more relevant to consider the rate of negative cultures by calculating the number of negative culture results for each observed result within each visit window thus accounting for patients' varying time within each visit window. This is done by including an offset, which constrains the total number of observations for patients to 1, within the following mixed-effects Poisson regression model[88] for patient k having negative results in the t^{th} window:

$$ln\left(\frac{u_{0,k}}{u_{1,k}}\right) = \beta_{0,k,t} + \beta_{1,k,t}time_{k,t} + \beta_2 trt_k + \beta_3(time_{k,t} * trt_k)$$

$$\beta_{0,k,t} = \beta_0 + u_{0,t} + e_{0,k,t}$$

$$\beta_{1,k,t} = \beta_1 + u_{1,t}$$

$$\begin{pmatrix} u_0 \\ u_1 \end{pmatrix} \sim N(0, \Omega_u) : \Omega_u \begin{pmatrix} \sigma^2_{u_0} & \sigma_{u_{0,1}} \\ \sigma_{u_{0,1}} & \sigma^2_{u_1} \end{pmatrix}$$

$e_{0,k,t} \sim N(0, \Omega_e) : (0, \sigma^2_{0,e})$,

where, $u_{0,k}$ is the total number of negative culture results for each patient k at visit window t for t=1, 2, 3 or 4 and $u_{1,k}$ is the total number of observations for each patient k at each visit window t. To take into account that patients have repeated observations of negative culture results over time, we include random effects parameters $u_{0,k}$ to represent the departure of the k^{th} patient from the overall rate intercept, $u_{1,k}$ represents the overall departure of the k^{th} patient from time and $e_{0,k,t}$ represents random variability within time for the REMoxTB and RIFAQUIN studies.

We next explore using a mixed-effects Poisson regression model to the REMoxTB and RIFAQUIN studies.

4.17.1 Application to REMoxTB

First, we investigate using a mixed-effects Poisson regression model to the REMoxTB study before looking at the mean rate of negative cultures within each visit window over the whole study. The mean rate of negative culture results are taken rather than the mean rate of positive culture results due to the presence of fewer positive culture results (and therefore fewer failures overall) during the second half of follow up. This will avoid any issues of numerical underflow in the computations. Table 4.21 shows the rate ratios obtained from the multilevel mixed-effects Poisson regression model and Table 4.22 shows the random effects parameters.

Table 4.21: Multilevel mixed-effects Poisson regression for REMoxTB, by treatment arm.

	Rate ratio	SE	97.5% CI
Week 0 to 4			
Intercept	0.002	0.0002	(0.0016, 0.002)
Isoniazid	0.815	0.111	(0.600, 1.106)

Ethambutol	0.837	0.114	(0.616, 1.136)
Week 5 to 8			
Intercept	0.017	0.001	(0.015, 0.019)
Isoniazid	0.993	0.075	(0.839, 1.175)
Ethambutol	0.975	0.074	(0.823, 1.155)
Week 12 to 26			
Intercept	0.026	0.001	(0.024, 0.029)
Isoniazid	0.961	0.061	(0.835, 1.108)
Ethambutol	0.945	0.060	(0.820, 1.089)
Week 39 to 78			
Intercept	0.023	0.001	(0.021, 0.026)
Isoniazid	0.944	0.062	(0.815, 1.093)
Ethambutol	0.951	0.062	(0.821, 1.101)

Table 4.22: Random effects parameters from the multilevel mixed-effects Poisson regression for REMoxTB.

Parameters	Variance	SE
Visit window (years)	6.448 $\times 10^{-16}$	2.210 $\times 10^{-10}$
Intercept	2.163 $\times 10^{-13}$	2.101 $\times 10^{-7}$
Cov(year, intercept)	-1.170 $\times 10^{-14}$	7.332 $\times 10^{-9}$

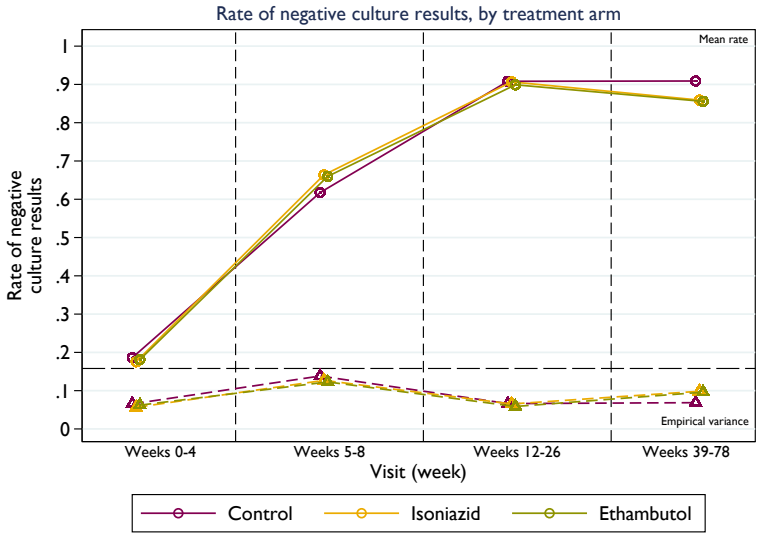
The rate of negative culture results increases up until 12-26 weeks and then decreases on the isoniazid and ethambutol treatment regimens in the next visit window between weeks 39 to 78, but remains constant on the control arm. Although the estimates are large, the results are broadly consistent with the results from the GEE models that imposed a monotone pattern (§4.4 to §4.6.1) and with results from using multiple imputation, where patients were more likely to be classed as failures over time for non-monotonic data (§4.6.2). Patients who are in the study for longer and have not culture converted in the later stage of the treatment phase would have been withdrawn from the study, and therefore their results would no longer be included in the analysis. This group of patients may be contributing to the decrease in the rate of negative culture results during the follow-up phase. Another group of patients who

might influence this decrease are patients who were randomised to one of the treatment arms but switched to the control arm because the treatment was failing for them. Given patients were analysed according to the treatment they were randomised on, the true effect of treatment for patients is masked.

Table 4.22 suggests underdispersion since there is less patient variation within the data than predicted. The patient specific random effect of the estimated variance component is very small at 2.163 $\times 10^{-13}$, which shows that treatment differs between patients.

The mean rate of negative culture results (Figure 4.17) is slightly less on the control arm between weeks 5-8, than for the treatment arms, but this is reversed by the time patients reach the follow-up phase between weeks 39-78 where there are more negative culture results for patients who received the control regimen than for those randomised to one of the treatment arms. The variance is at its highest between weeks 5 to 8 reflecting the patients who fluctuated between a positive and negative result within that visit window.

Figure 4.17: Mean rate and empirical variance of negative culture results for REMoxTB.



The results from using a mixed effects regression model are broadly consistent with the results found from the GEE models; the rate of having negative cultures increases between week 5 to 8 before decreasing over the next visit windows from weeks 12 to 26 weeks and 39 to 78 weeks. The Poisson regression models suggests underdispersion within the data. The underdispersion mirrors the trend of patients who are mostly positive at the start of the study which then switches to patients who are mostly negative towards the end of the study.

We now explore using Poisson regression for the RIFAQUIN study.

4.17.2 Application to the RIFAQUIN study

A multilevel mixed effects Poisson regression model is investigated for the RIFAQUIN study. Again, we account for the number of negative cultures present within each visit window rather than positive cultures to avoid any complications in the computation. Then we look at the mean rate of negative culture results over time. Table 4.23 shows the rate ratios obtained from the multilevel mixed-effects regression model and Table 4.24 shows the random effects parameters.

Table 4.23: Multilevel mixed-effects Poisson regression, by treatment arm for RIFAQUIN.

	Rate ratio	SE	95% CI
Months 0-3			
Intercept	0.032	0.003	(0.027, 0.038)
4 month regimen	1.012	0.123	(0.798, 1.285)
6 month regimen	1.015	0.122	(0.803, 1.284)
Months 4-6			
Intercept	0.056	0.004	(0.049, 0.064)
4 month regimen	1.036	0.099	(0.858, 1.250)
6 month regimen	1.043	0.098	(0.867, 1.254)
Month 7-10			
Intercept	0.024	0.002	(0.021, 0.027)
4 month regimen	0.896	0.091	(0.734, 1.094)

6 month regimen	1.000	0.097	(0.826, 1.211)
Month 11-18			
Intercept	0.029	0.002	(0.025, 0.033)
4 month regimen	1.008	0.103	(0.826, 1.231)
6 month regimen	0.959	0.095	(0.790, 1.163)

Table 4.24: Random effects parameters from the multilevel mixed-effects Poisson regression for RIFAQUIN.

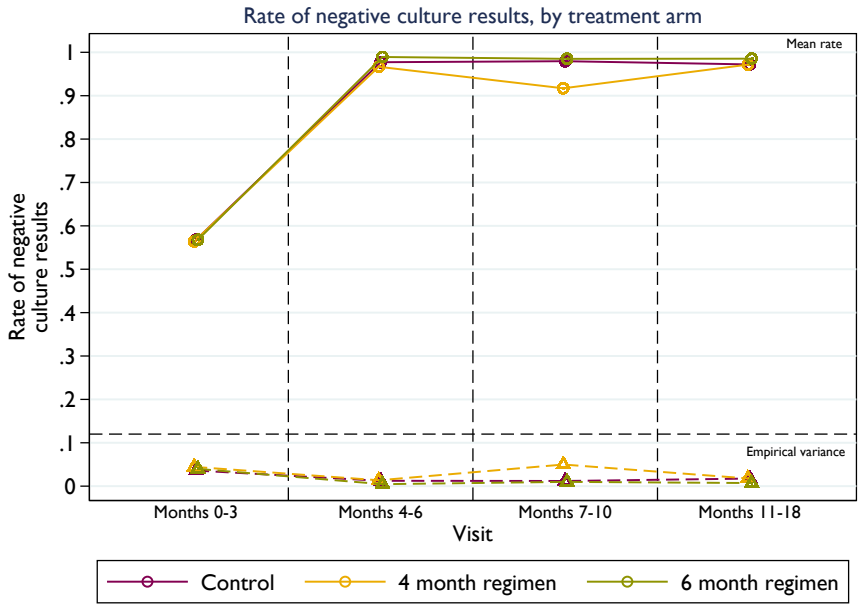
Parameters	Variance	SE
Visit window (years)	1.224×10^{-10}	1.740×10^{-6}
Intercept	1.272×10^{-9}	1.476×10^{-5}
Cov(year, intercept)	-3.960×10^{-10}	4.992×10^{-6}

The results in Table 4.23 show how small the resulting standard errors are. This suggests the data from the RIFAQUIN study are underdispersed. The random intercept of 1.272×10^{-9} (Table 4.24) shows that treatment regimens differs between patients.

The rate of negative culture results increases in the second visit window (month 4 to 6) and then decreases in the next visit window between month 7 to 10. In the final visit window, the rate of culture negative results increases again. By this point, patients who were failing to culture convert to having stable negative culture results would have switched treatment so that they are cured of TB. The rate of negative culture results increases in the second visit window (month 4 to 6) and then decreases in the next visit window between month 7 to 10. In the final visit window, the rate of negative culture results decreases slightly on the 6 month regimen in the final visit window. The rate of culture negative results increases again in the final 11 to 18 month window for patients who received the 4 month regimen. By this point, patients who were failing to culture convert to having stable negative culture results would have switched treatment so that they are cured of TB. These results are similar to those produced from the GEE models (§4.11 to §4.15.1) and to the multiple imputation model where data followed a non-monotone missingness pattern (§4.15.2).

Figure 4.18 shows the mean rate of negative culture results across visit windows. The four month regimen shows a departure from the control and 6 month regimens after 4 to 6 months, where the mean rate of negative culture results is less before increasing again by the end of the study. The rate of negative culture results is similar between the control arm and the 6 month regimen. These two observations correspond with the results of the trial where non-inferiority was demonstrated for the 6 month regimen on the PP analysis and the 4 month regimen failed to demonstrate non-inferiority.

Figure 4.18: Mean rate of negative culture results and empirical variance for RIFAQUIN.



To further investigate the large estimates produced by the Poisson model, Table 4.25 and Table 4.26 compares the proportion of positive results with and without applying our defined outcome for the REMoxTB and RIFAQUIN studies, in accordance with the trial protocol, where two consecutive negative culture results infer "success" (see §4.1). This is because the outcome defined requires two consecutive culture results to determine success. The Poisson model simply counts the number of positive or

negative culture results within each visit window without considering the location of the results, that is that the results must occur consecutively. We investigate whether this has any impact on the REMoxTB and RIFAQUIN studies by calculating the mean positive culture results within each visit window, disregarding the location of where positive or negative culture results occur and compare this to the mean treatment failure which does take into account that negative results occurring consecutively, to determine negative culture conversion.

Table 4.25: Comparison of mean positive culture results and treatment failure for the REMoxTB study.

Week	Treatment	Mean positive culture results	Mean treatment failure
Weeks 0-4	Control	0.805	0.806
	Isoniazid	0.815	0.807
	Ethambutol	0.801	0.789
Weeks 5-8	Control	0.344	0.316
	Isoniazid	0.308	0.284
	Ethambutol	0.304	0.262
Weeks 12-26	Control	0.039	0.035
	Isoniazid	0.043	0.038
	Ethambutol	0.067	0.051
Weeks 39-78	Control	0.040	0.032
	Isoniazid	0.088	0.068
	Ethambutol	0.086	0.059

Table 4.26: Comparison of mean positive culture results and treatment failure for the RIFAQUIN study.

Month	Treatment	Mean positive culture results	Mean treatment failure
Month 0-3	Control	0.432	0.048
	4 month regimen	0.436	0.067

	6 month regimen	0.432	0.059
Month 4-6	Control	0.023	0.019
	4 month regimen	0.033	0.019
	6 month regimen	0.011	0
Month 7-10	Control	0.020	0.005
	4 month regimen	0.083	0.051
	6 month regimen	0.015	0.009
Month 11-18	Control	0.028	0.022
	4 month regimen	0.028	0.017
	6 month regimen	0.015	0.005

The results presented in Tables 4.25 and 4.26 show that applying the protocol rule and classing patients as a success if two consecutive negative culture results occur at separate visits within the visit windows makes the probabilities smaller for both studies. Consecutive positive/negative results are not accounted for when using a Poisson regression model; only the number of positive (or negative) results available which occur in any particular order are within each visit window. It is for this reason that the resulting estimates from the Poisson regression model are larger than expected.

4.18 Discussion

In TB trials, patients are randomised if they have TB present within their lungs and therefore observed positive culture results are seen at the beginning of a study. When treatment is administered, patients produce negative sputum results and this stabilises towards the end of the study. This pattern was reflected in the mixed effects Poisson regression model which showed underdispersion. In both studies, patients randomised to the treatment arms seemed to improve within the first 4-6 months before worsening again. By the time of the final follow-up visit, patients seemed to have a higher rate of negative culture conversion. Patients who are in the study for longer and have not culture converted in the later stage of the treatment phase would have been withdrawn from the study, and therefore their results would no longer be included in the analysis. This group of patients may be contributing to this decrease in

the rate of negative culture results during the follow-up phase. Another group of patients who might influence this decrease are patients who were randomised to one of the treatment arms but switched to the control arm because the treatment was failing for them. Given patients were analysed according to the treatment they were randomised to, the true effect of treatment for patients is masked. For these analyses, the Poisson model inflated the resulting estimates by not applying the protocol rule where two negative results are required to determine stable negative culture conversion.

4.19 Summary

This chapter has investigated a simpler and alternative approach to impute the observations under an intention-to-treat analysis for patients whose outcome data were missing from TB studies, while looking at the proportion of patients who were classed as failures in more detail. The data were kept within clinically meaningful visit windows to create more stable weights rather than using weights for each follow-up visit. This meant we needed to impose an additional rule to class patients as a "success" or "failure" within each visit window. Doing so enabled us to look at whether the proportion of treatment failure changed the conclusions when determining non-inferiority over time using a 6% margin. We explored two valid analyses under the MAR assumption for the REMoxTB data using GEE models and multiple imputation. We have shown that by grouping visits into visit windows and imposing a monotone missingness pattern and using IPW with GEE models works well as verified by the multiple imputation where the pattern remained non-monotone for both the REMoxTB and RIFAQUIN studies. However, using IPW with GEE models in a longitudinal setting assumes the observations are independent which may not be a valid assumption if there is strong correlation within patient observations. In using the IPW methodology requires the data to follow a monotone missingness pattern, which was imposed for both the REMoxTB and RIFAQUIN datasets. Imposing a monotone missing pattern to the data produced similar results when reverting back to a non-monotone pattern and using multiple imputation, however the upper bound of the confidence intervals were slightly higher when a monotone pattern was imposed. This suggests that imposing a monotone pattern

results in deficiency in the estimates and confidence intervals. This is most likely because the data itself follows a non-monotone pattern, therefore by imposing a monotone pattern to this data removes information that a non-monotone pattern retains. Therefore, if this approach is taken by grouping visits into visit windows, multiple imputation should be used keeping the data in a non-monotone missingness pattern reflecting the true structure of the trial data.

To remove the extra rule imposed within each visit window, i.e. classing patients as a "success" or "failure" in each window, an alternative analysis was explored by counting the number of negative culture results within each window, using a Poisson regression model. The results from the Poisson model showed consistent results with the GEE analyses and was consistent with the analysis from multiple imputation where a non-monotone missing pattern was investigated. The Poisson regression model does not take into account that most patients have TB in the first few weeks of the study. At this time point, patients are less likely to produce negative culture results, towards the end of the study most patients do have negative culture results. This results in severe underdispersion of the data. By counting the number of negative cultures within each visit window, we lose the protocol defined outcome where patients are considered to reach negative stable culture conversion if they achieve two consecutive negative culture results at separate visits. As a descriptive analysis, this provides a nice indication of the trial as a whole. However as a formal analysis, we fail to capture the full story of what happens with these patients, and so the Poisson model is not recommended for these analysis.

Although performing multiple imputation where the missingness pattern is non-monotone worked well, more assumptions about the classification of the primary outcome are required, thus adding an extra layer of rules that may be unnecessary.

In the next chapter, we explore using multi-state models as an alternative analysis to multiple imputation, including all patient outcomes in the intention-to-treat analysis without pre-determination of false positive or false negative sputum culture results before performing the protocol defined primary analysis.

Chapter 5

Multi-state models

So far, we have looked at imputation methods and using weighted marginal models to estimate treatment effects when some trial participants' outcome data are missing. In the longitudinal data from the studies that motivate this work, there are long constant sequences of binary 1's at the start of follow-up where the majority of patients are positive and 0's towards the end of follow-up where most patients reach stable negative culture conversion. These sequences are tricky to model with logistic regression, conditioning on past and future observations (where available), as the fitted probabilities are often very close to either 1 or 0. This means the corresponding parameter estimates are often noisily estimated, with large standard errors. This is the perfect prediction problem[81].

To address this we needed to use an alternative multiple imputation method known as the two-fold fully conditional specification multiple imputation (see §3.6) that involved imputing patients' missing observations at each visit depending on observations in a window either side of that visit. A consequence of using the two-fold fully conditional specification (FCS) of multiple imputation is that we do not use the full sequence of available data on each patient. A possible disadvantage of this is that by not imputing using information from all visits, some information from past and future visits may be lost. For the marginal GEE models (Chapter 4), patient observations were grouped into visit windows to take observed information at closer time points into account. We also tried to form relatively stable models for the long term trend and smooth out the local noise. However using the models in this way,

analysing outcomes within each visit window, may be inefficient as this does not take into account any dependency of the history of observations at future time points.

Another approach, which arguably more closely reflects the clinical reality, is to think of the sputum tests as imperfect observations of an underlying disease state which is either positive (diseased) or negative (clear of disease). This is what multi-state models seek to do, so it is natural to explore their utility for modelling non-inferiority trials in TB.

In a longitudinal setting, using multi-state models means the entirety of each patient's available data is used to estimate their disease state at each time, t. Once fitted this can then be used to impute the missing sputum culture results which are needed to construct the primary clinical outcome in the study (see §3.1) under an intention-to-treat analysis. Therefore, multi-state Markov models appear to have potential for a more accurate and powerful primary analysis that includes information from all available data for all patients in the analysis. Hopefully, this will yield more appropriate and accurate inferences. The extent to which this potential can be realised is the focus of this chapter.

5.1 Motivation for multi-state models in tuberculosis clinical trials

In TB trials, patients tend to be excluded from analyses if they withdraw from the study under the intention-to-treat and per-protocol analyses depending on treatment completion. For example, if patients do not reach the end of treatment phase, or if they were not seen at the final follow up visit of the study and they were disease-free prior to withdrawal. Alongside this, during the course of a trial, patients may be intermittently observed. As outlined in §3.1, missing culture results (and contaminated results, classed as missing) are usually ignored according to the trial protocol. As previously discussed in §3.1, this can cause difficulties when determining whether or not a patient was "cured", according to the rules used to determine the primary outcome in the trial protocol. For example, for possible relapses where patients may have had two consecutive negative culture results followed by an

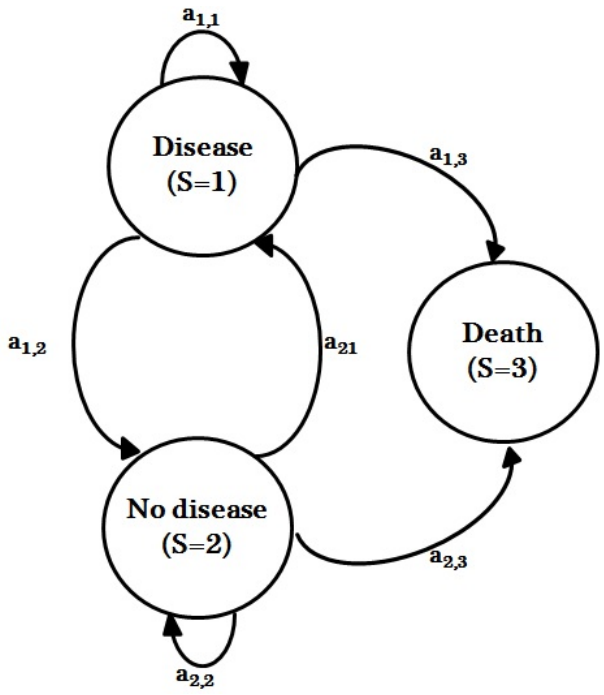
unattended visit, a positive result at the next visit and a negative result thereafter (i.e. $-$,$-$,missing,$+$,$-$), then under the trial protocol, it is unclear whether the underlying clinical state is positive or negative. In such cases, it is unclear whether the patient remained in a state of stable negative culture conversion (since the single observed positive result would be considered a "negative" according to the trial protocol), or whether they relapsed. Multi-state models can estimate the underlying state, impute the missing sputum test values, and hence readily allow us to calculate the primary outcome.

In the next section, we review the theory of multi-state Markov models as it relates to our application to TB trials. We begin by defining simple Markov chains before extending the theory to hidden Markov models and then we apply these methods to the REMoxTB study and the RIFAQUIN study.

5.2 Markov multi-state models

A multi-state model is a stochastic process, which generates a sequence, S_t, over time[89]. S_t is defined as an ordered set of *discrete* states at continuous time t[90]. Discrete time Markov chain models can be analogously defined[91], but are not our focus here.

Figure 5.1: Example of a 3 state Markov chain model.



Multi-state Markov models are a stochastic process and have the Markov property. This means that future states of the process are conditionally independent of the history of the process before time t given the state of the process at time t. If the sequence of states is observed, and therefore the state is known, a Markov chain can be modelled. Figure 5.1 shows an example of a Markov chain model with three states, disease ($S_t = 1$), no disease ($S_t = 2$) and death ($S_t = 3$).

Probability transitions are defined as the probability of being in state *j* at time t given state *i* at the previous time point $t - 1$ such that $a_{i,j}(t) = P(S_t = j|S_{t-1} = i)$ and can be interpreted as the instantaneous risk of transitioning from state i to state j. If there are only two states, this "risk" can be interpreted as a hazard, as in survival analysis. In the context of clinical trials, our exemplar model shows for a given sequence, patients can move in and out of state 1 (presence of disease) and 2 (absence of disease) over time with probability transitions $a_{1,2}(t)$, $a_{2,1}(t)$ and can move into state 3, death, with probability transitions $a_{1,3}(t)$, $a_{2,3}(t)$. For the purpose of this explanation, we assume the probabilities are constant over time so that $a_{i,j}(t) = a_{i,j}$. Death is an absorbing state

as patients cannot subsequently transition into states 1 or 2. In this example, the same state can occur consecutively and have probability transitions of $a_{1,1}$ or $a_{2,2}$. These state probability transitions can be presented in matrix form[92]:

$$A = \{a_{i,j}\} = \begin{array}{c} \\ \text{from state } (i) \end{array} \begin{array}{c} \text{to state } (j) \\ \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ 0 & 0 & 0 \end{bmatrix} \end{array}.$$

Note that probability transitions from state 3 all equal zero. This is because following death, patients cannot transition to any disease status nor death itself again.

For a fully observed sequence, transition probabilities are easily estimated by calculating the proportion of transitions from state to state. Let $s_0, s_1, ..., s_t, ..., s_T$ represent what state, S, a patient is in at times $t = 0, 1, ..., T$. The probability of a state sequence in a model can be found by:

$$\begin{aligned} P(S_0 = s_0, S_1 = s_1, S_2 = s_2, ..., S_{t-1} = s_{t-1}, S_t = s_t) = \\ \text{P}(S_T = s_T | \pi_i, S_1 = s_1, ..., S_{T-1} = s_{T-1}) P(S_1 = s_1, S_2 = s_2, ..., S_{T-1} = s_{T-1}). \end{aligned} \tag{5.1}$$

where $\pi i = P(S_0 = s_0)$, the initial state probability. In the example above (Figure 5.1), the possible states are $s \in 1, 2, 3$. By the Markov property this then becomes:

$$\begin{aligned} \text{P}(S_t = s_t | S_{t-1} = s_{t-1}) P(S_{t-1} = s_{t-1} | S_{t-2} = s_{t-2}) ... P(S_2 = s_2 | S_1 = s_1) P(S_0 = s_1) = \\ P(S_0) \prod_{t=1}^{T} a_{(S_{t-1}=s_{t-1})(S_t=s_t)}. \end{aligned} \tag{5.2}$$

This is the general form of the *likelihood* for a state sequence (S) given the model parameters ($\phi = (\pi_i, A)$) for the initial state probability and probability transitions $A = a_{i,j}$, where the likelihood of the model parameters ϕ given the data is the product of the probabilities of transitioning between states that are observed[93].

Using our 3 state model (Figure 5.1), suppose the patient is diseased during the first 4 visits, disease free at the fifth visit and then subsequently dies: $S_0 = 1$, $S_1 = 1$, $S_2 = 1$,

$S_3 = 1$, $S_4 = 2$, $S_5 = 3$. Then the probability of the sequence is:

$$\begin{aligned} P(S|\phi) &= P(\mathrm{S}_0 = 1, S_1 = 1, S_2 = 1, S_3 = 1, S_4 = 2, S_5 = 3) \\ &= P(\mathrm{S}_0 = 1)\mathrm{P}(S_1 = 1|S_0 = 1)\mathrm{P}(S_2 = 1|S_1 = 1)\mathrm{P}(S_3 = 1|S_2 = 1) \\ &P(\mathrm{S}_4 = 2|S_3 = 1)\mathrm{P}(S_5 = 3|S_4 = 2) \\ &= \pi_1 a_{1,1} a_{1,1} a_{1,1} a_{1,2} a_{2,3}, \end{aligned} \tag{5.3}$$

where π_i = P(S_0 = i) is the initial state probability (i.e. π_1 = P(S_0 = 1) in the above example).

In practice, often it is not possible to observe the states directly. In this case, we say the underlying Markov process is hidden. We now describe how the model can be extended to accommodate this in general settings before describing the special case of TB trials where the states are disease states and the observations are culture states. When we have a hidden Markov model (HMM), in simpler cases such as our TB setting, the true disease state, which is hidden, is measured with error. The measurements, positive or negative culture results, can be of the same type as the underlying disease state. Hidden Markov models can be applied much more generally when the observations have more values or when the observations are predictive of the underlying states that generate them. In the next few sections as we outline the HMM theory, we retain this generality.

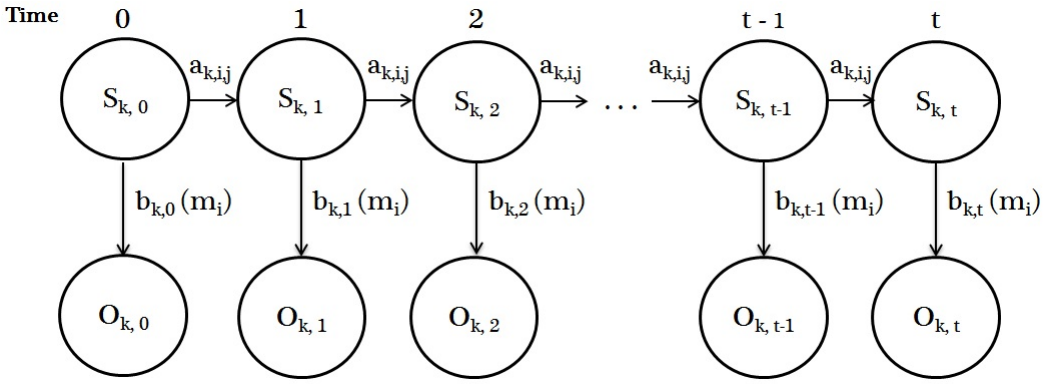
5.3 Hidden Markov models

A hidden Markov model (HMM) describes the setting where we have a set of observations but the underlying states of the Markov chain that generate the observations are unknown.

The Markov chain defined in §5.2 can be extended to HMMs. Figure 5.2 shows a graphical representation of a general HMM for states ($S_{k,0:t}$) and observations ($O_{k,0:t}$) for patients $k = 1, 2, ..., K$ over times $t = 0, 1, 2, ..., T$. As before (under the Markov assumption), for each patient k, probability transitions are defined as the probability of being in a state conditional only on the state at the previous time point:

$$a_{k,i,j} = P(\mathrm{S}_{k,t} = j|S_{k,t-1} = i). \tag{5.4}$$

Figure 5.2: Example of a hidden Markov model over time for patient k and i, j hidden states.



We denote the initial state probability by:

$$\pi_{k,i} \quad = P(\mathrm{S}_{k,0} = i) \tag{5.5}$$

let:

$$b_{k,t}(m_i) \quad = P(\mathrm{O}_{k,t} = m_i | S_{k,t} = i), \tag{5.6}$$

where $b_{k,t}(m_i)$ represents the probability of being observed as m within state i, where $m = 1, 2, ..., M$ at time t (where $t = 0, 1, ..., T$). Assuming that this does not vary over time, the observation probability matrix is then:

$$B_k = \{b_{k,t}(m_i)\} = \begin{matrix} \\ \text{state } (i \in 1,2,...,I) \end{matrix} \overset{\text{observations } (m \in 1,2,...,M)}{\begin{bmatrix} b_k(1_1) & b_k(2_1) & \cdots & b_k(M_1) \\ \vdots & \vdots & \ddots & \vdots \\ b_k(1_I) & b_k(2_I) & \cdots & b_k(M_I) \end{bmatrix}}$$

The underlying states ($i \in \{1, 2, ..., I\}$), that are not seen, emit the observation sequence ($m \in \{1, 2, ..., M\}$). For this HMM, the initial state, $s_{k,0} = i$, at time 0 follows the initial state probability distribution $\pi_{s_{k,0}}$ which is assumed to be known. In that initial state, an observation is emitted with probability $b_{k,0}(m_{s_{k,0}})$ for that state. We observe m_i when the true state is $s_{k,0}$. A new state ($s_{k,1}$) is then chosen according to the probability transition $a_{k,i,j}$ from time 0 to time 1, and the process is repeated until the final observation is observed[94].

This probability in (5.6) naturally models the *misclassification*, where observations are test results, given an unobservable (true) binary disease state[95]. For example, the probability the test is observed (positive) when the disease is present is the sensitivity and the probability that the test is not observed (negative) when the disease is not present is the specificity. These probabilities represent the *sensitivity* and *specificity* of the instrument and we estimate these for our TB trials. The probability of a negative culture observed when the true underlying state is positive and the probability of a positive result observed when the true underlying state is negative are also estimated.

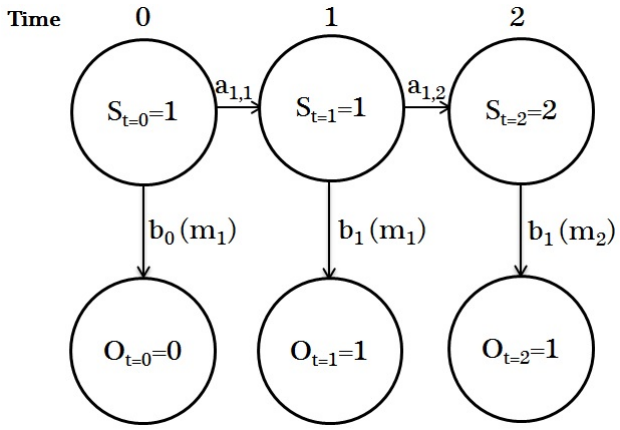
The joint probability of observations and states given the model can be found by extending (5.2) to include observation sequences as follows:

$$P_k(O, S) = \pi_{s_{t=0}} b_0(m_{s_{t=0}}) \prod_{t=1}^{T} b_{k,t}(m_{s_t}) a_{s_{t-1}, s_t} \tag{5.7}$$

for state sequence of length T+1. This is the product of the probability of i) the initial probability state (π_i), ii) the probability of the observation in the initial state ($b_0(m_{s_0})$) and iii) the transition probabilities from state i to state j (a_{s_{t-1}, s_t}) and the corresponding observation probability ($b_{t,k}(m_{s_t})$).

Now consider a HMM where the states ($i \in 1, 2, ..., I$) are of a different type to the observations $m \in \{1, 2, ..., M\}$ (i.e. what is observed is not the same as the true underlying state). For example, for one patient assume we observe them to be on ($O_t = 1$) or off ($O_t = 0$) treatment at each of three successive visits as follows: $O_0 = 0$, $O_1 = 1$, $O_2 = 1$ (Figure 5.3). The same patient can either be in the disease state ($S_t = 1$) or disease free state ($S_t = 2$) such that their sequence of states is: s_0 = 1, s_1 = 1, s_2 = 2.

Figure 5.3: Example of a hidden Markov model for one patient's observed treatment and hidden disease state.



Then, the probability of the state sequence being emitted by the observation sequence for a patient $k \in 1$ is[96]:

$$\begin{aligned} P(O,S) = P(S_0=1)P(O_0=0|S_0=1)P(S_1=1|S_0=1)P(O_1=1|S_1=1) \\ P(S_2=2|S_1=1)P(O_2=1|S_2=2) \\ =\pi_1 b_0(m_1) a_{11} b_1(m_1) a_{12} b_1(m_2). \end{aligned} \tag{5.8}$$

In order to find the most likely set of underlying, hidden, state sequences, there are four problems outlined as follows. The first and second problems (§5.4.1 and §5.4.2) describe the approach of finding the preferred model (ϕ) for the observed data. The third and fourth (§5.4.3 and §5.4.4) describe approaches to find the most probable pathway of the hidden states, given the chosen model.

5.4 Four HMM problems

As detailed in Rabiner's tutorial[92] and in the HMM literature there are a number of problems (described below unusually in the order of step 1, step 4 and step 2) that we need to address to fit HMMs to our TB datasets:

1. The first problem, which we address in §5.4.1, is to find the likelihood of the parameters under the HMM. This is challenging because the actual states are unobservable (i.e. "hidden"), so they have to be summed out of the likelihood

function. However, because each individual can have multiple state pathways given their data, this is computationally very intensive. Therefore, we outline the approximate approach used by the software.

2. Once we obtain the likelihood, we describe how to maximise it in §5.4.2. For this the BFGS optimiser is used.

Once the preferred model is chosen using steps 1 and 2, we can impute the missing TB sputum test data. There are two possible ways to do this. The first, uses the forwards/backwards algorithm to estimate the probability of a patient being in a particular state at each time point, from which we can directly impute the missing values. The second, uses the Viterbi algorithm to predict the most probable sequence of states for a patient overall. The critical difference between these two algorithms for our data is that the states predicted by the forwards/backwards algorithm and Viterbi algorithm can disagree. For example, if the transition probability is low between the most likely observations at time t and $t+1$, the Viterbi algorithm will give a lower weight to the underlying state at that time (even if we have an observation at that time). Both approaches are presented, and used in the application to check for consistency of results. Therefore:

3. to find the probability of a patient being in a particular state at a particular time point using the history of information available prior to that time point, by the Markov property, we use the forwards/backwards algorithm outlined in §5.4.3, and

4. to find the most probable overall pathway of a state sequence for each patient we use the Viterbi algorithm, outlined in §5.4.4.

5.4.1 Problem 1: Evaluation of the likelihood using the forward algorithm

The first problem to resolve is to find the likelihood. That is, the probability of an observation sequence, $O_{k,1:t} = O_{k,1}, O_{k,2}, O_{k,3}, ..., O_{k,t-1}, O_{k,t}$, given the state space model parametrised by ϕ. This probability is found by calculating every possible state

sequence. To do this, the joint probability of the observations and the state sequence given the model can be calculated. Let the model be denoted by ϕ, such that ϕ = (A, B, π_i) where A represents state transition probabilities ($a_{k,i,j}$) from $t-1$ to t, B represents the observation probabilities of a patient (k) being in a particular state, $b_{k,t}(m_i)$ at time t, and π_i is the initial probability in each state. By (5.4) and (5.5):

$$\mathrm{P}(S_k|\phi) \quad = \pi a_{k,s_0,s_1} a_{k,s_1,s_2} a_{k,s_2,s_3} ... a_{k,s_{t-1},s_t} \tag{5.9}$$

and by (5.6):

$$\mathrm{P}(O_k|S_k,\phi) \quad = \mathrm{b}_{k,0}(m_{s_0}) b_{k,1}(m_{s_1}) b_{k,2}(m_{s_2}) ... b_{k,t}(m_{s_T}), \tag{5.10}$$

where $m \in 1, 2, ..., M$ is observed when in a particular state (s_T) at time T. Then, by (5.9) and (5.10) the joint probability of observation sequence and the state sequence given the model is:

$$P(O_k, S_k|\phi) = P(O|S,\phi)P(S|\phi). \tag{5.11}$$

If the S_k states in (5.11) were observed then the total likelihood is:

$$L(\phi; \underline{S}, \underline{O}) = \prod_{k=1}^{K} P(O_k, S_k|\phi) \tag{5.12}$$

As the states are unobserved (and unobservable), to obtain the likelihood of the observed data, we need to sum over all the possible state sequences, S_k, that are consistent with the observed data. This is a requirement for every patient to obtain the observed data likelihood.

$$\begin{aligned} P(O_k|\phi) &= \sum_{\forall S_k} P(O_k, S_k|\phi) \\ &= \sum_{\forall S_k} P(O_k|S_k,\phi) P(S_k|\phi) \end{aligned} \tag{5.13}$$

Equation (5.13) can be found by (5.9) and (5.10). However, this is an exhaustive calculation and becomes computationally infeasible the larger the number of observations and (in our setting) as the number of follow up visits increases. Instead, the forward procedure is used to compute an approximation to the likelihood, (5.13), of the observed sequence in a computationally efficient manner[94].

This calculation used by the forward procedure is much simpler than summing over all possible state sequences for each patient. Given the current value of the model parameter, ϕ, for patient k, the forwards algorithm calculates the probability of the partial observation sequence, $O_{k,1:t}$, when state $S_{k,t} = s_{k,t}$. The initial probability of any observation sequence is:

$$\alpha_{k,t=0}(i) = \pi_i b_{k,0}(m_i) \tag{5.14}$$

for state *i* (where π_i=P($s_0 = i$)) at time 0 for patient k. The probability at each time for $t > 0$ can then be calculated recursively:

$$\alpha_{k,t+1}(j) = \left[\sum_{t=1}^{t} \alpha_k, t(s_k, t) a_{k,s_{t-1},s_t} \right] b_{k,t+1}(m_j) \tag{5.15}$$

until the end:

$$P(O|\phi) = \sum_{t=0}^{T} \alpha_{k,t}(s_t). \tag{5.16}$$

As an example, suppose we observe the weather which can either be sunny (S), cloudy (C) or rainy (R) over time t = 0, 1 and 2; let the hidden state be whether a patch of grass is dry (D) or wet (W). Assume that at the current value of the model parameters, ϕ we have:

$$\pi_i = \begin{bmatrix} 0.0 & 1.0 \end{bmatrix}, A_k = \begin{bmatrix} 0.8 & 0.2 \\ 0.9 & 0.1 \end{bmatrix}, B_k = \begin{bmatrix} 0.3 & 0.5 & 0.2 \\ 0.7 & 0.2 & 0.1 \end{bmatrix},$$

where the initial state $i \in Dry\ or\ Wet$ For one patch's sequence (i.e. $k \in 1$), let the observations be R, S, C. To calculate $P(O|\phi)$, the forward algorithm is used. The initial probability for a dry state when rain is observed, (i.e. $\alpha_{R,t=0}(D)$) and the initial probability for a wet state when rain is observed (i.e. $\alpha_{R,t=0}(W)$) is:

$$\alpha_{R,t=0}(D) = \pi_D \times b_0(R_D) = 0 \times 0.2 = 0.$$
$$\alpha_{R,t=0}(W) = \pi_W \times b_0(R_W) = 1 \times 0.1 = 0.1.$$

These initial probabilities can then be carried forwards to find the probabilities for a

dry state and a wet state when the weather is sunny at time=1:

$$\alpha_{S,t=1}(D) = [\alpha_{t=0}(D)a_{D,D} + \alpha_{t=0}(W)a_{WD}]b_{t=1}(S_D) = [(0 \times 0.8) + (0.1 \times 0.9)] \times 0.3 = 0.027.$$

$$\alpha_{S,t=1}(W) = [\alpha_{t=0}(D)a_{D,W} + \alpha_{t=0}(W)a_{WW}]b_{t=1}(S_W) = [(0 \times 0.2) + (0.1 \times 0.1)] \times 0.7 = 0.007.$$

The final probabilities for a dry state and wet state when the weather is cloudy are then:

$$\begin{aligned}\alpha_{C,t=2}(D) &= [\alpha_{t=1}(D)a_{D,D} + \alpha_{t=1}(W)a_{WD}]b_{t=2}(C_D) \\ &= [(0.027 \times 0.8) + (0.007 \times 0.9)] \times 0.5 = 0.014. \\ \alpha_{C,t=2}(W) &= [\alpha_{t=1}(D)a_{D,W} + \alpha_{t=1}(W)a_{WW}]b_{t=2}(C_W) \\ &= [(0.027 \times 0.2) + (0.007 \times 0.1)] \times 0.2 = 0.001.\end{aligned}$$

The probability of the observation sequence R, S, C given the model parameters ϕ is calculated by the sum of the above probabilities:

$$P(O|\phi) = 0 + 0.1 + 0.027 + 0.007 + 0.014 + 0.001 = 0.149.$$

Due to computational underflow, the log of the likelihood (equation 5.13) is usually calculated, i.e. $log(P(O|\phi))$. For our example taking the natural logarithm would result in a probability of -1.904.

Having found the likelihood, we need to train the model parameters to maximise the probability of the observations given the current model parameters. This is so that the observations seen are represented by the model ($\phi = (\pi_i, A_k, B_k)$) in the best way for application to the dataset.

5.4.2 Problem 2: Maximising the likelihood

This requires adjusting the model parameters (ϕ) to maximise the probability of observations given the model. The complete data log-likelihood can be found iteratively using (5.13). Let ϕ^{init} be the initial or previous estimates of the parameters[97]:

$$\begin{aligned}P(O,S|\phi) = \sum_{\forall S_k} log(\pi)P(O,S|\phi^{init}) + \sum_{\forall S_k}\left[\sum_{t=1}^{T} log(a_{k,i,j})\right]P(O,S|\phi^{init}) \\ + \sum_{\forall S_k}\left[\sum_{t=1}^{T} log(b_{k,t}(m_i))\right]P(O,S|\phi^{init})\end{aligned} \tag{5.17}$$

Each of the three terms can then be maximised using maximum likelihood estimation[98]. Alternative algorithms such as the Baum-Welch algorithm[99], used to guarantee convergence of the model and Viterbi training[100] used as an approximation to the likelihood (at a loss of efficiency, but gaining in speed) may be used. The Baum-Welch algorithm is an Expectation-Maximisation algorithm which uses the forwards/backwards algorithm (described in the next section §5.4.3) to choose model parameters so that $P(O|\phi)$ is locally maximised[92]. Viterbi training uses the Viterbi algorithm (defined in §5.4.4), which chooses the probability of the most likely state at a particular time going forwards. This results in a less computationally intensive algorithm. Transition probabilities (A) and observation probabilities (B) are initialised to random numbers and the most probable pathway for the underlying state can be calculated based on a set of observations. The most likely state sequence found is then used to re-estimate the hidden parameters. This is then cycled through repeatedly until the underlying hidden states are unchanged.

5.4.3 Problem 3: Smoothing using the forward/backward algorithm

Often it is of interest to find the probability an observation in a sequence came from a particular state at a particular time, i.e. $P(S_{k,t} = i|O_{k,t})$. The forward/backward algorithm is used to efficiently calculate the probability of being in a particular state at each time point in a hidden Markov model, given the entire observation sequence for each patient given the current parameter values ϕ. The forwards/backwards algorithm uses the forwards algorithm, defined in §5.4.1, (equations 5.14 to 5.16) and the backwards algorithm (5.18) to (5.20) for smoothing, as explained below.

The backwards algorithm is analogous to the forwards procedure, where the last observation is taken and iterated backwards. Assume the backward probability equals 1 to start with, such that if $\beta_t(i)$ = P($O_{k,0:t}|S_{k,t} = i, \phi$) then:

$$\beta_t(s_T) = 1 \tag{5.18}$$

Each recursive probability is calculated:

$$\beta_{k,t}(l) = \sum_{t=l}^{T} a_{k,s_{t-1},S_t} b_{k,t+1}(m_{s_t}) \beta_{t+1}(s_T) \tag{5.19}$$

for $l = t-1, t-2, t-3, ..., 0$ until termination:

$$P(O|\phi) = \sum_{t=0}^{t} \beta_{k,t}(l). \tag{5.20}$$

The probabilities from the forwards algorithm and backwards algorithm are often scaled to sum to 1 because as length of follow up time increases, the probabilities tend to 0 exponentially.

Smoothing is accomplished by multiplying the probabilities from the forwards algorithm and the backwards algorithm together in order of time, so estimating the marginal probability of transitioning from state to state at each time point.

Again we use the above weather example in §5.4.1 to determine if a patch of grass is dry or wet: the aim is to find the best hidden sequence of the grass state given R, S, C was observed and our model parameters. The forwards probabilities have already been calculated, so we now calculate the backwards probabilities assuming probabilities of 1 in each state:

$$\beta_{C,t=2}(C_D) = 1.0$$
$$\beta_{C,t=2}(C_W) = 1.0.$$

For t=1:

$$\beta_{S,t=1}(D) = [\beta_{t=2}(D)a_{DD} + \beta_{C,t=2}(W)a_{WD}]b_{t=2}(C_D) = [(0.8 \times 0.5 \times 1) + (0.9 \times 0.2 \times 1)] = 0.58.$$
$$\beta_{S,t=1}(W) = [\beta_{t=2}(D)a_{DW} + \beta_{C,t=2}(W)a_{WW}]b_{t=2}(C_W) = [(0.2 \times 0.5 \times 1) + (0.1 \times 0.2 \times 1)] = 0.12.$$

The final probabilities for t=0 are:

$$\begin{aligned}\alpha_{R,t=0}(D) &= [\beta_{t=1}(D)a_{DD} + \beta_{S,t=1}(W)a_{WD}]b_{t=1}(S_D) \\ &= [(0.8 \times 0.3 \times 0.58) + (0.9 \times 0.7 \times 0.58)] = 0.50. \\ \alpha_{R,t=0}(W) &= [\beta_{t=1}(D)a_{DD} + \beta_{S,t=1}(W)a_{WD}]b_{t=1}(S_D) \\ &= [(0.2 \times 0.3 \times 0.12) + (0.1 \times 0.7 \times 0.12)] = 0.02.\end{aligned}$$

Table 5.1: Marginal (scaled) probabilities at each time point for observing rain, sun, cloud using the forwards/backwards algorithm.

Time (t)	Forwards algorithm		Backwards algorithm		Forwards/ backwards algorithm	
	$\alpha = D$	$\alpha = W$	$\beta = D$	$\beta = W$	(scaled[1]) $\alpha\beta = D$	$\alpha\beta = W$
0	0	0.1	0.51	0.02	0 (0)	0.002 (1)
1	0.03	0.01	0.58	0.12	0.02 (0.995)	0.001 (0.05)
2	0.01	0.001	1	1	0.014 (0.93)	0.001 (0.07)

[1] Probabilities are scaled to sum to 1.

Table 5.1 shows the resulting probabilities in order of time and multiplies them to give the marginal probabilities of being in a dry or wet state over time using the forwards/backwards algorithm.

The next HMM problem is to find the hidden part of the model by calculating the optimal state sequence of the model given the current parameter values. The Viterbi algorithm, proposed by Viterbi[101] is typically used for this. The algorithm works by finding the most probable sequence of hidden states for patients overall that results in the observed states.

5.4.4 Problem 4: Decoding using the Viterbi algorithm

The Viterbi algorithm finds the most likely sequence of the underlying hidden states given the entire observation sequence and the current model parameters, ϕ. Using the above weather example to determine if a patch of grass is dry or wet, the aim would be to find the best hidden sequence of the grass state given rain (R), sun (S), cloud (C) was observed and the current value of the model parameters, ϕ. In other words, we seek to find for each patient k the state sequence ($\{S_{k,t}\}_{t=0}^{t=T}$) that maximises P($S|O, \phi$) or equivalently P($S, O|\phi$). Let:

$$V_{t-1}(i) = \max_{\forall s_{k,t=0:t}} P(s_{k,0}, S_{k,1}, ..., s_{k,t=i} = i, O_{k,0}(s_{k,0}), O_{k,1}(s_{k,1}), ..., O_{k,t}(s_{k,t} = i)|\phi), \tag{5.21}$$

where $V_t(i)$ is the maximum probability along a single hidden pathway for patient k at time t and $\max_{\forall S_{k,t=0:t}}$ is the most probable path taking the maximum over all possible previous state sequences. Given the probability of a patient being in every state at time t has already been calculated, the Viterbi probability is calculated by taking only the most probable state sequence that leads to the next state. For a state $S_{k,t}$ at time t:

$$V_{t+1}(j) = [\max_{k,s_t} V_t(s_t) a_{k,s_t,s_{t+1}}] b_{k,t+1}(m_j). \tag{5.22}$$

In addition to the probability $V_t(s_t)$, the Viterbi algorithm also produces the most likely state sequence for each patient, k, and hence for the dataset overall. We define this as $W_{k,t}(j)$, where j represents the current state a patient is in, to keep track of the sequence of hidden states that led to each state. In other words, this quantity remembers what current state a patient is in before finding the next most likely state at time $t+1$. The optimal state sequence can then be found by[92]:

1. Initialise:

$$\begin{aligned} V_{k,0}(s_0) &= \pi_{k,s_0} b_{k,0}(m_{s_0}) \\ W_{k,0}(s_0) &= 0. \end{aligned} \tag{5.23}$$

2. Recursion for t=2, ..., T:

$$\begin{aligned} V_{k,t}(s_t) &= \max_{k,s_{t-1}} [V_{k,t-1}(s_{t-1}) a_{s_{t-1},s_t}] b_{k,t}(m_{s_t}) \\ W_{k,t}(s_t) &= \operatorname*{argmax}_{k,s_{t-1}} [V_{k,t-1}(s_{t-1}) a_{s_{t-1},s_t}]. \end{aligned} \tag{5.24}$$

3. Termination:

$$\begin{aligned} P(S^*, O|\phi) &= \max_{k,s_T} [V_{k,T}(s_T)] \\ S^*_{k,T} &= \operatorname*{argmax}_{k,s_T} [V_{k,T}(s_T)] \end{aligned} \tag{5.25}$$

4. Backtracing the best state sequence to the beginning:

$$S^*_{k,t} = W_{k,t+1}(S^*_{k,t+1}), \tag{5.26}$$

for time $t = T-1, t-2, ..., 2, 1$ for the most probable path S^*. Due to computational underflow, the logarithms are usually taken for $\log(\pi_{k,s_0})$, $\log(a_{s_{t+1},s_t})$ and $\log[b_{k,t}(m_{s_t})]$. Note that this calculation is similar to that of the forwards algorithm (equation 5.14 to 5.16), where the summation is replaced with a maximum.

We now illustrate the above by returning to our example set out in §5.4.1. We can use the model parameters to find the most probable state sequence for a patch of grass given rain (R), sun (S) and cloud (C) are observed over time t = 0, 1 and 2. The initial probabilities are:

$$P(D) = \pi b_{t=0}(D) = 0 \times 0.2 = 0$$

$$P(W) = \pi b_{t=0}(W) = 1 \times 0.1 = 0.1$$

The key to the Viterbi algorithm is that only the highest scoring pathways are kept at each possible state rather than a list of all possible states. Since, at time $t = 0$, $P(W) > P(D)$, we proceed with the probability of being in a wet state at $t = 0$, to calculate the probability of the next state being wet or dry given that sun is observed:

$$P(W, D) = P(D)a_{W,D}b_{t=1}(D) = 0.1 \times 0.9 \times 0.3 = 0.027$$

$$P(W, W) = P(D)a_{W,D}b_{t=1}(W) = 0.1 \times 0.1 \times 0.7 = 0.007$$

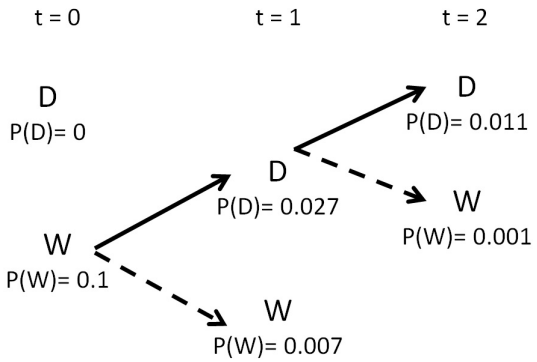
Taking the most probable states forward, P(W,D), the probability of the final state being wet or dry given that we observe cloudy weather is:

$$P(W, D, D) = P(W, D)a_{D,D}b_{t=2}(D) = 0.027 \times 0.8 \times 0.5 = 0.011$$

$$P(W, D, W) = P(W, D)a_{D,W}b_{t=2}(W) = 0.027 \times 0.2 \times 0.2 = 0.001$$

The most probable sequence of states for a patch of grass given rain, sun and cloud are observed and the model parameters is: wet, dry and dry (Figure 5.4). Note that the probabilities at the end are similar to those produced from the forwards/backwards algorithm in Table 5.1.

Figure 5.4: Graphical representation of Viterbi algorithm to find the most likely sequence of states for a patch of grass.



Having described multi-state Markov models and in particular outlined how the likelihood is obtained, and the role of the forwards/backwards algorithm and the Viterbi algorithm, we now return to our TB application.

5.5 Multi-state models in tuberculosis clinical trials

In TB clinical trials, part of the algorithm to determine the composite primary outcome of treatment failure (see §3.1) states that patients diagnosed with TB are considered "cured" at the point when two consecutive negative sputum culture results are observed (at separate visits). Using multi-state models enables us to impute the missing data (details to follow) resulting in a "completed" dataset which can be used to determine each patient's outcome. Hence, we can define the primary outcome of treatment failure over 18 months of follow-up as:

1. Relapse; two consecutive positive culture results at separate visits after the treatment phase following "cure",

2. Patients who are never "cured"

As in §4.1, the final classification which classes patients who had a positive culture result followed by a negative culture result, where the positive result was preceded by at least two consecutive negative culture results, i.e. $-,-,+,-$ is considered a treatment success. This definition differs from the REMoxTB and RIFAQUIN studies where if the last scheduled observed positive result was not followed by at least two possibly scheduled negative results, it was considered as "unfavourable". Our definition in step 3 has been relaxed from the protocol defined outcome to remove any false positive culture results that occur at the end of follow-up. This is because in the original analysis, unscheduled visits post the week 78 scheduled follow-up visit were included to determine the primary outcome. Therefore, any isolated positive cultures towards the end of scheduled follow-up in the main study would have been classed as "negative" by two negative culture results at these extra unscheduled visits post week 78. For analyses performed in this chapter, we only look at results collected at the scheduled follow-up visits to impute the missing data. Unscheduled results where positive or negative culture results were observed (and therefore known) between randomisation and week 78 are then included to determine patients' overall outcome.

To determine treatment failure, as defined above, requires data from each patient at each scheduled visit. The aim is to include all patients in the analysis (intent-to-treat), excluding patients for reasons unrelated to treatment such as drug resistance, protocol violations at enrolment or no positive culture results seen in the first 2 weeks of randomisation. When data are missing, we are going to use the underlying state to impute the missing observation at each scheduled follow-up visit for all patients. Although looking at the entirety of a pathway of state sequences given the observations seen in the data is useful to get an overall picture of a study, using the whole pathway per patient predicted by Viterbi may not always match the states that were observed in the study (see §5.4.4). This is because for misclassification models where the observations are states, a long sequence of observations seen may not always exactly match the sequence predicted by the Viterbi algorithm since Viterbi calculates the *expected* pathway given the entirety of what is seen. Instead, we use the forwards/backwards algorithm (see §5.4.1) to find the hidden states which are then used to impute the missing culture data, resulting in a "completed" dataset. Following imputation, any unscheduled, observed, visits are then included if those visits occurred over the course of scheduled follow-up to determine whether or not a patient was a treatment failure.

A sequence of underlying (unobservable) states based on taking the maximum probability of independently calculated states will not always equal the most probable overall sequence of states predicted by the Viterbi algorithm. This is because the Viterbi algorithm calculates the most probable set of state sequences overall, and so states predicted by this algorithm may not always match an observed (known) state. Consequently, to ensure that the known (observed) states within the dataset are used, the forwards/backwards algorithm is preferred for imputation of the missing data. As described above (§5.4.3), this algorithm finds the probability of a patient being in a particular state $(S_{k,t})$ at each time t.

5.5.1 Imputation for missing data in multi-state models

The forwards/backwards algorithm (see §5.4.1) predicts what underlying state a patient is in at a particular time point. We extend this algorithm to impute missing state data (i.e. sputum test results) taking into account the variability of the state predicted by the forwards/backwards algorithm as follows:

1. Draw a random parameter (say, $\hat{\phi}^*$) from the multi-variate normal distribution: $\hat{\phi}^* \sim N(\hat{\phi}, Var(\hat{\phi}))$. Then we use the forwards/backwards algorithm to find the marginal probabilities per patient k at time t.
   Then to impute the states, for each observation that is missing ($\tilde{z}_{k,t}$) per patient k:
2. Set $\phi = \hat{\phi}^*$ and calculate the probability of state, $S_{k,t}$, at time t.
3. Draw independent random numbers, $Rand_k$, between 0 and 1 from the uniform probability distribution, $U_{k,t}(0, 1)$.
4. For binary states (where the states are positive or negative culture results in our TB datasets), if $u_{k,t} < P(S_{k,t} = 1)$, impute the sputum test results as 1, otherwise as 0 (corresponding to $S_{k,t} = 2$).

Following this process gives us one imputed set of observations for states that are missing. Steps 1-4 can be repeated to produce I imputed datasets. For each of these I imputed datasets, the primary outcome can be determined, treating the states as positive/negative sputum test results. Each of these imputed sets gives the initial or previous parameter estimates (ϕ^{init}) and standard error. The final step is then to combine these I imputed data sets, following the calculation of the primary outcome for each imputed data set using Rubin's Rules[34] as defined in Chapter 3 (see §3.5.1).

We use the *msm* package in R to fit HMMs[95]. Within this package, an extra feature was added by Chris Jackson upon our request to draw a random parameter from the multi-variate normal distribution to implement step 1 (§5.5.1).

5.5.2 Calculation of probability transitions

To calculate the probability of transitioning from state to state, we need to estimate the probability transitions ($a_{k,i,j}$). The calculation of the probability transitions from the

HMM can be used to assess how well the HMM fits to our datasets over time. In order to find these probabilities, we use transition intensities. These transition intensities, $\lambda_{i,j}$, are defined as the instantaneous risk of moving from state i to state j[95]. Transition intensities are defined as:

$$\lambda_{i,j} = \lim_{dt \to 0} \frac{P(S_{k,t+dt} = j | S_{k,t} = i)}{dt}. \tag{5.27}$$

For our TB datasets, we use a two state Markov chain where a patient can either be in a positive state or a negative state. For this simple two state Markov chain, the intensity matrix Q takes the form:

$$Q(t) = \begin{array}{c} \\ \text{from state } (i) \end{array} \begin{array}{c} \text{to state } (j) \\ \begin{bmatrix} -\lambda_{12}(t) & \lambda_{12}(t) \\ \lambda_{21}(t) & -\lambda_{21}(t) \end{bmatrix} \end{array}.$$

By definition, the rows sum to 0[93].

Time constant probability transitions for a set of covariates

For continuous-time multi-state Markov models, possible predictors of the outcome (i.e. treatment failure) can be added to the model as covariates. In this section, we assume time is homogeneous. For a stationary Markov process, the matrix exponential[102] of the scaled transition intensity matrix is used to calculate the probability of transitioning from state i to state j (i.e. $a_{i,j}$). In the simple case where the baseline intensities are constant, $P_{i,j}(s,t) = 1 - e^{\lambda_{i,j}(t-s)}$ is used to estimate the probability of transitioning from state i to state j. This probability assumes the baseline intensities are constant and that the model is time-homogeneous.

Simple simulation study

To check our understanding of the relationship between transition intensities and transition probabilities where the effect of time is assumed to be constant, we performed a simulation study for a two state Markov model. The transition intensity from state 1 to state 2 was set at 0.1 and that from state 2 to state 1 at 0.3, and these are constant over time. The probability transitions were calculated using the matrix exponential of these known intensities. Data were simulated over 10 time points for

10,000 patients from the calculated probability transitions. The HMM was fitted using the *msm* package and the parameters of the estimated intensities from this model were used to estimate the transition probabilities, to compare with those used to generate the data. The probability of changing state over time was always defined on the initial probability of state 1 or 2 (i.e. at the first time point). For example, the probability of a change from state 1 to state 2 at time t (for t=2, 3, 4, 5, 6, 7, 8, 9 and 10) depends on the initial probability of state 1 at time t. The transition from state 1 to state 2 fitted by the HMM to the simulated data gave an estimated intensity of 0.098 (95% CI; 0.096 to 0.101) which includes the true value of 0.1. The transition from state 2 to state 1 had an intensity of 0.301 (95% CI; 0.294 to 0.308) which again includes the true value of 0.3.

Table 5.2 shows that the transition probabilities estimated from the simulated data closely agree with the true values of the probability transitions. The full simulation program, presenting these results and the estimates of the probability transitions are in Appendix G.

Table 5.2: Comparison of simulated probability transitions to the true values, where time is constant.

Time (t)	True values $P(S_t = 2 \mid S_{t-1} = 1)$	Estimated values[1] $P(S_t = 2 \mid S_{t-1} = 1)$	True values $P(S_t = 1 \mid S_{t-1} = 2)$	Estimated values[1] $P(S_t = 1 \mid S_{t-1} = 2)$
1	0.082	0.081	0.247	0.248
2	0.138	0.136	0.413	0.414
3	0.175	0.172	0.524	0.526
4	0.200	0.197	0.599	0.601
5	0.216	0.213	0.648	0.651
6	0.227	0.224	0.682	0.685
7	0.235	0.232	0.704	0.707
8	0.240	0.236	0.719	0.723
9	0.243	0.240	0.730	0.733
10	0.245	0.242	0.736	0.740

[1]Values estimated from the simulated data.

Time-varying probability transitions and a time-varying covariate

Probability transitions can also be calculated in cases where the covariate in the Markov model is time-dependent, such as age or time itself. With time-varying covariates, transition intensities Q(t) are dependent on time. Probability transitions can be estimated in smaller windows of time across the whole time. This assumes that transitions intensities (Q) are piecewise constant between time intervals. Probability transition matrices can then be multiplied individually, $P(t_z, t_{z+1}) = exp((t_{z+1} - t_z)Q(t_z))$, over time.

Simple simulation study

Data were simulated over 10 time points for 10,000 patients from the calculated probability transitions. The HMM was fitted using the *msm* package and the parameters from this model was compared with the time used to generate the data.

To check our understanding of the relationship between the transition intensities and the transition probabilities where time is not constant, we performed a simulation study including time itself as a time-varying covariate using the *msm* package. Data were simulated for 10,000 patients over 10 time points for a two state Markov model. The initial transition intensity from state 1 to state 2 was set at 0.22 and increased by increments of 0.02 over time t (from time=1, 2, 3, 4, 5, 6, 7, 8, 9 and 10). The initial transition intensity from state 2 to state 1 was set at 0.57 and increased by increments of 0.07 until the 10^{th} time point. From these intensities, probability transitions were calculated assuming transition intensities were piecewise constant from $t-1$ to t. The HMM was fitted using the data simulated from these true values and the parameters from this HMM were compared with the intensities used to generate the data. Table 5.3 shows the true value of the probability transitions and the estimates from the simulated data. The full simulation program, presenting these results and the estimates of the probability transitions are in Appendix G.

Table 5.3: Comparison of simulated probability transitions to the true values, where time is constant.

Time (t)	True values $P(S_t = 2 \mid S_{t-1} = 1)$	Estimated values[1] $P(S_t = 2 \mid S_{t-1} = 1)$	True values $P(S_t = 1 \mid S_{t-1} = 2)$	Estimated values[1] $P(S_t = 1 \mid S_{t-1} = 2)$
1	0.152	0.153	0.394	0.404
2	0.160	0.159	0.426	0.426
3	0.166	0.165	0.454	0.449
4	0.173	0.170	0.481	0.473
5	0.178	0.176	0.505	0.496
6	0.183	0.181	0.527	0.520
7	0.188	0.186	0.547	0.543
8	0.192	0.191	0.566	0.566
9	0.196	0.195	0.583	0.588
10	0.200	0.199	0.599	0.610

[1] Values estimated from the simulated data.

Checking the fit of the model

In addition to finding the probability transitions from the transition intensities, we can estimate how well these models fit to the observed raw data. Here, we consider having imperfect observations of an underlying disease state. We calculate the proportion of patients transitioning from state i to state j. This calculation is simply:

$$\frac{\text{Total number of transitions from state } i \text{ at time 0 to state } j \text{ at time } t}{\text{Total number of patients observed at time } t}. \tag{5.28}$$

Using our simulated example conducted above (§5.5.2), the probability of transitioning from state 1 to state 2 when time=1 is calculated by taking the total number of patients who transition from state 1 at time 0 to state 2 at time 1 divided by the total number of patients observed at time 1. For time=2 the probability of transitioning from state 1 to state 2 is calculated by taking the total number of patients who transition from state 1 at time 0 to state 2 at time 2 divided by the total number of patients observed at time 2, and so on until time=10. The calculation is similar for state 2 to state 1 transitions.

For time varying probabilities, the proportion of patients transitioning from state to state is calculated by taking the proportion of patients who transition from state i at time t to state j at $t+1$:

$$\frac{\text{Total number of transitions from state } i \text{ at time } t \text{ to state } j \text{ at time } t+1}{\text{Total number of patients observed in state } i \text{ at time } t}. \tag{5.29}$$

To demonstrate this, we take the simulated example above in §5.5.2 and calculate the proportion of patients with observed data transitioning from state to state.

Figure 5.5: Comparison of raw probability transitions to probability transitions from software for time-varying covariate from state 1 to state 2 and state 2 to state 1.

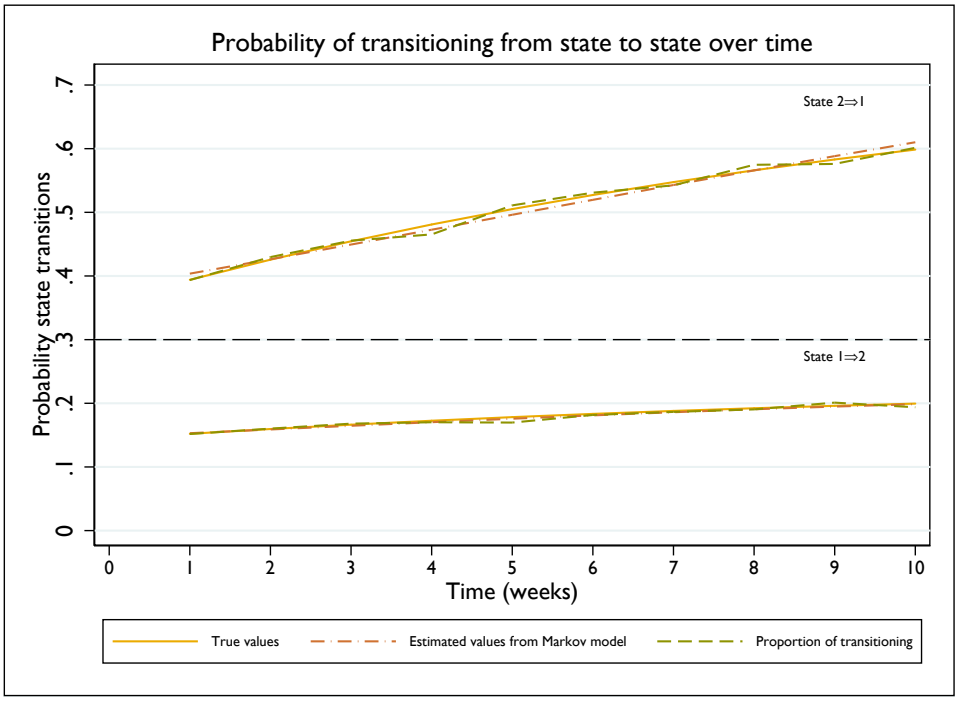


Figure 5.5 compares these proportions (indicated by the green line) to the true values (indicated in yellow) and to the probability transitions calculated from the *msm* package (indicated by the orange line). The model fitted by the *msm* package fits relatively well to the probability transitions calculated from the raw data as well as the true values. Estimating fit of these models produced by the software in this way to our exemplar TB data sets will give us an approximate idea of how well these models may be fitting in practice to the raw data.

5.6 Application of hidden Markov model to the REMoxTB and RIFAQUIN studies

The REMoxTB (§3.2.1) and RIFAQUIN (§3.2.2) studies will be used as examples to evaluate the practical utility of HMMs for TB trials, to handle missing data issues (under MAR) for an intention-to-treat type analysis. As we have already noted, this is a key issue: approximately 10% of patients were excluded from the REMoxTB study and 15% were excluded from the RIFAQUIN study due to loss of follow up or withdrawal, and the non-inferiority margin is 6%. The proportion of patients excluded in these studies therefore exceeds the margin, and so including these observations may have a non-trivial impact on the conclusions.

The average baseline intensities are estimated from the *msm* package. Since we only have two states in our model, these intensities can be interpreted as hazards. From now on, this is the term we will use to describe transition intensities. This HMM will include randomised treatment arm (*trt*) and *time* as covariates and an interaction between the two covariates if needed to allow the hazards to be modelled differently in the treatment arms. The β parameters are interpreted as log hazard ratios.

Different smoothing methods are compared since the intensities are clearly expected to vary over time. Sections 5.6.1 to 5.6.4 describe these methods in detail. These smoothing methods may be able to better capture the probability of transitioning from state to state over time, particularly in the early part of follow-up where patients may fluctuate from state to state. Smoothing methods include piecewise constants, linear splines, cubic splines and fractional polynomials.

5.6.1 Piecewise constant

A piecewise constant splits the covariate into different sections depending on the placement of ζ_z knots where, $z = 1, 2, ..., Z$. A constant hazard between the knots is assumed which allows us to model a non-linear hazard. To find the transition intensity matrix estimates ($\lambda_{i,j}(t)$), transitioning from state i to state j (at time t) for the REMoxTB and RIFAQUIN studies, if patients are observed on the control arm then $trt_1 = 0$ and $trt_2 = 0$. For the REMoxTB study if patients are observed on the

isoniazid arm then $trt_1 = 1$ and $trt_2 = 0$, if observed on the ethambutol arm then $trt_1 = 0$ and $trt_2 = 1$. For the RIFAQUIN study if patients are observed on the 6 month regimen then $trt_1 = 1$ and $trt_2 = 0$, if observed on the 4 month regimen then $trt_1 = 0$ and $trt_2 = 1$. Continuous $time$ represents the follow up time of the scheduled visit, for patients k:

$$\lambda_{i,j}(t) = \beta_0 + \beta_1 trt_1 + \beta_2 trt_2 + \beta_3 time + \beta_4 trt_1 * time + \beta_5 trt_2 * time + ...+ \\ \beta_f \left[(time > \zeta_z)_+) + (trt(time > \zeta_z)_+) \right], \begin{cases} \text{where } (time > \zeta_z)_+ = 1 \\ 0, \text{otherwise} \end{cases} \tag{5.30}$$

and β_f represents the f^{th} parameter. The subscript of "+" means that the resulting values from using the specified knot will always be greater than 0.

5.6.2 Linear splines

Linear splines assume that the relationship between the covariate and the hazard is linear, joining at the knots where the covariate has been split into sections. The model is:

$$\lambda_{i,j}(t) = \beta_0 + \beta_1 trt_1 + \beta_2 trt_2 + \beta_3 time + \beta_4 trt_1 * time + \beta_5 trt_2 * time + ... + \\ \beta_f [(time - \zeta_z)_+ + trt_1(time - \zeta_z) + trt_2(time - \zeta_z)] \tag{5.31}$$

for $\zeta_z = 1, 2, 3, ..., z$ knots. The interaction terms are denoted by $trt_1(time - \zeta_z)$ and $trt_2(time - \zeta_z)$.

5.6.3 Restricted cubic splines

Restricted cubic splines (RCS) assume a cubic polynomial relationship between the covariate and the hazard where the cubic terms change at the knots, and below the first knot and above the final knot. The lower and upper range of the relationship is restricted to linear. These models are more flexible than linear splines as they allow for a smoothly varying cubic relationship between the covariate and the outcome. The model is:

$$\lambda_{i,j}(t) = \beta_0 + \beta_1 trt_1 + \beta_2 trt_2 + \beta_3 * time + \beta_4 trt_1 time + \beta_5 trt_2 time + ... + \\ R[time] + ... + R[trt_1 time] + R[trt_2 time] \tag{5.32}$$

where $R[time]$, $R[trt_1 time]$ and $R[trt_2 time]$ represent the cubic polynomial part of the model. For $R[time]$ and z knots at times $\zeta_q = 1, 2, ..., z$[97]:

$$\beta_f time + \sum_{q=2}^{z-1} \beta_f \left[(time - \zeta_{q-1})_+^3 - (time - \zeta_{z-1})_+^3 \frac{\zeta_z - \zeta_{q-1}}{\zeta_z - \zeta_{z-1}} + (time - \zeta_z)_+^3 \frac{\zeta_{z-1} - \zeta_{q-1}}{\zeta_z - \zeta_{z-1}} \right], \tag{5.33}$$

$$\begin{cases} \text{where } (time - \zeta)_+^3 = (time - \zeta)^3, & \text{if } time \geq \zeta \\ 0, & \text{if } time < \zeta. \end{cases}$$

For the $R[trt_1 time]$ and $R[trt_2 time]$ covariates, (5.33) applies replacing $time$ with the interaction: $trt_1 time$ or $trt_2 time_d$ respectively.

5.6.4 Fractional polynomials

Fractional polynomials are estimated from a set of polynomials of order -2, -1, -0.5, 0, 0.5, 1, 2, 3, which transform the covariate of interest into a polynomial term[103]. Fractional polynomials of order 2 (i.e. we restrict the choice to two powers,FP_1, FP_2, from the set) will be used to attain two transformations of the $time$, $trt_1 time$ and $trt_2 time$ covariates. The model is:

$$\begin{aligned} \lambda_{i,j}(t) = \beta_0 + \beta_1 trt_1 + \beta_2 trt_2 + \beta_3 time^{FP_1} + \beta_4 time^{FP_2} + \beta_5 trt_1 time^{FP_1} \\ + \beta_6 trt_2 time^{FP_1} + \beta_7 trt_1 time^{FP_2} + \beta_8 trt_2 time^{FP_2}. \end{aligned} \tag{5.34}$$

We will now apply these smoothing methods to find the best fitting model to our observed data for the REMoxTB and RIFAQUIN studies, using the -2 log-likelihood ratios as a guide. Having chosen our preferred model, we use the forwards/backwards algorithm to impute each missing observation and then we use the Viterbi algorithm to investigate what the most probable underlying state sequence is overall (§5.3 to §5.5.1).

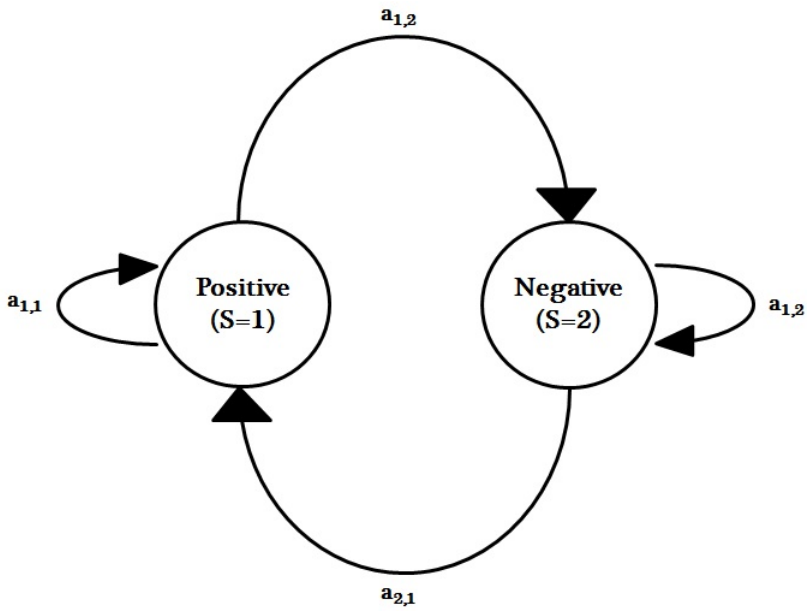
5.7 Application to the REMoxTB study

We now explore fitting a HMM to the REMoxTB study where the negative and positive culture results observed are measures (made with error) on the underlying disease which can either be positive or negative. As already discussed, the aim is to

model all the observed data, and then use the model to multiply impute patients' missing sputum test results under the intention-to-treat analysis. Patients are excluded for reasons unrelated to treatment (Table 3.1). This will enable calculation of the primary outcome for each patient (in each imputed dataset).

For our HMM analyses, the 2% (n=40) of patients who died during the study were removed since the number of deaths is small and balanced among the treatment groups. Therefore removing these patients has no material effect on any inferences made. This makes the modelling substantially less complex since we look at 2 states instead of 3. Recall that REMoxTB compared two treatment arms to the standard of care regimen and patients were followed up over 17 scheduled visits (see §3.2.1). At each of these visits a sputum test was taken and cultured for TB. As previously discussed (§5.5) culture results from sputum samples collected at unscheduled visits are not included in the modelling. As the timing of these observations differs by patient, including them complicates the model and increases the challenge of fitting it.

Figure 5.6: Two state HMM for TB[1].



[1] for continuous time t.

Figure 5.6 shows a two-state Markov chain model for TB trials where a patient's state is either "positive" or "negative". Here, patients can transition in and out of each state over time t. The HMM, showing the observation process, follows as for Figure 5.2 with time t running from time = 0 (i.e. randomisation) to 78 weeks. The state, S_t, is the actual positive or negative TB state at time t. The probability transitions, $a_{k,i,j}(t)$ where $i, j \in 1, 2$, is the instantaneous transition intensity interpreted as hazards at time t. Patients can remain in or transition to and from each state at any time during follow-up. However the underlying TB state cannot be directly observed. Instead, at each scheduled follow-up visit patient sputum samples are taken and are transferred to laboratories. A Mycobacteria Growth Indicator Tube (MGIT) machines is then used to automatically detect whether patients are culture negative or positive[63]. The REMoxTB and RIFAQUIN studies also collected results from a Löwenstein-Jensen medium (LJ) which detects positive or negative results manually. However, these processes are not error-free. Therefore it is possible that false negative or false positive results are returned. Therefore, the observation process (the culture results of the sputum samples at the scheduled visits) needs to be accommodated in the HMM. This is done by specifying a misclassification matrix (see §5.3).

Next, we describe the model we will build for our HMM before choosing our preferred model. The model chosen will then be used to impute missing observations to determine the overall outcome for each patient and compare these results to that from the main study. We then predict the most likely overall state sequence per patient, use this to determine the overall outcome per patient and compare the results of this to the results of the main study.

5.7.1 Model building

Recall from §3.6.1 that patients can still be included in the study if they have a positive culture result during the first 2 weeks of the study if not at baseline. Therefore, to initialise the state transitions, we take an initial working assumption that 80% of patients have a positive culture result (i.e. have TB) and 1% have a negative culture result at baseline. We also assume initial values of 95% sensitivity (i.e the probability of having a true positive result) and specificity (i.e. the probability of having a true negative result) based on expert opinion. Following the first 8 weeks where patients

are assessed weekly, patients are assessed less frequently over the following 70 weeks (§3.8.1). During the first 8 weeks we assume the observation times represent the exact times of the transition, and patients are assumed to be in the same state between these follow-up visits for continuous time[95]. In other words, the observed state remains constant between scheduled follow-up visits. This is reasonable to assume since it is unlikely that the result of the sputum culture per patient would frequently fluctuate between a positive and a negative culture result between the 7 days of weekly follow-up visits. Subsequent visits have a much larger time gap, and therefore after 8 weeks it is assumed that transitions fluctuate from positive to negative culture results and vice versa between observed scheduled visits until the final 78 week follow up visit.

Our strategy to find our preferred model is to begin with constant hazards (no covariates included) and then increase the flexibility of the hazard model until improvements in the goodness of fit are negligible. There may be instances where the resulting -2 log-likelihood suggests a model is a good fit, but the resulting parameters have great uncertainty surrounding them. The chosen model will be determined on a combination of both the log-likelihood and reasonable confidence intervals that surround the resulting parameters. This also applies when choosing our preferred model overall.

Our first addition to the model is to include treatment and time as covariates, together with the interaction between them allow the hazards to vary linearly with time and allow this to differ by randomised treatment arm. The next step is to investigate increasingly complex smoothing models for the hazard, working through the approaches described in §5.6.1 to §5.6.4.

In each case, we fit a full interaction with treatment, to allow the hazards to vary by treatment group. Due to the complexity of the smoothing models, with additional knots and trying to estimate the sensitivity and specificity of positive and negative results can result in unrealistic estimates. In this case, we can use the estimated matrix of the hazard, misclassification probabilities and hazard ratios of the covariates as initial values[95] thus restarting the HMM. Additionally, the computation of these

complex models may break down. Here, we can take a simpler model (e.g. without estimating the misclassifications) and use the estimated matrix of the hazard and the hazard ratios of the covariates as initial values, until we are able to estimate the misclassifications. This process is performed iteratively until the preferred model is found thus improving the choice of the probabilities of the transition and misclassification matrices.

For each model we investigate, a graphical representation of the prevalence will be produced to visualise the goodness of fit of the model. Prevalence is defined as the proportion of patients observed to be in a positive or negative state over time, and the results from the Viterbi algorithm are used to assess prevalence if patient observation results are missing at any time point. This is compared against the expected prevalence which is forecasted from our preferred model. Since misclassifications are assessed, the expected prevalence of the observed (and therefore known) states are estimated from the assumed proportion occupying each state at the initial time using the fitted probability transitions (see §5.5.2 and §5.5.2)[104]. The expected prevalences of the known states is multiplied by the misclassification probabilities to obtain the expected prevalences of the observed states[104].

In a second stage, we derive estimated positive to negative probability transitions and negative to positive probability transitions from our preferred model. These probability transitions are then compared to the raw positive to negative probability transitions and negative to positive transitions from the raw data. We also compare these estimated probability transitions and raw transitions to probability transitions (positive to negative and negative to positive) from the two-fold fully conditional specification multiple imputation model (§3.6) to estimate the fit of these models to the raw data. Although the raw data have missing values, we hope that the chosen model will take these missing observations into account and so large departures from the raw data are unexpected.

Next, having chosen the preferred model, we use the forwards/backwards algorithm from §5.4.3 to find the probability of being in a positive state or a negative state at each time point. We use multiple imputation as set out in §5.5.1 to account for the uncertainty of the estimated probabilities. A total of 20 imputations will be used

(§5.5.1). Having used this algorithm, the outcome (treatment failure) of each patient is determined for each imputation set. These are then combined using Rubin's Rules[34]. The resulting estimates are then compared to those of the original study.

For completeness, we also use the Viterbi algorithm (see §5.4.4) which finds the most probably pathway of hidden states for each patient. The outcome of each patient can then be determined (see §5.5). These results are then compared to the estimates found in the original study. We then compare these results and the results from the forwards/backwards algorithm to the two-fold fully conditional specification multiple imputation model.

All analyses were implemented in R version 3.3.2 using Chris Jackson's *msm* package[95].

5.7.2 Results

In the original analysis, a total of 237 patients were excluded from the PP analysis and 111 were excluded from the mITT analysis. For these analyses, a total of 1785 patients were included in the analysis (see Table 3.1). Aggregating the number of transitions over the follow-up time and individual patients, we find most observations (n=11, 325) were negative to negative transitions, and around half of that (n=5480) were positive to positive transitions (Table 5.4). As noted earlier (see §5.7.1), most patients were in a positive state at the start of the study and most become negative over the first 3 months of follow-up. There were fewer negative to positive transitions (n=856) than positive to negative transitions (n=2078) and a non-trivial number of culture results were missing.

Table 5.5 shows the results from fitting different models (see 5.6.1 to 5.6.4) to the data and Figures 5.7, and 5.10 to 5.14 show the observed and expected prevalence for each fitted HMM.

Table 5.4: Total number of state transitions for all patients across all visits.

		$[To(S_t = j)]$		
		Positive	Negative	Missing
$[From(S_{t-1} = i)]$	Positive	5480	2078	1125
	Negative	856	11325	1897
	Missing	613	1975	3211

The first model is a simple HMM with no covariates included. The second model includes treatment, time and an interaction between the two as covariates. Models 3 to 6 investigate smoothing methods (a piecewise constant model, a linear splines model, a restricted cubic splines model and a second order fractional polynomial model) as described from §5.6.1 to §5.6.4. The knots chosen for the piecewise constant, linear splines and restricted cubic splines models were placed at 4, 8 and 26 weeks. This is a natural choice given that these are where visit windows were imposed in Chapter 3 and Chapter 4.

Model 1: No covariates

The first HMM with no covariates included in the model shows that on average, the probability of a patient being in a negative state in the next instant if they are currently in a positive state (i.e. presence of TB) is 0.14 and the probability of a patient having a positive culture result in the next instant if they are currently in a negative state (i.e. no presence of TB) is much lower at 0.005. The probability the true underlying state is positive given the observed state was negative is 7.6% (0.076; 95% CI: 0.069 to 0.085) and the probability that the true underlying state is negative given the observed state is positive is 1.6% (0.016; 95% CI: 0.013 to 0.019). As shown in Table 5.5, these misclassification probabilities are mostly consistent for all HMMs explored. The fitted, marginal prevalence from this model (Figure 5.7) shows that the model severely underestimates the proportion of patients in a negative state between 10 and 20 weeks and overestimates the proportion between 39 and 58 weeks.

Figure 5.8 shows the probability transitions with no covariates, assuming a constant hazard. This is compared to probabilities from the raw data for positive to negative and negative to positive transitions. This model overestimates the positive to negative transitions between weeks 6 and 17 and is underestimated from week 22 until the final follow-up visit at week 78. The model underestimates the probability of transitioning from a negative state to a positive state throughout the study. This suggests that this model is not a good fit to our data.

Figure 5.7: Estimated and observed marginal prevalence: no covariates included for REMoxTB.

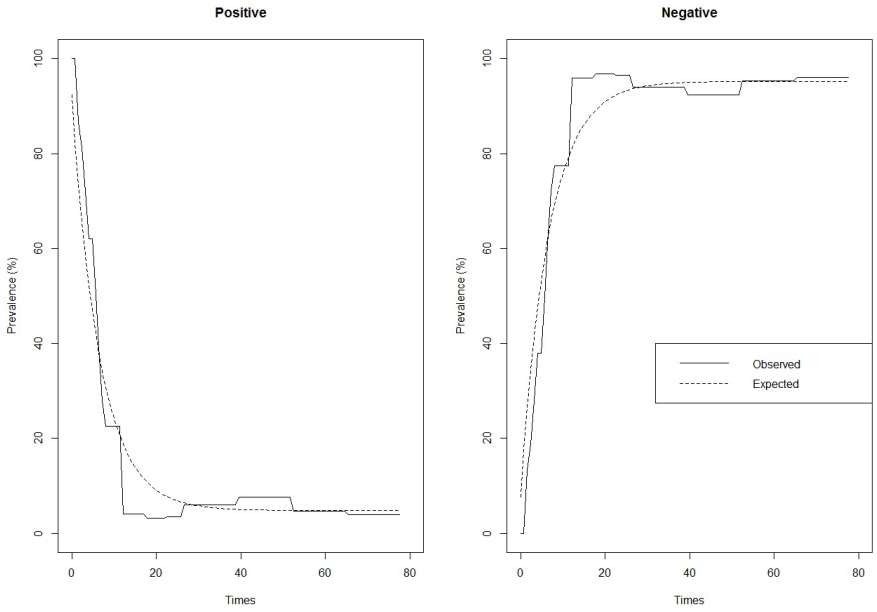
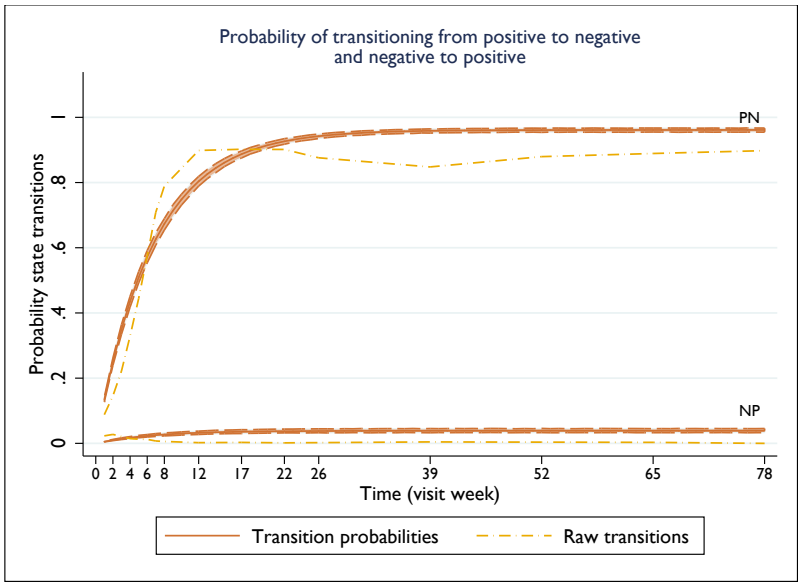


Figure 5.8: Estimated probability transitions, $P(S_t = j|S_0 = 1)$, with no covariates for REMoxTB.



PN: positive to negative transitions;

NP: negative to positive transitions.

Table 5.5: Different HMMs for REMoxTB.

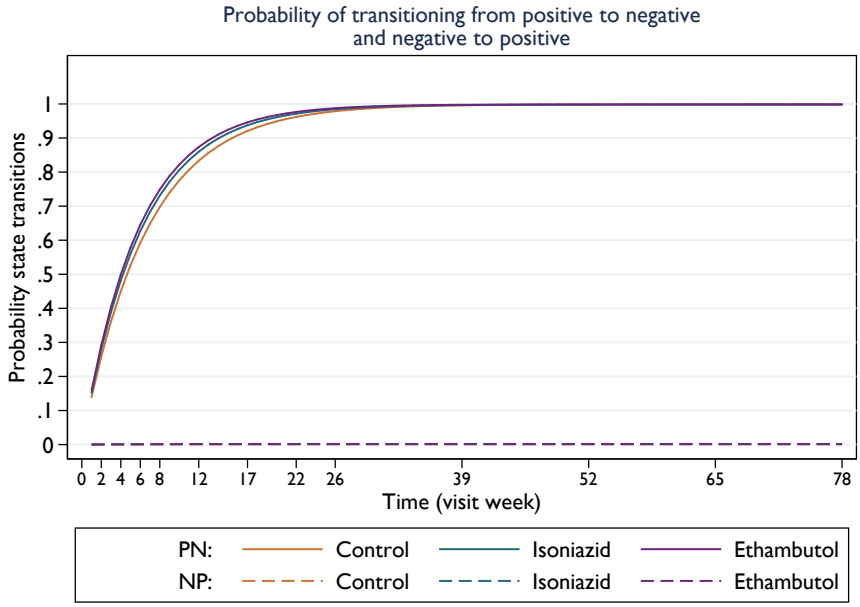
Model	Transition states $P(S_t = j \mid S_{t-1} = i)$ (95% CI) $P(Neg \mid Pos)$	$P(Pos \mid Neg)$	Misclassifications $P(O \mid S)$ (95% CI) $P(O_t = Pos \mid S_t = Pos)$	$P(O_t = Neg \mid S_t = Pos)$	$P(O_t = Pos \mid S_t = Neg)$	$P(O_t = Neg \mid S_t = Neg)$
No covariates						
Baseline hazard	0.145 (0.138, 0.153)	0.005 (0.004, 0.006)	0.923 (0.915, 0.931)	0.077 (0.069, 0.085)	0.016 (0.013, 0.019)	0.984 (0.981, 0.987)
-2 log-likelihood: 16550.9						
Treatment, week interaction						
Baseline hazard	9.812 (7.057, 13.642)	0.366 (0.238, 0.563)	0.920 (0.912,0.927)	0.080 (0.073,0.088)	0.020 (0.016,0.025)	0.980 (0.975,0.984)
Isoniazid	0.908 (0.681, 1.212)	0.746 (0.316, 1.763)				
Ethambutol	1.006 (0.753,1.345)	1.586 (0.6980,3.604)				
Week	1.326 (1.268,1.386)	1.293 (1.226,1.365)				
Isoniazid*Week	1.052 (0.989,1.119)	1.027 (0.965,1.094)				
Ethambutol*Week	1.089 (1.017,1.167)	1.051 (0.980,1.127)				
-2 log-likelihood: 16229.22						
Piecewise Constant (see 5.6.1)						
Baseline hazard	0.080 (0.056, 0.114)	0.006 (0.003, 0.011)	0.929 (0.920,0.937)	0.071 (0.063,0.080)	0.016 (0.012,0.022)	0.984 (0.978,0.988)
Isoniazid	0.886 (0.588, 1.334)	0.353 (0.1065, 1.172)				
Ethambutol	1.020 (0.798, 1.304)	0.570 (0.237, 1.371)				
Week	1.034 (0.956, 1.118)	0.980 (0.888, 1.082)				
Week_4	2.589 (1.712, 3.916)	0.060 (0.002, 1.814)				
Week_8	0.173 (0.018, 1.652)	0.552 (0.019, 16.334)				
Week_{26}	0.097 (0.006, 1.466)	0.292 (0.007, 12.481)				
Isoniazid*Week	1.068 (0.875, 1.304)	1.026 (0.920, 1.146)				
Ethambutol*Week	0.994 (0.915, 1.078)	1.030 (0.924, 1.148)				
Isoniazid*Week_4	0.879 (0.361, 2.141)	0.606 (0.006, 61.40)				
Ethambutol*Week_4	1.189 (0.746, 1.893)	0.495 (0.011, 22.82)				
Isoniazid*Week_8	0.032 (0.001, 0.8629)	4.765 (0.058, 388.92)				
Ethambutol*Week_8	0.231 (0.024, 2.246)	5.771 (0.125, 265.50)				
Isoniazid*Week_{26}	1.003 (0.0001, 8216.39)	0.634 (0.010, 41.54)				
Ethambutol*Week_{26}	10.028 (0.502, 200.14)	0.344 (0.005, 24.71)				
-2 log-likelihood: 15597.36						
Linear splines (see 5.6.2)						
Baseline hazard[1]	0.107 (0.095, 0.119)	0.008 (0.006, 0.011)	0.940	0.060	0.020	0.980
Isoniazid	0.791 (0.531, 1.180)	0.323 (0.014, 7.209)				
Ethambutol	0.889 (0.600, 1.319)	0.060 (0.001, 7.478)				
Week	1.3621 (1.239, 1.498)	0.7702 (0.490, 1.210)				
Week_4	0.887 (0.758, 1.038)	0.789 (0.405, 1.535)				
Week_8	0.698 (0.620, 0.785)	1.434 (0.974, 2.109)				
Week_{26}	1.186 (1.068, 1.317)	1.100 (0.941, 1.285)				

Isoniazid*Week	1.083 (0.950, 1.235)	1.094 (0.434, 2.756)				
Ethambutol*Week	1.054 (0.924, 1.202)	1.998 (0.534, 7.475)				
Isoniazid*Week$_4$	0.963 (0.774, 1.199)	1.016 (0.298, 3.461)				
Ethambutol*Week$_4$	0.973 (0.782, 1.212)	0.506 (0.110, 2.335)				
Isoniazid*Week$_8$	0.939 (0.801, 1.101)	1.019 (0.585, 1.777)				
Ethambutol*Week$_8$	0.970 (0.829, 1.134)	1.120 (0.685, 1.828)				
Isoniazid*Week$_{26}$	1.034 (0.910, 1.175)	0.841 (0.703, 1.006)				
Ethambutol*Week$_{26}$	1.028 (0.912, 1.159)	0.822 (0.687, 0.984)				
-2 log-likelihood: 15416.98						
Cubic splines (see 5.6.3)						
Baseline hazard[1]	0.116 (0.101, 0.133)	0.001 (0.0001, 0.006)	0.948	0.052	0.016	0.984
Isoniazid	0.778 (0.516, 1.174)	0.196 (0.011, 3.589)				
Ethambutol	0.947 (0.633, 1.417)	0.631 (0.092, 4.338)				
Week	1.420 (1.262, 1.600)	0.645 (0.434, 0.957)				
Week$_4$	7.407×10^{-7} (1.370×10^{-12}, 0.401)	41.25 (1.268×10^{-15}, 1.342×10^{18})				
Week$_8$	1.023×10^{8} (3.876×10^{-6}, 2.700×10^{21})	0.946 (3.875×10^{-37}, 2.307×10^{36})				
Week$_{26}$	0.239 (1.035×10^{-9}, 5.504×10^{7})	0.005 (6.673×10^{-24}, 3.806×10^{18})				
Isoniazid*Week	1.091 (0.925, 1.286)	1.231 (0.500, 3.031)				
Ethambutol*Week	1.039 (0.883, 1.223)	0.948 (0.437, 2.055)				
Isoniazid*Week$_4$	(2.103×10^{-10}, 2.685×10^{6})	(2.570×10^{-32}, 5.645×10^{30})				
Ethambutol*Week$_4$	0.027 (3.836×10^{-10}, 1.891×10^{6})	1882.0 (8.085×10^{-29}, 4.381×10^{34})				
Isoniazid*Week$_8$	876.147 (1.778×10^{-16}, 4.316×10^{21})	0.626 (2.793×10^{-67}, 1.402×10^{66})				
Ethambutol*Week$_8$	8229 (6.560×10^{-15}, 1.032×10^{22})	0.025 (2.153×10^{-72}, 2.861×10^{68})				
Isoniazid*Week$_{26}$	0.05617 (1.842×10^{-13}, 1.713×10^{10})	4.54062 (1.487×10^{-36}, 1.387×10^{37})				
Ethambutol*Week$_{26}$	0.003 (2.425×10^{-14}, 2.843×10^{8})	5.477×10^{-6} (8.153×10^{-48}, 3.679×10^{36})				
-2 log-likelihood: 15524.37						
Fractional polynomials (see 5.6.4)						
Baseline hazard[1]	0.007 (0.004, 0.015)	0.005 (0.002, 0.011)	0.96	0.04	0.04	0.96
Isoniazid	0.497 (0.301,0.821)	0.821 (1.751×10^{-9}, 3.854×10^{8})				
Ethambutol	0.695 (0.423, 1.143)	3.641 (2.543×10^{-9}, 5.213×10^{9})				
Week	0.295 (0.209, 0.415)	1.142 (0.529, 2.461)				
Week$^{FP_1(0.5)}$	1.923 (1.424, 2.598)	0.070 (0.0004, 13.445)				
Week$^{FP_2(0.5)}$	4.734 (3.033, 7.389)	1.049 (0.092, 11.897)				
Isoniazid*Week	1.158 (0.736, 1.823)	0.541 (0.142, 2.065)				
Ethambutol*Week	0.870 (0.516, 1.468)	0.504 (0.049, 5.242)				
Isoniazid*Week$^{FP_1(0.5)}$	1.760 (1.111, 2.787)	0.232 (1.527×10^{-7}, 3.535×10^{5})				
Ethambutol*Week$^{FP_1(0.5)}$	1.643 (1.011, 2.672)	0.158 (1.182×10^{-8}, 2.104×10^{6})				
Isoniazid*Week$^{FP_2(0.5)}$	0.719 (0.391, 1.324)	4.607 (0.026, 825.129)				
Ethambutol*Week$^{FP_2(0.5)}$	1.024 (0.526, 1.993)	5.084 (0.003, 8142.642)				
-2 log-likelihood: 15610.74						

[1] Misclassifications were fixed at for sensitivity and specificity.

In terms of covariates, we can see that time should be included in the model by the differing probability transitions when HMMs are fitted separately by treatment arm (Figure 5.9).

Figure 5.9: Estimated probability transitions, $P(S_t = j|S_0 = 1)$, modelled separately by treatment for REMoxTB.



PN: positive to negative transitions;

NP: negative to positive transitions.

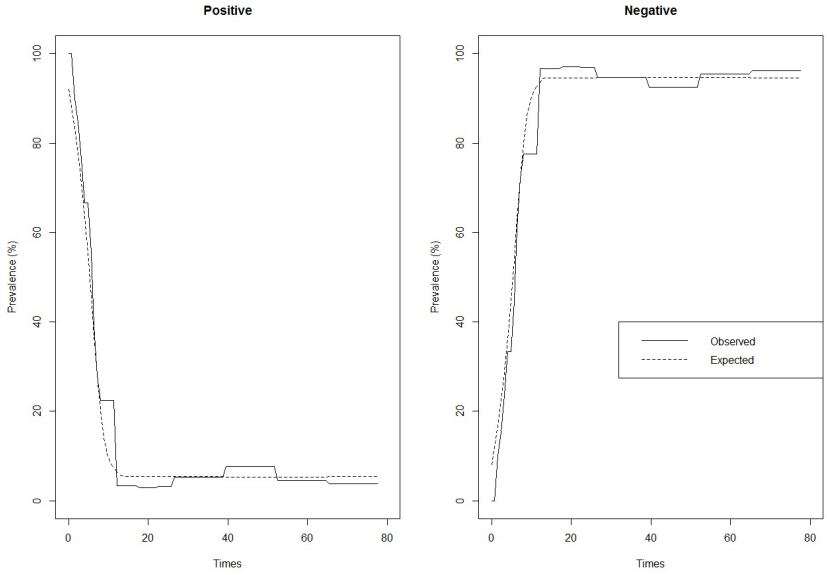
Model 2: Treatment, time and an interaction between treatment and time

The second HMM includes treatment, time as a time-varying covariate and an interaction between the two as covariates. The average risk of transitioning from a positive state to a negative state is quite large at 9.8, but the wider 95% confidence intervals (7.06 to 13.6) reflect some uncertainty around this (Table 5.5). It is unclear why the baseline hazard is so large, but may be an indication that this is not a suitable model. The average risk of transitioning from a negative state to a positive state is, on average, 0.37 (95% CI, 0.24 to 0.56). Patients who were randomised to the isoniazid treatment arm on average have a 9% (HR: 0.91; 95% CI, 0.68 to 1.2) reduction in

hazard of being in a negative state given their current state is positive and an even lower hazard (HR: 0.75; 95% CI, 0.32 to 1.76) of being in a positive state given their current state is negative. On average, patients who received ethambutol have a small increase in risk of transitioning from a positive state to a negative state and a 59% increase in hazard (1.59; 95% CI, 0.70 to 3.60) transitioning from a negative state to a positive state in the next instant, although the wide confidence intervals reflect some uncertainty around this. The hazard of transitioning from a positive to negative state increases by 33% (HR: 1.33, 95% CI 1.27 to 1.39) over time and the hazard increases by 29% (HR: 1.29, 95% CI, 1.23 to 1.37) for patients in a negative state who transition to a positive state over time ("Week" covariate; Table 5.5). The interaction terms show a 5% increase (1.05, 95% CI, 0.99 to 1.12) in hazard of transitioning from a positive state to a negative state for patients who received isoniazid over time and a 3% increase in hazard (1.03, 95% CI, 0.96 to 1.09) of transitioning from a negative to positive state over the duration of follow up. The interaction between the isoniazid arm and time (week) shows that the hazard of transitioning from a positive state to a negative state for patients who receive ethambutol increases by 9% over follow up time and the confidence intervals show that this is a significant interaction (95% CI: 1.02 to 1.17). The interaction between the ethambutol treatment arm and time (week) shows that the hazard of transitioning from a negative state to a positive state for patients who receive ethambutol increases by 5% over time (1.05, 95% CI: 0.98 to 1.13).

The prevalence when treatment, time and the interaction between the two is included (Figure 5.10) shows that this model is a better fit to the observed data than not including any covariates. However, the proportion of negative cultures being observed after 17 weeks from the forecasted model remains constant thus failing to capture the decreasing trend of having a negative culture between 26 weeks and 52 weeks.

Figure 5.10: Estimated and observed prevalence when treatment, time and their interaction are included for REMoxTB.



Model 3: Piecewise constant model

The results from the piecewise constant model suggest the model is a better fit than the interaction model as the -2 log-likelihood is lower at 15597.36 (Table 5.5), however the interaction terms between treatment and at week 8 for negative to positive transitions (HR: 4.765; 95% CI: 0.058 to 388.92 for isoniazid at week 8 and HR: 5.77; 95% CI: 0.125 to 265.50 for ethambutol at week 8) and the interaction terms between treatment and week 26 are also a poor fit reflected by the 95% confidence intervals (HR: 0.634; 95% CI: 0.010 to 41.54 for isoniazid at week 26 and HR: 0.344; 95% CI: 0.005 to 24.71 for ethambutol at week 26). The prevalence for the piecewise constant model (Figure 5.11) captures the decreasing incidence of having a negative culture between 12 and 25 weeks, but it is markedly underestimated.

Model 4: Linear splines model

Figure 5.12 shows the incidence of having a positive culture or negative culture over time for the linear splines model. This model does capture the decreasing trend of negative culture results which then levels out again, matching the observed data, to around 98% by the end of the study at week 78.

Figure 5.11: Estimated and observed prevalence for piecewise constant model with knots included at 4, 8 and 26 weeks for REMoxTB.

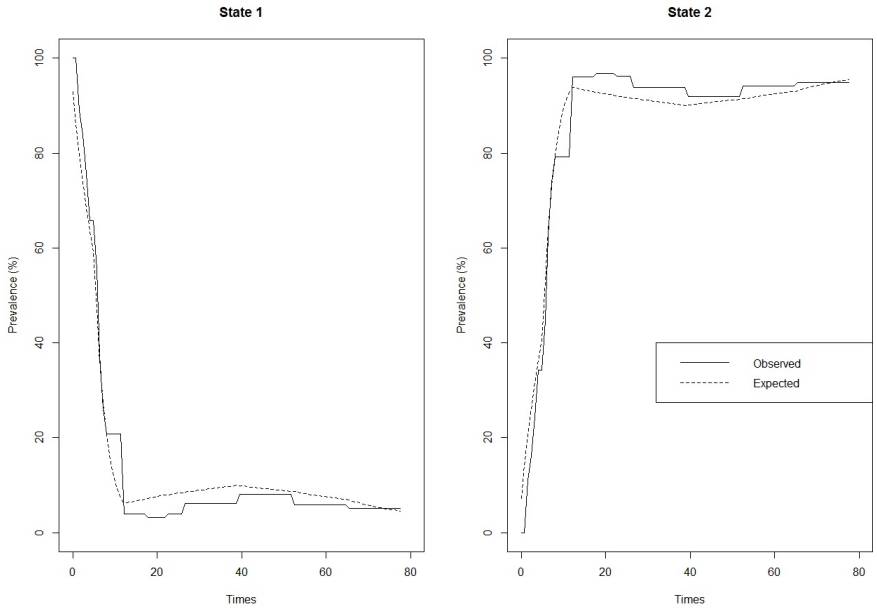
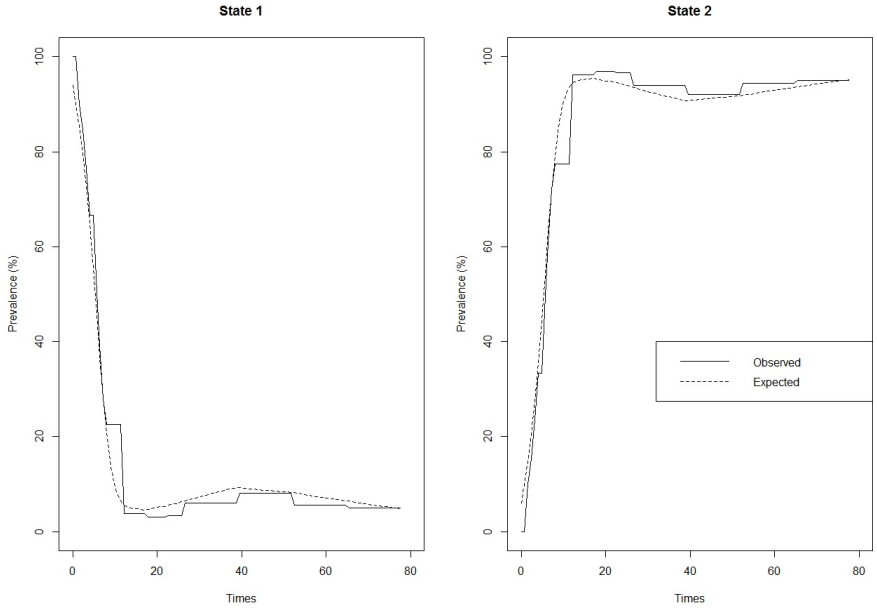


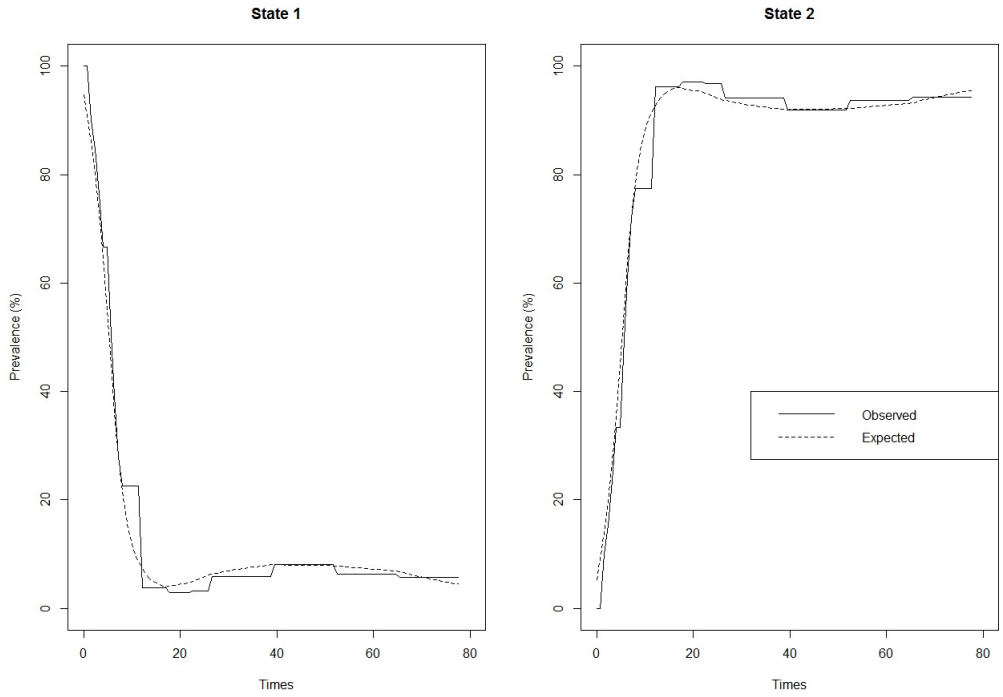
Figure 5.12: Estimated and observed prevalence for linear splines model with knots included at 4, 8 and 26 weeks for REMoxTB.



Model 5: Restricted cubic splines (RCS) model

The prevalence for the RCS model in Figure 5.13 shows the expected prevalence from the HMM also gives a similar fit to the observed data when RCS is used. However, looking at the results from this model shows unreasonable estimates suggesting the model is not as good as the linear splines model (Table 5.5). This is reflected in the resulting -2 log likelihoods which is 15524.37 with the RCS included and is lower for the linear splines model at 15416.98 (Table 5.5).

Figure 5.13: Estimated and observed prevalence for restricted cubic splines with knots included at 4, 8 and 26 weeks for REMoxTB.



Model 6: Fractional polynomials model

The second order fractional polynomials model was challenging to fit and so a fourth order fractional polynomials model was explored. However, this model failed to converge. Therefore we continued to use a second order fractional polynomial beginning with a simpler model, gradually building the complexity by fixing the misclassification to specific values and using the estimates (i.e. baseline hazard and

hazard ratios) from the working model as initial values. The width of the confidence intervals suggest that there is simply not enough information within the dataset to accurately estimate the baseline hazard and hazard ratios using fractional polynomials. The wide 95% confidence intervals resulting from the fitted fractional polynomials model (Table 5.5) show that the hazard of transitioning to a positive state when patients are currently in a negative state are not well estimated, since most of the confidence intervals are wide for the covariates. The most uncertainty surrounds the treatment covariates, where for the isoniazid arm there is a decrease in hazard of 18% (HR: 0.822; 95% CI: 1.751×10^{-9} to 3.854×10^{8}) and for the ethambutol arm the hazard of transitioning to a positive state given a patient is in a negative state increases by 3.641 (95% CI: 2.543×10^{-9} to 5.213×10^{9}). The expected prevalence (Figure 5.14) does not capture what happens at the end of the study as it underestimates the observed data at the end of the study at 78 weeks.

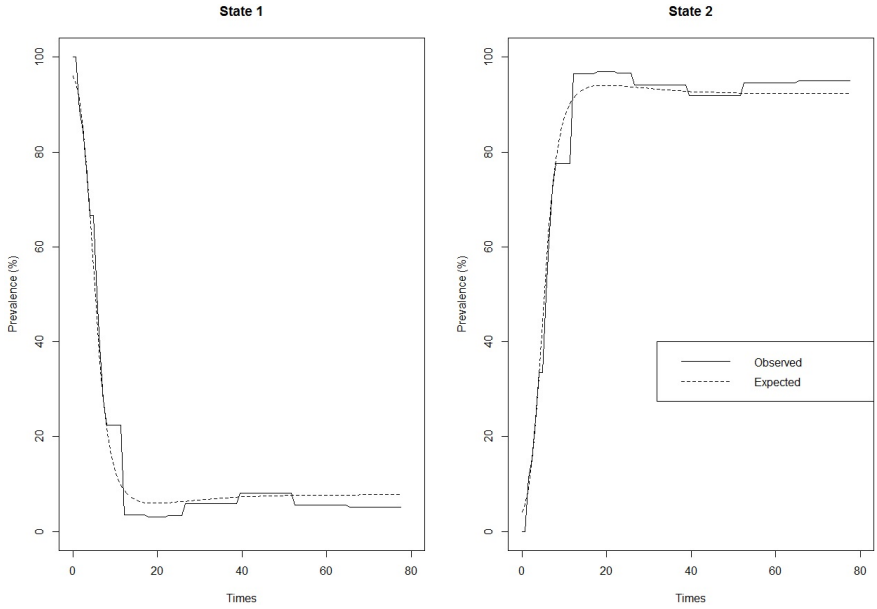
Summary of the model for REMoxTB

The linear splines model is the preferred HMM and is significantly better compared with just treatment, time and the interaction between them as covariates ($P<0.001$). Therefore the linear splines model is our chosen hidden Markov model for the REMoxTB study.

Probability transitions for the linear splines model

The linear splines model is our preferred model for the REMoxTB data. Figures 5.15 and 5.16 shows the probability of transitioning from a positive to negative state and a negative to positive state over time since randomisation from *t* weeks to *t+1* using the linear splines model from the HMM, including time as a time-varying covariate with knots at 4, 8 and 26 weeks. The probability transitions from this model are compared to the raw probability transitions from the REMoxTB dataset and the two-fold fully conditional specification multiple imputation model (see 3.5.3) over time from *t* weeks to *t+1*. The HMM shows that between 4 and 17 weeks there is a higher probability of transitioning from positive to negative on the treatment arms than on the control arm before levelling out (Figure 5.15). The shaded regions within these figures represent 95% confidence intervals for the probability transitions in each treatment arm over time.

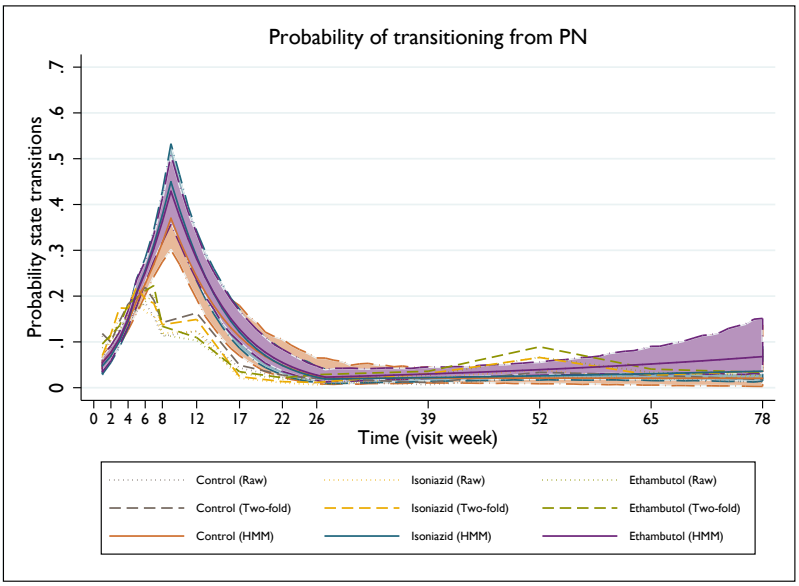
Figure 5.14: Estimated and observed prevalence for the fractional polynomials model for REMoxTB.



The probability transitions from the HMM overestimates the probabilities around 8 weeks, but fits a little better to the raw data than the two-fold imputation does towards the end of the study at around 52 weeks. The two-fold imputation fits better to the raw data for the negative to positive transition probabilities whereas the linear splines HMM is slightly underestimated in comparison to the raw data between weeks 26 to 52, but look reasonable towards the end of the study. At the very beginning, the HMM shows wide confidence intervals for negative to positive transitions. This could be because not many patients are negative in the first couple of weeks in the study, so there is little to no information in those first one to two weeks which the HMM has picked up on.

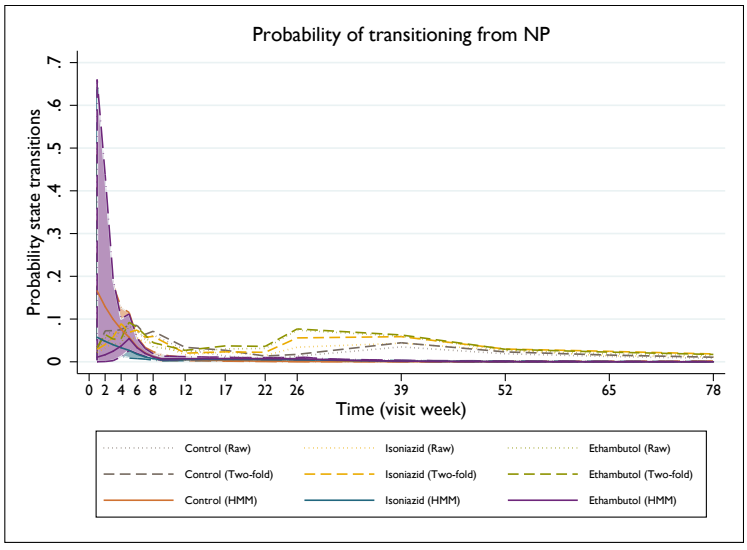
The preferred linear splines HMM was further investigated by fitting knots in different places for positive to negative transitions (see Appendix H). This was done to see if the choice of knots had a large influence on the estimates of the probability transitions thus reducing the overestimated positive to negative probability transitions between 8 to 12 weeks of follow-up. Our original linear splines model with knots at 4, 8 and 26 weeks still proved to be preferable to describe the REMoxTB data.

Figure 5.15: Positive to negative probability transitions (PN) for linear splines model with knots at 4, 8 and 26 weeks for REMoxTB.



PN: positive (P) to negative (N) transitions where $P(S_t = N|S_{t-1} = P)$.

Figure 5.16: Negative to positive probability transitions (NP) for linear splines model with knots at 4, 8 and 26 weeks for REMoxTB.

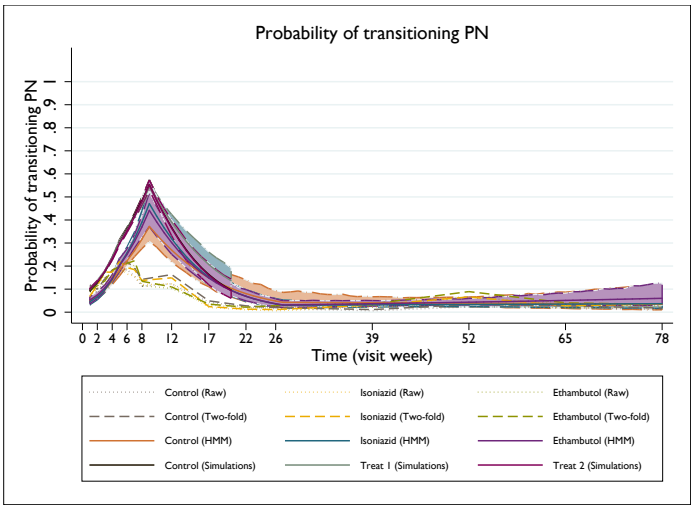


NP: negative (N) to positive (P) transitions where $P(S_t = P|S_{t-1} = N)$.

One possible reason for this apparent poor fit in the early stage of the follow-up visit for positive to negative transitions is that there is insufficient data to estimate the state transitions around this point. To explore this further, we proceeded to simulate data based on the results produced from the linear splines HMM with knots at 4, 8 and 26 weeks. Data were simulated for 30,000 patients, 10,000 patients in each treatment arm, taking time up to 20 weeks using the estimated hazard, hazard ratios, knots and misclassification probabilities from the linear splines model in Table 5.5. We then fitted the model using the data simulated, and include the results of the transition probabilities from these data in the transition probability figures. The results from fitting the HMM were similar to those used to simulate the data where the hazard of transitioning from a positive to negative state peaks around 0.5 around 8 weeks before reducing to around 0.1 by 20 weeks (Figure 5.17).

As a knot was already placed at 4 weeks, an extra knot at an earlier time point was chosen at 2 weeks to try and bring the estimates of the probability transitions closer to the raw data at the future 8 week follow up visit. The data were simulated again for 30,000 patients, 10,000 in each treatment arm. This brought the probability transitions between 6 to 17 weeks down matching closer to the probability transitions shown by the raw data (Figure 5.19). As this additional knot brought down the probability transitions, an extra knot was added at 2 weeks on the raw REMoxTB data. The resulting model matches closer to the raw probability transitions around 8 weeks but is still not as closely matched as the transition probabilities are from the two-fold multiple imputation model (see Table H1 and Figures H2 to H3 in Appendix H). The confidence intervals from the hazards and hazard ratios adding a fourth knot at 2 weeks on the REMoxTB data suggests that this is not the best model (Table H1), in particular for the negative to positive transitions, also reflected in the higher -2 log-likelihood of 15416.98. For negative to positive probability transitions, the simulations support the HMM fitted with and without the extra knot placed at 2 weeks (Figures 5.18 and 5.20). In both of these cases, the HMM and the simulated data underestimate the probability transitions during the continuation phase, remaining constant with a probability of 0 transitions.

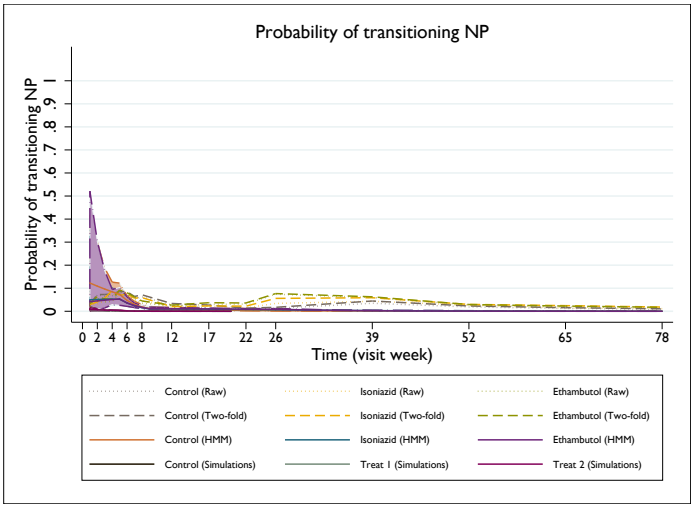
Figure 5.17: Simulated positive to negative probability transitions (PN) for linear splines HMM with knots at 4, 8 and 26 weeks for REMoxTB.



PN: positive (P) to negative (N) transitions where $P(S_t = N|S_{t-1} = P)$.

Data simulated from estimated hazards, hazard ratios and misclassifications from the linear splines HMM.

Figure 5.18: Simulated negative to positive probability transitions (NP) for linear splines HMM with knots at 4, 8 and 26 weeks for REMoxTB.

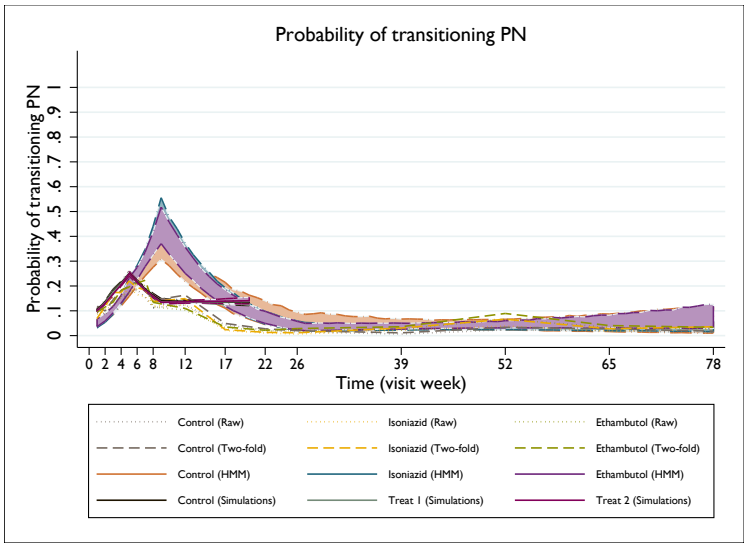


NP: negative (N) to positive (P) transitions where $P(S_t = P|S_{t-1} = N)$.

Data simulated from estimated hazards, hazard ratios and misclassifications from the linear splines HMM.

The addition of a knot at 2 weeks for the simulated data which were based on the linear splines model for the REMoxTB data worked well. Adding an extra knot at 2 weeks to the original REMoxTB data did not make a huge impact to reduce the estimated probability transitions around 8 to 12 weeks. This suggests that the HMM does not fit quite so well around 8 weeks of follow up. Additionally as an extra knot at 2 weeks on the simulated data was a closer fit to the raw data suggests that there is not enough data available to fit our complex model between 8 to 12 weeks of follow-up. For negative to positive transitions, data that were simulated suggested there were no transitions during follow-up. This further suggests that there is insufficient data available for the HMM to fit our data.

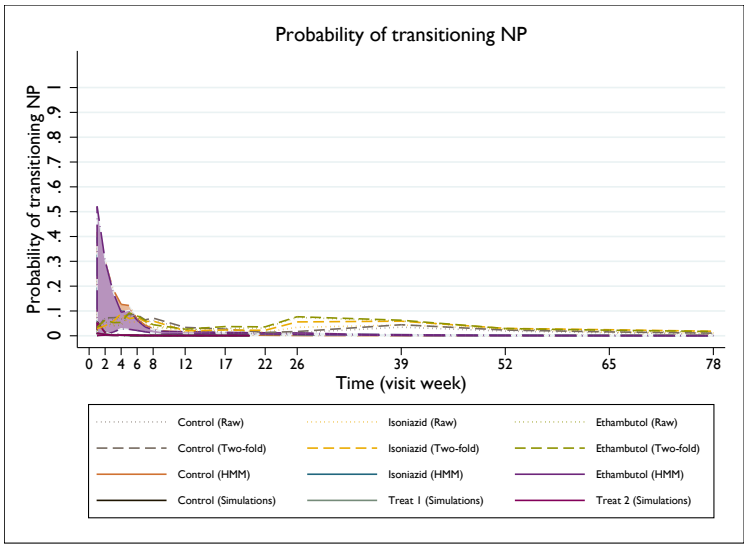
Figure 5.19: Simulated positive to negative probability transitions (PN) for linear splines HMM with knots at 2, 4, 8 and 26 weeks for REMoxTB.



PN: positive (P) to negative (N) transitions where $P(S_t = N|S_{t-1} = P)$.

Data simulated from estimated hazards, hazard ratios and misclassifications from a linear splines HMM with an additional knot at 2 weeks.

Figure 5.20: Including simulated negative to positive probability transitions (NP) for linear splines HMM with knots at 2, 4, 8 and 26 weeks for REMoxTB.



NP: negative (N) to positive (P) transitions where $P(S_t = P|S_{t-1} = N)$.

Data simulated from estimated hazards, hazard ratios and misclassifications from the linear splines HMM with an additional knot at 2 weeks.

We also explored adding an offset as a covariate to the linear splines model with knots placed at 4, 8 and 26 weeks. This is because given an expression for the hazard, $\lambda_{i,j}(t)$, for the transition intensities, the transition probabilities (e.g. from week 0 to week 1) are $e^{\lambda_{i,j}(t)t}$. This is not a linear expression in t_1, which may be a source of the relatively poor fit. A possible way to alleviate this is to include $log(t)$ as an offset in the log hazard ratio model:

$$\begin{aligned} \log[\lambda_{i,j}(t)] &= \log \lambda_{i,j}(t=0) - \log(t) + x^T\beta \\ i.e. \lambda_{i,j}(t) &= \frac{\lambda_{i,j}(t=0)e^{x^T\beta}}{t}. \end{aligned} \tag{5.35}$$

This is the matrix exponent expression to obtain the transition probabilities, which takes the extra t out. To remove this, potentially improving the fit of the model, we took time on a log scale and constrained it to -1. This made little difference to the overall fit of this model so we did not explore this further. Our final model to predict the missing culture data is therefore the HMM including a linear splines with 3 knots at weeks 4, 8, and 26 and no offset.

Prediction of states for the REMoxTB study

Table 5.6 shows the results of the primary outcome (failure) from using the forwards/backwards algorithm (see §5.4.1 and §5.4.2) and Viterbi algorithm (see §5.4.4) to predict states for missing observations, resulting in a "completed" dataset. We were then able to determine each patient's outcome. We calculated the difference in proportions of treatment failure (defined in §5.5) using a binomial model with an identity link. This model adjusted for weight and centre. The model struggled to converge when missing states were predicted using the Viterbi algorithm and so a cut-off of 1000 iterations was used.

The results from this model and from using the forwards/backwards algorithm were compared to the results from that of the two-fold fully conditional specification multiple imputation algorithm and the results produced from the authors of the study (Figures 5.21 and 5.22).

The results of the forwards/backwards algorithm and Viterbi algorithm are compared to the original results of the REMoxTB study (Table 5.6). There is a small gain in information using the forwards/backwards algorithm reflected by the slightly narrower confidence intervals. The results are consistent with the PP and mITT analyses. They fail to demonstrate non-inferiority since the upper bound of the 97.5% CI lies above the 6% non-inferiority margin for patients randomised to the isoniazid arm (11.14%) and ethambutol arm (12.25%). There is a larger gain in information using the Viterbi algorithm judged by the 97.5% confidence intervals. The results from this model are consistent with the forwards/backwards algorithm and the PP/mITT analyses, failing to demonstrate non-inferiority (upper bound of 97.5% CI:10.42% for isoniazid arm and 10.91% for ethambutol arm). Appendix I presents the unadjusted results for these analyses.

Table 5.6: Adjusted risk differences using the forwards/backwards algorithm and the Viterbi algorithm for REMoxTB.

	Risk difference (97.5% CI)
PP analysis (N = 1548)	
Isoniazid	6.10% (1.70% to 10.5%)
Ethambutol	11.40% (6.70% to 16.1%)
mITT analysis (N = 1674)	
Isoniazid	7.80% (2.70% to 13.00%)
Ethambutol	9.00% (3.80% to 14.20%)
Forwards/backwards algorithm	
Isoniazid	7.04% (2.94% to 11.14%)
Ethambutol	7.86% (3.46 to 12.25)
Viterbi algorithm	
Isoniazid	7.12% (3.82 to 10.42)
Ethambutol	7.46% (4.01% to 10.91%)

The results from the Viterbi algorithm are similar to the forwards/backwards algorithm and are consistent with the findings from the study where non-inferiority could not be concluded. To know which of the two performs best, a simulation study would be needed to find the bias and coverage of these two algorithms. However, as the models do not fit that well, we did not do this. The narrower confidence intervals from the Viterbi algorithm suggest a larger gain in information in comparison to the PP and mITT analyses. Predictions made from the Viterbi algorithm always identified relapses from the original data during the 78 weeks of scheduled follow-up.

The majority of patients who were excluded from the original mITT and PP analyses were imputed as achieving stable negative culture conversion from using the forwards/backwards algorithm (see Table I1, Appendix I).

Comparison of states predicted with the two-fold fully conditional specification multiple imputation model

The results from the forwards/backwards algorithm and Viterbi algorithm are compared to the two-fold fully conditional specification multiple imputation method (Figures 5.21 and 5.22). The results of these models are also consistent with the two-fold FCS multiple imputation model. The estimates from the two-fold imputation (7.07%; 97.5% CI: 1.84% to 12.30%) and the HMM (7.86%, 97.5% CI: 3.46% to 12.25%) for the ethambutol arm suggest those on the ethambutol regimen did better than that shown from the PP (11.4%, 97.5% CI: 6.70% to 16.10%) and mITT (9.00%, 97.5% CI: 3.80% to 14.20%) analyses. The estimates from the two-fold imputation suggest the isoniazid arm performed slightly better than that shown from the PP and mITT analyses and the estimates from the HMM are somewhere in between those of the PP and mITT analyses.

Figure 5.21: HMM (adjusted) estimates of primary endpoint using the forwards/backwards algorithm for the REMoxTB study.

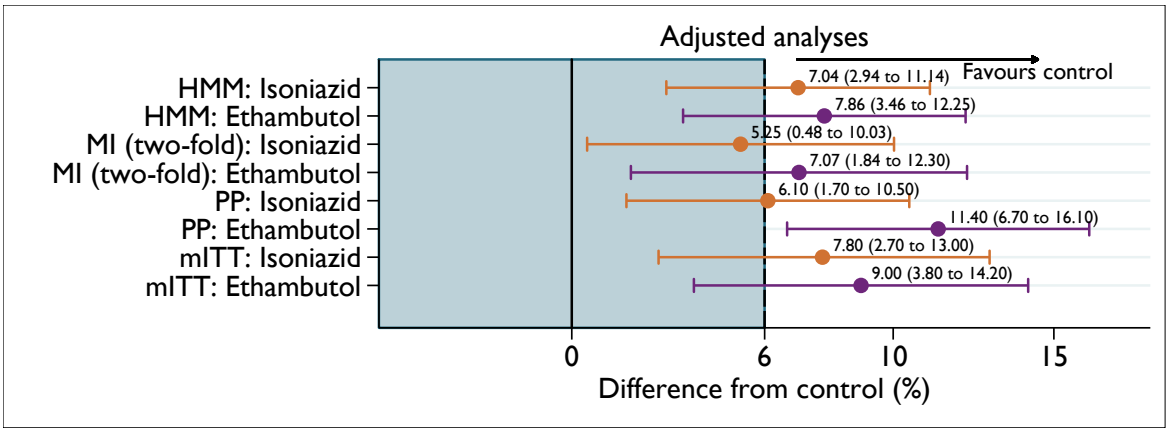
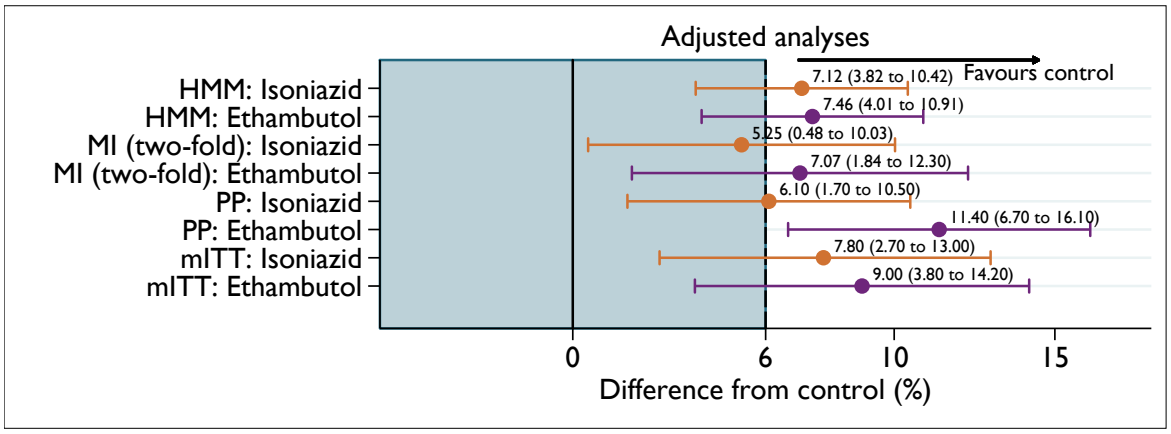


Figure 5.22: HMM (adjusted[1]) estimates of primary endpoint using the Viterbi algorithm for the REMoxTB study.



[1]1000 iterations

5.7.3 Discussion

So far, we have investigated whether using multi-state Markov models for the REMoxTB study works well to impute the missing observations resulting in a complete data set. A complete data set can then be used to determine each patient's clinical outcome in this study. Using multi-state models for the REMoxTB trial has worked reasonably well. As per the definition of the primary outcome, single positive results after reaching stable negative culture conversion are classed as negative and single negative results are ignored if stable negative culture conversion has not be achieved. Here we estimated the probability of false positive results and false negative results (i.e. the misclassifications). An alternative approach would be to have re-classed these false positive and negative results before using these multi-state models. However, this approach does not estimate the error of false positive or false negative culture results, which are of interest within the TB community, as a consequence of the MGIT machine or using LJ selection to determine a culture result. Following the modelling strategy defined in §5.7.1, our preferred model was the linear splines model. This model had the lowest -2 log-likelihood and this also gave the closest agreement between the model estimates of the expected prevalence and the direct comparison to the raw data.

However, when we calculated the transition probabilities obtained from the fitted model, and compared them with the raw data, we found the model overestimated the probability of transitioning from a positive to negative culture result around 8-12 weeks compared to the raw data. Clearly, the fit of the HMM depends on where the chosen knots are placed and how much information is provided into the HMM. However, we found little improvement by varying the knot positions. Nevertheless, the HMM still gives a good fit to the data during the follow-up phase (26 to 78 weeks) where most of the missing data occur and we rely on the HMM to impute this.

Using the forwards/backwards algorithm, accounting for the uncertainty of its predictions and using the Viterbi algorithm failed to demonstrate non-inferiority. While these different methods are consistent with that of results produced from the PP and mITT analyses, there is some, albeit small, gain in including the 10% of patients who were excluded from the mITT analyses due to withdrawal or lost to follow-up from the study. Given the context of the study, any gain in information is worthwhile.

Although the probability transitions from our chosen HMM was overestimated in the early part of follow-up, the approach appears promising. Therefore we now apply the same methods used here for the RIFAQUIN study. REMoxTB was unique in its design by including weekly follow-up visits in the first 8 weeks of the study. The RIFAQUIN study has fewer follow up visits and is a more typical representation of the amount of data collected in Phase III TB clinical trials.

5.8 Application to the RIFAQUIN study

The methods used for the REMoxTB study in §5.7 are now applied to the RIFAQUIN study. Patients are excluded from this intention-to-treat analysis for reasons unrelated to treatment (Table 3.7). Scheduled follow up visits differ here in comparison to REMoxTB; patients were assessed 2 months after baseline and monthly up to 12 months with two final visits at 15 and 18 months. For this study, patients were followed up less frequently in the first 2 months. We therefore assumed that the culture test results could fluctuate between observed follow up visits throughout the whole study, which reflects the true Markov process. The two states are a positive

culture result and a negative culture result. There were nearly twice as many deaths in the 4 month regimen treatment arm (n=12), however there was no association between the occurrence of death and treatment ($\chi^2 test = 2.582; p = 0.275$). Given that there were so few deaths overall, death is not included as a state (n=27).

To initialise the state transitions, as a working assumption we assume 80% of patients are in a positive state (i.e. have TB) and 1% are in a negative culture result at baseline, and assume a 95% sensitivity (i.e. true positive result) and 95% specificity (i.e. true negative result) as starting values in our model fitting. These values are used to initialise the state transitions and the misclassification matrix. The methods applied for RIFAQUIN then follow that of the REMoxTB study in §5.7.1, where first we choose our preferred model after smoothing the data and assessing the goodness of fit from prevalence plots. Then we compare the probability state transitions of this chosen HMM to the probability of the raw data before imputing missing cultures using the forwards/backwards algorithm. Finally, we use the Viterbi algorithm to find the most likely sequence of hidden states.

5.8.1 Results

The original results of the RIFAQUIN study excluded 313 patients from the PP analysis and excluded 304 patients from the mITT analysis. A total of 730 patients were included in the analyses for this study (see Table 3.7). Aggregating the number of state transitions over the follow-up time and individual patients, 5859 transitions were from a negative to a negative state (Table 5.7). There is a non-trivial number of missing culture results (n=1346). There were few positive to positive transitions (n=147) for all patients across all 14 scheduled follow up visits, although all patients were in a positive state at the start of the study (see §5.7.1). There are more positive to negative transitions (n=711) than negative to positive transitions (n=90), although there were more negative to missing transitions (n=707) than positive to missing transitions (n=111).

Table 5.7: Total number of state transitions for all patients across all visits.

		$[To(S_t = j)]$		
		Positive	Negative	Missing
$[From(S_{t-1} = i)]$	Positive	147	711	111
	Negative	90	5859	707
	Missing	14	505	1346

Table 5.8 shows the results from fitting different HMMs with increasing complexity of models (see §5.6.1 to §5.6.4) to the data. Figures 5.23 and 5.26 to 5.30 compares the forecasted prevalence with the expected prevalence to visualise the goodness of fit of these models. The knots chosen for the piecewise constant, linear splines and restricted cubic splines models were placed at 3, 6 and 10 months. This is a natural choice given that these are where visit windows were imposed in Chapter 3 and Chapter 4.

Table 5.8 shows that for most HMMs fitted the hazard of a patient being in a negative state if they are currently in a positive state is approximately 0.83 and the hazard of a patient being in a positive state at the next instant if they are currently in a negative state is low at around 0.02. In general, most HMMs explored suggest patients had an increase in hazard of transitioning from a negative state to a positive state for the 4 month regimen and for the 6 month regimen since the hazard ratios for these treatments are greater than 1. Similarly, there is an increase in hazard for patients transitioning from a positive to a negative state, since the estimates from the hazard ratios are greater than 1 for the 4 month and 6 month treatment regimens.

The misclassifications for all models explored were very low since the probability that the true underlying state is positive given the observed state was negative and the probability that the true underlying state is negative given the observed state is positive is nearly 0. This suggests there were few false negative culture results and false negative results detected. Models were re-run without assessing misclassifications and the estimates of the hazards and hazard ratios were

approximately the same for all models explored. Given that we are interested in the sensitivity, specificity and misclassifications of each for the RIFAQUIN study, we present these results.

Model 1: No covariates

The first HMM with no covariates added to the model shows that the probability of transitioning from a positive to negative state at the next instant is high at 1.108 (95% CI: 1.012 to 1.214) and the probability of transitioning from a negative to positive state is low at 0.024 (95% CI: 0.020 to 0.030).

The fitted, marginal prevalence from this model (Figure 5.23) shows that the model severely underestimates the proportion of patients in a positive state over the first 2 to 3 months of follow-up suggesting this model is not such a good fit to our data.

Figure 5.23: Estimated and observed marginal prevalence: no covariates included for RIFAQUIN.

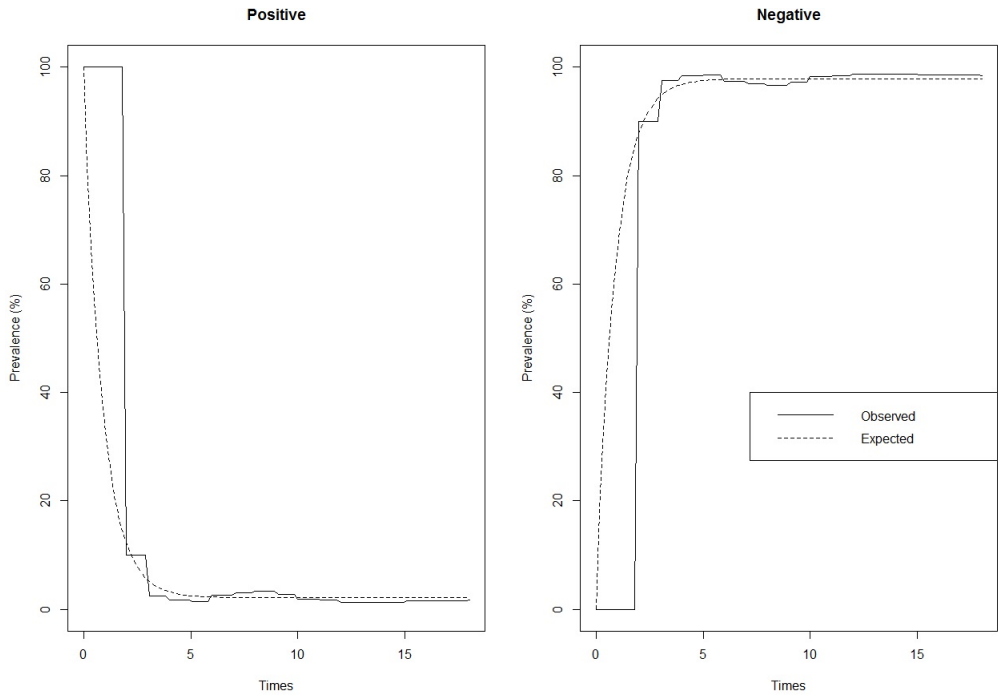
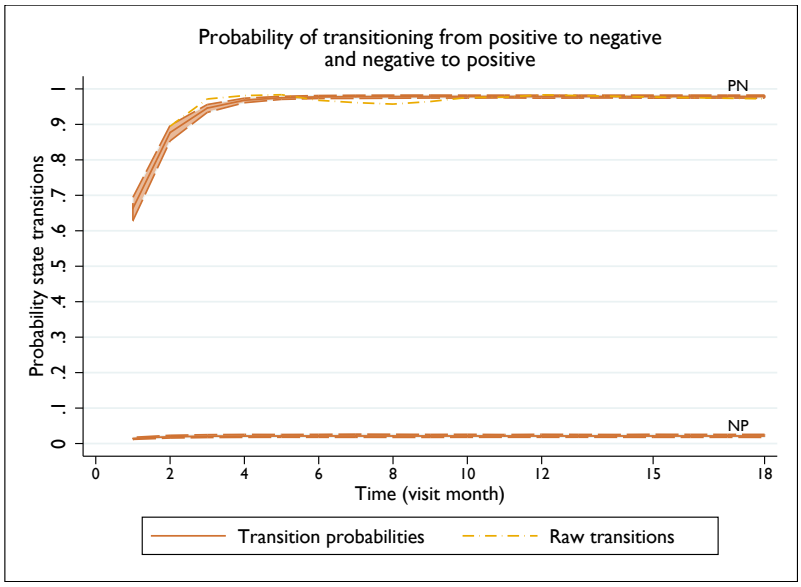


Figure 5.24 shows the probability transitions with no covariates, assuming a constant hazard. This HMM is compared to probabilities from the raw data for positive to negative and negative to positive transitions. The probability transitions from this model are well matched to the raw data.

Figure 5.24: Estimated probability transitions, $P(S_t = j|S_0 = 1)$, with no covariates for RIFAQUIN.



PN: positive to negative transitions;

NP: negative to positive transitions.

Table 5.8: Different HMMs for RIFAQUIN.

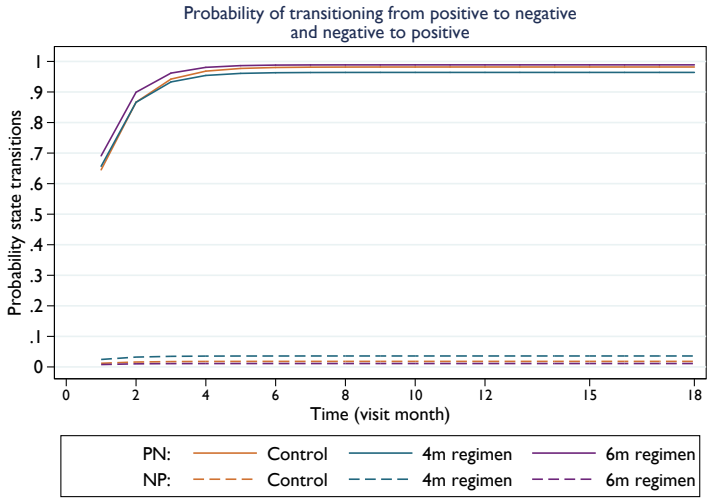
Model	Transition states $P(S_t = j \mid S_{t-1} = i)$ (95% CI)		Misclassifications $P(O \mid S)$ (95% CI)			
	$P(Neg \mid Pos)$	$P(Pos \mid Neg)$	$P(O_t = Pos \mid S_t = Pos)$	$P(O_t = Neg \mid S_t = Pos)$	$P(O_t = Pos \mid S_t = Neg)$	$P(O_t = Neg \mid S_t = Neg)$
No covariates						
Baseline hazard	1.108 (1.012, 1.214)	0.024 (0.020, 0.030)	1.000 (0.961, 1.000)	0.0002 (9.457×10^{-7}, 0.039)	4.989×10^{5} (2.419×10^{-7}, 0.0102)	1.000 (0.999, 1.000)
-2 log-likelihood: 1781.297						
Treatment, month interaction						
Baseline hazard	0.824 (0.679, 1.000)	0.019 (0.015, 0.024)	1.000 (0.960, 1.00)	0.0002 (1.14×10^{-6}, 0.041)	0.0001 (4.485×10^{-7}, 0.016)	1.000 (0.984, 1.000)
4m regimen	1.476 (1.039, 2.098)	5.259 (1.708, 16.19)				
6m regimen	1.163 (0.903, 1.500)	1.234 (0.304, 5.01)				
Month	0.969 (0.92, 1.02)	1.016 (0.914, 1.130)				
4m regimen*Month	0.919 (0.854, 0.989)	0.870 (0.760, 0.996)				
6m regimen*Month	0.984 (0.909, 1.064)	0.921 (0.778, 1.092)				
-2 log-likelihood: 1727.305						
Piecewise Constant (see 5.6.1)						
Baseline hazard	0.834 (0.624, 1.114)	0.019 (0.014, 0.026)	1.000 (0.939, 1.000)	0.0002 (5.768×10^{-7}, 0.061)	0.0001 (2.332×10^{-7}, 0.042)	0.999 (0.958, 1.000)
4m regimen	1.225 (0.937, 1.601)	1.651 (0.243, 11.20)				
6m regimen	1.176 (0.901, 1.535)	1.540 (0.181, 13.086)				
Month	1.242 (1.017, 1.517)	1.101 (0.766, 1.583)				
$Month_3$	0.257 (0.074, 0.888)	1.093 (0.194, 6.142)				
$Month_6$	0.277 (0.065, 1.188)	0.380 (0.066, 2.200)				
$Month_{10}$	0.676 (0.152, 3.013)	1.401 (0.245, 8.009)				
4m regimen*Month	0.832 (0.625, 1.108)	1.012 (0.637, 1.608)				
6m regimen*Month	0.919 (0.683, 1.236)	1.012 (0.569, 1.803)				
4m regimen*$Month_3$	2.841 (0.468, 17.24)	2.120 (0.238, 18.85)				
6m regimen*$Month_3$	7.579 (0.554, 103.62)	1.066 (0.041, 27.93)				
4m regimen*$Month_6$	1.935 (0.361, 10.366)	0.725 (0.0833, 6.303)				
6m regimen*$Month_6$	0.127 (0.007, 2.417)	0.358 (0.012, 10.384)				
4m regimen*$Month_{10}$	0.398 (0.042, 3.738)	0.120 (0.011, 1.288)				
6m regimen*$Month_{10}$	4.479 (0.427, 46.950)	0.951 (0.056, 16.280)				
-2 log-likelihood: 1687.292						
Linear splines (see 5.6.2)						
Baseline hazard	0.83577 (0.685, 1.020)	0.016 (0.012, 0.022)	1.000 (0.963, 1.000)	2.394×10^{-4} (1.514×10^{-6}, 0.037)	9.893×10^{-5} (6.410×10^{-7}, 0.015)	1.000 (0.985, 1.00)
4m regimen	1.163 (0.901, 1.503)	0.217 (0.008, 5.569)				
6m regimen	1.247 (0.819, 1.901)	3.737 (0.117, 119.7)				
Month	1.015 (0.897, 1.149)	1.102 (0.618, 1.966)				
$Month_5$	0.911 (0.725, 1.145)	0.888 (0.454, 1.737)				
4m regimen*Month	0.942 (0.802, 1.107)	1.863 (0.888, 3.91)				
6m regimen*Month	1.015 (0.812, 1.268)	0.700 (0.297, 1.650)				

4m regimen*Month$_5$	0.990 (0.728, 1.348)	0.406 (0.173, 0.951)				
6m regimen*Month$_5$	0.917 (0.607, 1.383)	1.369 (0.492, 3.809)				
-2 log-likelihood: 1705.719						
Cubic splines (see 5.6.3)						
Baseline hazard	0.830 (0.639, 1.078)	0.019 (0.014, 0.027)	1.000 (0.974, 1.000)	0.0003 (3.506$\times10^{-6}$, 0.026)	0.0001 (1.420$\times10^{-6}$, 0.010)	1.000 (0.991, 1.000)
4m regimen	1.348 (0.966, 1.882)	3.105 (0.758, 12.714)				
6m regimen	1.189 (0.891, 1.587)	1.385 (0.229, 8.393)				
Month	0.946 (0.878, 1.020)	0.961 (0.806, 1.145)				
Month$_5$	1.446 (0.577, 3.624)	1.610 (0.544, 4.766)				
4m regimen*Month	0.958 (0.871, 1.054)	0.971 (0.783, 1.205)				
6m regimen*Month	0.978 (0.866, 1.104)	0.899 (0.673, 1.200)				
4m regimen*Month$_5$	0.542 (0.185, 1.590)	0.391 (0.100, 1.535)				
6m regimen*Month$_5$	1.345 (0.101, 17.870)	1.495 (0.088, 25.510)				
-2 log-likelihood: 1723.325						
Fractional polynomials (see 5.6.4)						
Baseline hazard[1]	0.081 (4.206$\times10^{-7}$, 1.561$\times10^{4}$)	0.0002 (1.9$\times10^{-26}$, 2.4$\times10^{18}$)	0.980	0.020	0.020	0.980
4m regimen	1.310 (0.988, 1.737)	0.142 (0.000, ∞)				
6m regimen	1.235 (0.941, 1.622)	0.078 (0.000, ∞)				
Month	0.33 (2.69$\times10^{-9}$, 41330834)	0.088 (1.08$\times10^{-8}$, 714008)				
Month$^{FP_1(-0.5)}$	82.902 (1.004, 6844)	0.008467 (0.000, ∞)				
Month$^{FP_2(0)}$	0.9519 (1.2$\times10^{-24}$, 7.5$\times10^{23}$)	6.6$\times10^{5}$ (3.8$\times10^{-120}$, 1.2$\times10^{131}$)				
4m regimen*Month	2.750 (2.2$\times10^{-8}$, 3.4$\times10^{8}$)	0.120 (2.8$\times10^{-14}$, 4.9$\times10^{11}$)				
6m regimen*Month	3.547 (2.8$\times10^{-8}$, 4.6$\times10^{8}$)	0.867 (3.1$\times10^{-13}$, 2.4$\times10^{12}$)				
4m regimen*Month$^{FP_1(-0.5)}$	0.010 (8.401$\times10^{-5}$, 1.104)	0.004 (0.000, ∞)				
6m regimen*Month$^{FP_1(-0.5)}$	0.050 (0.0003, 8.224)	0.046 (0.000, ∞)				
4m regimen*Month$^{FP_2(0)}$	0.870 (1.1$\times10^{-24}$, 7.0$\times10^{23}$)	0.001 (1.6$\times10^{-216}$, 2.1$\times10^{225}$)				
6m regimen*Month$^{FP_2(0)}$	0.230 (2.2$\times10^{-25}$, 2.3$\times10^{23}$)	14.100 (4.4$\times10^{-246}$, 4.5$\times10^{247}$)				
-2 log-likelihood: 1763.083						

[1] Misclassifications were fixed at 98% for sensitivity and specificity.

Figure 5.25 shows that time should be included in the model since the probability transitions differ by treatment arm over time when these HMMs are fitted separately by treatment arm.

Figure 5.25: Estimated probability transitions, $P(S_t = j|S_0 = 1)$ modelled separately by treatment for RIFAQUIN.



PN: positive to negative transitions;

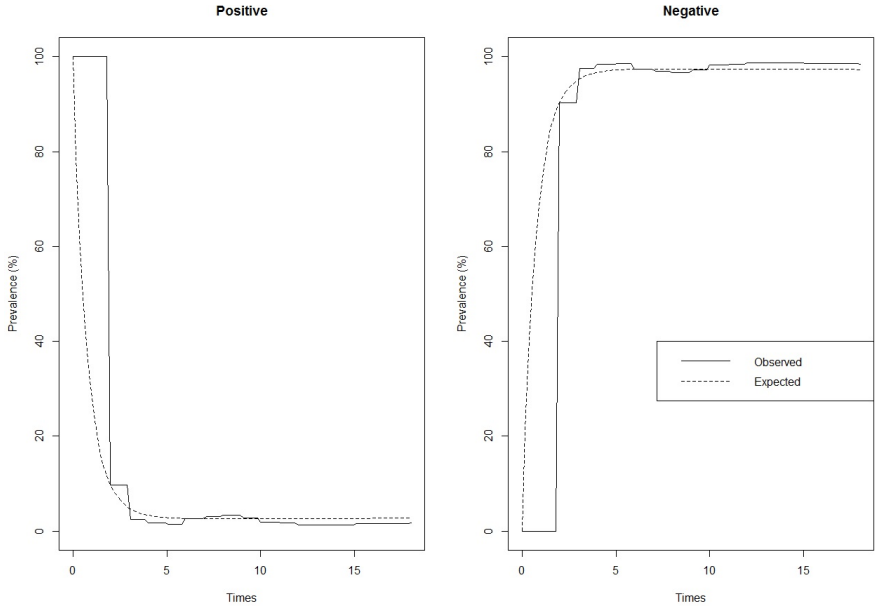
NP: negative to positive transitions.

Model 2: Including treatment, time and an interaction between them as covariates

The second model in Table 5.8 which includes treatment, time and the interaction between them as covariates shows that there is a significant interaction between the 4 month regimen and time. This is most likely due to the fact that patients do not receive a further 2 months of treatment compared to the control and 6 month treatment regimens. The risk of transitioning from a positive to negative state decreases by 8% (HR: 0.92; 95% CI: 0.85 to 0.99) as time increases. There is a 13% decrease in risk of transitioning from a negative to positive state (HR: 0.87; 95% CI: 0.76 to 1.00). Patients who received the 4 month regimen had a 48% increase in risk of transitioning from a positive to a negative state (HR: 1.48; 95% CI: 1.04 to 2.10) and a huge increase of risk transitioning from a negative to positive state (HR: 5.26; 95% CI: 1.71 to 16.19). However, there is greater uncertainty surrounding this as demonstrated

by the confidence intervals. Patients randomised to the 6 month treatment regimen have a 16% increase in risk of transitioning from a positive to a negative state (HR: 1.163; 95% CI: 0.90 to 1.50) and a 23% increase in risk of transitioning from a negative to a positive state (HR: 1.234; 95% CI: 0.304 to 5.01) on average. Figure 5.26 shows that the prevalence for this model slightly underestimates the proportion of patients in a negative state around 3 to 5 months and from 10 to 18 months of follow-up.

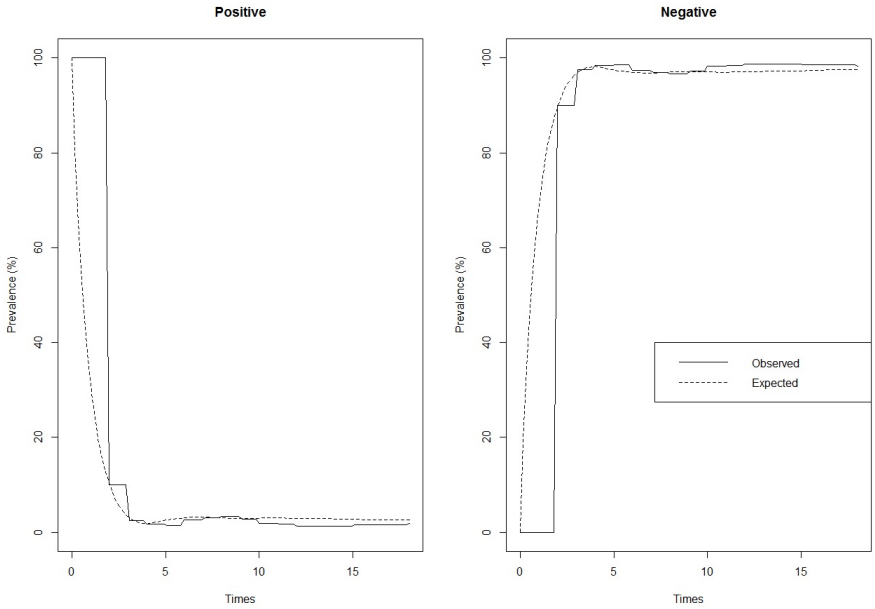
Figure 5.26: Estimated and observed prevalence when treatment, time and their interaction are included for RIFAQUIN.



Model 3: Piecewise constant

The prevalence for the piecewise constant model (Figure 5.27; negative state) fits slightly better around 5 months, supported by the -2 log-likelihood of 1687.29 (Table 5.8), compared to a simpler model with treatment, months and an interaction (Figure 5.26) between the two covariates. This is shown by the likelihood ratio test comparing this model to the HMM with treatment, months and an interaction. This test showed that the piecewise constant model is significantly better ($P<0.002$).

Figure 5.27: Estimated and observed prevalence for piecewise constant model with knots included at 3, 6 and 10 months for RIFAQUIN.

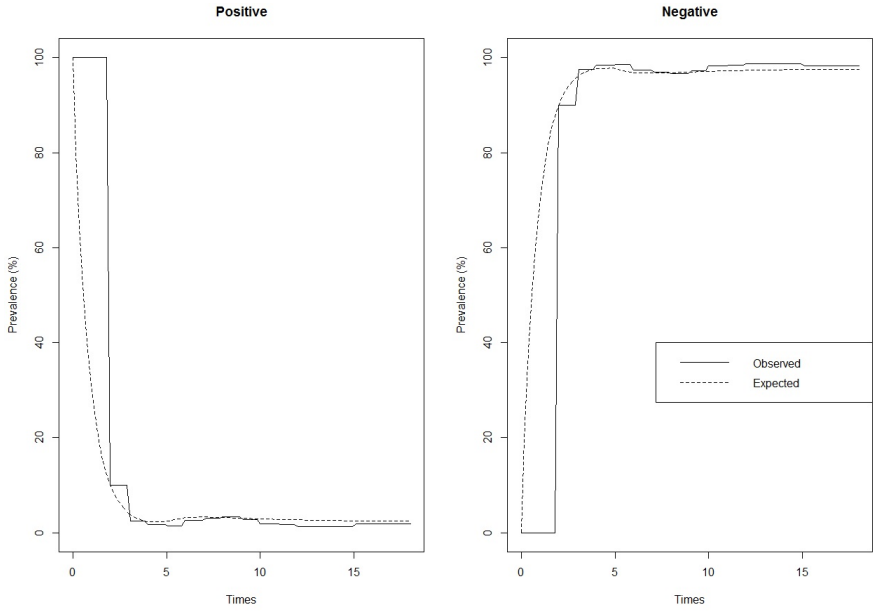


Model 4: Linear splines model

For the linear splines model, adding 3 knots at 3, 6 and 10 months failed to converge. Reducing the number of knots to 2 also proved challenging to fit. We therefore tried one knot at 3, 6 and 10 months separately. As a check, we tried re-fitting the model with one knot at each follow-up visit (monthly, from 2 to 12 months and at 15 and 18 months). From this process we found that adding one knot at 5 months was the better HMM for the RIFAQUIN data. For the 4 month regimen there is a 16% increase in risk transitioning from a positive to negative state (95% CI: 0.90 to 1.50) when a knot at 5 months is used (Table 5.8) and an increase of 25% in risk transitioning from a positive to negative state for the 6 month regimen. There is a 78% reduction in risk transitioning from a negative to positive state (HR: 0.22, 95% CI; 0.008 to 5.57) for the 4 month regimen and, although there is greater uncertainty surrounding the estimate, the hazard of transitioning from a negative to positive state on the 6 month regimen is far larger (HR: 3.74, 95% CI; 0.117 to 119.7). The expected prevalence from this HMM suggests this model provides a reasonable fit to the data after 2 months of follow-up (Figures 5.28). However, the estimates of the hazard ratios from this HMM contradict

the results of the original study, where the 4 month regimen showed more relapses and failed to demonstrate non-inferiority at the 6% margin. The expectation would be to have a higher increase in risk (i.e. a hazard greater than 1) of transitioning from a negative to a positive culture to reflect that the 4 month regimen did not perform as well as the control regimen and to have a higher hazard than the 6 month regimen given that the 4 month regimen failed to demonstrate non-inferiority but the 6 month regimen did. Given that the results from this HMM go against our intuition of what the RIFQUIN study showed, we reject this model.

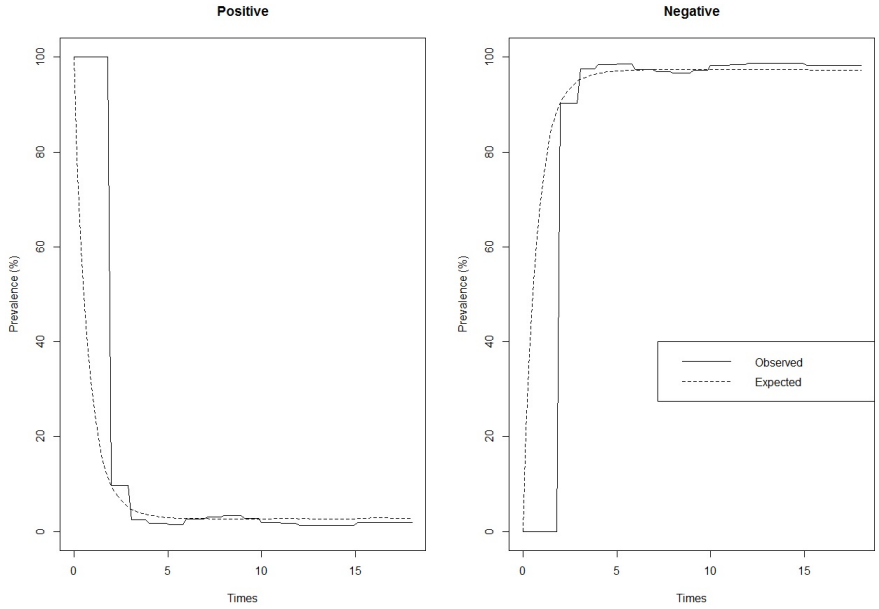
Figure 5.28: Estimated and observed prevalence for linear splines model with knots included at 5 months for RIFAQUIN.



Model 5: Restricted cubic splines model (RCS)

The prevalence for the RCS model in Figure 5.29 shows the expected prevalence from the HMM is not such a good fit to the observed data since, for the proportion of patients in a negative state, the model underestimates the observed data around 3 to 6 months and towards the end of follow-up (around 10 to 18 months). The resulting -2 log likelihood suggests that this model (-2 log likelihood=1723.3) is not as good as the piecewise constant model, which is 1687.3 (Table 5.8).

Figure 5.29: Estimated and observed prevalence for restricted cubic splines model with knots included at 5 months for RIFAQUIN.

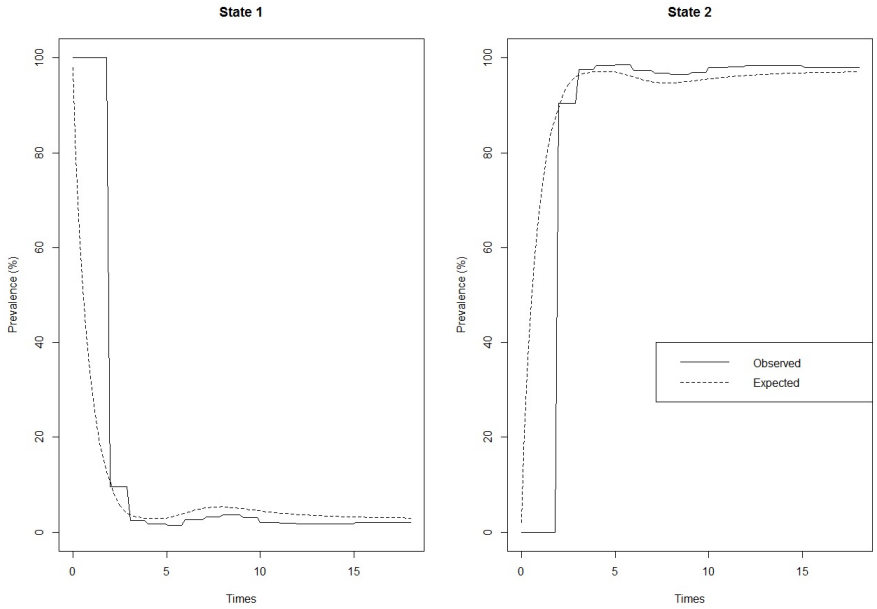


Model 6: Fractional polynomials model

The fractional polynomials model was unable to estimate the misclassifications and therefore these were fixed at 2%; the confidence intervals demonstrate the difficulty of fitting and interpreting this model. However the average intensity is consistent with other HMMs fitted. Figure 5.30 shows that the expected prevalence from this HMM is poorly fitted to the observed data for the proportion of patients in a negative state since the model underestimates the observed data from 3 months until the end of follow-up at 18 months. This in conjunction with the resulting estimates of the confidence intervals produced from the fractional polynomials HMM model suggest that this is a poor model.

We therefore choose the piecewise constant as our preferred HMM for the RIFAQUIN study.

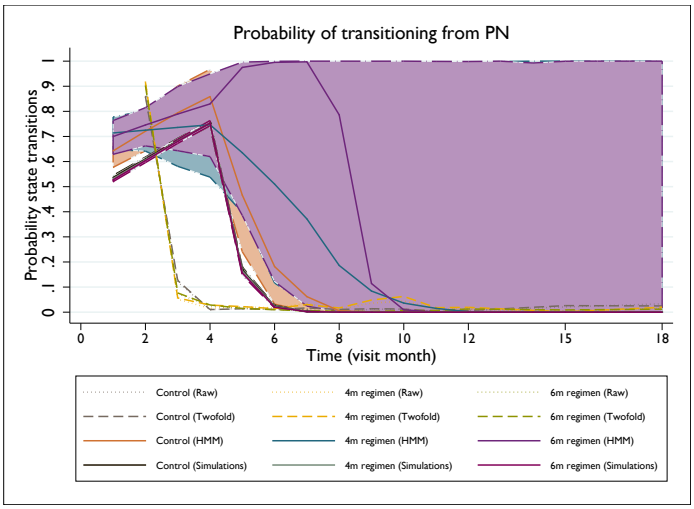
Figure 5.30: Estimated and observed prevalence for fractional polynomials model with knots included at 5 months for RIFAQUIN.



Probability transitions for the piecewise constant model

Probability transitions from the piecewise constant HMM are compared to probability transitions from the raw data (Figure 5.31 and 5.32). We also compare probability transitions from the two-fold fully conditional specification multiple imputation model. These figures suggest that the piecewise constant HMM is a poor fit in comparison to the raw data, particularly for the 4 month and 6 month treatment regimens. These figures show that the two-fold fully conditional specification multiple imputation model fits closer to the raw positive to negative and negative to positive probability transitions. The positive to negative transition probabilities from this HMM follow the same pattern as for the raw probability transitions, where the probability of transitioning sharply decreases and then levels out. Similarly, the probability of transitioning from negative to positive follows a similar pattern to that of the raw transitions where an increase in negative to positive transitions is captured around 5 to 10 months for the control and 4 month regimens. The shaded regions within these figures represent 95% confidence intervals for the probability transitions in each treatment arm over time.

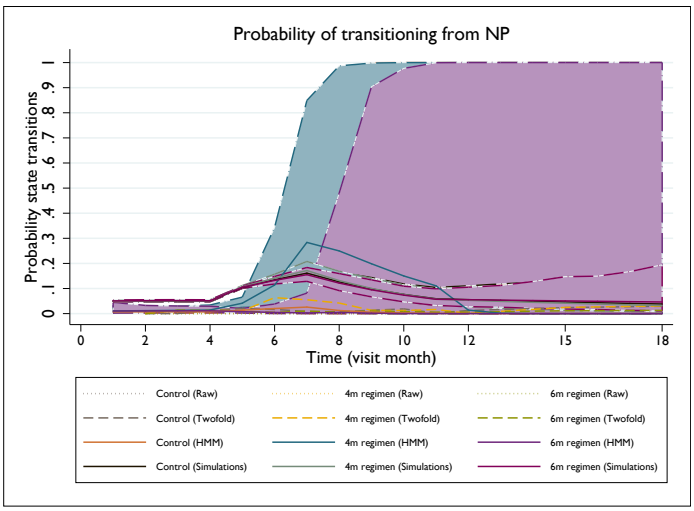
Figure 5.31: Comparison of simulated positive to negative probability transitions (PN) to the piecewise constant HMM at 3, 6, and 10 months for RIFAQUIN.



PN: positive (P) to negative (N) transitions where $P(S_t = N|S_{t-1} = P)$.

Data simulated from estimated hazards, hazard ratios and misclassifications from the piecewise constant HMM.

Figure 5.32: Negative to positive probability transitions (NP) with piecewise constants at 3, 6, and 10 months compared to data simulated for RIFAQUIN.



NP: negative (N) to positive (P) transitions where $P(S_t = P|S_{t-1} = N)$.

Data simulated from estimated hazards, hazard ratios and misclassifications from the piecewise constant HMM.

We further explored this piecewise constant model with alternative knots for positive to negative transitions (see Appendix J). This was done to investigate whether the choice of the knots made a huge difference to the estimated probability transitions. The piecewise constant model was still the preferred model.

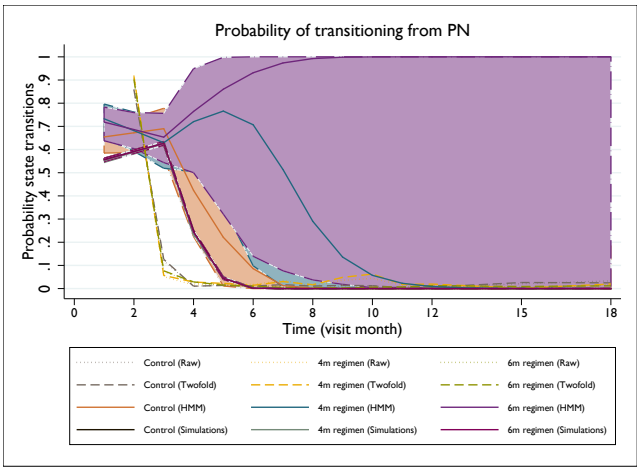
To investigate whether there was not enough data in this study to estimate the state transitions, we simulated data for 30,000 patients (10,000 in each arm) over 18 months. Data were simulated based on the results of the piecewise constant HMM using the estimated hazard, hazard ratio, knots and misclassifications from the model in Table 5.8. We then fitted a piecewise constant model to the simulated data and these results are included in the transition probability figures (Figures 5.31 and 5.32). By simulating data, and therefore having more data in the model, improved the fit of the HMM (Figure 5.31). The probability of transitioning from a positive to a negative culture result match closer to the raw data and mirror the pattern of positive to negative probability transitions, but this still seems to be a poor fit. The probability of transitioning from a negative to positive culture result was also improved using simulated data with more pronounced negative to positive transitions around 4 to 10 months.

We proceeded to a simpler piecewise constant model adding one knot at 2 months to see whether this would bring the probability transitions any closer to the raw data, since this is where there is a sharp decrease in positive to negative probability transitions. The resulting simulations from this model showed some improvement (Figure 5.33 and 5.34), shifting closer to where the raw probability transitions are. We then went back to the original RIFAQUIN data and fitted a piecewise constant HMM with one knot at 2 months. The results from this model (Table 5.9) were similar to the piecewise constant model with knots at 3, 6 and 10 months for positive to negative transitions. For negative to positive transitions, the estimates and confidence intervals were much larger suggesting this model would not be a good fit. The probability transitions from fitting this model to the data were poorer for the negative to positive probability transitions (Figure 5.34) and for the 6 month regimen for positive to negative transitions. This further suggests there is insufficient data (i.e. not enough patients randomised in the study) for the HMM to fit well to our data and that the model is not well-fitted to the data for the RIFAQUIN study.

Table 5.9: Piecewise constant imposed at 2 months only for RIFAQUIN.

Model	Transition states[1] $P(S_t = j \mid S_{t-1} = i)$ (95% CI) $P(Neg \mid Pos)$	$P(Pos \mid Neg)$	Misclassifications $P(O \mid S)$ (95% CI) $P(O_t = Pos \mid S_t = Pos)$	$P(O_t = Neg \mid S_t = Pos)$	$P(O_t = Pos \mid S_t = Neg)$	$P(O_t = Neg \mid S_t = Neg)$
Piecewise Constant (see 5.6.1)						
Baseline hazard	0.799 (0.649, 0.984)	0.017 (0.012, 0.024)	1.000 (0.350, 1.000)	6.0×10^{-5} (1.763×10^{-9}, 0.650)	2.0×10^{-5} (7.387×10^{-10}, 0.418)	0.999 (0.582, 1.000)
4m regimen	1.270 (0.952, 1.694)	13.4 (0.089, 2015.6)				
6m regimen	1.216 (0.924, 1.600)	12.12 (0.08, 1835.2)				
Month	1.05 (0.931,1.185)	1.030 (0.908,1.169)				
$Month_2$	0.452 (0.146, 1.398)	6.459 (0.045, 881.24)				
4m regimen*Month	0.822 (0.697, 0.970)	0.821 (0.700, 0.963)				
6m regimen*Month	0.869 (0.729, 1.036)	0.904 (0.741, 1.103)				
4m regimen*$Month_2$	3.438 (0.801, 14.75)	0.637 (0.004, 110.9)				
6m regimen*$Month_2$	3.297 (0.684, 15.89)	0.114 (0.001, 21.54)				
-2 log-likelihood: 1717.477						

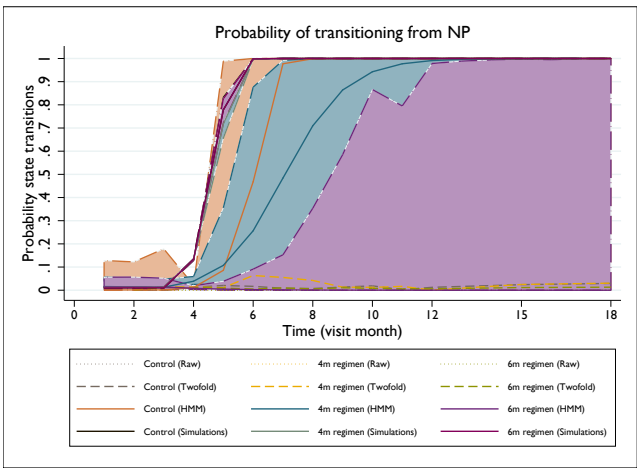
Figure 5.33: Positive to negative probability transitions (PN) with piecewise constants at 2 months compared to data simulated for RIFAQUIN.



PN: positive (P) to negative (N) transitions where $P(S_t = N|S_{t-1} = P)$.

Data simulated from estimated hazards, hazard ratios and misclassifications from a piecewise constant HMM with one knot at 2 months.

Figure 5.34: Negative to positive probability transitions (PN) with piecewise constants at 2 months compared to data simulated for RIFAQUIN.



NP: negative (N) to positive (P) transitions where $P(S_t = P|S_{t-1} = N)$.

Data simulated from estimated hazards, hazard ratios and misclassifications from a piecewise constant HMM with one knot at 2 months.

As for the REMoxTB study (see equation 5.35), we investigated whether the addition of an offset as a covariate to the model would improve the overall fit of the piecewise constant model with knots at 3, 6 and 10 months. The addition of an offset was computationally impossible to fit. We therefore choose the piecewise constant model with knots placed at 3, 6 and 10 months, with no offset, as the final HMM for the RIFAQUIN study to predict the missing culture data.

Prediction of states for the RIFAQUIN study

After using the forwards/backwards algorithm to impute the missing data, resulting in a complete data set, each patient's outcome was determined (see §5.5). This was also done using the Viterbi algorithm. The results from each of these algorithms are compared to the original study results of the RIFAQUIN study over 18 months. Table 5.10 shows the results from the forwards/backwards algorithm and Viterbi algorithm based on the piecewise constant HMM is consistent with the mITT analysis and is similar to the PP analysis. The 4 month regimen fails to demonstrate non-inferiority (upper bound of the 95% CI: 16.9%). As for the REMoxTB study, most patients from the forwards/backwards algorithm had their missing state imputed as negative.

Table 5.10: Adjusted risk differences using the forwards/backwards algorithm and the Viterbi algorithm for RIFAQUIN.

	Risk difference (97.5% CI)
PP analysis (N = 514)	
4 month regimen	13.60% (7.00% to 20.20%)
6 month regimen	-1.80% (-6.90% to 3.30%)
mITT analysis (N = 593)	
4 month regimen	13.10% (5.60% to 20.60%)
6 month regimen	0.40% (-5.70% to 6.60%)
Forwards/backwards algorithm	
4 month regimen	10.42% (3.91% to 16.92%)
6 month regimen	-1.80% (-6.10% to 2.51%)
Viterbi algorithm	

4 month regimen	7.66% (3.01 to 12.32)
6 month regimen	1.00% (-3.75% to 1.75%)

The Viterbi algorithm is also consistent with the analyses from the mITT analysis for the 4 month regimen failing to demonstrate non-inferiority (upper bound of the 95% CI: 12.3%). For both treatment regimens there is a much larger gain in information from the HMM using the Viterbi algorithm reflected in the narrower confidence intervals. However the Viterbi algorithm calculates the most probable pathway for a patient overall rather than imputing missing observations at each point based on the observed data, and so uncertainty of an imputed state cannot be calculated. Predictions made from the Viterbi algorithm could identify relapses from the original data over 18 months of follow-up.

Comparison of states predicted with the two-fold fully conditional specification multiple imputation model

Figures 5.35 and 5.36 show the results from comparing the forwards/backwards algorithm and Viterbi algorithm with the two-fold fully conditional multiple imputation model. These results were adjusted for centre of recruitment (see Appendix K for unadjusted results). The results from using the forwards/backwards algorithm are similar to the two-fold FCS multiple imputation and supports the PP analysis demonstrating non-inferiority on the 6 month regimen (upper bound of the 95% CI: 2.51%). The confidence intervals show a small gain in information using a piecewise HMM compared to the two-fold fully conditional specification multiple imputation model for the 4 month regimen and a slightly bigger gain in information for the 6 month regimen. The point estimates and confidence intervals from the HMM and two-fold FCS multiple imputation tend more towards favouring treatment compared to the PP and mITT analysis which tend towards the control regimen.

The results from using the Viterbi algorithm are also consistent with the two-fold fully conditional specification multiple imputation algorithm. The confidence intervals from the Viterbi algorithm are much narrower for the 6 month treatment regimen (-1.00%; upper bound of the 95% CI: 1.75). This suggests a large gain in information for this treatment arm.

Figure 5.35: Analysis of RIFAQUIN using the forwards/backwards algorithm (adjusted analysis).

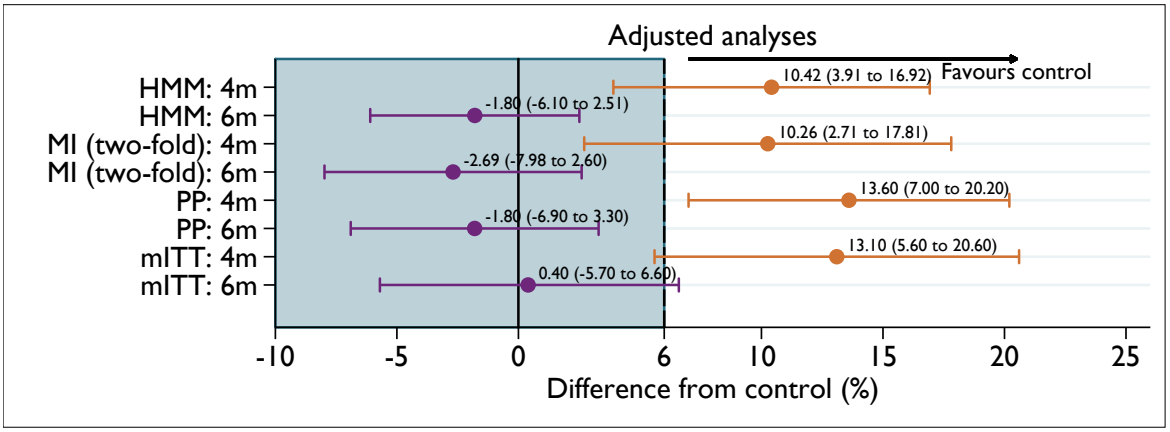
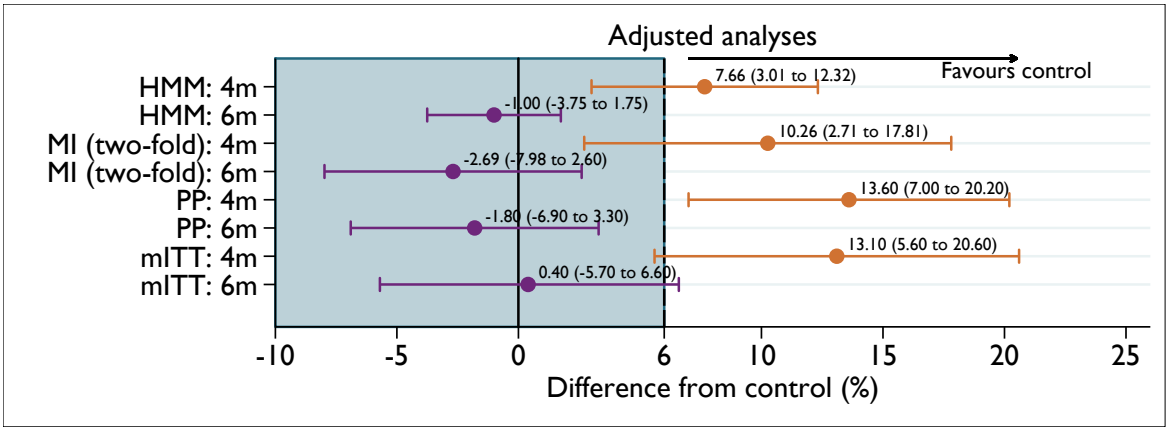


Figure 5.36: Analysis of RIFAQUIN using the Viterbi algorithm (adjusted analysis).



5.8.2 Discussion

For the RIFAQUIN study, under an ITT type analysis, we investigated whether using multi-state models to impute the missing observations resulting in a complete dataset works well. With an imputed, "complete", dataset each patient's clinical outcome can be readily determined. For the RIFAQUIN study which had fewer follow-up visits and a smaller sample size than the REMoxTB study, the application of these methods did not work so well.

The estimated sensitivity and specificity suggest that there were few, if any, false negative states or false positive states. Following the modelling strategy defined in §5.7.1, the preferred model was the piecewise constant model with knots placed at 3, 6 and 10 months. This model also had the lowest -2 log-likelihood supporting that this model was the better fit out of the models investigated. The expected prevalence of being in a positive or negative state compared to the observed data also showed that this model was reasonable.

Upon examining the probability transitions, calculating these probabilities from the fitted model and comparing them with the raw data, this model was not an altogether satisfactory fit. Even choosing different knots failed to make any large improvements to these probabilities. After simulating data based on the piecewise constant HMM, there was some improvement in the positive to negative and negative to positive state transitions, but not enough to match that of the raw transition probabilities. Using a simpler model and reducing the number of knots to one at 2 months, simulating the data again only showed a marginal improvement to the fit of raw transition probabilities. Increasing the number of knots to four was computationally impossible to fit. This suggests that there is not enough information from the RIFAQUIN study to fit the HMM and that the model is not such a good fit to the data.

The results from using the forwards/backwards algorithm accounting for the uncertainty in the predictions made and using the Viterbi algorithm were nonetheless broadly consistent with the results of the study. For both analyses, the 6 month regimen supported the PP analysis for demonstrating non-inferiority and this was also found using the two-fold fully conditional specification multiple imputation method. The results from the HMM favoured the treatment arm than the control when all patients were included in the analysis. This suggests that the loss of information when patients were excluded from the original analysis provides more conservative results.

The fitted probability transitions for the REMoxTB and RIFAQUIN studies suggest the Markov model does not seem to fit so well to the raw data. It is possible that the trial data is not solely dependent on the previous time point, but also on previous time points. We next investigate whether or not the data are in fact Markov.

5.9 Are the data truly Markov?

To further explore what may lie behind the relatively poor fit of the model estimates of the probability transitions to the raw data, we investigate whether or not the data are truly Markov. From a clinical perspective, we know that patients who achieve stable culture conversion (i.e. two consecutive negative results at separate visits) early on in the study (around 8-12 weeks) are expected to maintain that status until the final follow-up visit. However, those that have a mixture of positive and negative culture results within 8 to 12 weeks of follow-up or those who take longer to culture convert are expected to have a lower probability of maintaining negative status towards the end of the study. This clinical perspective, that what happens around 8 to 12 weeks is informative for what happens later in follow-up, suggests the data may not be Markov.

We investigate this in the REMoxTB study and then the RIFAQUIN study. We take the culture result at each visit from the second follow-up visit onwards and estimate the dependence on the culture result at the previous two visits, and the previous visit. Logistic regression of the culture result at visit $t-1$ and $t-2$ (from the second visit onwards) was performed separately for each treatment arm. Additionally, backwards stepwise logistic regression was also performed using the entire history of patient data available to predict the final 18 month visit. For this stepwise regression model, we include the immediate preceding visit (65 weeks for the REMoxTB study and 15 months for the RIFAQUIN study) and a threshold of 5% is used to determine whether dependency on the previous time point is statistically significant.

If the transition probabilities are first order Markov, we should find a strong dependence on the (positive or negative) culture result at $t-1$, but given this a relatively weak, generally non-significant, dependence on the culture result at $t-2$.

Table 5.11: Predictors of culture results from time t at observations $t-1$ and $t-2$, by treatment arm for REMoxTB.

Treatment	Week, OR (95% CI)	OR (95% CI) at t-1	P-value	OR (95% CI) at t-2	P-value
Control (N = 590)	2	9.06 (5.08 to 16.14)	< 0.001	4.69 (2.22 to 9.91)	< 0.001
	3	6.58 (3.80 to 11.40)	< 0.001	4.66 (2.69 to 8.08)	< 0.01
	4	7.13 (4.33 to 11.75)	< 0.001	4.62 (2.73 to 7.80)	< 0.001
	5	7.57 (4.70 to 12.19)	< 0.001	6.27 (3.63 to 10.84)	< 0.001
	6	6.22 (3.89 to 9.94)	< 0.001	5.66 (3.39 to 9.45)	< 0.001
	7	9.25 (5.43 to 15.76)	< 0.001	6.11 (3.48 to 10.73)	< 0.001
	8	7.07 (4.03 to 12.39)	< 0.001	3.79 (2.20 to 6.52)	< 0.001
	12	3.95 (1.65 to 9.46)	0.002	3.74 (1.59 to 8.81)	0.603
	17	11.82 (3.07 to 45.47)	< 0.001	0.94 (0.20 to 4.39)	0.937
	22	48.00 (9.47 to 243.40)	< 0.001	2.56 (0.30 to 21.85)	0.391
	26	17.29 (2.95 to 101.13)	0.002	14.87 (3.35 to 66.08)	< 0.001
	39	15.29 (2.39 to 98.03)	0.004	15.82 (4.03 to 62.05)	0.164
	52	33.45 (11.12 to 100.66)	< 0.001	5.06 (0.55 to 46.22)	< 0.001
	65	36.90 (10.11 to 134.62)	< 0.001	17.60 (5.06 to 61.22)	0.151
	78	96.00 (21.41 to 430.38)	< 0.001	22.95 (5.51 to 95.57)	< 0.001
Isoniazid (N = 609)	2	7.76 (3.70 to 16.27)	< 0.001	3.12 (1.20 to 8.10)	0.019
	3	6.61 (3.72 to 11.73)	< 0.001	7.42 (3.69 to 14.95)	< 0.001
	4	6.86 (4.36 to 10.79)	< 0.001	3.69 (2.16 to 6.32)	< 0.001
	5	8.78 (5.46 to 14.11)	< 0.001	4.30 (2.64 to 7.01)	< 0.001
	6	10.18 (6.27 to 16.54)	< 0.001	11.63 (6.30 to 21.48)	< 0.001
	7	10.58 (6.16 to 18.18)	< 0.001	11.39 (6.10 to 21.28)	< 0.001
	8	8.73 (4.87 to 15.65)	< 0.001	4.50 (2.52 to 8.04)	< 0.001
	12	1.46 (0.30 to 7.05)	< 0.001	0.636 (0.14 to 3.09)	0.801
	17	7.05 (1.38 to 36.09)	0.019	1.18 (0.25 to 5.54)	0.603
	22	63.26 (15.56 to 257.24)	0.009	14.45 (3.29 to 63.47)	0.832
	26	70.83 (17.74 to 282.75)	< 0.001	14.42 (3.39 to 61.24)	< 0.001
	39	28.05 (11.38 to 69.14)	< 0.001	7.61 (2.22 to 26.16)	< 0.001
	52	19.78 (8.16 to 47.93)	< 0.001	10.96 (4.05 to 29.67)	0.001

	65	89.40 (29.53 to 270.60)	< 0.001	10.97 (4.55 to 26.45)	< 0.001
	78	56.62 (18.37 to 174.44)	< 0.001	41.42 (14.47 to 118.60)	0.001
Ethambutol (N = 586)	2	5.90 (3.15 to 11.03)	< 0.001	4.48 (1.87 to 10.74)	0.001
	3	13.01 (7.23 to 23.42)	< 0.001	5.27 (2.80 to 9.92)	< 0.001
	4	12.45 (7.23 to 21.44)	< 0.001	8.77 (4.75 to 16.20)	< 0.001
	5	6.34 (4.07 to 9.88)	< 0.001	4.75 (2.88 to 7.81)	< 0.001
	6	7.08 (4.46 to 11.25)	< 0.001	9.37 (5.36 to 16.38)	< 0.001
	7	5.67 (3.44 to 9.34)	< 0.001	8.80 (4.70 to 16.49)	< 0.001
	8	11.80 (6.20 to 22.45)	< 0.001	6.77 (3.44 to 13.30)	< 0.001
	12	5.35 (2.05 to 13.96)	0.001	1.47 (0.55 to 3.93)	< 0.001
	17	5.54 (1.44 to 21.37)	0.013	4.53 (1.57 to 13.07)	0.446
	22	31.34 (10.73 to 91.57)	0.734	< 0.001 (0.96 to 13.69)	0.009
	26	12.07 (4.65 to 31.30)	< 0.001	6.20 (2.09 to 18.38)	0.057
	39	27.74 (12.69 to 60.64)	< 0.001	7.58 (2.92 to 19.67)	0.001
	52	13.94 (6.21 to 31.26)	< 0.001	10.81 (4.60 to 25.44)	< 0.001
	65	29.67 (10.93 to 80.56)	< 0.001	5.61 (2.16 to 14.56)	< 0.001
	78	31.65 (9.97 to 100.50)	< 0.001	28.65 (9.77 to 84.00)	< 0.001

For the REMoxTB study, Table 5.11 shows that, across all treatment arms, there is a steady dependence of culture results at t on culture results at $t-1$. However, for all treatment arms, around weeks 3 to 12, we find additional dependence on the culture results at t-2. This suggest that when patients transition from a positive state to a negative state, this is not well modelled by the Markov assumption. As this assumption underpins the HMM, this is a plausible explanation for the observed and fitted transition probabilities in weeks 6 to 12.

A total of 334 patients were included for the backwards stepwise regression model. Table 5.12 confirms that the data are not Markov since having a positive culture result at week 6 (P=0.003), 7 (P=0.041), 26 (P=0.002) is predictive of results at week 78 (P<0.003) as well as the previous 65 week scheduled visit (P<0.001) for all patients.

Table 5.12: Odds ratios (OR), and confidence intervals (CI) for predicting positive cultures at week 78 for REMoxTB.

Covariate	OR (95% CI)	P-value
Week 6	0.189 (0.038 to 0.933)	0.003
Week 7	9.739 (2.163 to 43.852)	0.041
Week 26	10.405 (2.371 to 45.654)	0.002
Week 65	63.674 (12.592 to 321.980)	< 0.001

For the RIFAQUIN study, the logistic regression models to assess time at $t-1$ and $t-2$ often failed due to perfect prediction or collinearity. Therefore, since the results for the REMoxTB study were similar across treatment groups, we combine treatment groups for the RIFAQUIN study.

Table 5.13 shows dependence of culture results at time t on culture results at $t-1$. There is also steady dependence of culture results at time t on culture results at $t-2$. This suggests that when patients transition from a positive to negative state and a negative to positive state, this is not well modelled by the Markov assumption.

Table 5.13: Predictors of culture results from time t at observations $t-1$ and $t-2$, by treatment arm for RIFAQUIN.

Month, OR	OR (95% CI) (95% CI)	P-value at t-1	OR (95% CI)	P-value at t-2
3	NA		27.40 (8.29 to 90.59)	< 0.001
4	12.32 (3.64 to 41.72)	< 0.001	32.91 (8.29 to 130.62)	< 0.001
5	10.23 (1.96 to 53.43)	0.006	16.17 (2.98 to 87.76)	0.001
6	10.23 (1.96 to 53.43)	0.006	16.04 (3.75 to 68.59)	< 0.001
7	7.20 (1.41 to 36.89)	0.018	34.05 (11.33 to 102.31)	< 0.001
8	11.95 (3.36 to 42.48)	< 0.001	340.90 (68.24 to 1702.89)	< 0.001
9	48.40 (15.13 to 154.85)	< 0.001	121.25 (36.39 to 403.99)	< 0.001
10	20.85 (6.05 to 71.86)	< 0.001	62.09 (17.36 to 222.06)	< 0.001
11	38.25 (10.00 to 146.38)	< 0.001	310.67 (59.54 to 1621.01)	< 0.001

12	83.04 (18.31 to 376.59)	< 0.001	182.00 (35.22 to 940.52)	< 0.001
15	25.06 (5.31 to 118.39)	< 0.001	23.78 (3.85 to 146.94)	0.001
18	24.53 (3.95 to 152.56)	0.001	12.79 (2.29 to 71.45)	0.004

*NA = results not presented due to perfect prediction or collinearity

Table 5.14: Odds ratios (OR), and confidence intervals (CI) predicting positive cultures at month 18 for RIFAQUIN.

Covariate	OR (95% CI)	P-value
Month 3	94.987 (5.984 to 1507.754)	0.001
Month 15	2.038 (0.070 to 59.205)	0.001

The stepwise logistic regression shows that positive cultures at month 3 (P=0.002) was predictive of positive cultures at month 18 in addition to the previous visit at month 15 (P=0.001; Table 5.14). This also suggests that positive culture results in the early part of follow-up are predictive of positive results at the final follow-up visit.

For both the REMoxTB and RIFAQUIN studies, at week t, some culture results are dependent on culture results at $t - 2$. This suggests that patients who transition from state to state are not well modelled by the Markov assumption. It most likely that this is the reason that the probability transitions from these multi-state models were not as well matched to the raw data of these studies as we would like. Nevertheless, the REMoxTB study did show a reasonable approximation to the data. However this study was unique in terms of the number of follow-up visits conducted. Tuberculosis trials usually follow that of the RIFAQUIN study with fewer follow up visits at the start of the study, and so overall HMMs do not seem to be well suited for these studies.

5.10 Summary and discussion

This chapter has investigated the use of multi-state Markov models in tuberculosis clinical trials using two data sets as examples. Although the data are not truly Markov, especially around 8 to 12 weeks when patients are transitioning from a positive to negative state, they are more so in the latter part of the follow-up when most data are missing. Therefore, if we use a flexible model for the log hazard over time this should be a reasonable approach. Results from the REMoxTB study confirm this.

For the REMoxTB study, we assumed in the first 8 weeks of follow-up, a patient's state remained constant from week to week. Post-week 8 we assumed patients could transition between scheduled follow-up visits over time. It was more accurate to assume a patients state remained largely unchanged between 7 days of follow-up as the transitions from state to state would be quite slow. This shows how flexible these models can be. Using HMMs enabled us to fit complex models which were able to forecast a better fit to the observed data. After finding a HMM for the observed data, the estimates from the hazard, hazard ratios and misclassifications can be used as starting values to re-fit the HMM. The advantage of doing this is to improve the estimates of a model.

We used a linear splines model with 3 knots at 4, 8 and 26 weeks and we were also able to add a fourth knot at 2 weeks increasing the model's complexity. Using the forwards/backwards algorithm and Viterbi algorithm produced results consistent with that of the published mITT and PP analyses. This is reassuring since there were no major departures from the original results of the study. In using the extended forwards/backwards algorithm to impute the missing observations resulted in a "completed" dataset that allowed each patient to be classed as a treatment failure or as reaching stable negative culture conversion. The results from this algorithm were also similar with the results from the two-fold fully conditional specification multiple imputation which provides further re-assurance that these models are able to estimate the data well.

Although the expected prevalence for the restricted cubic splines model suggested the model was a reasonable fit to the data, assessing the misclassification matrix for the HMM tended to withdraw a lot of information from the data. This in conjunction with fitting the more complex cubic splines meant that the model struggled to fit well. The second order fractional polynomials model did not fit so well and so we did attempt to fit a fourth order fractional polynomial model. The model for this however failed to converge. Perhaps if we had more data, we would have been able to fit this model better for these more complicated models. When assessing how closely matched the probability transitions were to the raw data, the probability of transitioning from a positive to a negative state over time using the linear splines HMM were not well matched around 8 to 12 weeks of follow-up. Although the HMM includes more information after imputing the missing data, large departures from the raw data are not expected. Nevertheless, overall the probability transitions from the HMM were still well matched to the raw data in the latter part of follow-up, suggesting most patients were in a negative state by the end of the study. Simulating the data for 30,000 patients based on estimates from the linear splines HMM when including a fourth knot at 2 weeks, and re-fitting the model using these simulated data, did improve the fit of the data in the early part of follow-up. However adding this fourth knot to our preferred HMM to the REMoxTB data did not make a large difference between 8 to 12 weeks, suggesting there was not enough information at this point to estimate the probability transitions well at that point.

For the RIFAQUIN study, the piecewise constant HMM was the preferred model which included 3 knots at 3, 6 and 10 months. Adding this number of knots seemed to work well, but in increasing the complexity of the model using splines and fractional polynomials meant fitting these models with fewer knots. This was done either so that the more complex HMM models converged or to obtain more sensible estimates from the chosen model. Even though these models were challenging to fit, the estimates from using our extended forwards/backwards algorithm to account for the uncertainty of estimates produced by the algorithm and the Viterbi algorithm were better matched to the original analyses of the PP estimates rather than the mITT analyses. This is most likely because the ITT type analysis proposed here does not make extreme assumptions about the missing data unlike that which is imposed in the original mITT analysis for the RIFAQUIN study.

The linear splines HMM with one knot at 5 months did not produce a sensible estimate for the 4 month treatment regimen for negative to positive transitions. The result of this HMM suggested that the risk of transitioning from a negative state to a positive state was low for the 4 month regimen and high for the 6 month regimen. This result goes against our intuition since there was good evidence from the original analyses of the RIFAQUIN study that the 4 month treatment regimen failed to demonstrated non-inferiority. Due to this observation, we would have expected the results from the fitted HMM to show a much larger hazard ratio to reflect that there were more positive culture results in the study for this treatment arm and therefore more treatment failures. Therefore when using these methods to find the preferred HMM for the data, careful interpretation of the results for each HMM investigated is also needed to ensure there are no conflicting results between the models. Having used the estimates produced from the piecewise constant HMM with knots at 3, 6, and 10 months to simulate data for 30,000 patients, and then re-fitting the HMM failed to improve the fit of the probability transitions. The most plausible explanation for this is that RIFAQUIN was a smaller study in comparison to the REMoxTB study, with fewer follow-up visits.

Assuming patients could transition between states outside of observations collected at scheduled follow-up visits over the duration of the whole study reflects the true Markov process. Although a strong and less plausible assumption for longer follow-up visits, it is possible to assume states only change at the time of follow-up. However doing so for these studies reduced the size of the estimates, but resulted in a higher -2 log-likelihood indicating a poorer fit to the data.

False positive cultures and false negative cultures were accounted for by assessing misclassifications using hidden Markov models. This meant we were able to estimate these probabilities and impute the culture states using the estimated hazards and the estimated sensitivity and specificity rather than using the trials' definition. If we were not estimating the misclassifications, we would have had to re-class single positive culture results as negative (if patients had reached stable negative culture conversion prior to the single positive result) or re-classing negative results as positive (if patients

had not reached stable negative culture conversion/relapsed) before running these models. This latter approach was not taken since a MGIT machine or manual LJ spectrum is used to detect presence or absence of TB. This means it is of interest to estimate such occurrences rather than imposing a rule. However, in most cases when choosing the preferred model for the data for our exemplar studies, we found that assessing the misclassification matrix drained out quite a lot of information from the data, making them more challenging to fit. If the misclassifications could be assessed (an added complexity to Markov models) when fitting the models, the missing culture results could be imputed, but the confidence intervals will be slightly wider inferring uncertainty. This is a consequence of assessing the misclassifications for the studies explored here. Therefore presenting these results to researchers, which appear to show that a loss of information as a consequence of the wider confidence intervals when the rationale for using HMMs was to gain some information is not ideal. This is why fixing the misclassifications may be preferable, since the intention is not to lose information. The goal is to gain information from these models and with less importance on precisely estimating what the misclassifications are. This is no excuse to define and fix nonsensical values for sensitivity and specificity; rather caution to not estimate these nuisance parameters at the cost of losing information. The process for assessing these models is to try to fit the HMM resulting in a misclassification model or to coax the model, eventually estimating the misclassification matrix and then using those results as initial values fixing the misclassifications at those values found by the previous model. Fixing the misclassifications even if the misclassification matrix could be directly estimated from the model made very little difference to the resulting probabilities for the REMoxTB and RIFAQUIN studies, but did inflate the confidence intervals slightly when they were not fixed.

For the studies we used in this chapter, simpler models using a linear spline or using a piecewise constant model worked better than a restricted cubic spline or fractional polynomial model. For the fractional polynomial model and cubic splines model, more data is included in comparison to the linear splines and piecewise constant model, where a value of zero is taken up to the point of the specified knot. This in conjunction with the added variables create a more complicated model overall reflected in the estimates produced for the cubic splines and fractional polynomial models. This indicates that these models are too complex for the software to fit for the

data we have. These models may however work well if the collection of sputum samples were better placed in terms of timing or in other studies with a larger sample size. The simulations conducted improved the fit of probability transitions from the chosen HMM to the raw data for the REMoxTB study, but not for the RIFAQUIN study. This further supports that these models may work better for trials with larger sample sizes. Additionally, the weekly visits collected over the first two months in the REMoxTB study may have contributed to having a better Markov model that provided a closer fit to the data than for the RIFAQUIN study where most patients suddenly switch to a negative state after initially being diagnosed with TB around 8 weeks.

Having compared positive to negative and negative to positive transition probabilities over time to the raw data, it appears that the two-fold fully conditional specification method, which is not dependent on the Markov assumption, provides a closer fit to the data than multi-state Markov models. Having investigated whether or not the data are Markov, it appears that the culture results are dependent on earlier time points as well and therefore this assumption does not hold so well for the data we investigated here. Although it is possible that a second order Markov model may work, doing so will increase the complexity of fitting these models as the number of parameters increases exponentially with order[105]. Given that the models fitted here were already challenging to fit, we did not proceed with this.

The two-fold FCS multiple imputation method is therefore the preferred approach when further investigating the impact of missing data for TB studies using an ITT analysis, excluding patients for reasons unrelated to treatment. This is evident in the RIFAQUIN study which had fewer follow-up visits and fewer patients in the study than for REMoxTB. This discrepancy is important since the RIFAQUIN study is representative of other studies within the TB field with fewer follow-up visits. To better capture the trend of the data in the HMM analysis, we fitted more complex models using piecewise constants, linear splines, cubic splines and fractional polynomials. The most likely explanation for the differences in the probability transitions to the raw data is that the multi-state models are computationally intensive to fit where there are few visits (and therefore fewer observed states) over a long

period of follow-up. This is made even more complicated by using smoothing methods for these type of data.

In the next chapter we use a range of sensitivity analyses to test the robustness of our analyses and conclusions thus far. We look at departures from the MAR assumption under MNAR (see §3.3). As the two-fold FCS method provided a closer fit of probability transitions to the data than the HMMs explored in this chapter for both the REMoxTB and RIFAQUIN studies, and so is the preferred choice, we explore reference-based sensitivity analyses using multiple imputation.

Chapter 6

Sensitivity analyses

Chapters 3-5 have explored different methods to handle the missing data, allowing us to include in the analysis patients with interim missing values and patients who had reached a stable negative culture conversion when last seen but withdrew before the end of follow-up. Any analysis with missing data makes inherently untestable assumptions about the distribution of the unobserved data. Consequently, where missing data arises, analysis should not only consist of a primary analysis under the most plausible assumption for the missing data but should ideally include a range of sensitivity analyses under alternative missing data assumptions to test the robustness of conclusions. The importance of conducting such sensitivity analysis is highlighted in the 2010 EMA guidelines for missing data in confirmatory clinical trials[106]. These guidelines state: "When the results of the sensitivity analyses are consistent with the primary analysis and lead to reasonably similar estimates of the treatment effect, this provides some assurance that neither the lost information nor the methods used to handle missing data had an important effect on the overall study conclusions". Different results obtained when the assumptions of the missing data are varied are just as important since this reveals under what conditions different results would be obtained.

The systematic review in Chapter 2 showed only 16% (27/168) of articles reported sensitivity analyses in which the assumption made for the missing data was changed. This illustrates a need for accessible methods of sensitivity analysis busy trialists can utilise. For the REMoxTB study best case/worse case scenarios were performed in the

primary analysis for all analyses across all treatment arms. For the RIFAQUIN study patients who died during the study were classed as unfavourable, reinfections were classed as unfavourable for PP and mITT analyses and a worst case analysis was performed for all patients in the mITT analysis who were excluded provided they were not a late screening failure.

In §3.6.2 and §3.10 we investigated a best case scenario for the REMoxTB and RIFAQUIN studies where missing observations in the standard of care regimen were imputed with positive culture results and missing observations on treatment arms were imputed with negative culture results. We also explored worst case scenarios where missing observations on the standard of care arm were imputed with negative culture results and missing observations on treatment arms were imputed with positive cultures.

In this chapter we propose new, alternative sensitivity analyses to assess the impact of changing patients' post-deviation behaviour (i.e. after being lost to follow up) on trial results. These are less extreme than the simple best case/worst case scenario approaches, hence have the potential to be more realistic and useful. Specifically, we investigate reference-based sensitivity analyses using multiple imputation to explore departures from the MAR assumption, under MNAR (3.3). We first outline the methodology behind reference-based sensitivity analyses using multiple imputation for a continuous outcome. An extension which enables these methods to be applied for binary outcomes is then proposed. The methods are then applied to the REMoxTB and RIFAQUIN studies.

6.1 Reference-based sensitivity analyses via multiple imputation

The aim for any primary analysis should be to estimate the primary objective, or the *estimand* i.e. that which is being estimated. Sensitivity analyses should also be designed to address the estimand of interest. Therefore when framing sensitivity analyses we must carefully consider precisely what is being estimated under the selected assumptions. Universal terminology is proposed by Carpenter, Roger and

Kenward[107]; the *de jure* and *de facto* estimands. The de jure estimand estimates the expected treatment effect if eligible patients randomised into a study adhered to their randomised treatment as specified in the trial protocol and the de facto estimand estimates the treatment effect seen in practice if this treatment were assigned to the target population of eligible patients, as defined by the trial inclusion criteria[107]. These terms are arguably similar to PP and ITT analyses respectively, but as found in the systematic review in Chapter 2, PP and ITT can be interpreted in several ways. The de jure and de facto estimands reduce these ambiguous definitions focusing on the specific assumption used to impute the missing data and now relate to the estimand.

Reference based-sensitivity analyses use a multiple imputation model constructed using data observed from a designated reference (typically control) arm to impute missing outcome data in the treatment arm. This enables one to assess the impact of deviators behaving like a reference patient post-deviation on trial results. These methods were developed by Carpenter, Roger and Kenward[107], based on ideas from Little and Yau[108] and shown to be statistically valid by Cro[109]. Appealingly, reference-based multiple imputation procedures enable the estimation of both de jure and de facto estimands. That is, they assess the impact of assuming all patients adhered to their randomised treatment and the impact seen in practise where patients may switch treatment arms and subsequently behave as if allocated to a treatment reference arm.

For reference-based sensitivity analyses via multiple imputation, data are split according to patient withdrawal so that each patient's data can be divided into pre-deviation data and post-deviation data. Different options to construct a joint distribution between pre-deviation and post-deviation can be used, corresponding to alternative assumptions for the unobserved data. The joint distribution is then used to create several imputed data sets that are then combined for analyses using Rubin's rules[34]. We now describe in detail how Carpenter et al[107] propose how this should be done for a longitudinal continuous outcome, under the assumption of multivariate normality.

6.1.1 Algorithm for reference-based sensitivity analyses

The algorithm proceeds as for standard multiple imputation under MAR (see §3.5.1) while accounting for pre- and post-deviation data. As defined by Carpenter et al[107]:

1. Assuming MAR, for each randomised treatment arm a MVN distribution with unstructured mean and unstructured variance-covariance matrix is fitted for all patients' pre-deviation observations. Adopting a Bayesian approach, an improper prior for the mean and an uninformative Jeffreys prior for the variance-covariance matrix is used.

2. For each randomised treatment arm a mean vector and variance-covariance matrix is drawn from the posterior distribution. These draws are used to construct the joint distribution of each deviating patient's pre- and post-deviation outcome data using one of the options presented in §6.1.2. This joint distribution is used to form the conditional distribution of the post-deviation responses given pre-deviation responses. The post-deviation data is then sampled from this constructed conditional distribution resulting in one complete data set.

3. Repeat step 2 to create I imputed datasets.

Having created I imputed datasets, the estimates are combined to get an overall estimate and variance using Rubin's rules[34] (see §3.5.1).

6.1.2 Options to construct the joint MVN distribution

The distribution of each patient's post-deviation responses given their pre-deviation responses and deviation time, required for imputation is defined as:

$$MVN \sim (Y_{mis(k)}|Y_{obs(k)}, D_k, trt_k, \eta), \tag{6.1}$$

for patient k for the randomised treatment arm trt deviating at time D. Y_{obs} represents pre-deviation responses from baseline until the point of deviation, $(Y_{obs(k)} = Y_{k,0}, ..., Y_{k,D_k})$, $Y_{mis(k)}$ represents post-deviation responses until the end of scheduled follow-up $(Y_{mis(k)} = Y_{k(D_k+1),...,Y_{k,J}})$ and η represents noise.

The heart of the reference-based approach in step 2 of the algorithm presented in §6.1.1 is the construction of each patient's joint MVN distribution of pre-deviation and post-deviation data. In longitudinal data, such as arises from TB trials, this can involve specifying many parameters. Reference-based methods do this implicitly, by reference to other treatment arms and/or group of patients. The reason for doing this is to make it easier for non-statistical experts to understand the assumption, and (as mentioned) to avoid specifying many sensitivity parameters explicitly.

Following the formation of the required joint distribution for each patient who deviates, the conditional distribution of post-deviation data given pre-deviation data can then be constructed for imputation. Each option corresponds to an alternative underlying missing data assumption. These scenarios apply to studies where patients are randomised to one or more active interventions, alongside a reference (e.g. control or placebo) intervention. The reference-based options for constructing of the joint MVN distribution presented by Carpenter and Kenward include[110,111]:

1. Jump to reference: The joint distribution of a deviating patient's observed and missing data is formed as MVN where the mean and variance is taken from their randomised treatment arm up until the last pre-deviation observation. Post-deviation, the mean response distribution and the variance follow that of the reference arm, i.e. the control regimen. This corresponds to the assumption that post-deviation, the deviator ceased their randomised treatment and started treatment similar to that available in one of the other arms (the reference arm).

2. Copy increments in reference: This is similar to the jump to reference option, where the joint distribution takes the mean from the randomised arm up to the last pre-deviation observation. Post-deviation the mean increments copy those from the reference arm and the variance follows that of the reference arm. This corresponds to the assumption that post-deviation the deviators response resumes the course observed in the reference arm.

3. Copy reference: The whole distribution of a patient pre-deviation and post-deviation is assumed to be the same as that of the reference arm. This corresponds to the assumption that the deviator followed the reference treatment throughout the trial.

These three options fit under the de facto terminology umbrella since they explore scenarios regardless of adherence to the protocol defined treatment. Under the de jure estimand, where the assumption is that patients follow the trial protocol continuing to adhere to treatment, the following options may be applied[110,111]:

4. Randomised-arm MAR: Patients' pre-deviation and post-deviation follows a MVN distribution with mean and variance from the randomised treatment arm.

5. Last mean carried forward: The marginal treatment group mean at the final observed visit is held at this value and the variance comes from the randomised treatment arm. This option is appropriate when the effect of randomised treatment is assumed to be maintained on average post-deviation.

Appendix L shows the technical details for how each of these options are formed.

The methods of Carpenter, Roger and Kenward[107] described here are valid where patient outcomes follow a continuous MVN distribution. However, in TB studies, the outcomes are binary. We now describe options that may be used to extend reference-based sensitivity analyses via multiple imputation so that they are applicable to binary outcome data. The first describes simple rounding, the second approach describes the coin flip algorithm and the final proposal is the adaptive rounding algorithm methodology proposed by Horton et al[112]. Other extensions of the multiple imputation procedures for counts[113] and time-to-event outcomes have been proposed[114], but to our knowledge no one has yet proposed an extension for binary outcomes.

6.2 Adaptive rounding algorithm

For reference-based sensitivity analyses in a binary setting, we proceed to model the binary data as if it were continuous and use multiple imputation as described in §6.1.1. Following the imputation, missing observations imputed as continuous are then back-transformed to binary observations. Each imputed data set with all outcome values on the binary scale can then be analysed using the analysis model of interest and results combined using Rubin's rules[34] for inference.

To back-transform to binary data when multiple imputation is performed under the assumption of MAR, the following methods proposed by Horton et al[112], described by Carpenter and Kenward[79] and Bernaards et al[115] can be used. The first involves using a simplistic method (simple rounding) that rounds the imputed value for the missing observation to the nearest 0 or 1. The second method is the coin flip algorithm, where any imputed value ≤ 0 is replaced with a 0 and any value ≥ 1 is replaced with a 1. Values that fall in between 0 and 1 (Y_v) are imputed with a binary response of 1 with probability Y_v. The third method is the adaptive rounding algorithm:

(a) For binary variable w in imputed dataset $I = 1, ..., I$ let $\bar{Y}_{w,\tilde{z},t}$ be the mean of the observed (binary) and imputed $\tilde{z}$ values at time t.

(b) The binomial distribution is approximated to the normal distribution:

$$\tilde{r}_{v,w,t} = \frac{\bar{Y}_{v,w,t} - C_{v,w,t}}{\sqrt{\bar{Y}_{w,z,t}(1 - \bar{Y}_{w,\tilde{z},t}).}}$$

Let $\varphi(.)$ be the cumulative distribution function of the standard normal.
We set $\varphi(r_{v,w,t}) = \bar{Y}_{w,\tilde{z},t}$ and then construct a threshold such that:

$$C_{w,\tilde{z},t} = \bar{Y}_{w,\tilde{z},t} - \varphi^{-1}(\bar{Y}_{w,\tilde{z},t})\sqrt{\bar{Y}_{w,\tilde{z},t}(1 - \bar{Y}_{w,\tilde{z},t})} \tag{6.2}$$

(c) Imputed values are re-coded as 0 if $Y_{w,\tilde{z},t} \leq C_{w,\tilde{z},t}$ and 1 if $Y_{w,\tilde{z},t} > C_{w,\tilde{z},t}$

Bernaads et al performed a simulation study using these three methods and found that the adaptive rounding algorithm performed best under standard MAR multiple imputation[115]. Although this was only slightly better than the simple rounding method, the adaptive rounding algorithm is preferred to increase the variability for values that are imputed close to 0 or 1[79]. This is clearly an important component in the context of imputing for TB studies.

We therefore use the adaptive rounding algorithm with reference-based sensitivity analysis for the REMoxTB and RIFAQUIN studies to test for departures from the MAR assumption made about the missing data under MNAR.

6.3 Application to the REMoxTB and RIFAQUIN studies

In non-inferiority studies, new treatments are compared to the standard of care. It is plausible to assume patients who are lost to follow-up receive some form of standard care. We also use de jure methods assuming patients continued with their randomised treatment arm post-deviation. These analyses will capture patients who completed intensive treatment and may not have had any further treatment since it is assumed that their outcomes will follow the distribution of the randomised treatment arm even though we are unable to verify this.

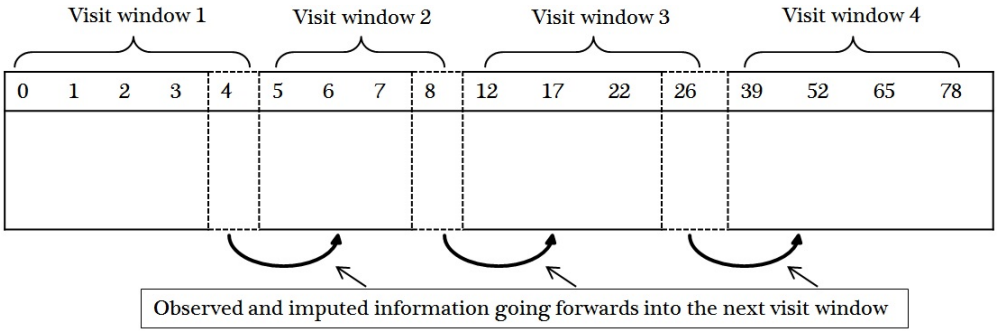
As outlined in the previous subsection (see §6.2), for imputation it is assumed that the data in the REMoxTB and RIFAQUIN study can be modelled as continuous, and a total of 50 imputations are created for each de jure and de facto scenario. For missing outcomes, data are back-transformed after imputation by applying the adaptive rounding algorithm defined in §6.2. The results are then combined using Rubin's rules[34]. The results are interpreted according to the definition of treatment failure as in §3.1 at the 1.25% (one-sided) significance level for the REMoxTB study and 2.5% (one-sided) significance level for the RIFAQUIN study. Analyses for the REMoxTB study adjusts for weight band and centre of recruitment and the RIFAQUIN study adjusts for centre of recruitment. To determine whether non-inferiority could be concluded, the 6% margin is used as defined in the original studies.

As for §3.2, any observed (and therefore known) sputum test results that occur during an unscheduled visit will be included after imputation and after the application of the adaptive rounding algorithm between randomisation and the final 18 month scheduled follow-up visit. Any unscheduled visits after this final scheduled follow-up visit are ignored.

6.3.1 Results from the REMoxTB study

Conducting reference-based sensitivity analyses via multiple imputation across all visits proved impossible because there is insufficient information within the dataset to estimate any variability between patient outcomes. This means an underlying MVN distribution to the observed data could not be fitted across all time points at once. Instead, the general approach taken was to split the data into 4 visit windows. Imputations were performed using Suzie Cro's mimix command in Stata[111]. Assuming patient outcomes are continuous, for one set of imputations the first visit window was imputed across all visits within that visit window. The information (i.e. the observed and imputed data) from the final follow-up visit within that visit window was taken forwards to impute the next visit window. This mimics the two-fold algorithm (see §3.5.3) using one pass. This was done until the final visit window was imputed (see Figure 6.1).

Figure 6.1: Diagram showing "one pass" of the two-fold algorithm for the REMoxTB study.



As in §3.8.2 when investigating patterns of missing data, we took visits from weeks 0 to 4, 5 to 8, 12 to 26 and 39 to 78 and imputed separately within each of those windows, repeated for 50 imputations. The imputed datasets for each visit window were then combined into one large dataset (with multiple imputations) before continuing with the analysis. Figures 6.2 to 6.6 show the results from using jump to reference (j2r), copy increments in reference (cir) and copy reference (cr) under the de facto analysis and from using last mean carried forward (lmcf) and randomised-arm missing at random (mar) under the de jure analysis to assess for departures from the MAR assumption. Unadjusted results are presented in Appendix M.

Figure 6.2: Jump to reference sensitivity analysis for the REMoxTB study (adjusted analyses).

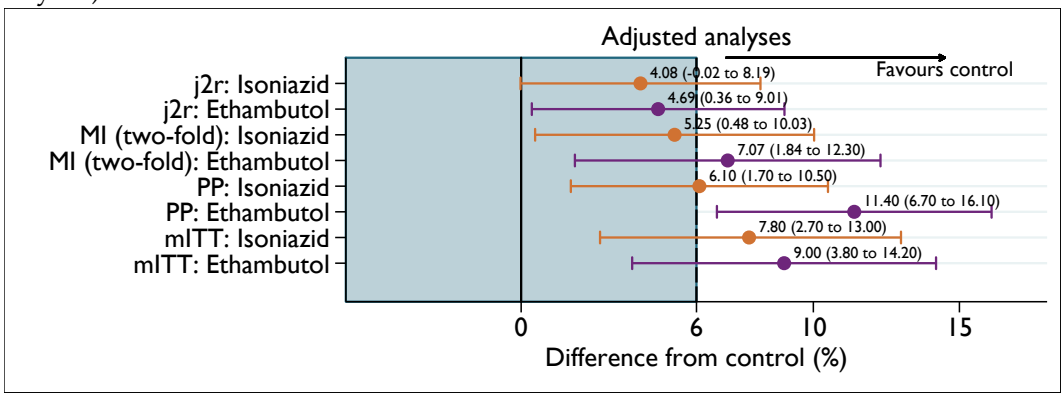


Figure 6.3: Copy increments in reference sensitivity analysis for the REMoxTB study (adjusted analyses).

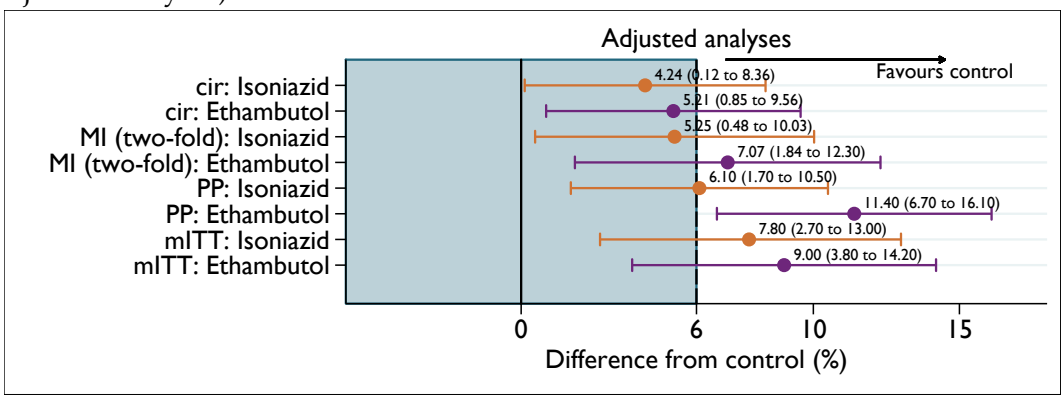


Figure 6.4: Copy reference sensitivity analysis for the REMoxTB study (adjusted analyses).

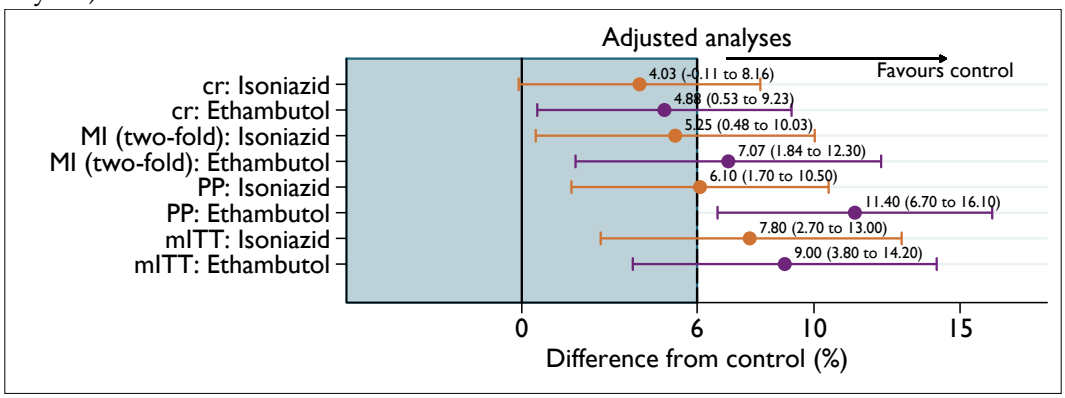


Figure 6.5: Last mean carried forward sensitivity analysis for the REMoxTB study (adjusted analyses).

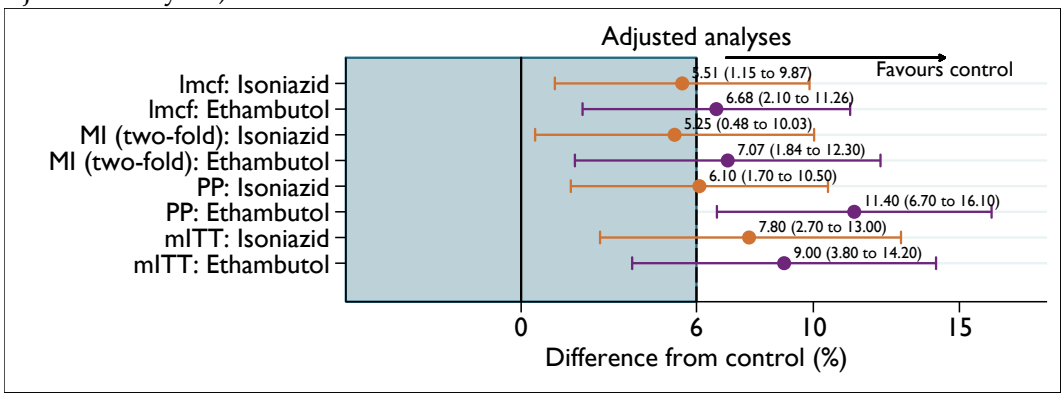
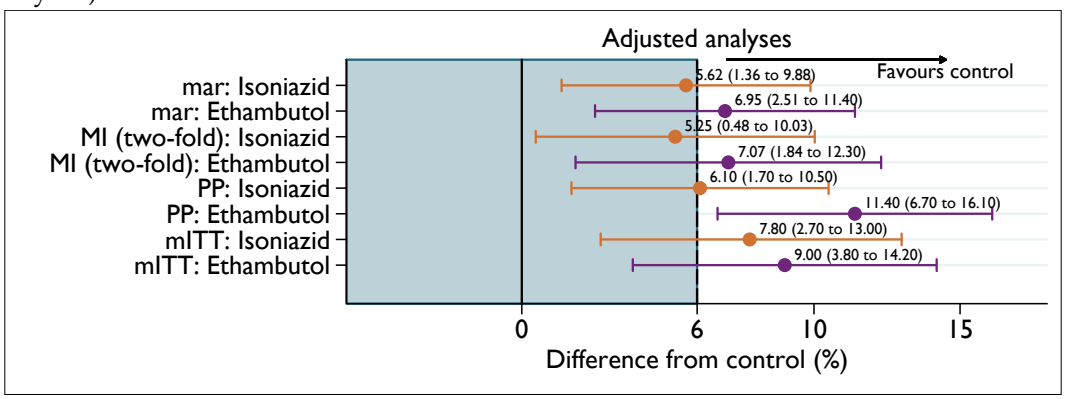


Figure 6.6: Missing at random sensitivity analysis for the REMoxTB study (adjusted analyses).



For the jump to reference sensitivity analysis, where post-deviation it is assumed patients follow the mean distribution of the control arm, the upper bound of the 97.5% confidence interval was 8.19% for the isoniazid arm (4.08%; 97.5% CI: -0.02 to 8.19%) and 9.01% for the ethambutol arm (4.69%; 97.5% CI: 0.36% to 9.01%) failing to demonstrate non-inferiority (Figure 6.2). The results from the copy increments in reference, copy reference, last mean carried forward and missing at random sensitivity analyses are similar (Figures 6.3 to 6.6). All fail to demonstrate non-inferiority since the upper bound of the 97.5% confidence intervals lie above the 6% non-inferiority margin. The randomised-arm missing at random and last mean carried forwards sensitivity analyses, where it is assumed patients continue on their randomised treatment arm post-deviation, had larger estimates of the upper bound of the 97.5% confidence interval which favoured the control arm.

6.3.2 Discussion

Reference-based sensitivity analyses showed consistent results with those of the primary analysis and with that of the two-fold FCS multiple imputation method where non-inferiority was not demonstrated on either the isoniazid or ethambutol treatment arms. The results from all analyses explored suggest that the results are not as extreme as those shown from the PP and mITT analyses, moving slightly closer towards non-inferiority. For this study, the PP and mITT analyses suggest the treatments were not performing as well as the control regimen and so, as expected the sensitivity analyses show greater support for non-inferiority if patients are assumed to move to the effective standard of care regimen under the de facto analyses. By contrast, the de jure analyses assume patients with missing outcomes continue with their randomised treatment and so the results for the last mean carried forward and missing at random sensitivity analyses are closer to the PP analysis than the mITT analysis and move towards favouring the control regimen. This is expected since these treatments did not perform so well; if it is assumed patients continue on a treatment arm that is not as effective as the control regimen then the effect of the treatment regimen will favour the control regimen supporting the conclusions of failing to demonstrate non-inferiority.

The sensitivity analyses explored here seem to work well for the REMoxTB study and provide reasonable estimates for our data. Overall, we conclude that the conclusions from the analysis under the MAR assumption are robust to the plausible assumptions captured in the reference-based analyses. The differences between the MAR and reference-based analyses are in the direction our intuition would expect. Next, we apply the same methods here to the RIFAQUIN study which reflects more closely to how tuberculosis trials are designed.

6.3.3 Results from the RIFAQUIN study

As for the REMoxTB study, there was not enough information within the data to estimate the variance using reference-based multiple-imputation across all visits for the RIFAQUIN study. Therefore, once again, data were imputed within each visit window taking the observed and imputed observations in the last visit within each

visit window forwards. Again, 50 imputations were generated. The four windows were defined as follows: months 0 to 3, 4 to 7, 8 to 10, 11 to 18. These visit windows differ slightly to those in §3.11 where month 7 is included in the second window for imputation. This is due to collinearity when month 7 was included in the third visit window between 7-10 months.

Figure 6.7 shows that the jump to reference sensitivity analysis failed to demonstrate non-inferiority for the 4 month regimen (10.33%; 95% CI: 4.72% to 15.94%). These results are consistent with the copy increments in reference and copy reference sensitivity analyses (Figures 6.8 and 6.9). On the 6 month regimen the upper bound of the 95% CI was 4.75% for the jump to reference sensitivity analysis (0.49%; 95% CI: -3.78% to 4.75%) which was consistent with the copy reference sensitivity analysis demonstrating non-inferiority. These results are closer to the results from the PP analysis than the mITT analysis and are consistent with the two-fold FCS multiple imputation. The copy increments in reference sensitivity analysis suggested a slightly stronger case for non-inferiority where the upper bound of the 95% confidence interval was 2.82% (-1.75%; 95% CI: -6.32 to 2.82).

Figure 6.7: Jump to reference sensitivity analyses for the RIFAQUIN study (adjusted analysis).

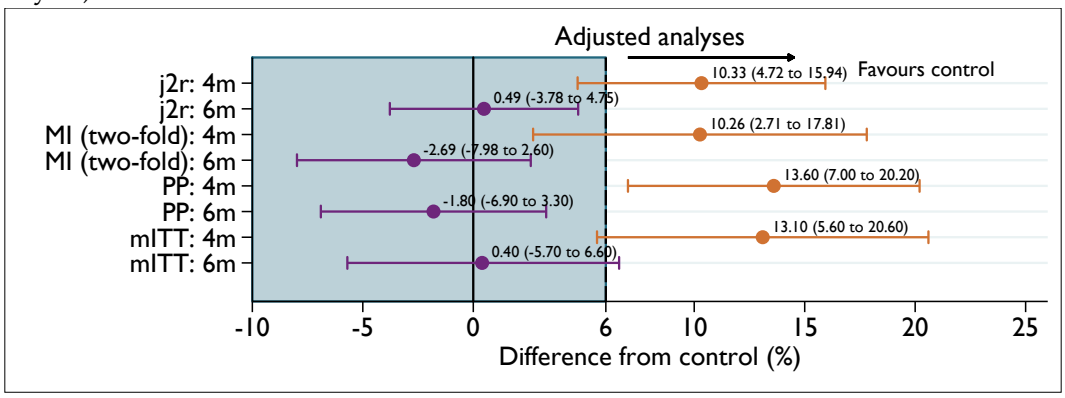


Figure 6.8: Copy increments in reference sensitivity analyses for the RIFAQUIN study (adjusted analysis).

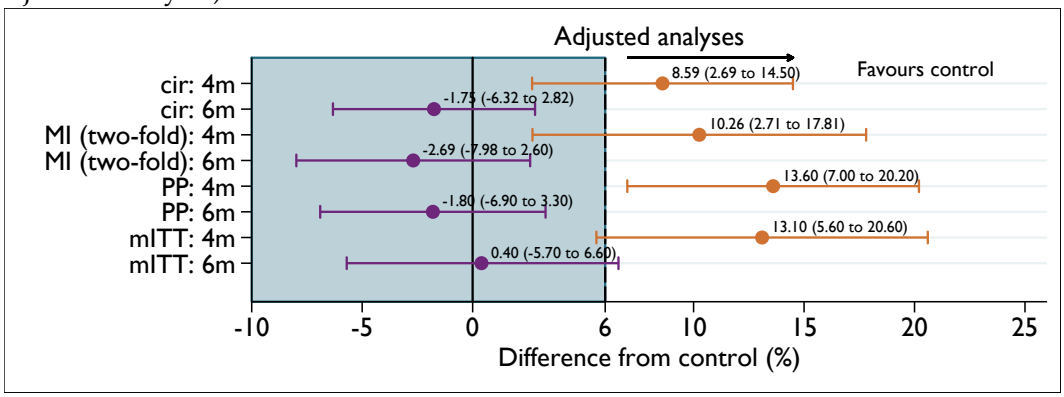


Figure 6.9: Copy reference sensitivity analyses for the RIFAQUIN study (adjusted analysis).

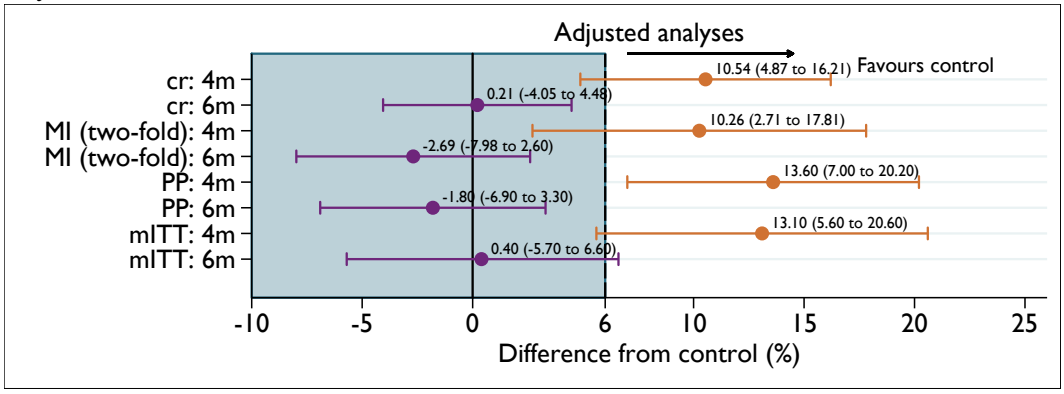


Figure 6.10: Last mean carried forward sensitivity analyses for the RIFAQUIN study (adjusted analysis).

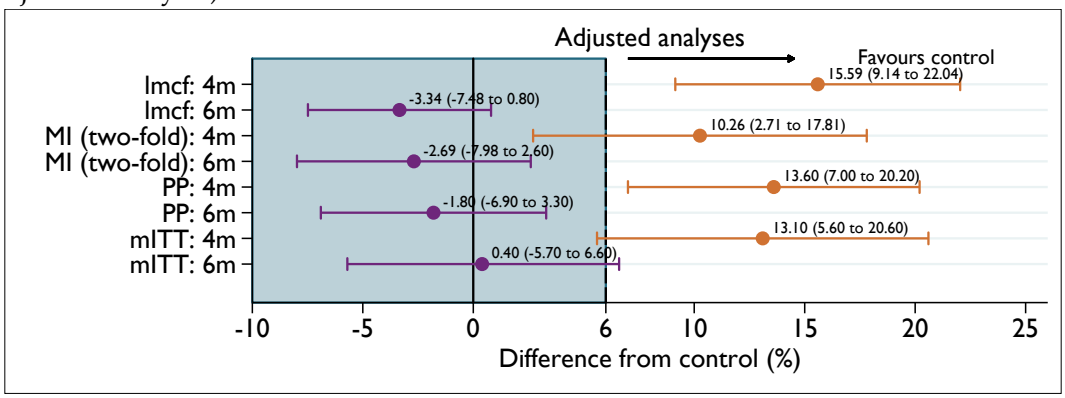
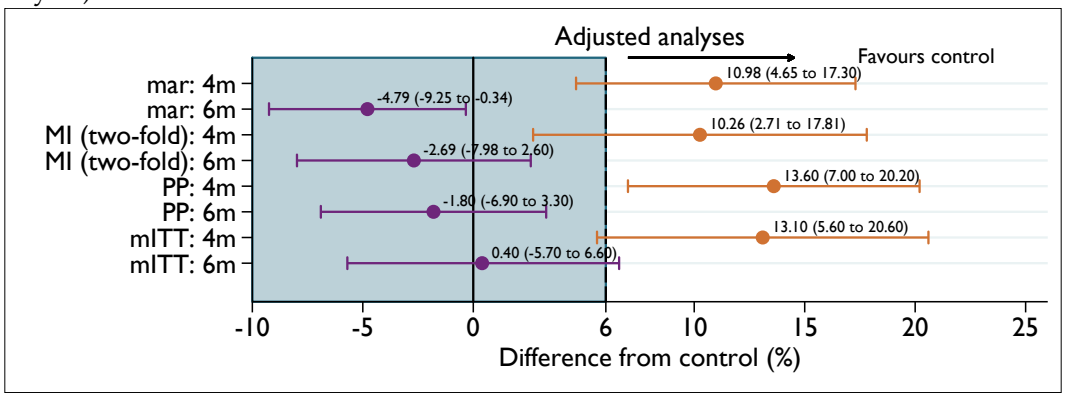


Figure 6.11: Missing at random sensitivity analyses for the RIFAQUIN study (adjusted analysis).



The results from the last mean carried forward sensitivity analysis, where patients are analysed according to their randomised treatment arm shows a stronger case for failing to demonstrate non-inferiority on the 4 month regimen (upper bound of 95% CI: 22.04%) than the PP and mITT analyses (upper bound of 95% CI: 20.2% and 20.6% respectively). The 6 month regimen demonstrates non-inferiority where the upper bound of the 95% CI is 0.8% (-3.34%; 95% CI: -7.48% to 0.8%), and this result was similar for the missing at random sensitivity analysis. The 4 month regimen for the missing at random sensitivity analysis also favours the control arm (upper bound of 95% CI: 17.3%) but not as strongly as the last mean carried forward analysis. See Appendix N for unadjusted results which give broadly similar conclusions.

6.3.4 Discussion

The results from all sensitivity analyses performed were consistent with the PP analysis. Overall, and as expected, the de facto point estimates (where the jump to reference, copy increments in reference and copy reference options assume patients follow the distribution of the control arm) move the treatment estimates in the opposite direction to the de jure estimates (where the last mean carried forwards and missing at random assume patients followed the distribution of their randomised arm) depending on where the starting point is. Consistent with this, the results from the 4 month regimen using analyses under the de facto estimand were not as inferior as the PP and mITT analyses. For analyses that fit under the de facto estimand, the PP

and mITT analyses from the study showed that the 4 month regimen did not perform as well as anticipated, and therefore the sensitivity analyses show some benefit if it is reasonable to assume that patients were administered the standard of care regimen after withdrawal. The results from the 6 month regimen suggest the results are non-inferior when assuming de facto and de jure.

Under the de jure assumption analyses where patients are assumed to continue with their randomised treatment post-deviation both the last mean carried forward and missing at random sensitivity analyses were consistent with the PP analysis demonstrating non-inferiority on the 6 month regimen. The missing at random sensitivity analysis was consistent with the PP and mITT analysis but was not as strongly inferior for both treatment arms and the last mean carried forward analysis tended more so towards the control arm for the 4 month regimen. The last mean carried forward analysis does however carry a strong assumption that, over time the marginal mean result (from randomisation until the final observed visit) is carried forwards for the imputation of later time points until the last scheduled follow-up visit of the study. It is likely that for this treatment arm, those who were last seen prior to withdrawal were already performing poorly on that treatment arm and so in reality would have had their treatment regimen changed.

The original analyses from the study showed conflicting results for the 6 month regimen; the PP analysis demonstrated non-inferiority and the mITT analysis failed to demonstrate non-inferiority. All sensitivity analyses supported the conclusions of the PP analysis. This suggests that the original mITT analyses performed for the RIFAQUIN study makes extreme assumptions, where patients who are lost to follow-up are considered to be failures. The de facto analyses showed the estimates tended towards the control regimen. If the 6 month treatment regimen is in fact the better regimen then analyses under de facto, which predicts results for patients who are missing using the information from observed patients who are in the control arm (i.e. not the better regimen) and so patients behave as if on the control regimen post-deviation, will show less benefit for patients randomised to the 6 month regimen. As a result, the estimates will tend towards failing to demonstrate non-inferiority. Whereas, if it is assumed patients with missing data continued on the better 6 month

regimen and therefore information to predict the missing values were borrowed from those observed on that treatment arm, a stronger benefit of the regimen would be shown. Therefore the estimates calculated from this type of analysis will tend more so towards demonstrating non-inferiority. Similarly, if the 4 month regimen is the worse regimen compared to control, assuming de facto will show a benefit as the results will shift towards demonstrating non-inferiority and for analyses under de jure will continue to show a lack of benefit. This is reflected in the estimates of the sensitivity analyses.

6.4 Summary

In previous chapters, we showed that multiple imputation (assuming MAR) is a robust and practical way to handle missing data in TB trials. Having shown this, in this chapter, we have developed methodology for conducting sensitivity analyses under MNAR assumptions for trials with unobserved binary outcomes. The proposed methodology allows the impact of departures from MAR on trial results to be assessed. Specifically, we have extended the reference-based multiple imputation methodology of Carpenter, Roger and Kenward[107] for use with a binary outcome. This was achieved by assuming data were continuous and then using the adaptive rounding algorithm we back-transformed the data to binary outcomes.

Instead of assuming a worst case scenario as recommended by regulators in TB trials, this allows us to make the more plausible assumption that patients who deferred from treatment were subsequently administered the standard of care regimen. For completeness we also investigated de jure methods, where we assume patients continued on their randomised treatment arm after deviation. The sensitivity analyses explored here show that, assuming patients continued with their randomised treatment arm, results are supportive of the conclusions of the PP analysis. That is non-inferiority was demonstrated. Arguably, this is a reasonable assumption, since post-withdrawal patients are most likely to be administered the standard of care rather than no treatment because patients are quite ill from this disease. However, assuming patient outcomes for those on the alternative treatment regimens follow the distribution of the standard of care arm post-deviation will make the treatment look

similar to the standard of care arm thus biasing towards the null (i.e. the non-inferiority margin) demonstrating non-inferiority. It is therefore equally important to use the de jure methods where patient outcomes are assumed to continue following the distribution of their randomised arm. Conducting sensitivity analyses under a range of alternative assumptions for the unobserved data is important to get a handle under what conditions the results vary, if any.

The REMoxTB and RIFAQUIN studies both highlight what is likely to be a recurring issue in the application of TB studies. This is that there is insufficient information to estimate the full variance-covariance matrix, resulting in collinearity when applying reference-based sensitivity analyses via imputation over all observation times. We therefore adopted our earlier proposal of splitting the data into visit windows, imputing each window assuming a multivariate joint distribution between the mean and variance of the control arm sequentially taking the last imputed visit, and therefore "completed" data, within each window forwards to the next window for imputation. This was done to remove imputing across all visits within each visit window independently, retaining the information from the last visit at the previous visit window forwards over time, as a "one-path" of the two-fold algorithm. This was done assuming the data were continuous using the *mimix* command in Stata software[111]. The adaptive rounding algorithm was then used, back-transforming these imputed values to binary values for analysis. We modified the software to accommodate for the adaptive rounding algorithm. We have shown that this is a feasible approach that works well for our TB datasets.

Although the imputations were retained using the last visit within each visit window, the results from the sensitivity analyses were consistent with that of the primary analyses and were closely matched to the two-fold FCS multiple imputation performed in Chapter 3. These methods also matched our intuition for patients on the treatment regimens which were not performing so well. This is because, assuming after withdrawal they received the standard of care regimen, it was expected the estimates move in the direction of demonstrating non-inferiority as discussed (§6.3.2). Analogous to this is if a treatment does perform "acceptably worse" in comparison to the control regimen, the estimates would move in the opposite direction (failing to demonstrate non-inferiority). Assuming patients continued with their randomised

treatment regimen also matched our intuition; treatment regimens that performed poorly continued to move in the direction of favouring the control regimen and treatment regimens that performed well continued to favour the treatment.
The de jure analysis for both the REMoxTB and RIFAQUIN studies, assuming randomised-arm missing at random, showed that the estimates were broadly similar to the two-fold fully conditional specification multiple imputation algorithm but that the confidence intervals were narrower. The expectation would be that the two analyses produce similar results as they almost make the same assumption. This could be a consequence of using one forwards path for imputation within the four visit windows. A natural extension given that the results look promising might be to implement these sensitivity analyses using a two-fold approach.

For the RIFAQUIN study, the PP and mITT analyses were conflicting. The sensitivity analyses used in this chapter provided more support for the PP analysis even when we assumed patients deviated to the standard of care regimen. The mITT analysis appears to make a rather strong assumption relative to the reference-based methods applied here. In some cases an extreme assumption can be made and the conclusions still be sustained. This is acceptable in situations where if a more plausible assumption is made then the overall conclusions will also hold. However, there will be cases where the most plausible assumptions through the reference-based methods will still retain the conclusions where as for a relatively extreme example it might not. In this situation we argue that the reference-based methods, being that they make more plausible assumptions, would provide more useful information to decision makers.

Using the adaptive rounding algorithm to extend the existing reference-based sensitivity methods to binary outcomes has never been applied before. Here, we have shown that in principle the adaptive rounding algorithm works well providing consistent results with all other analyses. However, while we have demonstrated proof of concept, further work is required to validate the methods proposed here, and we discuss this aspect in §7.3. By using more plausible assumptions, sensitivity analyses will inevitably result in a more accurate interpretation of the whole of the study. It is for this reason that we argue reference-based sensitivity analyses can be and should be considered for use for future TB non-inferiority studies.

Chapter 7

Discussion

In medical research, non-inferiority trials aim to find an alternative to the standard treatment that may be less efficacious, but has an advantage over the standard of care, such as fewer side effects or reduced cost. The clinical advantage of non-inferior treatment regimens is that clinicians have more than one treatment regimen option to administer to patients[116], should the standard of care cause side effects to a patient or if the patient is allergic to a particular drug. Arguably, as more treatments are found to be superior to placebo, the use of non-inferiority designs will continue to increase. It is therefore all the more essential that non-inferiority trials are both well-designed and well-conducted, with appropriate, transparent methodology used.

The lack of clear guidance for designing these studies demonstrates a need for more appropriate guidelines. The first aim in this thesis was to highlight these issues, reviewing current practice in design and analysis of these studies. The second aim was to find better methods for analysing the primary outcome in non-inferiority clinical trials, as there is a real need to identify and disseminate a valid, practical, approach to deal with the missing data. The third aim of this thesis was to investigate a better, yet still accessible approach for performing sensitivity analysis. Each of these aims are discussed in turn and we end with a final discussion of this thesis overall.

The guidelines reviewed in this thesis were often conflicting, making it challenging for researchers to implement a well-designed non-inferiority trial. Since our systematic review in Chapter 2 was completed, the U.S. FDA guidelines for non-inferiority were

finalised in November 2016[13]. The finalised version had little improvement in terms of clarity to the draft version and still remained inconsistent with other guidelines reviewed. Over the last six years, between the draft version of the U.S. FDA non-inferiority guidelines and the final version, researchers have become and continue to be increasingly aware of imputation methods developed to address issues raised by missing data. This was perhaps one of the more noticeable additions to the final version of the U.S. FDA guidance which recommend using imputation methods to account for attrition bias[13]. They however fail to distinguish between single imputation and multiple imputation methods. The guidance also fails to highlight that imputation methods carry untestable assumptions and completely neglect that using sensitivity analyses, under a range of plausible assumptions, is also important to test whether conclusions are robust.

The inconsistency in the guidance given to researchers is reflected to what is being done in practice, as shown by our systematic review reported in Chapter 2. One of the most concerning things we found was the lack of robust justification(s) of the non-inferiority margin. This is because there is a direct link from this to the clinical impact on patient well-being. We hope that this finding in our published review[44] and our warning that editors and other researchers within the community must be satisfied with the justification of the choice of the margin itself has an impact within the field.

The review performed found that ITT and PP analyses were often performed, and the general consensus is that if both analyses provide similar conclusions we can be reassured. Often, differences are due to patient withdrawal (from treatment, follow-up or both). Commonly, one of the ITT or PP analysis was taken to be the primary analysis, with the other considered a sensitivity analysis. These two analyses actually answer quite different questions about the behaviour of the population and do not explore the robustness of the conclusions to well-defined assumptions about the missingness mechanism. Therefore, they are not ideal sensitivity analyses.

Beyond this, the different interpretations of an ITT analysis and in particular a PP analysis were also concerning. There has been debate, and will continue to be debate,

surrounding the two analyses and whether one ought to take precedence over the other[116–121]. A recent review performed for non-inferiority trials by Aberegg et al found that there was little difference between results of ITT and PP analyses and where differences arose, the ITT was actually a more conservative analysis[122], (although, this research only reviewed articles from 5 journals). Ultimately, as with any research, the population chosen for an analysis and the analysis itself should be based on the real question to be answered, rather than performing an analyses without any careful thought about the question. This distinction is something currently being emphasised within the medical statistics field[123], where the focus is on what the estimand (i.e. the outcome to be estimated) should be. Nonetheless, the ITT and PP terminology appear to be the norm when designing non-inferiority trials, evident in all non-inferiority guidelines. The lack of clarity on how PP is defined in current guidelines has the potential to give researchers some leeway to perform a wide range of analyses, and as a consequence introduce bias into a study. In doing so the focus of what truly is the estimand is lost.

Another key finding within the review was that imputation techniques to test the missing data assumptions were rarely considered. In clinical trials, there inevitably will be some missing data[35]. The best advice to minimise the ambiguity that unavoidably arise from a non-trivial proportion of missing data is simply to minimise the amount of missing data in the first place, through creative preventative measures. For example, training staff involved during the development of a study about the impact missing data can have, should encourage them to collect as much data as possible and in particular persist in efforts to follow up patients to the end of a study. One study published that was included in our systematic review included case report forms (CRF) in the protocol (within the supplementary content)[124]. There is a specific question that asks whether all the patient information was collected and if not then an explanation for why the information was not collected. This is something that could easily be implemented in all clinical trials, and should help assess the plausibility of the assumptions made about the distribution of the missing data (i.e. whether MCAR, MAR or MNAR are reasonable) for the analyses. Logistically, this question can also remind researchers to collect vital information relating to patient outcomes that could have otherwise been forgotten.

For TB studies, the mITT and PP analyses dealt with missing data by excluding patients depending on treatment completion. The second aim of this thesis was to compare and contrast more sophisticated statistical approaches which allow the inclusion of these patients who were excluded and illustrate their use, using two TB datasets as examples. In terms of missing data, around 10-15% patients who are lost to follow-up are completely excluded from analyses in these trials. This is unsatisfactory when the non-inferiority margin is around 6% since the exclusion of these data may affect the overall conclusions made for these studies.

Our datasets were from TB non-inferiority trials where the goal was to find a shorter 4 month intensive treatment regimen compared to standard 6 months of care. For patients to be classified as "cured" after treatment, the results of patients' sputum samples need to be classed as negative at two consecutive, separate follow-up visits over the 18 months of total follow-up. The requirement of this confirmatory result means missing data can be problematic when attempting to determine the overall outcome of a patient. The missing data, and contaminated results which are regarded as missing, are ignored for analysis purposes. Our goal was to also include in the analysis information from patients with these missing observations in the analysis, to provide a more powerful, clinically meaningful analysis. We investigated different statistical methods to impute this missing data, resulting in a "completed" dataset that then allowed us to determine the primary outcome of treatment failure for each patient under the MAR assumption. We then explored a method, known as reference-based imputation, for exploring the robustness of conclusions obtained under MAR to plausible MNAR mechanisms.

7.1 Summary of results under MAR

Chapter 3 explored various single imputation methods and multiple imputation methods to include the missing observations of culture data from patients randomised to the REMoxTB (Table 3.2) and RIFAQUIN (Table 3.10) studies. We found that single imputation methods made extreme assumptions about the missing data. The complete case analysis resulted in a huge loss of data, therefore providing a less

powerful analysis, and best case/worst case scenarios produced extreme results in favour of demonstrating and failing to demonstrate non-inferiority respectively.

Due to the long, binary sequence of positive culture results at the start of follow-up and negative culture results towards the end of follow-up in both the REMoxTB and RIFAQUIN datasets, performing standard multiple imputation (including all visits in the imputation) was computationally infeasible. Instead we used the two-fold fully conditional specification multiple imputation algorithm. This method was used to impute the missing observations at each follow-up visit sequentially using observed outcomes on either side of that visit, propagating the imputed information forwards, until the final 18 month visit. The 10-15% of patients who withdrew were included in this analysis and imputations were performed separately in each treatment arm. This method seemed to work well for the data, producing consistent results with those of the primary PP analysis in both studies where the four month regimens failed to demonstrate non-inferiority and the 6 month regimen used in the RIFAQUIN study demonstrated non-inferiority.

Given the computational difficulties we faced applying multiple imputation in Chapter 3, we smoothed the data by creating visit windows, partitioning the data prior to conducting any further analyses. Using these visit windows, we investigated different patterns of missing data in both the REMoxTB and RIFAQUIN studies. The proportion of negative culture results in these different missing patterns challenges regulatory guidelines that recommend performing a worst case scenario. This is because we find this is a very conservative analysis. Sensible interpretation of results relies on sensible methods and analyses, and so we do not recommend using the worst case scenario for TB trials.

In Chapter 4, data remained partitioned in visit windows meaning that we were able to look at the data focusing on specific time points over the course of follow-up and also perform analyses on a simpler dataset. This is a key step for estimating stable weights for inverse probability weighted analysis. Marginal models such as Generalised Estimating Equations worked well when including weights in the model to account for the missing data. Weights were determined by investigating predictors

of treatment failure and withdrawals. Production of sputum and time to not producing sputum were key predictors and were included in the weights. Looking at risk differences within each visit window rather than overall gave more insight into how treatment failure changed over time for the REMoxTB and RIFAQUIN studies. For the REMoxTB study, we found that the treatments performed best within the first 8 weeks of treatment and performed well in the first 6 months of treatment. By the time of the continuation phase (6 to 18 months) non-inferiority could no longer be demonstrated. This suggested that the treatment used during the intensive phase is not strong enough or that treatment is not administered to patients long enough to suppress any latent TB bacteria remaining in the lungs. As a consequence, any remaining bacteria can multiply causing the patient to have a recurrence of the disease. This was a similar story for the RIFAQUIN study where treatment was most effective in the first 4-6 months of treatment. The 6 month regimen containing a higher dosage of rifapentine demonstrated non-inferiority for all visit windows. Although the 4 month regimen demonstrated non-inferiority in the final 11 to 18 months, it failed to demonstrate non-inferiority between 7 to 10 months. Towards the end of the study, patients not on treatment who were not becoming cured of TB would have had their treatment switched to a standard of care regimen as a consequence of the treatment failing. Therefore we can be confident that the 4 month regimen failed to demonstrate non-inferiority. This was supportive of the original primary analysis.

Observations over time were kept as visit windows, and as an alternative analysis, we used a mixed effects Poisson regression model to count the number of negative results within each window. There was less patient variation within the data predicted by the model suggesting data were underdispersed. Further, simply counting the number of culture results within a visit window, we lose the protocol defined outcome where patients are considered to reach negative stable culture conversion if they achieve two consecutive negative culture results at separate visits. In the Generalised Estimating Equation analysis this was handled by including weights within the model. The Poisson regression analysis is therefore not recommended for TB studies.

Chapter 5 explored multi-state Markov models, and in particular used hidden Markov models where the positive and negative cultures results were of the same type as the

hidden underlying disease state. This is a potentially attractive approach since these models can estimate the sensitivity and specificity. This is of interest within the TB community since the automated MGIT machine or manual LJ spectrum used to detected TB are not known to be 100% accurate. We explored a range of models such as piecewise constants, splines and fractional polynomials to smooth the data and provide a better fit of the Markov model to our data. A linear spline worked best for the REMoxTB study and a piecewise constant worked best for the RIFAQUIN study.

For the REMoxTB study, a lack of fit was marked around 6 to 12 weeks where patients were changing from mostly positive to mostly negative culture results. It was in this region that the data were not truly first order Markov. However, the fit to the latter part of the follow-up, when the data were nearly Markov and where most of the missing values were, was good. Various models of increasing complexity were explored to improve the fit, but a major constraint is the relatively limited information in the data to fit such models. This was even more noticeable when using these models for the RIFAQUIN study which had fewer patients and fewer follow-up visits. Using simpler models, reducing the complexity of the chosen model for the RIFAQUIN study resulted in a much poorer fit of the data.

To impute the missing states (considering the states as sputum test results thereafter) for each patient using our TB datasets, the forwards/backwards algorithm was used to calculate the probability of being in a positive or negative state at each time point. This algorithm was extended to enable proper imputation of missing sputum results. The Viterbi algorithm which calculates the overall sequence of states per patient was also used to check the consistency of conclusions made. The forwards/backwards algorithm is less computationally intensive than the Viterbi algorithm when predicting the true underlying disease state. Note that the forwards/backwards algorithm and Viterbi algorithm are both approximate, but can disagree in cases where the probability transitions are low between the most likely observations at time t and the next time point. Consequently, the Viterbi algorithm may calculate a lower probability for the underlying disease state at that time, even if the state observed is known.

Even though the fit of the final model for these studies did not match exactly to the raw data, is reassuring that the results from using the forwards/backwards algorithm and from using the Viterbi algorithm were consistent with mITT, PP analyses and two-fold fully conditional specification multiple imputation for both the REMoxTB and RIFAQUIN studies. This is what we hoped, as the HMM and multiple imputation approach both assume MAR, and impute within arms assuming patients continue to follow the protocol. The HMMs were easier to fit for the REMoxTB study where there were 1000 more patients, than for the RIFAQUIN study. REMoxTB also had more early follow-up visits resulting in richer data. The HMMs for RIFAQUIN also needed additional covariates in order to try and better model the fit of the observed transitional probabilities. Given that most TB clinical trials are of a similar size to the RIFAQUIN study, with a similar follow-up schedule, this suggests fitting these models may be relatively impractical for TB studies with smaller populations. These methods may however still prove to be very useful for other disease areas.

Since the REMoxTB and RIFAQUIN studies were published, there has been much discussion surrounding the choice of treatment regimens for phase III TB trials. In phase II studies, potential treatment regimens to shorten the duration of TB treatment look at the surrogate endpoint of culture conversion at 8 weeks. Although a useful biomarker for these trials as a whole, as Ruan et al assert[125], this marker is not strong enough to reliably predict patient outcomes in phase III trials. This was a lesson learnt from these studies and perhaps a strong contributor for these 4 month regimens failing to demonstrate non-inferiority.

We have shown how multiple imputation, weighted Generalised Estimating Equations and HMMs, can be used under an ITT type analysis (excluding reasons unrelated to treatment) to impute missing outcome data under the MAR assumption. Of these, we found that the two-fold fully conditional specification multiple imputation algorithm was the most practical, robust, approach. For the studies explored in this thesis, the results were similar to the per-protocol analysis. This is because the ITT analysis we propose does not make the extreme assumptions used in the original analysis for these studies. That is, assuming patients who do not not reach the end of follow-up are considered to be failures. Well-defined imputation based

approaches are preferable to ad-hoc approaches for dealing with missing values. We therefore argue that two-fold fully conditional specification multiple imputation should be used to handle missing outcomes for the primary analysis. Of course techniques like multiple imputation or inverse probability weighting should not be used without first carefully exploring the impact of the missing data, predictors of outcomes being missing and the likely correlation of the actual missing values. Our results confirm that the treatment regimens used were simply not strong enough to kill any remaining latent bacteria in the lungs during the continuation phase. Given the high costs of the study, there can be little justification for not using the most appropriate statistical approach which makes full use of information in the data.

7.2 Summary of results under MNAR

The results from the systematic review found that sensitivity analyses that tested for departures from the assumption made about the distribution of the missing data were rarely performed. When they were, simplistic methods that made strong assumptions about the missing data were used such as best case/worst case scenarios or last observation carried forward. In TB trials, it is often recommended to use a worst case scenario. In the original analyses for the REMoxTB and RIFAQUIN studies, the worst case assumption used as a sensitivity analysis was performed across all treatment arms, showing consistent results with the primary analyses performed. An actual worst case scenario, performed in Chapter 3, where the missing data for patients randomised to the control arm is imputed with the best result and the missing data for patients randomised to the treatment arms are imputed with the worst result shows just how extreme this analysis is. A better and more accessible approach to use sensitivity analyses was also a requirement following this review. This motivated the work in Chapter 6 looking at departures from the MAR assumption to MNAR.

For the reasons explained above, we proceeded with a multiple imputation approach. As discussed in Chapter 6, specifying a full MNAR distribution for the data requires specifying the distribution of a large number of parameters. Reference-based sensitivity analysis was designed to address this issue by instead specifying the MNAR distribution by reference to other groups of patients in the study, typically the

reference (often control) arm[79]. Accordingly, these studies are a natural setting to apply this approach. However, this is the first time reference-based sensitivity analyses have been explored in a non-inferiority setting. While current methods and software exist for continuous longitudinal outcomes[111], the approach has not been used for binary data before. Our approach was to assume the data were continuous for imputation, back transforming imputed values to binary values using the adaptive rounding algorithm. This algorithm has been shown to have good statistical properties[115]. Additionally, we had to address the computational issue of there being insufficient data to estimate the unstructured variance-covariance matrix of the outcome data. We handled this by applying the reference-based imputation approach within overlapping windows, analogous to the two-fold algorithm, although we only used one (forwards) pass through the data.

We performed two types of analysis. The first, de facto, analyses imputed missing data from patients in the control (i.e. reference) arm. This type of analyses assume (and may often be plausible) that post-withdrawal patients took a control-like treatment regimen, and this shifted the estimates of the treatment effect in the opposite direction. Depending on the starting point, this can either decrease or increase the evidence for non-inferiority. By contrast, the de jure analyses, imputes the missing data from patients in the same trial arm, in effect as if they continued to follow the protocol. As expected, the de jure MAR analysis, which makes the same assumption to the two-fold fully conditional specification multiple imputation MAR analysis, gave similar results.

The assumptions underpinning this approach are much more plausible than those for the simplistic best case/worst case scenario which produces extreme results. Therefore we believe this approach should be adopted routinely. To facilitate this, the program we extended for the analysis of these binary outcomes is to be developed for applicability in general settings as an extension to the currently available *mimix* software in Stata[111] to make this program available for all researchers.

7.3 Future work

The methods applied in Chapters 3-5 in this thesis, to impute observations for patients whose outcome data were missing (and who were therefore excluded if they did not reach the end of the study), were very similar to the estimates found in the pre-specified primary analyses. Although this is reassuring, there are still gaps remaining in the methods used in these trials. The trial protocol defined primary outcome of treatment failure can occur at any point over several scheduled follow-up visits. Analytically the long sequence of data per patient causes computational problems making it difficult to apply robust methods with meaningful results. This is because there are long constant sequences of positive culture results at the start of follow-up and long constant sequences of negative culture results towards the end of follow-up. There are some ways around this problem. The most practical of which seems to be partitioning the data into visit windows, but a simpler analysis that focuses on the treatment effects at the end of the intensive phase and at the end of the continuation phase may be sufficient to draw valid inferences. If so, this will result to a simpler, clearer design for these trials.

Chapter 3 investigated hot-deck multiple imputation but we found this was computationally impossible. An extension of this to longitudinal data using the principals behind the two-fold approach may work well, but would be expected to gain little information, if at all, relative to parametric imputation.

Chapter 5 investigated multi-state hidden Markov models, which allow the misclassification probabilities to be estimated. First order Markov chains were investigated, so that the previous visit was accounted for. However, the REMoxTB study showed that during the period between 6 to 12 weeks, when patients are moving from a positive culture results to mostly negative culture results, depend on the state at the previous two visits. This was a similar story for the RIFAQUIN study. Second order or higher order Markov chains could be explored. However, we did not do this, because of the issues we faced with the first order Markov models — specifically the limited amount of available information in the data to fit the models. This did not suggest this was a promising practical approach. As we only have two

states, the estimated intensities can be interpreted as (log) hazards and (log) hazard ratios; this assumes local proportionality. Beyond this, the forwards/backwards algorithm used to predict patient outcomes for multi-state Markov models could be extended to incorporate sensitivity analyses in a similar way to the reference-based multiple imputation explored in Chapter 6. These sensitivity analyses could then test for departures of the MAR assumption made for these multi-state models under MNAR. We note that, while the statistical performance (in terms of bias and coverage) of the two-fold algorithm has been extensively explored using simulations[78], if we wanted to use HMMs routinely in the primary analysis, a simulation study to confirm their performance would be desirable.

Chapter 6 used the adaptive rounding algorithm, extending the methods to binary outcomes using the reference-based multiple-imputation methodology of Carpenter, Roger and Kenward[107]. While we demonstrated proof of concept, further work is required to validate the "information anchoring" property of the method in this setting. Collinearity was an issue for multiple imputation under MAR, which was why we used a version of the two-fold approach using reference-based sensitivity analyses. We took the approach of grouping visits into visits windows, partially mimicking the two-fold FCS multiple imputation method imputing missing observations within each visit window conditional on the last visit at the previous visit window. The encouraging results with this approach support further developing reference-based sensitivity analyses using a two-fold FCS multiple imputation approach for non-inferiority trials with binary outcomes, such as in TB. We note that this approach has the attraction of being "information anchored". That is to say the information lost due to missing data is held constant across the primary and sensitivity analyses. This is an attractive property for regulators and trialists.

7.4 Conclusion

This thesis began with three aims: to review current practice in design and analysis of non-inferiority trials; to identify the most practical, accessible approach for handling missing data under MAR; and to identify a promising approach to sensitivity analyses. The systematic review performed highlighted the need for a more practical

approach to handle missing data for non-inferiority trials. Our proposed imputation approach is likely to be increasingly acceptable to the research community; research shows the use of multiple imputation is rising[126] with more researchers using the methods. This is perhaps because it is now easily accessible in most statistical software. The increasing awareness that the method exists has lead to missing data and multiple imputation guidelines. The methods explored included other methods to handle missing data, in particular such as inverse probability weighting and multi-state models. We found that for longitudinal binary data, although multi-state models can provide reasonable predictions of what state a patient is in if the observation is missing, it was much harder to get them to fit, and the fit was not entirely satisfactory. It is likely that this is due to violation of the first-order Markov assumption. Therefore, two-fold fully conditional specification multiple imputation is the preferred choice for TB studies as the method allows us to choose a non-Markov dependency. The results from these analyses were closer to the original PP analysis from the studies explored. The ITT analysis proposed here makes less extreme assumptions than the typically used mITT and PP definitions in TB studies. Turning to our final aim of performing sensitivity analyses, our extension of reference-based sensitivity analysis for binary outcomes worked well for the REMoxTB and RIFAQUIN studies. These sensitivity analyses are robust and more plausible than the recommended worst case analysis and we believe this should be the first choice for non-inferiority trials.

While the focus of this thesis was on TB non-inferiority trials, in theory the analyses used here can be applied to all clinical trials with binary longitudinal data that require a confirmatory result to determine a patient's outcome. In conclusion, we have successfully addressed the three motivating aims, providing a practical way forward in TB non-inferiority trials.

Appendices

A Data extraction form

GENERAL INFORMATION

1. Journal:
2. Article filename:

STUDY DESIGN

3. What was the effect measure of the primary outcome?
   (a) Odds ratio
   (b) Risk ratio
   (c) Hazard ratio
   (d) Rate ratio (for counts)
   (e) Difference in proportions
   (f) Difference in means
   (g) Ratio of means
   (h) Other
4. What was the margin?
   (a) Was the choice of the margin justified? y/n
   (b) How was the margin justified?
5. What type of intervention was used?
   (a) Drug

(b) Surgery

(c) Other

6. What type of trial was this?

   (a) Participant randomisation

   (b) Cluster randomised

   (c) Crossover randomisation

   (d) Other

Sample size

7. What was the type I error rate used in the sample size calculation?
   a. Was this one-sided or two-sided? One-sided/two-sided/unclear

8. What was the power used in the sample size calculation?

9. Was the treatment effect assumed to be zero? y/n/unclear

PRIMARY OUTCOME

10. Copy and paste the primary outcome:

11. What population was chosen (fill all that apply)?
    a. Intention-to-treat (ITT)? y/n

       i) If yes how was this defined (implicitly or explicitly)?

       ii) Was this primary or secondary (implicitly or explicitly)? primary/secondary/NA

    b. Per protocol (PP)? y/n

       i) If yes how was this defined (implicitly or explicitly)?

       ii) Was this primary or secondary (implicitly or explicitly)? primary/secondary/NA

c. Modified intention-to-treat (mITT)? y/n

i) If yes how was this defined (implicitly or explicitly)?

ii) Was this primary or secondary (implicitly or explicitly)? primary/secondary/NA

d. As-treated? y/n

i) If yes how was this defined (implicitly or explicitly)?

ii) Was this primary or secondary (implicitly or explicitly)? primary/secondary/NA

e. Other? y/n

i) If yes how was this defined?

ii) Was this primary or secondary? primary/secondary/NA

f. Unclear? y/n

12. Was the primary outcome a composite outcome? y/n

13. What disease is the primary outcome answering?

Study results

14. What level is the confidence interval being reported? 90%/95%/other
    a. If other:

15. Which bound of the confidence interval is being reported? one-sided/two-sided/unclear
    a. Was the direction pre-specified? Upper bound/lower bound/not specified

16. Is the confidence interval consistent with the type I error rate? y/n/unclear

17. What was the p-value?
    a. What side of the p-value has been reported? One-sided/two-sided/NA

Missing data

18. What percentage of missing outcome data was reported?

19. Were any imputation techniques used? y/n/NA
    a. If yes, what method was used?

Sensitivity analysis

20. Were any sensitivity analyses on the primary outcome conducted? y/n
    a. If yes, what were they?

Conclusions

21. Was non-inferiority declared? y/n
    a. Copy and paste conclusions made on non-inferiority

Other

22. Do any questions need to be checked by another reviewer (detail in the comments section)? y/n

Comments:

B Missing data patterns for REMoxTB

Table B1: Number of negative culture results and proportion of patients who achieved negative culture conversion for patients with most culture results observed (i.e. completers') over visit windows for REMoxTB on control arm[1].

Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78	Total (N=590)	Number of negative culture results Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78	Total number of negative culture results	Treatment success n/no. patient per pattern
O	O	O	O	272 (46.10%)	267/1269 = 21.04%	692/1019 = 67.91%	985/1011 = 97.43%	998/1033 = 96.61%	2942	253 (93.01%)
O	O	Δ	O	33 (5.59%)	26/149 = 17.45%	64/118 = 54.24%	62/66 = 93.94%	118/121 = 97.52%	270	29 (87.88%)
O	O	O	.	19 (3.22%)	13/93 = 13.98%	41/72 = 56.94%	64/70 = 91.43%	10/11 = 90.91%	128	3 (15.79%)
O	O	O	Δ	27 (4.58%)	28/128 = 21.88%	64/101 = 63.37%	96/99 = 96.97%	54/54 = 100.00%	242	24 (88.89%)
O	Δ	O	O	19 (3.22%)	16/80 = 20.00%	25/38 = 65.79%	65/66 = 98.48%	69/73 = 94.52%	175	16 (84.21%)
Δ	O	O	O	20 (3.39%)	7/53 = 13.21%	52/71 = 73.24%	72/74 = 97.30%	68/75 = 90.67%	199	18 (90.00%)
.	.	.	.	7 (1.19%)	0/7 = 0.00%	0/0 = 0%	1/1 = 100.00%	1/1 = 100.00%	2	0 (0.00%)
O	O	.	.	10 (1.69%)	16/50 = 32.00%	30/36 = 83.33%	2/2 = 100.00%	2/2 = 100.00%	50	0 (0.00%)
.	.	O	O	9 (1.53%)	3/10 = 30.00%	3/4 = 75.00%	29/29 = 100.00%	31/32 = 96.88%	66	8 (88.89%)
O	.	O	O	10 (1.69%)	9/43 = 20.93%	7/9 = 77.78%	37/38 = 97.37%	37/37 = 100.00%	90	10 (100.00%)
O	O	.	O	8 (1.36%)	10/38 = 26.32%	17/27 = 62.96%	6/6 = 100.00%	29/29 = 100.00%	62	6 (75.00%)
O	.	.	.	9 (1.53%)	7/33 = 21.21%	0/1 = 0.00%	0/0 = 0%	0/0 = 0%	7	0 (0.00%)
.	O	O	O	7 (1.19%)	1/8 = 12.50%	20/24 = 83.33%	26/27 = 96.30%	26/26 = 100.00%	73	6 (85.71%)
O	D	D	D	2 (0.34%)	5/8 = 62.50%	1/1 = 100.00%	0/0 = 0%	0/0 = 0%	6	0 (0.00%)
O	.	.	O	4 (0.68%)	3/18 = 16.67%	1/3 = 33.33%	2/2 = 100.00%	9/15 = 60.00%	15	1 (25.00%)
O	O	O	D	4 (0.68%)	1/19 = 5.26%	3/15 = 20.00%	13/16 = 81.25%	2/3 = 66.67%	19	0 (0.00%)
.	.	.	O	3 (0.51%)	0/6 = 0.00%	0/1 = 0.00%	2/2 = 100.00%	11/11 = 100.00%	13	1 (33.33%)
O	.	O	.	2 (0.34%)	1/7 = 14.29%	0/1 = 0.00%	6/7 = 85.71%	0/0 = 0%	7	0 (0.00%)
D	D	D	D	2 (0.34%)	1/1 = 100.00%	0/0 = 0%	0/0 = 0%	0/0 = 0%	1	0 (0.00%)
.	.	O	.	1 (0.17%)	0/2 = 0.00%	0/0 = 0%	3/3 = 100.00%	0/0 = 0%	3	0 (0.00%)
.	O	O	.	1 (0.17%)	1/1 = 100.00%	3/4 = 75.00%	3/3 = 100.00%	0/0 = 0%	7	0 (0.00%)
.	O	.	D	1 (0.17%)	0/2 = 0.00%	4/4 = 100.00%	1/1 = 100.00%	0/0 = 0%	5	0 (0.00%)
O	.	.	D	1 (0.17%)	0/3 = 0.00%	0/0 = 0%	1/1 = 100.00%	0/0 = 0%	1	0 (0.00%)

[1]Where O=Most results observed within a window; Δ=Intermittent results observed within a window; D=Death or .=Missing.

Table B2: Number of negative culture results and proportion of patients who achieved negative culture conversion for patients with most culture results intermittently observed over visit windows for REMoxTB on control arm[1].

Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78	Total (N=590)	Number of negative culture results Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78	Total number of negative culture results	Treatment success n/no. patient per pattern
Δ	Δ	O	O	12 (2.03%)	5/32 = 15.63%	13/24 = 54.17%	41/42 = 97.62%	42/42 = 100.00%	101	12 (100.00%)
O	O	Δ	Δ	7 (1.19%)	5/34 = 14.71%	19/25 = 76.00%	14/14 = 100.00%	14/14 = 100.00%	52	7 (100.00%)
Δ	O	Δ	O	11 (1.86%)	6/32 = 18.75%	20/34 = 58.82%	22/22 = 100.00%	42/42 = 100.00%	90	11 (100.00%)
O	Δ	Δ	O	2 (0.34%)	0/7 = 0.00%	2/4 = 50.00%	3/4 = 75.00%	6/6 = 100.00%	11	1 (50.00%)
O	Δ	O	Δ	5 (0.85%)	6/23 = 26.09%	9/10 = 90.00%	15/16 = 93.75%	8/10 = 80.00%	38	4 (80.00%)
Δ	.	.	.	3 (0.51%)	1/8 = 12.50%	0/1 = 0.00%	0/1 = 0.00%	0/0 = 0%	1	0 (0.00%)
Δ	Δ	Δ	O	4 (0.68%)	0/11 = 0.00%	6/8 = 75.00%	8/8 = 100.00%	15/15 = 100.00%	29	4 (100.00%)
Δ	Δ	Δ	.	3 (0.51%)	0/8 = 0.00%	0/6 = 0.00%	6/6 = 100.00%	2/2 = 100.00%	8	1 (33.33%)
O	Δ	Δ	Δ	2 (0.34%)	2/9 = 22.22%	3/4 = 75.00%	3/4 = 75.00%	4/4 = 100.00%	12	2 (100.00%)
Δ	.	Δ	.	1 (0.17%)	0/2 = 0.00%	0/0 = 0%	0/2 = 0.00%	0/0 = 0%	0	0 (0.00%)
.	.	Δ	Δ	1 (0.17%)	1/2 = 50.00%	0/1 = 0.00%	2/2 = 100.00%	2/2 = 100.00%	5	1 (100.00%)
Δ	Δ	.	.	1 (0.17%)	1/2 = 50.00%	1/2 = 50.00%	1/1 = 100.00%	1/1 = 100.00%	4	1 (100.00%)
Δ	O	O	Δ	1 (0.17%)	0/2 = 0.00%	3/3 = 100.00%	3/3 = 100.00%	2/2 = 100.00%	8	1 (100.00%)
Δ	.	Δ	Δ	1 (0.17%)	0/3 = 0.00%	1/1 = 100.00%	2/2 = 100.00%	2/2 = 100.00%	5	1 (100.00%)
Δ	Δ	Δ	Δ	1 (0.17%)	1/3 = 33.33%	2/2 = 100.00%	2/2 = 100.00%	2/2 = 100.00%	7	1 (100.00%)
Δ	.	.	Δ	1 (0.17%)	1/3 = 33.33%	0/0 = 0.00%	0/0 = 0.00%	2/2 = 100.00%	3	0 (0.00%)

[1]Where O=Most results observed within a window; Δ=Intermittent results observed within a window or .=Missing.

Table B3: Number of negative culture results and proportion of patients who achieved negative culture conversion for patients with a mixture of observed, intermittent and missing culture results within visit windows for REMoxTB on control arm[1].

Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78	Total (N=590)	Number of negative culture results Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78	Total number of negative culture results	Treatment success n/no. patient per pattern
O	O	Δ	.	8 (1.36%)	4/32 = 12.50%	15/27 = 55.56%	16/16 = 100.00%	3/4 = 75.00%	38	1 (12.50%)
O	.	Δ	O	10 (1.69%)	6/41 = 14.63%	4/6 = 66.67%	18/20 = 90.00%	31/35 = 88.57%	59	7 (70.00%)
O	Δ	.	.	2 (0.34%)	1/8 = 12.50%	2/4 = 50.00%	0/1 = 0.00%	0/0 = 0%	3	0 (0.00%)
O	Δ	O	.	3 (0.51%)	1/15 = 6.67%	3/6 = 50.00%	9/10 = 90.00%	3/3 = 100.00%	16	1 (33.33%)
.	Δ	O	O	2 (0.34%)	0/2 = 0.00%	4/4 = 100.00%	6/6 = 100.00%	7/7 = 100.00%	17	2 (100.00%)
O	Δ	.	O	4 (0.68%)	2/17 = 11.76%	3/8 = 37.50%	4/4 = 100.00%	16/16 = 100.00%	25	1 (25.00%)
O	O	.	Δ	2 (0.34%)	0/9 = 0.00%	6/8 = 75.00%	2/2 = 100.00%	4/4 = 100.00%	12	2 (100.00%)
O	O	Δ	D	2 (0.34%)	1/8 = 12.50%	7/7 = 100.00%	4/4 = 100.00%	1/1 = 100.00%	13	0 (0.00%)
O	.	Δ	.	2 (0.34%)	0/9 = 0.00%	1/1 = 100.00%	4/4 = 100.00%	2/2 = 100.00%	7	0 (0.00%)
Δ	Δ	O	.	1 (0.17%)	2/2 = 100.00%	2/2 = 100.00%	3/3 = 100.00%	1/1 = 100.00%	8	0 (0.00%)
.	.	Δ	O	3 (0.51%)	0/4 = 0.00%	0/1 = 0.00%	6/6 = 100.00%	11/11 = 100.00%	17	3 (100.00%)
Δ	.	O	O	3 (0.51%)	3/8 = 37.50%	2/2 = 100.00%	11/11 = 100.00%	11/11 = 100.00%	27	3 (100.00%)
O	.	.	Δ	2 (0.34%)	1/9 = 11.11%	1/2 = 50.00%	0/0 = 0%	3/4 = 75.00%	5	0 (0.00%)
Δ	Δ	.	O	2 (0.34%)	0/5 = 0.00%	2/4 = 50.00%	2/2 = 100.00%	7/7 = 100.00%	11	2 (100.00%)
O	Δ	Δ	.	4 (0.68%)	5/18 = 27.78%	5/8 = 62.50%	7/8 = 87.50%	1/1 = 100.00%	18	0 (0.00%)
Δ	O	.	O	1 (0.17%)	0/3 = 0.00%	2/4 = 50.00%	0/0 = 0%	2/3 = 66.67%	4	1 (100.00%)
.	Δ	O	.	1 (0.17%)	0/1 = 0.00%	1/2 = 50.00%	4/4 = 100.00%	1/1 = 100.00%	6	0 (0.00%)
Δ	O	Δ	.	1 (0.17%)	1/3 = 33.33%	4/4 = 100.00%	2/2 = 100.00%	0/0 = 0%	7	0 (0.00%)
Δ	O	.	.	1 (0.17%)	0/2 = 0.00%	4/4 = 100.00%	0/0 = 0%	0/0 = 0%	4	0 (0.00%)
Δ	O	.	Δ	1 (0.17%)	1/3 = 33.33%	2/3 = 66.67%	1/1 = 100.00%	1/2 = 50.00%	5	0 (0.00%)
.	.	O	Δ	1 (0.17%)	0/1 = 0.00%	0/1 = 0.00%	1/3 = 33.33%	2/2 = 100.00%	3	0 (0.00%)
Δ	.	.	O	2 (0.34%)	0/4 = 0.00%	0/0 = 0%	2/2 = 100.00%	7/7 = 100.00%	9	0 (0.00%)
O	Δ	.	Δ	2 (0.34%)	6/9 = 66.67%	4/4 = 100.00%	2/2 = 100.00%	4/4 = 100.00%	16	2 (100.00%)
Δ	O	O	D	2 (0.34%)	4/5 = 80.00%	5/6 = 83.33%	6/8 = 75.00%	1/1 = 100.00%	16	0 (0.00%)
Δ	.	Δ	O	1 (0.17%)	0/2 = 0.00%	1/1 = 100.00%	2/2 = 100.00%	3/3 = 100.00%	6	1 (100.00%)

[1]Where O=Most results observed within a window; Δ=Intermittent results observed within a window; D=Death or .=Missing.

Table B4: Number of negative culture results and proportion of patients who achieved negative culture conversion for patients with most culture results observed (i.e. completers') over visit windows for REMoxTB on isoniazid arm[1].

Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78	Total (N=609)	Number of negative culture results Weeks 0-4	Weeks 5-8	Weeks 12-26	Months 39-78	Total number of negative culture results	Treatment success n/no. patient per pattern
O	O	O	O	306 (50.25%)	267/1470 = 18.16%	796/1148 = 69.34%	1107/1142 = 96.94%	1094/1161 = 94.23%	3264	274 (89.54%)
O	O	Δ	O	38 (6.24%)	19/174 = 10.92%	89/136 = 65.44%	72/76 = 94.74%	123/136 = 90.44%	303	33 (86.84%)
O	O	O	.	31 (5.09%)	24/150 = 16.00%	84/115 = 73.04%	104/113 = 92.04%	18/23 = 78.26%	230	3 (9.68%)
O	O	O	Δ	22 (3.61%)	15/102 = 14.71%	54/85 = 63.53%	72/75 = 96.00%	40/44 = 90.91%	181	18 (81.82%)
O	Δ	O	O	23 (3.78%)	14/109 = 12.84%	29/46 = 63.04%	73/82 = 89.02%	73/85 = 85.88%	189	18 (78.26%)
Δ	O	O	O	13 (2.13%)	8/39 = 20.51%	39/48 = 81.25%	44/45 = 97.78%	49/49 = 100.00%	140	13 (100.00%)
.	.	.	.	21 (3.45%)	2/29 = 6.90%	0/0 = 0%	0/0 = 0%	0/0 = 0%	2	0 (0.00%)
O	O	.	.	15 (2.46%)	14/68 = 20.59%	32/54 = 59.26%	5/5 = 100.00%	2/2 = 100.00%	53	0 (0.00%)
.	.	O	O	7 (1.15%)	0/7 = 0.00%	0/1 = 0.00%	26/26 = 100.00%	23/25 = 92.00%	49	6 (85.71%)
O	.	O	O	9 (1.48%)	11/38 = 28.95%	6/7 = 85.71%	33/34 = 97.06%	28/34 = 82.35%	78	6 (66.67%)
O	O	.	O	13 (2.13%)	12/63 = 19.05%	28/46 = 60.87%	11/12 = 91.67%	43/45 = 95.56%	94	9 (69.23%)
O	.	.	.	8 (1.31%)	11/30 = 36.67%	2/3 = 66.67%	1/1 = 100.00%	0/0 = 0%	14	0 (0.00%)
.	O	O	O	3 (0.49%)	2/5 = 40.00%	10/10 = 100.00%	12/12 = 100.00%	12/12 = 100.00%	36	3 (100.00%)
O	D	D	D	2 (0.33%)	2/10 = 20.00%	2/2 = 100.00%	0/0 = 0%	0/0 = 0%	4	0 (0.00%)
O	.	.	O	2 (0.33%)	3/9 = 33.33%	2/2 = 100.00%	2/2 = 100.00%	7/7 = 100.00%	14	2 (100.00%)
O	O	O	D	1 (0.16%)	1/5 = 20.00%	2/4 = 50.00%	3/3 = 100.00%	0/1 = 0.00%	6	0 (0.00%)
.	.	.	O	1 (0.16%)	0/0 = 0%	0/0 = 0%	0/0 = 0%	4/4 = 100.00%	4	0 (0.00%)
.	.	O	.	3 (0.49%)	0/3 = 0.00%	1/1 = 100.00%	8/9 = 88.89%	0/2 = 0.00%	9	0 (0.00%)
D	D	D	D	2 (0.33%)	1/4 = 25.00%	0/0 = 0%	0/0 = 0%	0/0 = 0%	1	0 (0.00%)
O	O	D	D	2 (0.33%)	0/9 = 0.00%	7/8 = 87.50%	2/2 = 100.00%	0/0 = 0%	9	0 (0.00%)
.	.	D	D	1 (0.16%)	0/2 = 0.00%	0/0 = 0%	0/0 = 0%	0/0 = 0%	0	0 (0.00%)

[1]Where O=Most results observed within a window; Δ=Intermittent results observed within a window; D=Death or .=Missing.

Table B5: Number of negative culture results and proportion of patients who achieved negative culture conversion for patients with most culture results intermittently observed over visit windows for REMoxTB on isoniazid arm[1].

Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78	Total (N=609)	Number of negative culture results Weeks 0-4	Weeks 5-8	Weeks 12-26	Months 39-78	Total number of negative culture results	Treatment success n/no. patient per pattern
Δ	Δ	O	O	4 (0.66%)	2/10 = 20.00%	8/8 = 100.00%	14/14 = 100.00%	16/16 = 100.00%	40	4 (100.00%)
O	O	Δ	Δ	8 (1.31%)	4/39 = 10.26%	19/28 = 67.86%	15/16 = 93.75%	14/16 = 87.50%	52	6 (75.00%)
Δ	O	Δ	O	2 (0.33%)	0/6 = 0.00%	3/7 = 42.86%	4/4 = 100.00%	7/7 = 100.00%	14	2 (100.00%)
O	Δ	Δ	O	6 (0.99%)	2/27 = 7.41%	8/12 = 66.67%	12/12 = 100.00%	21/21 = 100.00%	43	6 (100.00%)
O	Δ	O	Δ	2 (0.33%)	1/8 = 12.50%	4/4 = 100.00%	6/6 = 100.00%	4/4 = 100.00%	15	2 (100.00%)
Δ	.	.	.	2 (0.33%)	1/6 = 16.67%	0/1 = 0.00%	0/0 = 0%	0/0 = 0%	1	0 (0.00%)
Δ	Δ	Δ	O	1 (0.16%)	3/3 = 100.00%	2/2 = 100.00%	2/2 = 100.00%	4/4 = 100.00%	11	0 (0.00%)
.	.	Δ	Δ	1 (0.16%)	1/2 = 50.00%	0/0 = 0%	2/2 = 100.00%	1/2 = 50.00%	4	0 (0.00%)
Δ	Δ	.	Δ	1 (0.16%)	2/3 = 66.67%	0/2 = 0.00%	0/0 = 0%	1/2 = 50.00%	3	0 (0.00%)
Δ	Δ	O	Δ	1 (0.16%)	1/3 = 33.33%	1/2 = 50.00%	3/3 = 100.00%	2/2 = 100.00%	7	1 (100.00%)

[1]Where O=Most results observed within a window; Δ=Intermittent results observed within a window or .=Missing.

Table B6: Number of negative culture results and proportion of patients who achieved negative culture conversion for patients with a mixture of observed, intermittent and missing culture results within visit windows for REMoxTB on isoniazid arm[1].

Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78	Total (N=609)	Number of negative culture results Weeks 0-4	Weeks 5-8	Weeks 12-26	Months 39-78	Total number of negative culture results	Treatment success n/no. patient per pattern
O	O	Δ	.	9 (1.48%)	1/42 = 2.38%	14/34 = 41.18%	18/18 = 100.00%	4/5 = 80.00%	37	1 (11.11%)
O	.	Δ	O	6 (0.99%)	9/26 = 34.62%	2/2 = 100.00%	12/12 = 100.00%	18/22 = 81.82%	41	5 (83.33%)
O	Δ	.	.	6 (0.99%)	1/28 = 3.57%	7/12 = 58.33%	2/2 = 100.00%	1/1 = 100.00%	11	0 (0.00%)
O	Δ	O	.	3 (0.49%)	4/13 = 30.77%	4/6 = 66.67%	9/11 = 81.82%	0/1 = 0.00%	17	0 (0.00%)
.	Δ	O	O	3 (0.49%)	1/1 = 100.00%	6/6 = 100.00%	12/12 = 100.00%	12/12 = 100.00%	31	3 (100.00%)
O	Δ	.	O	3 (0.49%)	3/13 = 23.08%	5/6 = 83.33%	3/3 = 100.00%	10/10 = 100.00%	21	3 (100.00%)
O	O	.	Δ	1 (0.16%)	0/4 = 0.00%	1/4 = 25.00%	0/1 = 0.00%	2/2 = 100.00%	3	0 (0.00%)
Δ	O	O	.	5 (0.82%)	7/15 = 46.67%	17/17 = 100.00%	16/19 = 84.21%	0/2 = 0.00%	40	0 (0.00%)
O	O	Δ	D	4 (0.66%)	3/19 = 15.79%	11/14 = 78.57%	8/8 = 100.00%	1/1 = 100.00%	23	0 (0.00%)
.	.	Δ	O	1 (0.16%)	0/0 = 0%	0/0 = 0%	2/2 = 100.00%	4/4 = 100.00%	6	1 (100.00%)
O	.	Δ	.	2 (0.33%)	1/8 = 12.50%	2/2 = 100.00%	4/4 = 100.00%	0/1 = 0.00%	7	0 (0.00%)
Δ	Δ	O	.	2 (0.33%)	1/5 = 20.00%	3/4 = 75.00%	5/7 = 71.43%	1/2 = 50.00%	10	0 (0.00%)
O	.	O	Δ	4 (0.66%)	2/15 = 13.33%	1/3 = 33.33%	14/14 = 100.00%	8/8 = 100.00%	25	4 (100.00%)
Δ	Δ	.	O	2 (0.33%)	3/5 = 60.00%	3/4 = 75.00%	2/2 = 100.00%	5/6 = 83.33%	13	1 (50.00%)
Δ	O	.	O	1 (0.16%)	1/3 = 33.33%	4/4 = 100.00%	1/1 = 100.00%	4/4 = 100.00%	10	1 (100.00%)
Δ	O	.	.	1 (0.16%)	3/3 = 100.00%	4/4 = 100.00%	1/1 = 100.00%	0/0 = 0%	8	0 (0.00%)
.	Δ	O	.	2 (0.33%)	1/4 = 25.00%	4/4 = 100.00%	4/7 = 57.14%	2/2 = 100.00%	11	1 (50.00%)
Δ	.	Δ	O	1 (0.16%)	2/3 = 66.67%	0/0 = 0%	2/2 = 100.00%	3/3 = 100.00%	7	1 (100.00%)
O	.	Δ	Δ	2 (0.33%)	7/9 = 77.78%	0/0 = 0%	4/4 = 100.00%	4/4 = 100.00%	15	2 (100.00%)

[1]Where O=Most results observed within a window; Δ=Intermittent results observed within a window or .=Missing.

Table B7: Number of negative culture results and proportion of patients who achieved negative culture conversion for patients with most culture results observed (i.e. completers') over visit windows for REMoxTB on ethambutol arm[1].

Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78	Total (N=609)	Number of negative culture results				Total number of negative culture results	Treatment success n/no. patient per pattern
					Weeks 0-4	Weeks 5-8	Weeks 12-26	Months 39-78		
O	O	O	O	301 (51.37%)	265/1449 = 18.29%	773/1140 = 67.81%	1050/1107 = 94.85%	1055/1142 = 92.38%	3143	263 (87.38%)
O	O	Δ	O	35 (5.97%)	34/160 = 21.25%	85/125 = 68.00%	64/70 = 91.43%	112/124 = 90.32%	295	30 (85.71%)
O	O	O	.	36 (6.14%)	33/174 = 18.97%	89/131 = 67.94%	104/126 = 82.54%	17/26 = 65.38%	243	9 (25.00%)
O	O	O	Δ	30 (5.12%)	26/147 = 17.69%	74/110 = 67.27%	104/110 = 94.55%	54/60 = 90.00%	258	26 (86.67%)
O	Δ	O	O	25 (4.27%)	22/113 = 19.47%	40/50 = 80.00%	80/83 = 96.39%	87/89 = 97.75%	229	23 (92.00%)
Δ	O	O	O	17 (2.90%)	7/49 = 14.29%	43/60 = 71.67%	54/59 = 91.53%	62/64 = 96.88%	166	15 (88.24%)
O	O	.	.	10 (1.71%)	10/48 = 20.83%	29/40 = 72.50%	6/6 = 100.00%	4/4 = 100.00%	49	0 (0.00%)
.	.	.	.	7 (1.19%)	4/8 = 50.00%	0/0 = 0%	0/0 = 0%	0/0 = 0%	4	0 (0.00%)
.	.	O	O	15 (2.56%)	3/13 = 23.08%	5/5 = 100.00%	53/55 = 96.36%	54/56 = 96.43%	115	12 (80.00%)
O	.	O	O	9 (1.54%)	9/36 = 25.00%	4/6 = 66.67%	31/32 = 96.88%	31/36 = 86.11%	75	6 (66.67%)
O	O	.	O	6 (1.02%)	1/26 = 3.85%	16/23 = 69.57%	6/6 = 100.00%	20/20 = 100.00%	43	6 (100.00%)
O	.	.	.	5 (0.85%)	5/19 = 26.32%	1/1 = 100.00%	0/0 = 0%	1/1 = 100.00%	7	0 (0.00%)
.	O	O	O	5 (0.85%)	4/9 = 44.44%	18/19 = 94.74%	18/20 = 90.00%	19/20 = 95.00%	59	4 (80.00%)
O	D	D	D	3 (0.51%)	3/14 = 21.43%	0/1 = 0.00%	0/0 = 0%	0/0 = 0%	3	0 (0.00%)
.	.	.	O	1 (0.17%)	1/1 = 100.00%	0/0 = 0%	0/0 = 0%	4/4 = 100.00%	5	0 (0.00%)
O	.	O	.	2 (0.34%)	3/7 = 42.86%	0/0 = 0%	5/7 = 71.43%	1/1 = 100.00%	9	1 (50.00%)
.	O	O	.	1 (0.17%)	0/2 = 0.00%	2/3 = 66.67%	3/3 = 100.00%	1/1 = 100.00%	6	0 (0.00%)
.	.	O	D	1 (0.17%)	0/1 = 0.00%	1/1 = 100.00%	3/3 = 100.00%	0/0 = 0%	4	0 (0.00%)
O	O	.	D	1 (0.17%)	1/4 = 25.00%	2/3 = 66.67%	1/1 = 100.00%	0/0 = 0%	4	0 (0.00%)
.	O	.	.	1 (0.17%)	1/2 = 50.00%	4/4 = 100.00%	0/0 = 0%	1/1 = 100.00%	6	1 (100.00%)

[1]Where O=Most results observed within a window; Δ=Intermittent results observed within a window; D=Death or .=Missing.

Table B8: Number of negative culture results and proportion of patients who achieved negative culture conversion for patients with most culture results intermittently observed over visit windows for REMoxTB on ethambutol arm[1].

Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78	Total (N=609)	Number of negative culture results Weeks 0-4	Weeks 5-8	Weeks 12-26	Months 39-78	Total number of negative culture results	Treatment success n/no. patient per pattern
Δ	Δ	O	O	8 (1.37%)	14/22 = 63.64%	16/16 = 100.00%	26/28 = 92.86%	27/31 = 87.10%	83	6 (75.00%)
O	O	Δ	Δ	7 (1.19%)	3/35 = 8.57%	17/25 = 68.00%	13/14 = 92.86%	13/14 = 92.86%	46	6 (85.71%)
O	Δ	Δ	O	4 (0.68%)	1/18 = 5.56%	4/8 = 50.00%	7/8 = 87.50%	14/15 = 93.33%	26	3 (75.00%)
O	Δ	O	Δ	2 (0.34%)	0/9 = 0.00%	4/4 = 100.00%	7/7 = 100.00%	4/4 = 100.00%	15	2 (100.00%)
Δ	.	.	.	3 (0.51%)	1/8 = 12.50%	1/1 = 100.00%	0/0 = 0%	0/0 = 0%	2	0 (0.00%)
Δ	Δ	Δ	O	1 (0.17%)	0/3 = 0.00%	1/2 = 50.00%	2/2 = 100.00%	4/4 = 100.00%	7	1 (100.00%)
O	Δ	Δ	Δ	1 (0.17%)	1/5 = 20.00%	1/2 = 50.00%	2/2 = 100.00%	2/2 = 100.00%	6	1 (100.00%)
Δ	.	Δ	.	1 (0.17%)	0/2 = 0.00%	1/1 = 100.00%	2/2 = 100.00%	0/0 = 0%	3	0 (0.00%)
Δ	Δ	.	.	1 (0.17%)	1/2 = 50.00%	2/2 = 100.00%	1/1 = 100.00%	0/0 = 0%	4	0 (0.00%)
Δ	.	Δ	Δ	1 (0.17%)	0/3 = 0.00%	0/0 = 0%	2/2 = 100.00%	2/2 = 100.00%	4	1 (100.00%)
Δ	O	O	Δ	1 (0.17%)	0/3 = 0.00%	2/4 = 50.00%	4/4 = 100.00%	2/2 = 100.00%	8	1 (100.00%)

[1]Where O=Most results observed within a window; Δ=Intermittent results observed within a window or .=Missing.

Table B9: Number of negative culture results and proportion of patients who achieved negative culture conversion for patients with a mixture of observed, intermittent and missing culture results within visit windows for REMoxTB on ethambutol arm[1].

O	O	Δ	.	6 (1.02%)	1/28 = 3.57%	6/22 = 27.27%	9/12 = 75.00%	3/3 = 100.00%	19	0 (0.00%)
O	.	Δ	O	2 (0.34%)	0/10 = 0.00%	0/1 = 0.00%	4/4 = 100.00%	6/6 = 100.00%	10	2 (100.00%)
O	Δ	.	.	2 (0.34%)	2/9 = 22.22%	2/4 = 50.00%	0/0 = 0%	0/0 = 0%	4	0 (0.00%)
O	Δ	.	O	1 (0.17%)	0/4 = 0.00%	1/2 = 50.00%	0/1 = 0.00%	1/3 = 33.33%	2	0 (0.00%)
O	Δ	O	.	2 (0.34%)	2/8 = 25.00%	3/4 = 75.00%	6/7 = 85.71%	1/1 = 100.00%	12	1 (50.00%)
.	Δ	O	O	3 (0.51%)	0/3 = 0.00%	6/6 = 100.00%	11/11 = 100.00%	11/11 = 100.00%	28	3 (100.00%)
O	O	.	Δ	4 (0.68%)	5/20 = 25.00%	14/14 = 100.00%	2/2 = 100.00%	8/8 = 100.00%	29	4 (100.00%)
O	O	Δ	D	1 (0.17%)	2/5 = 40.00%	2/4 = 50.00%	2/2 = 100.00%	0/0 = 0%	6	0 (0.00%)
Δ	O	O	.	2 (0.34%)	1/5 = 20.00%	3/7 = 42.86%	6/7 = 85.71%	0/0 = 0%	10	0 (0.00%)
Δ	.	O	O	3 (0.51%)	3/8 = 37.50%	2/3 = 66.67%	10/10 = 100.00%	12/12 = 100.00%	27	3 (100.00%)
.	.	Δ	O	2 (0.34%)	3/4 = 75.00%	0/0 = 0%	4/4 = 100.00%	7/7 = 100.00%	14	1 (50.00%)
Δ	Δ	O	.	3 (0.51%)	1/9 = 11.11%	3/6 = 50.00%	8/10 = 80.00%	0/2 = 0.00%	12	0 (0.00%)
O	.	Δ	.	2 (0.34%)	1/8 = 12.50%	2/2 = 100.00%	4/4 = 100.00%	1/1 = 100.00%	8	0 (0.00%)
O	.	O	Δ	1 (0.17%)	1/5 = 20.00%	1/1 = 100.00%	4/4 = 100.00%	2/2 = 100.00%	8	1 (100.00%)
O	.	.	Δ	2 (0.34%)	1/8 = 12.50%	0/0 = 0%	1/1 = 100.00%	4/4 = 100.00%	6	0 (0.00%)
Δ	O	Δ	.	2 (0.34%)	2/6 = 33.33%	4/6 = 66.67%	3/4 = 75.00%	1/1 = 100.00%	10	1 (50.00%)
Δ	O	.	.	1 (0.17%)	2/3 = 66.67%	2/3 = 66.67%	0/0 = 0%	0/0 = 0%	4	0 (0.00%)
Δ	O	.	O	1 (0.17%)	0/3 = 0.00%	2/3 = 66.67%	1/1 = 100.00%	3/3 = 100.00%	6	1 (100.00%)
Δ	.	D	D	2 (0.34%)	1/5 = 20.00%	2/2 = 100.00%	0/0 = 0%	0/0 = 0%	3	0 (0.00%)
Δ	O	.	Δ	1 (0.17%)	0/3 = 0.00%	3/3 = 100.00%	1/1 = 100.00%	2/2 = 100.00%	6	1 (100.00%)
.	.	O	Δ	1 (0.17%)	0/1 = 0.00%	0/0 = 0%	3/4 = 75.00%	2/2 = 100.00%	5	1 (100.00%)
Δ	.	O	Δ	1 (0.17%)	0/2 = 0.00%	0/0 = 0%	4/4 = 100.00%	2/2 = 100.00%	6	1 (100.00%)

[1]Where O=Most results observed within a window; Δ=Intermittent results observed within a window or .=Missing.

C Missing data patterns for RIFAQUIN

Table C10: Number of negative culture results and proportion of patients who achieved negative culture conversion for patients with most culture results observed (i.e. completers') over visit windows for RIFAQUIN on control arm[1].

Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78	Total (N=590)	Number of negative culture results Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78	Total number of negative culture results	Treatment success n/no. patient per pattern
O	O	O	O	127 (52.92%)	227/372 = 61.02%	369/374 = 98.66%	478/484 = 98.76%	456/469 = 97.23%	1530	121 (95.28%)
O	O	O	Δ	17 (7.08%)	30/50 = 60.00%	47/50 = 94.00%	56/59 = 94.92%	32/34 = 94.12%	165	16 (94.12%)
O	O	O	.	16 (6.67%)	28/45 = 62.22%	46/46 = 100.00%	57/57 = 100.00%	13/13 = 100.00%	144	16 (100.00%)
O	O	.	.	17 (7.08%)	29/51 = 56.86%	40/46 = 86.96%	4/5 = 80.00%	4/4 = 100.00%	77	3 (17.65%)
O	.	.	.	8 (3.33%)	13/21 = 61.90%	4/4 = 100.00%	0/0 = 0%	1/1 = 100.00%	18	1 (12.50%)
.	.	.	.	7 (2.92%)	0/7 = 0.00%	0/0 = 0%	0/0 = 0%	1/2 = 50.00%	1	0 (0.00%)
O	Δ	O	O	7 (2.92%)	11/21 = 52.38%	13/14 = 92.86%	26/26 = 100.00%	25/26 = 96.15%	75	6 (85.71%)
Δ	O	O	O	8 (3.33%)	8/16 = 50.00%	22/22 = 100.00%	30/30 = 100.00%	28/28 = 100.00%	88	8 (100.00%)
O	O	Δ	O	4 (1.67%)	8/12 = 66.67%	11/11 = 100.00%	6/8 = 75.00%	15/15 = 100.00%	40	4 (100.00%)
O	.	O	O	4 (1.67%)	6/11 = 54.55%	4/4 = 100.00%	13/13 = 100.00%	14/14 = 100.00%	37	3 (75.00%)
.	O	O	O	1 (0.42%)	0/1 = 0.00%	2/2 = 100.00%	4/4 = 100.00%	3/3 = 100.00%	9	1 (100.00%)
O	.	.	O	1 (0.42%)	0/2 = 0.00%	0/0 = 0%	1/1 = 100.00%	3/3 = 100.00%	4	0 (0.00%)
O	O	O	D	1 (0.42%)	2/3 = 66.67%	3/3 = 100.00%	3/3 = 100.00%	0/0 = 0%	8	1 (100.00%)
O	O	D	D	1 (0.42%)	2/3 = 66.67%	2/2 = 100.00%	0/0 = 0%	0/0 = 0%	4	0 (0.00%)
O	D	D	D	1 (0.42%)	2/3 = 66.67%	1/1 = 100.00%	0/0 = 0%	0/0 = 0%	3	0 (0.00%)
.	.	O	O	1 (0.42%)	0/1 = 0.00%	1/1 = 100.00%	4/4 = 100.00%	3/3 = 100.00%	8	0 (0.00%)
D	D	D	D	1 (0.42%)	0/1 = 0.00%	0/0 = 0%	0/0 = 0%	0/0 = 0%	0	0 (0.00%)

[1]Where O=Most results observed within a window; Δ=Intermittent results observed within a window; D=Death or .=Missing.

Table C11: Number of negative culture results and proportion of patients who achieved negative culture conversion for patients with most culture results intermittently observed over visit windows for RIFAQUIN on control arm[1].

Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78	Total (N=590)	Number of negative culture results				Total number of negative culture results	Treatment success n/no. patient per pattern
					Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78		
O	O	Δ	Δ	3 (1.25%)	6/9 = 66.67%	8/8 = 100.00%	6/6 = 100.00%	6/6 = 100.00%	26	3 (100.00%)
Δ	O	O	Δ	2 (0.83%)	2/4 = 50.00%	6/6 = 100.00%	8/8 = 100.00%	4/4 = 100.00%	20	2 (100.00%)
O	Δ	O	Δ	1 (0.42%)	2/3 = 66.67%	2/2 = 100.00%	4/4 = 100.00%	2/2 = 100.00%	10	1 (100.00%)
Δ	O	Δ	O	1 (0.42%)	1/2 = 50.00%	3/3 = 100.00%	2/2 = 100.00%	3/3 = 100.00%	9	1 (100.00%)
O	Δ	Δ	O	1 (0.42%)	1/3 = 33.33%	2/2 = 100.00%	2/2 = 100.00%	3/3 = 100.00%	8	1 (100.00%)

[1]Where O=Most results observed within a window; Δ=Intermittent results observed within a window or .=Missing.

Table C12: Number of negative culture results and proportion of patients who achieved negative culture conversion for patients with a mixture of observed, intermittent and missing culture results within visit windows for RIFAQUIN on control arm[1].

Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78	Total (N=590)	Number of negative culture results Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78	Total number of negative culture results	Treatment success n/no. patient per pattern
O	O	Δ	.	6 (2.50%)	11/18 = 61.11%	16/16 = 100.00%	12/12 = 100.00%	2/2 = 100.00%	41	6 (100.00%)
O	O	.	Δ	1 (0.42%)	2/3 = 66.67%	3/3 = 100.00%	1/1 = 100.00%	2/2 = 100.00%	8	1 (100.00%)
O	.	.	Δ	1 (0.42%)	2/3 = 66.67%	1/1 = 100.00%	0/0 = 0%	2/2 = 100.00%	5	1 (100.00%)
O	O	Δ	D	1 (0.42%)	1/2 = 50.00%	3/3 = 100.00%	2/2 = 100.00%	0/0 = 0%	6	1 (100.00%)
O	Δ	.	Δ	1 (0.42%)	2/3 = 66.67%	2/2 = 100.00%	0/0 = 0%	2/2 = 100.00%	6	1 (100.00%)

[1]Where O=Most results observed within a window; Δ=Intermittent results observed within a window; D=Death or .=Missing.

Table C13: Number of negative culture results and proportion of patients who achieved negative culture conversion for patients with most culture results observed (i.e. completers') over visit windows for RIFAQUIN on isoniazid arm[1].

Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78	Total (N=609)	Number of negative culture results Weeks 0-4	Weeks 5-8	Weeks 12-26	Months 39-78	Total number of negative culture results	Treatment success n/no. patient per pattern
O	O	O	O	123 (51.46%)	222/358 = 62.01%	338/352 = 96.02%	434/464 = 93.53%	436/450 = 96.89%	1430	105 (85.37%)
O	O	O	Δ	25 (10.46%)	47/73 = 64.38%	73/74 = 98.65%	93/97 = 95.88%	49/50 = 98.00%	262	23 (92.00%)
O	O	O	.	12 (5.02%)	23/36 = 63.89%	32/33 = 96.97%	37/44 = 84.09%	4/4 = 100.00%	96	8 (66.67%)
O	O	.	.	12 (5.02%)	24/36 = 66.67%	30/31 = 96.77%	2/3 = 66.67%	3/3 = 100.00%	59	10 (83.33%)
O	.	.	.	13 (5.44%)	16/35 = 45.71%	7/7 = 100.00%	0/0 = 0%	2/3 = 66.67%	25	2 (15.38%)
.	.	.	.	10 (4.18%)	0/10 = 0.00%	0/0 = 0%	0/0 = 0%	1/1 = 100.00%	1	0 (0.00%)
O	Δ	O	O	8 (3.35%)	15/23 = 65.22%	15/16 = 93.75%	30/31 = 96.77%	29/29 = 100.00%	89	8 (100.00%)
Δ	O	O	O	5 (2.09%)	4/10 = 40.00%	13/14 = 92.86%	16/20 = 80.00%	18/18 = 100.00%	51	4 (80.00%)
O	O	Δ	O	2 (0.84%)	4/6 = 66.67%	6/6 = 100.00%	3/4 = 75.00%	8/8 = 100.00%	21	2 (100.00%)
O	.	O	O	1 (0.42%)	2/3 = 66.67%	0/0 = 0%	3/3 = 100.00%	2/3 = 66.67%	7	0 (0.00%)
O	O	O	D	1 (0.42%)	2/3 = 66.67%	3/3 = 100.00%	4/4 = 100.00%	0/0 = 0%	9	1 (100.00%)
O	.	.	O	1 (0.42%)	2/3 = 66.67%	1/1 = 100.00%	0/0 = 0%	3/3 = 100.00%	6	1 (100.00%)
.	.	O	O	1 (0.42%)	0/1 = 0.00%	1/1 = 100.00%	3/3 = 100.00%	4/4 = 100.00%	8	0 (0.00%)
O	O	D	D	1 (0.42%)	2/3 = 66.67%	2/2 = 100.00%	0/0 = 0%	0/0 = 0%	4	0 (0.00%)
O	D	D	D	1 (0.42%)	2/3 = 66.67%	1/1 = 100.00%	0/0 = 0%	0/0 = 0%	3	0 (0.00%)
.	.	.	O	1 (0.42%)	0/1 = 0.00%	0/0 = 0%	1/1 = 100.00%	4/4 = 100.00%	5	0 (0.00%)
O	O	.	D	1 (0.42%)	2/3 = 66.67%	3/3 = 100.00%	1/1 = 100.00%	0/0 = 0%	6	1 (100.00%)
.	.	.	D	1 (0.42%)	0/1 = 0.00%	0/0 = 0%	0/0 = 0%	0/0 = 0%	0	0 (0.00%)
.	O	O	.	1 (0.42%)	0/1 = 0.00%	3/3 = 100.00%	3/3 = 100.00%	1/1 = 100.00%	7	0 (0.00%)

[1]Where O=Most results observed within a window; Δ=Intermittent results observed within a window; D=Death or .=Missing.

Table C14: Number of negative culture results and proportion of patients who achieved negative culture conversion for patients with most culture results intermittently observed over visit windows for RIFAQUIN on isoniazid arm[1].

Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78	Total (N=609)	Number of negative culture results Weeks 0-4	Weeks 5-8	Weeks 12-26	Months 39-78	Total number of negative culture results	Treatment success n/no. patient per pattern
O	O	Δ	Δ	3 (1.26%)	6/9 = 66.67%	9/9 = 100.00%	4/6 = 66.67%	6/6 = 100.00%	25	2 (66.67%)
Δ	O	O	Δ	2 (0.84%)	1/4 = 25.00%	5/6 = 83.33%	5/6 = 83.33%	4/4 = 100.00%	15	1 (50.00%)
Δ	Δ	O	Δ	1 (0.42%)	1/2 = 50.00%	2/2 = 100.00%	4/4 = 100.00%	2/2 = 100.00%	9	1 (100.00%)
.	.	Δ	Δ	1 (0.42%)	0/1 = 0.00%	1/1 = 100.00%	2/2 = 100.00%	2/2 = 100.00%	5	0 (0.00%)
Δ	Δ	Δ	O	1 (0.42%)	1/2 = 50.00%	2/2 = 100.00%	2/2 = 100.00%	4/4 = 100.00%	9	1 (100.00%)

[1]Where O=Most results observed within a window; Δ=Intermittent results observed within a window or .=Missing.

Table C15: Number of negative culture results and proportion of patients who achieved negative culture conversion for patients with a mixture of observed, intermittent and missing culture results within visit windows for RIFAQUIN on isoniazid arm[1].

Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78	Total (N=609)	Number of negative culture results Weeks 0-4	Weeks 5-8	Weeks 12-26	Months 39-78	Total number of negative culture results	Treatment success n/no. patient per pattern
O	O	Δ	.	3 (1.26%)	6/9 = 66.67%	9/9 = 100.00%	6/6 = 100.00%	1/1 = 100.00%	22	2 (66.67%)
O	.	Δ	O	2 (0.84%)	3/6 = 50.00%	2/2 = 100.00%	3/4 = 75.00%	8/8 = 100.00%	16	1 (50.00%)
O	O	.	Δ	1 (0.42%)	2/3 = 66.67%	3/3 = 100.00%	0/0 = 0%	2/2 = 100.00%	7	1 (100.00%)
O	.	O	Δ	1 (0.42%)	1/3 = 33.33%	1/1 = 100.00%	3/3 = 100.00%	2/2 = 100.00%	7	1 (100.00%)
.	O	O	Δ	1 (0.42%)	0/1 = 0.00%	3/3 = 100.00%	1/4 = 25.00%	2/2 = 100.00%	6	0 (0.00%)
O	Δ	O	D	1 (0.42%)	2/3 = 66.67%	2/2 = 100.00%	3/3 = 100.00%	0/0 = 0%	7	1 (100.00%)
Δ	Δ	O	.	1 (0.42%)	1/2 = 50.00%	2/2 = 100.00%	4/4 = 100.00%	1/1 = 100.00%	8	1 (100.00%)
O	Δ	D	D	1 (0.42%)	2/3 = 66.67%	2/2 = 100.00%	0/0 = 0%	0/0 = 0%	4	0 (0.00%)

[1]Where O=Most results observed within a window; Δ=Intermittent results observed within a window or .=Missing.

Table C16: Number of negative culture results and proportion of patients who achieved negative culture conversion for patients with most culture results observed (i.e. completers') over visit windows for RIFAQUIN on ethambutol arm[1].

Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78	Total (N=609)	Number of negative culture results Weeks 0-4	Weeks 5-8	Weeks 12-26	Months 39-78	Total number of negative culture results	Treatment success n/no. patient per pattern
O	O	O	O	134 (53.39%)	246/394 = 62.44%	381/386 = 98.70%	503/510 = 98.63%	491/496 = 98.99%	1621	127 (94.78%)
O	O	O	Δ	18 (7.17%)	31/53 = 58.49%	53/53 = 100.00%	66/69 = 95.65%	34/36 = 94.44%	184	17 (94.44%)
O	O	O	.	14 (5.58%)	25/41 = 60.98%	41/41 = 100.00%	45/45 = 100.00%	12/12 = 100.00%	123	14 (100.00%)
O	O	.	.	11 (4.38%)	20/31 = 64.52%	29/29 = 100.00%	3/3 = 100.00%	2/2 = 100.00%	54	4 (36.36%)
O	.	.	.	10 (3.98%)	16/28 = 57.14%	4/4 = 100.00%	1/1 = 100.00%	1/1 = 100.00%	22	0 (0.00%)
.	.	.	.	9 (3.59%)	0/9 = 0.00%	0/0 = 0%	0/0 = 0%	0/0 = 0%	0	0 (0.00%)
O	Δ	O	O	7 (2.79%)	14/21 = 66.67%	14/14 = 100.00%	27/27 = 100.00%	28/28 = 100.00%	83	7 (100.00%)
Δ	O	O	O	6 (2.39%)	5/12 = 41.67%	18/18 = 100.00%	22/24 = 91.67%	21/23 = 91.30%	66	4 (66.67%)
O	O	Δ	O	7 (2.79%)	10/19 = 52.63%	17/18 = 94.44%	14/14 = 100.00%	23/24 = 95.83%	64	6 (85.71%)
O	.	O	O	3 (1.20%)	6/9 = 66.67%	2/2 = 100.00%	11/11 = 100.00%	12/12 = 100.00%	31	3 (100.00%)
.	O	O	O	3 (1.20%)	0/3 = 0.00%	8/8 = 100.00%	11/11 = 100.00%	11/11 = 100.00%	30	3 (100.00%)
O	.	.	O	1 (0.40%)	2/3 = 66.67%	1/1 = 100.00%	1/1 = 100.00%	3/3 = 100.00%	7	1 (100.00%)
O	O	O	D	1 (0.40%)	2/3 = 66.67%	2/2 = 100.00%	3/3 = 100.00%	0/0 = 0%	7	1 (100.00%)
.	.	.	O	1 (0.40%)	0/1 = 0.00%	0/0 = 0%	0/0 = 0%	4/4 = 100.00%	4	0 (0.00%)
O	O	.	D	1 (0.40%)	1/2 = 50.00%	3/3 = 100.00%	1/1 = 100.00%	0/0 = 0%	5	0 (0.00%)
O	O	.	O	2 (0.80%)	4/6 = 66.67%	6/6 = 100.00%	2/2 = 100.00%	6/6 = 100.00%	18	2 (100.00%)
D	D	D	D	1 (0.40%)	0/1 = 0.00%	0/0 = 0%	0/0 = 0%	0/0 = 0%	0	0 (0.00%)

[1]Where O=Most results observed within a window; Δ=Intermittent results observed within a window; D=Death or .=Missing.

Table C17: Number of negative culture results and proportion of patients who achieved negative culture conversion for patients with most culture results intermittently observed over visit windows for RIFAQUIN on ethambutol arm[1].

Weeks 0-4	Weeks 5-8	Weeks 12-26	Weeks 39-78	Total (N=609)	Number of negative culture results Weeks 0-4	Weeks 5-8	Weeks 12-26	Months 39-78	Total number of negative culture results	Treatment success n/no. patient per pattern
O	O	Δ	Δ	3 (1.20%)	4/8 = 50.00%	9/9 = 100.00%	6/6 = 100.00%	6/6 = 100.00%	25	3 (100.00%)
Δ	O	O	Δ	4 (1.59%)	3/8 = 37.50%	10/10 = 100.00%	15/15 = 100.00%	8/8 = 100.00%	36	4 (100.00%)
O	Δ	O	Δ	2 (0.80%)	4/6 = 66.67%	4/4 = 100.00%	8/8 = 100.00%	4/4 = 100.00%	20	2 (100.00%)
Δ	Δ	O	O	1 (0.40%)	1/2 = 50.00%	2/2 = 100.00%	4/4 = 100.00%	4/4 = 100.00%	11	1 (100.00%)

[1]Where O=Most results observed within a window; Δ=Intermittent results observed within a window or .=Missing.

Table C18: Number of negative culture results and proportion of patients who achieved negative culture conversion for patients with a mixture of observed, intermittent and missing culture results within visit windows for RIFAQUIN on ethambutol arm[1].

O	O	Δ	.	3 (1.20%)	6/9 = 66.67%	8/8 = 100.00%	6/6 = 100.00%	1/1 = 100.00%	21	3 (100.00%)
O	.	Δ	O	3 (1.20%)	5/9 = 55.56%	2/2 = 100.00%	6/6 = 100.00%	11/11 = 100.00%	24	3 (100.00%)
O	O	.	Δ	1 (0.40%)	2/3 = 66.67%	3/3 = 100.00%	1/1 = 100.00%	2/2 = 100.00%	8	1 (100.00%)
O	.	.	Δ	1 (0.40%)	1/2 = 50.00%	1/1 = 100.00%	1/1 = 100.00%	2/2 = 100.00%	5	1 (100.00%)
Δ	.	O	O	1 (0.40%)	1/2 = 50.00%	1/1 = 100.00%	3/3 = 100.00%	3/3 = 100.00%	8	1 (100.00%)
O	Δ	O	.	1 (0.40%)	2/3 = 66.67%	2/2 = 100.00%	4/4 = 100.00%	1/1 = 100.00%	9	1 (100.00%)
Δ	.	.	O	1 (0.40%)	1/2 = 50.00%	1/1 = 100.00%	1/1 = 100.00%	3/3 = 100.00%	6	1 (100.00%)
O	Δ	.	.	1 (0.40%)	2/3 = 66.67%	2/2 = 100.00%	0/0 = 0%	0/0 = 0%	4	0 (0.00%)

[1]Where O=Most results observed within a window; Δ=Intermittent results observed within a window or .=Missing.

D Predictions of outcome failure and withdrawals for REMoxTB

Table D1: Unadjusted odds ratios (OR), standard errors (SE) and confidence intervals (CI) for predicting outcome failure for all covariates included in the model for the REMoxTB study

Covariate	OR	SE	95% CI	P-value
Treatment				
Isoniazid	1.770	0.297	(1.274, 2.458)	0.001
Ethambutol	1.897	0.318	(1.366, 2.633)	<0.001
Baseline DTP	0.995	0.012	(0.972, 1.019)	0.686
Weight band (adjusted)				
40-45 kg	1.359	0.328	(0.846, 2.182)	0.204
>45-55 kg	0.919	0.204	(0.594, 1.421)	0.703
>55 kg	0.754	0.174	(0.480, 1.184)	0.220
Age	1.013	0.005	(1.003, 1.023)	0.013
Chest X-ray cavities				
Yes	1.279	0.219	(0.914, 1.789)	0.151
Smoker				
Past	1.546	0.242	(1.137, 2.102)	0.005
Current	1.182	0.182	(0.874, 1.599)	0.277
Race				
Black	0.653	0.094	(0.492, 0.866)	0.003
Mixed Race or Coloured	0.538	0.096	(0.379, 0.763)	0.001
Other (N=3)	1.877	2.307	(0.169, 20.880)	0.608
HIV				
Positive	1.914	0.408	(1.260, 2.908)	0.002
Sex				
Female	0.547	0.085	(0.403, 0.741)	<0.001
Centre (adjusted)				
Cape Town	1.050	0.250	(0.658, 1.674)	0.838

Other South Africa	1.223	0.310	(0.744, 2.011)	0.428
India	2.269	0.446	(1.544, 3.334)	<0.001
Kenya/Zambia/Tanzania	1.314	0.274	(1.544, 3.334)	<0.001
Other (East Asia)	1.139	0.295	(0.686, 1.891)	0.616
Sputum production				
No	0.711	0.092	(0.552, 0.915)	0.008
Time to not producing sputum	0.550	0.139	(0.335, 0.902)	0.018

Table D2: Adjusted odds ratios (OR), standard errors (SE) and confidence intervals (CI) for predicting outcome failure for all covariates included in the model for the REMoxTB study

Covariate	Adjusted OR	SE	95% CI	P-value
Treatment				
Isoniazid	1.796	-0.346	(1.231, 2.619)	0.002
Ethambutol	2.237	-0.423	(1.544, 3.240)	<0.001
Baseline DTP	0.976	-0.015	(0.947, 1.006)	0.115
Weight band (adjusted)				
40-45 kg	1.323	-0.385	(0.748, 2.340)	0.336
>45-55 kg	0.875	-0.256	(0.494, 1.552)	0.649
>55 kg	0.708	-0.218	(0.388, 1.293)	0.261
Age	1.010	-0.006	(0.998, 1.023)	0.095
Chest X-ray cavities				
Yes	1.122	-0.211	(0.776, 1.623)	0.541
Smoker				
Past	0.784	-0.371	(1.186, 2.682)	0.005
Current	1.872	-0.432	(1.191, 2.941)	0.007
Race				
Black	0.104	-0.139	(0.008, 1.432)	0.091
Mixed Race or Coloured	0.107	-0.143	(0.008, 1.478)	0.095
Other (N=3)	0.31	-0.564	(0.009, 10.99)	0.52

HIV				
Positive	3.053	-0.88	(1.736, 5.370)	<0.001
Sex				
Female	0.506	-0.105	(0.337, 0.760)	0.001
Centre (adjusted)				
Cape Town	0.902	-0.246	(0.529, 1.539)	0.705
Other South Africa	1.63	-0.538	(0.853, 3.115)	0.139
India	0.28	-0.386	(0.019, 4.153)	0.355
Kenya/Zambia/Tanzania	1.605	-0.501	(0.870, 2.960)	0.13
Other (East Asia)	0.117	-0.16	(0.008, 1.720)	0.118
Sputum production				
No	0.19	-0.046	(0.117, 0.306)	<0.001
Time to not producing sputum	0.047	-0.021	(0.020, 0.114)	<0.001

E Working correlation matrices

1. First order autoregressive which assumes observations which are closer together are more similar than observations further apart.

$$R_{k,t}(\alpha) = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$$

2. Exchangeable which assumes repeated observations have the same correlation.

$$R_{k,t}(\alpha) = \begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix}$$

F Predictions of outcome failure and withdrawals for RIFAQUIN

Table F1: Unadjusted odds ratios (OR), standard errors (SE) and confidence intervals (CI) for predicting withdrawals for all covariates included in the model for the RIFAQUIN study

Covariate	OR	SE	95% CI	P-value
Treatment				
4m regimen	1.336	0.326	(0.828, 2.156)	0.236
6m regimen	0.804	0.221	(0.468, 1.379)	0.427
Baseline DTP	1.020	0.015	(0.990, 1.051)	0.186
Weight band (adjusted)				
40-45 kg	0.577	0.206	(0.287, 1.162)	0.124
>45-55 kg	0.476	0.148	(0.259, 0.875)	0.017
>55 kg	0.525	0.165	(0.259, 0.972)	0.04
Age	1.011	0.008	0.995, 1.028)	0.174
Chest X-ray cavities				
Yes	0.810	0.207	(0.491, 1.335)	0.408
Smoker				
Past	0.982	0.255	(0.590, 1.635)	0.945
Current	0.940	0.230	(0.582, 1.519)	0.800
Race				
Black	0.766	0.171	(0.495, 1.187)	0.233
Mixed Race or Coloured	0.361	0.121	(0.187, 0.695)	0.002
HIV				
Positive	1.004	0.405	(0.455, 2.215)	0.991
Sex				
Female	0.817	0.192	(0.516, 1.295)	0.390
Centre (adjusted)				
Cape Town	2.293	0.907	(1.056, 4.979)	0.036
Other South Africa	1.451	0.687	(0.574, 3.668)	0.423

India	3.231	1.155	(1.603, 6.512)	0.001
Kenya/Zambia/Tanzania	1.951	0.740	(0.927, 4.105)	0.008
Other (East Asia)	1.659	0.960	(0.675, 4.074)	0.616
Sputum production				
No	0.486	0.106	(0.317, 0.745)	0.001
Time to not producing sputum	0.079	0.032	(0.035, 0.177)	<0.001

Table F2: Adjusted odds ratios (OR), standard errors (SE) and confidence intervals (CI) for predicting withdrawals for all covariates included in the model for the RIFAQUIN study

Covariate	Adjusted OR	SE	95% CI	P-value
Treatment				
4m regimen	0.9611	0.300	(0.522, 1.771)	0.899
6m regimen	0.7012	0.232	(0.367, 1.340)	0.282
Baseline DTP	1.0146	0.019	(0.977, 1.053)	0.449
Weight band (adjusted)				
40-45 kg	0.673	0.321	(0.264, 1.713)	0.406
>45-55 kg	0.457	0.223	(0.176, 1.190)	0.109
>55 kg	0.465	0.234	(0.173, 1.249)	0.129
Age	1.015	0.011	(0.994, 1.036)	0.172
Chest X-ray cavities				
Yes	0.747	0.234	(0.405, 1.379)	0.352
Smoker				
Past	2.005	0.764	(0.950, 4.229)	0.068
Current	2.482	1.04	(1.092, 5.642)	0.03
Race				
Black	7.10×10^4	52161832	0	0.988
Mixed Race or Coloured	2.66×10^4	19552427	0	0.989
HIV				
Positive	1.136	0.617	(0.392, 3.291)	0.814

Sex				
Female	0.921	0.33	(0.456, 1.859)	0.817
Centre (adjusted)				
Cape Town	2.343	1.132	(0.909, 6.038)	0.078
Other South Africa	1.723	1.056	(0.518, 5.729)	0.375
India	9.85×10^4	72397857	0	0.988
Kenya/Zambia/Tanzania	1.077	0.597	(0.364, 3.190)	0.893
Other (East Asia)	5.48×10^4	40253957	0	0.988
Sputum production				
No	0.022	0.01	(0.009, 0.055)	<0.001
Time to not producing sputum	0.001	0	(0.000, 0.002)	<0.001

G Simulation of transition probabilities in R

```
> # A. Simulation of transition probabilities assuming transition intensites
    remain constant over time
>
> # Purpose: Simulate data from probability transitions after choosing transition
     intensities and trace back to the known transition intensities and
    probabilities.
>
> library(msm)
>
> # 1. Generate a matrix of constant intensities.
> Q<- matrix( c(-.1,.1,.3,-.3),ncol=2,byrow=T)
>
> # Under this intensity matrix, we get the following transmission probabilities,
     at time t = 0, 1, 2, 3, 4.
>
>  MatrixExp(Q*0)
     State 1   State 2
State 1    1    0
State 2    0    1
> #  State 1   State 2
> #State 1   1    0
> #State 2    0    1
>
>  MatrixExp(Q*1)
     State 1   State 2
State 1 0.91758 0.08241999
State 2 0.24726 0.75274003
>
>  MatrixExp(Q*2)
       State 1   State 2
State 1 0.8623322 0.1376678
State 2 0.4130033 0.5869967
>
>  MatrixExp(Q*3)
       State 1   State 2
State 1 0.8252986 0.1747014
State 2 0.5241043 0.4758957
>
>  MatrixExp(Q*4)
       State 1   State 2
State 1 0.8004741 0.1995259
State 2 0.5985776 0.4014224

>
>
```

```
> # 2. Simulate states under the above transmission probabilities for 10 time
    points.
> ntimes<-10
> Y.1.2<-rep(0,ntimes)
> Y.2.1<-rep(0,ntimes)
>
> # 2a. Step 1: Extract transmission probabilities from
> # intensity matrix using matrix exponential.
>
> for( i in 1:ntimes)  {
+
+ Y.1.2[i]<- MatrixExp(Q*i)[1,2] # State 1 to State 2 from t=i-1 to t=i
+ Y.2.1[i]<- MatrixExp(Q*i)[2,1] # State 2 to State 1 from t=i-1 to t=1
+
+ }
>
> # True probability transition values
> Y.P.N
 [1] 0.08241999 0.13766776 0.17470145 0.19952587 0.21616618 0.22732051 0.23479748
 [8] 0.23980945 0.24316907 0.24542109
>
> Y.N.P
 [1] 0.2472600 0.4130033 0.5241043 0.5985776 0.6484985 0.6819615 0.7043925
     0.7194283
 [9] 0.7295072 0.7362633
>
> # Step 2: Simulate data for 10,000 patients:
> > npat<-10000
>
> Y<-matrix(NA,nrow=npat,ncol=ntimes+1)
>
> labs<-rep("",ntimes+1)
> for(i in 1: (ntimes+1) ) { labs[i]<-paste("t=",i-1,sep="")  }
>
> dimnames(Y)[[2]]<-labs
> dimnames(Y)[[1]]<-1:dim(Y)[1]
>
> Set a random seed number so simulations are re-producible.
> set.seed(1875263)
>
> Y[,1]<-rep(c(0,1),npat/2)
>
> # Y.1.2[1] and Y.2.1[1] denote the transition probabilities where the
    transitions are calculated
> # at the first time point.
>
> for (i in 1:npat) {
```

```
+
+ for (j in 2:(ntimes+1) ) {
+
+ Y[i,j] <- rbinom(1,1,Y.1.2[1])*(Y[i,j-1]==0) + rbinom(1,1,1-Y.2.1[1])*(Y[i,j
    -1]==1)
+
+ }
+
+ }
>
>
> ##################
> # msm modelling #
> ##################
> library(reshape2)
>
> # Format data for msm modelling
> sim1<- melt(Y, id.vars = c("t="))
>
> # Rename variables for msm to run
> names(sim1)[1] <- "subject"
>
> # States are simulated as 0's and 1's. Transform to 1's and 2's for msm to run.
> sim1\$state<-sim1\$value+1
>
> # Denote the state variable as binary
> sim1\$state<-as.factor(sim1\$state)
>
> # Remove this variable produced from reshaping the data.
> sim1\$value <- NULL
>
> # Recode variable to show time and rename Var2.
> sim1\$Var2=gsub("t=*","",sim1\$Var2)
> names(sim1)[2] <- "time"
>
> # Bind the variable names to the simulated dataset.
> sim1\$subject <- sim1\$subject
> sim1\$time <- sim1\$time
> sim1\$state <- sim1\$state
>
> # Denote time as a continuous variable.
> sim1\$time<-as.numeric(sim1\$time)
>
> # Sort data.
> sim1<-sim1[with(sim1, order(subject, time)), ]
>
> # denote initial intensities (start at true values)
```

```
> qm <- rbind(c(0.1, 0.1),
+             c(0.3, 0.3))
>
> # Fit the msm model assuming constant intensities and not including any
    covariates.
> sim1.msm <- msm(state ~ time, subject = subject, data = sim1,
+                     qmatrix = qm, exacttimes=FALSE,
+                     method = ''BFGS'',
+                     control = list(fnscale = 4000, maxit = 10000))
>
> sim1.msm

> # Transition intensity matrix:
> qmatrix.msm(sim1.msm)
        State 1                       State 2
State 1 -0.09849 (-0.10118,-0.09587)  0.09849 ( 0.09587, 0.10118)
State 2  0.30093 ( 0.29429, 0.30772) -0.30093 (-0.30772,-0.29429)

> # Transition probability matrix for t=1:
> pmatrix.msm(sim1.msm, t=1)
          State 1    State 2
State 1 0.9188051 0.08119493
State 2 0.2480946 0.75190538

> # Transition probability matrix for t=2:
> pmatrix.msm(sim1.msm, t=2)
          State 1   State 2
State 1 0.8643468 0.1356532
State 2 0.4144943 0.5855057

> # Transition probability matrix for t=3:
> pmatrix.msm(sim1.msm, t=3)
          State 1   State 2
State 1 0.8278210 0.1721790
State 2 0.5261003 0.4738997

> # Transition probability matrix for t=4:
> pmatrix.msm(sim1.msm, t=4)
          State 1   State 2
State 1 0.8033228 0.1966772
State 2 0.6009556 0.3990444

> # Transition probability matrix for t=5:
> pmatrix.msm(sim1.msm, t=5)
          State 1   State 2
State 1 0.7868917 0.2131083
State 2 0.6511618 0.3488382
```

```
> # Transition probability matrix for t=6:
> pmatrix.msm(sim1.msm, t=6)
          State 1   State 2
State 1 0.7758711 0.2241289
State 2 0.6848356 0.3151644

> # Transition probability matrix for t=7:
> pmatrix.msm(sim1.msm, t=7)
          State 1   State 2
State 1 0.7684795 0.2315205
State 2 0.7074210 0.2925790

> # Transition probability matrix for t=8:
> pmatrix.msm(sim1.msm, t=8)
          State 1   State 2
State 1 0.7635218 0.2364782
State 2 0.7225693 0.2774307

> # Transition probability matrix for t=9:
> pmatrix.msm(sim1.msm, t=9)
          State 1   State 2
State 1 0.7601967 0.2398033
State 2 0.7327294 0.2672706

> # Transition probability matrix for t=10:
> pmatrix.msm(sim1.msm, t=10)
          State 1   State 2
State 1 0.7579665 0.2420335
State 2 0.7395439 0.2604561

> # The transition intensities and probabilities closely match those to the true
    values.
> # END
> ###############



> # B. Simulation of transition probabilities assuming transition intensities
    change over time
>
> # Purpose: Simulate data from probability transitions after choosing transition
     intensities that vary over time and trace back to the known transition
    intensities and probabilities.
>
> library(msm)
> # Begin: simulate observed data under these
```

```
> # time-varying transmission probabilities
>
> # Choose 10 time points
> ntimes<-10
>
> Y.1.2<-rep(0,ntimes)
> Y.2.1<-rep(0,ntimes)
>
> Q<-vector('list',ntimes)
>
> # create time varying intensities
> for( i in 1:ntimes)  {
+
+ Q[[i]]<-matrix( c(-(0.2+(0.02*i)),(0.2+(0.02*i)),(0.5+(0.07*i)),(-(0.5+(0.07*i)
    ))),ncol=2,byrow=T)
+
+ }

> # Check true transition intensities for all 10 time points
> Q[1]
[[1]]
      [,1]  [,2]
[1,] -0.22  0.22
[2,]  0.57 -0.57

> Q[2]
[[2]]
      [,1]  [,2]
[1,] -0.24  0.24
[2,]  0.64 -0.64

> Q[3]
[[3]]
      [,1]  [,2]
[1,] -0.26  0.26
[2,]  0.71 -0.71

> Q[4]
[[4]]
      [,1]  [,2]
[1,] -0.28  0.28
[2,]  0.78 -0.78

> Q[5]
      [,1]  [,2]
[1,] -0.30  0.30
[2,]  0.85 -0.85
```

```
> Q[6]
      [,1]  [,2]
[1,] -0.32  0.32
[2,]  0.92 -0.92

> Q[7]
      [,1]  [,2]
[1,] -0.34  0.34
[2,]  0.99 -0.99

> Q[8]
      [,1]  [,2]
[1,] -0.36  0.36
[2,]  1.06 -1.06

> Q[9]
      [,1]  [,2]
[1,] -0.38  0.38
[2,]  1.13 -1.13

> Q[10]
     [,1] [,2]
[1,] -0.4  0.4
[2,]  1.2 -1.2


> #Step 1: extract transmission probabilities from
> # intensity matrices using matrix exponential
>
> for( i in 1:ntimes)  {
+
+ Y.1.2[i]<- MatrixExp(Q[[i]])[1,2] # from t=i-1 to t=i
+ Y.2.1[i]<- MatrixExp(Q[[i]])[2,1] # from t=i-1 to t=1
+
+ }
>
> # True probability transition values
> Y.1.2
 [1] 0.1520939 0.1596047 0.1664314 0.1726343 0.1782687 0.1833847 0.1880284
     0.1922415
 [9] 0.1960624 0.1995259
> Y.2.1
 [1] 0.3940614 0.4256124 0.4544856 0.4809099 0.5050946 0.5272311 0.5474944
     0.5660445
 [9] 0.5830276 0.5985776
>
```

```
> # Step 2: Simulate data for 10,000 patients based on these probability
    transitions:
> > npat<-10000
>
> Y<-matrix(NA,nrow=npat,ncol=ntimes+1)
>
> labs<-rep("",ntimes+1)
> for(i in 1: (ntimes+1) ) { labs[i]<-paste("t=",i-1,sep="")  }
>
>
> dimnames(Y)[[2]]<-labs
> dimnames(Y)[[1]]<-1:dim(Y)[1]
>
> set.seed(1875263)
>
> Y[,1]<-rep(c(0,1),npat/2)
>
>
> for (i in 1:npat) {
+
+ for (j in 2:(ntimes+1) ) {
+
+ Y[i,j] <- rbinom(1,1,Y.1.2[j-1])*(Y[i,j-1]==0) + rbinom(1,1,1-Y.2.1[j-1])*(Y[i,
    j-1]==1)
+
+ }
+
+ }
>
>
> ##################
> # msm modelling #
> ##################
>
> # load packages to use for msm
>
> library(msm)
> library(reshape2)
>
> # format data for msm modelling
> sim1<- melt(Y, id.vars = c("t="))
>
> names(sim1)[1] <- "subject"
>
> sim1\$state<-sim1\$value+1
> sim1\$state<-as.factor(sim1\$state)
> sim1\$value <- NULL
```

```
>
> sim1\$Var2=gsub("t=*","",sim1\$Var2)
>
> names(sim1)[2] <- "time"
>
> sim1\$subject <- sim1\$subject
> sim1\$time <- sim1\$time
> sim1\$state <- sim1\$state
>
> sim1\$time<-as.numeric(sim1\$time)
>
> # Sort data
> sim1<-sim1[with(sim1, order(subject, time)), ]
>
> # Denote some intial intensities
> qm <- rbind(c(0.2,   0.2),
+             c(0.5,  0.5))
>
> # Fit msm model: including time as a covariate
> sim1.time.msm <- msm(state ~ time, subject = subject, data = sim1,
+                      qmatrix = qm, exacttimes=FALSE,
+                      covariates = ~time,
+                      method = "BFGS",
+                      control = list(fnscale = 4000, maxit = 10000))
>
>
> # NB: outputs the average intensity #
> sim1.time.msm

Call:
msm(formula = state ~ time, subject = subject, data = sim1, qmatrix = qm,
    covariates = ~time, exacttimes = FALSE, method = "BFGS",     control = list(
    fnscale = 4000, maxit = 10000))

Maximum likelihood estimates
Baselines are with covariates set to their means

Transition intensities with hazard ratios for each covariate
                  Baseline                   time
State 1 - State 1 -0.3015 (-0.3080,-0.2950)
State 1 - State 2  0.3015 ( 0.2950, 0.3080) 1.069 (1.06,1.077)
State 2 - State 1  0.8580 ( 0.8414, 0.8749) 1.087 (1.08,1.094)
State 2 - State 2 -0.8580 (-0.8749,-0.8414)

-2 * log-likelihood:  106776.2
[Note, to obtain old print format, use "printold.msm"]
>
```

```
> # recover individual intensities using qmatrix:
> qmatrix.msm(sim1.time.msm, covariate=list(time=0))
        State 1                    State 2
State 1 -0.2236 (-0.2329,-0.2147)  0.2236 ( 0.2147, 0.2329)
State 2  0.5899 ( 0.5714, 0.6089) -0.5899 (-0.6089,-0.5714)
>
> qmatrix.msm(sim1.time.msm, covariate=list(time=1))
        State 1                    State 2
State 1 -0.2390 (-0.2473,-0.2309)  0.2390 ( 0.2309, 0.2473)
State 2  0.6411 ( 0.6241, 0.6585) -0.6411 (-0.6585,-0.6241)
>
> qmatrix.msm(sim1.time.msm, covariate=list(time=2))
        State 1                    State 2
State 1 -0.2554 (-0.2628,-0.2481)  0.2554 ( 0.2481, 0.2628)
State 2  0.6968 ( 0.6811, 0.7128) -0.6968 (-0.7128,-0.6811)
>
> qmatrix.msm(sim1.time.msm, covariate=list(time=3))
        State 1                    State 2
State 1 -0.2729 (-0.2796,-0.2663)  0.2729 ( 0.2663, 0.2796)
State 2  0.7573 ( 0.7423, 0.7725) -0.7573 (-0.7725,-0.7423)
>
> qmatrix.msm(sim1.time.msm, covariate=list(time=4))
        State 1                    State 2
State 1 -0.2916 (-0.2981,-0.2853)  0.2916 ( 0.2853, 0.2981)
State 2  0.8230 ( 0.8074, 0.8389) -0.8230 (-0.8389,-0.8074)

> qmatrix.msm(sim1.time.msm, covariate=list(time=5))
        State 1                    State 2
State 1 -0.3116 (-0.3185,-0.3049)  0.3116 ( 0.3049, 0.3185)
State 2  0.8945 ( 0.8764, 0.9130) -0.8945 (-0.9130,-0.8764)

> qmatrix.msm(sim1.time.msm, covariate=list(time=6))
        State 1                    State 2
State 1 -0.3330 (-0.3413,-0.3249)  0.3330 ( 0.3249, 0.3413)
State 2  0.9722 ( 0.9495, 0.9954) -0.9722 (-0.9954,-0.9495)

> qmatrix.msm(sim1.time.msm, covariate=list(time=7))
        State 1                    State 2
State 1 -0.3559 (-0.3664,-0.3457)  0.3559 ( 0.3457, 0.3664)
State 2  1.0566 ( 1.0275, 1.0865) -1.0566 (-1.0865,-1.0275)

> qmatrix.msm(sim1.time.msm, covariate=list(time=8))
        State 1                    State 2
State 1 -0.3803 (-0.3938,-0.3673)  0.3803 ( 0.3673, 0.3938)
State 2  1.1484 ( 1.1111, 1.1869) -1.1484 (-1.1869,-1.1111)

> qmatrix.msm(sim1.time.msm, covariate=list(time=9))
```

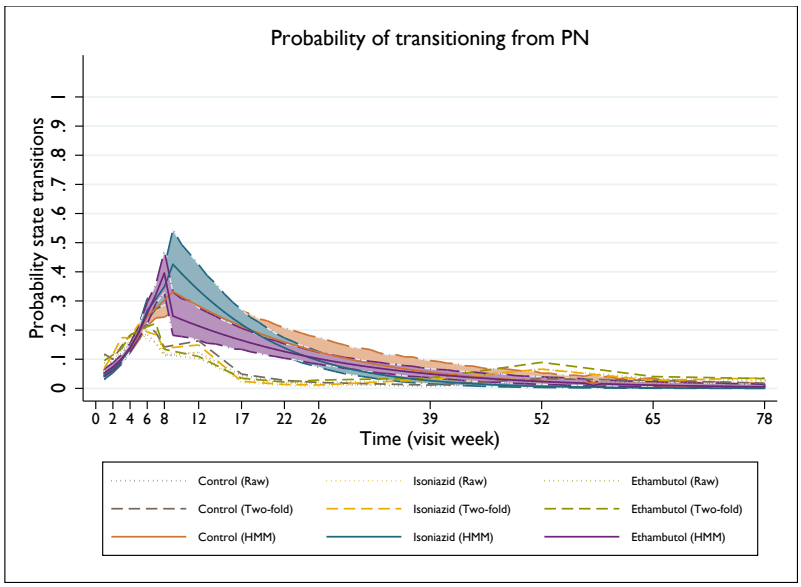
```
        State 1                   State 2
State 1 -0.4064 (-0.4235,-0.3901)  0.4064 ( 0.3901, 0.4235)
State 2  1.2481 ( 1.2009, 1.2971) -1.2481 (-1.2971,-1.2009)

> qmatrix.msm(sim1.time.msm, covariate=list(time=10))
        State 1                   State 2
State 1 -0.4343 (-0.4556,-0.4140)  0.4343 ( 0.4140, 0.4556)
State 2  1.3565 ( 1.2977, 1.4179) -1.3565 (-1.4179,-1.2977)
>
> # Close enough to the known transition intensities above
>
> # Calculate transition probabilities, assuming time is
> # piecewise constant
> time_p<-1:10
> times<-c(1:10)
>
> transcov.1.2<-rep(0,10)
> transcov.2.1<-rep(0,10)
>
> for (i in 1:10) {
+   sim1.cov<-pmatrix.piecewise.msm(sim1.time.msm, i-1, i, times, ci=c("none"),
+      covariates= (list(list (time = 0),
+                        list (time = 1),
+                        list (time = 2),
+                        list (time = 3),
+                        list (time = 4),
+                        list (time = 5),
+                        list (time = 6),
+                        list (time = 7),
+                        list (time = 8),
+                        list (time = 9),
+                        list (time = 10)) ))
+
+   transcov.1.2[i]<-sim1.cov[1,2]
+   transcov.2.1[i]<-sim1.cov[2,1]
+ }
>
> # Probability transitions from state 1 to state 2 over 10 time points:
> transcov.1.2
 [1] 0.1530220 0.1589080 0.1646963 0.1703432 0.1758034 0.1810302 0.1859765
     0.1905959
 [9] 0.1948441 0.1986801
>
> # Probability transitions from state 2 to state 1 over 10 time points:
> transcov.2.1
 [1] 0.4036580 0.4263246 0.4493793 0.4727036 0.4961651 0.5196186 0.5429082
     0.5658701
```

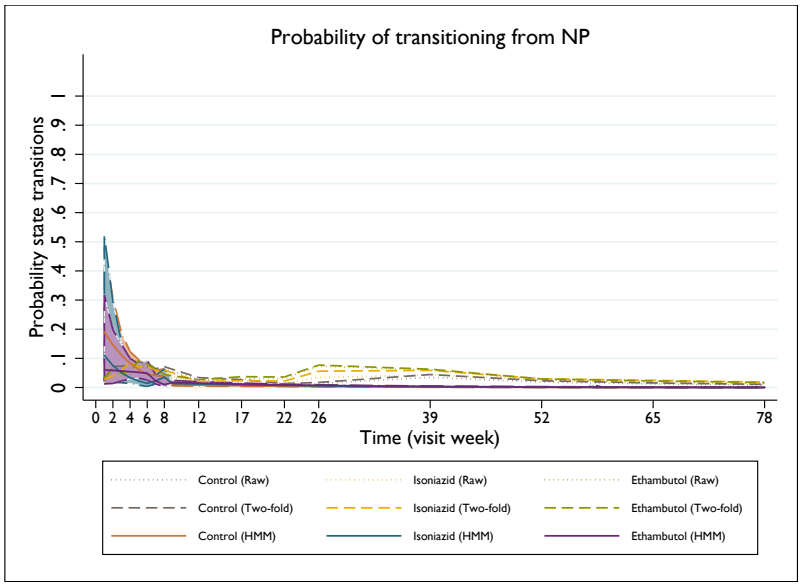
```
 [9] 0.5883358 0.6101365
>
>
> # END
> ###############
```

H Probability transitions for REMoxTB

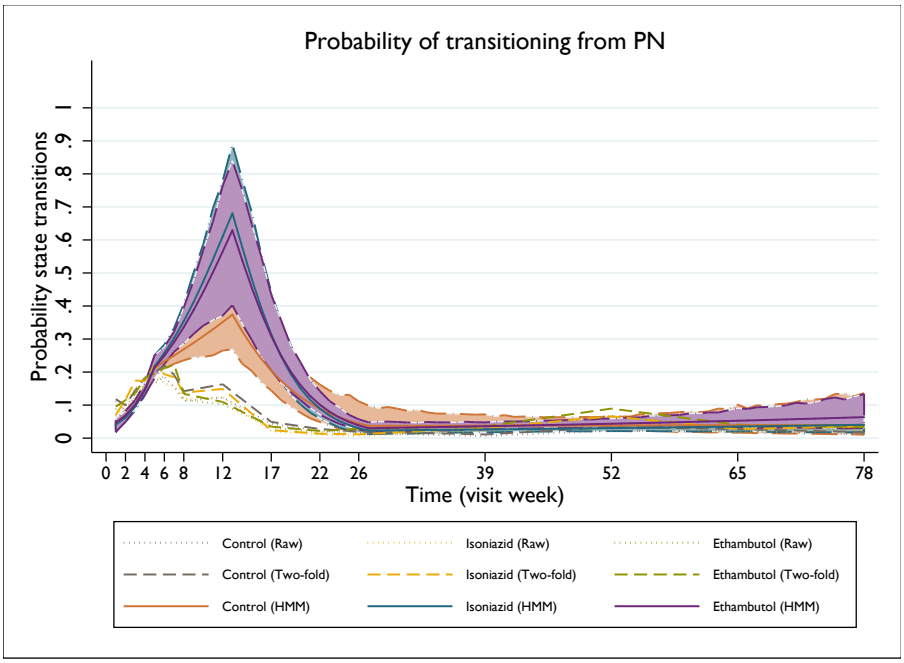
Figure H1: Positive to negative probability transitions with linear splines at 5, 7 and 8 weeks for REMoxTB.



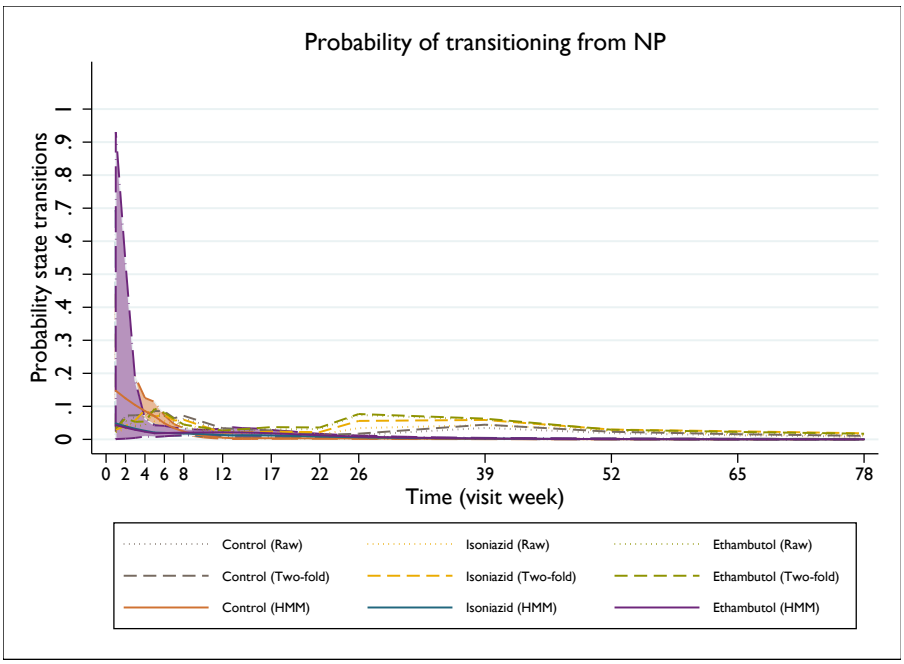
Negative to positive probability transitions with linear splines at 5, 7 and 8 weeks for REMoxTB.



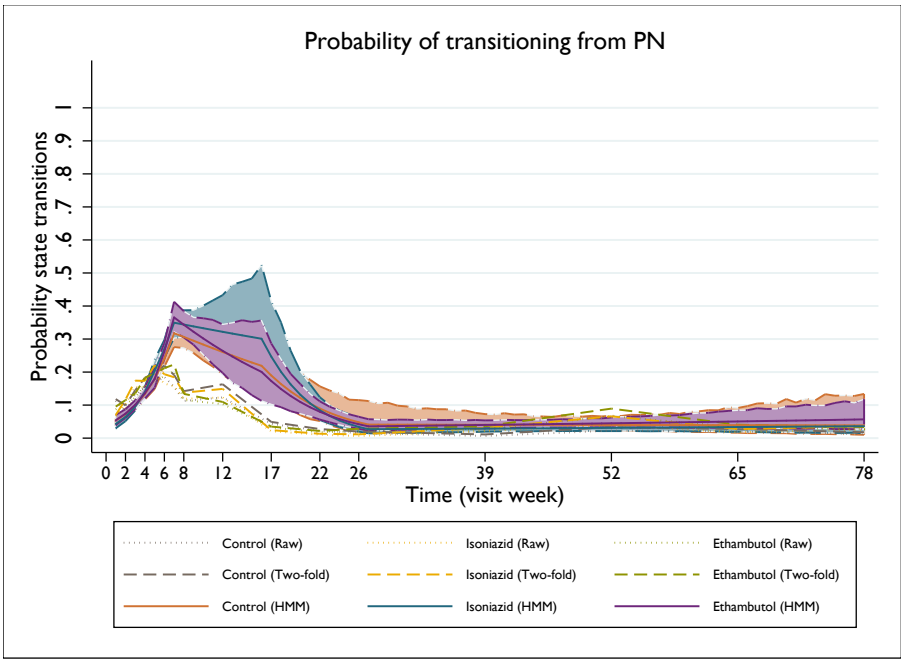
Positive to negative probability transitions with linear splines at 4, 12 and 26 weeks for REMoxTB.



Negative to positive probability transitions with linear splines at 4, 12 and 26 weeks for REMoxTB.



Positive to negative probability transitions with linear splines at 4, 6, 15 and 26 weeks for REMoxTB.



Negative to positive probability transitions with linear splines at 4, 6, 15 and 26 weeks for REMoxTB.

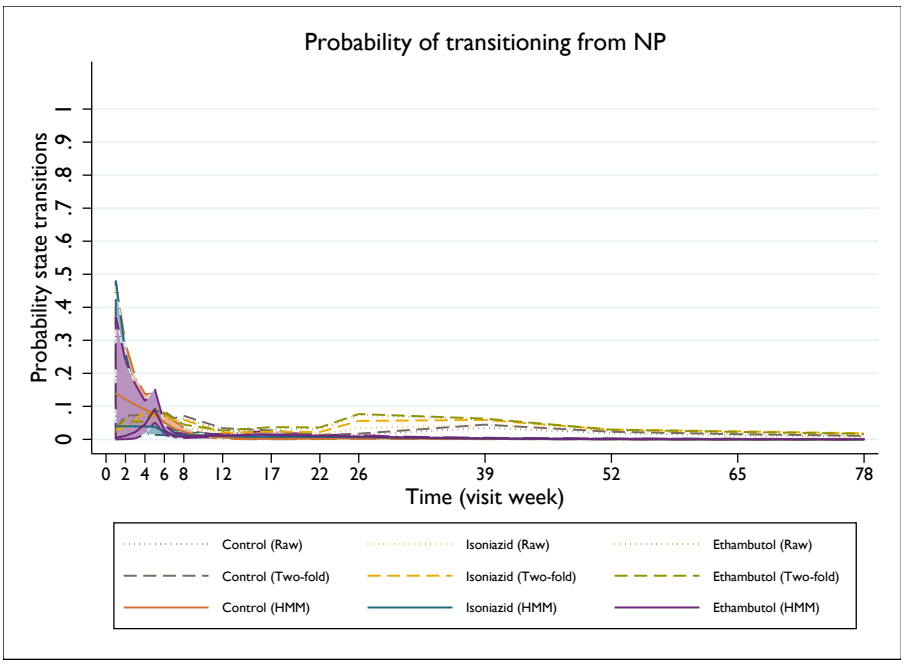


Table H1: Linear splines HMM with at knot at 2, 4, 8 and 26 weeks for REMoxTB.

Model	Transition states[1] $P(S_t = j \mid S_{t-1} = i)$ (95% CI) $P(Neg \mid Pos)$	$P(Pos \mid Neg)$	Misclassifications $P(O \mid S)$ (95% CI) $P(O_t = Pos \mid S_t = Pos)$	$P(O_t = Neg \mid S_t = Pos)$	$P(O_t = Pos \mid S_t = Neg)$	$P(O_t = Neg \mid S_t = Neg)$
Linear splines (see 5.6.2)						
Baseline hazard[2]	0.12247 (0.10981, 0.13658)	0.01507 (0.01102, 0.02062)	0.930829	0.069171	0.007196	0.992804
Isoniazid	0.5881 (0.339597, 1.018)	0.6580 (0.005947,72.798)				
Ethambutol	0.8673 (0.51732, 1.454)	1.9180 (0.01547,237.878)				
Week	1.1784 (0.9139,1.519)	0.7602 (0.1060,5.453)				
$Week_2$	1.318 (0.88804, 1.955)	1.203 (0.09766,14.825)				
$Week_4$	0.7695 (0.6072,0.9752)	0.7169 (0.3035,1.6934)				
$Week_8$	0.7374 (0.6616,0.8219)	1.3599 (1.0394,1.7792)				
$Week_{26}$	1.126 (1.0414,1.216)	1.072 (0.9679,1.188)				
Isoniazid*Week	1.446 (1.00082, 2.088)	1.082 (0.06336,18.479)				
Ethambutol*Week	1.1334 (0.79357,1.619)	0.5013 (0.02618,9.599)				
Isoniazid*$Week_2$	0.6180 (0.35572, 1.074)	0.7803 (0.02188,27.833)				
Ethambutol*$Week_2$	0.8426 (0.48779, 1.455)	2.5118 (0.05828,108.261)				
Isoniazid*$Week_4$	1.216 (0.8774,1.686)	1.285 (0.3579,4.611)				
Ethambutol*$Week_4$	1.1045 (0.7955,1.533)	0.8493 (0.2294,3.144)				
Isoniazid*$Week_8$	0.8779 (0.7555,1.020)	0.9946 (0.6694,1.478)				
Ethambutol*$Week_8$	0.9198 (0.7926,1.067)	1.0138 (0.7013,1.465)				
Isoniazid*$Week_{26}$	1.0628 (0.9662,1.169)	0.9036 (0.7967,1.025)				
Ethambutol*$Week_{26}$	1.0508 (0.9612,1.149)	0.8853 (0.7819,1.002)				
-2 log-likelihood: 15416.98						

Figure H2: Positive to negative probability transitions with linear splines at 2, 4, 8 and 26 weeks for REMoxTB.

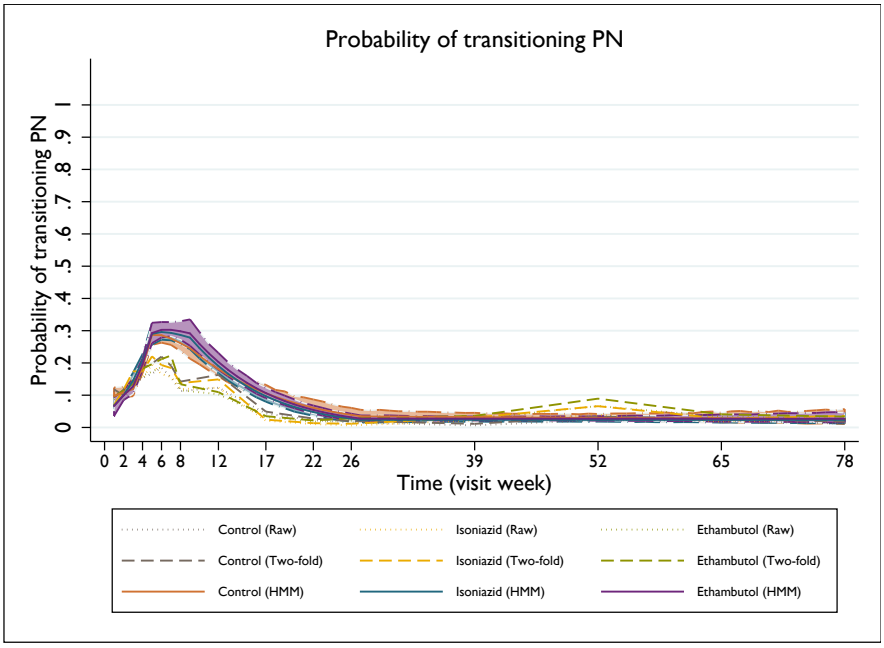
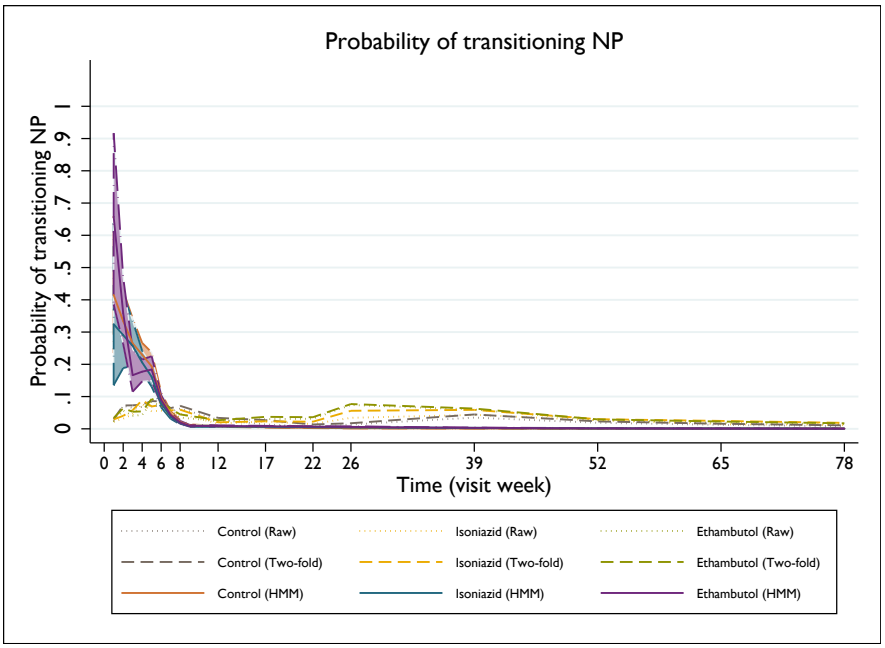


Figure H3: Negative to positive probability transitions with linear splines at 2, 4, 8 and 26 weeks for REMoxTB.



I Analyses Viterbi forwards/backwards for REMoxTB

Figure I1: Analysis of REMoxTB using the forwards/backwards algorithm (unadjusted analysis).

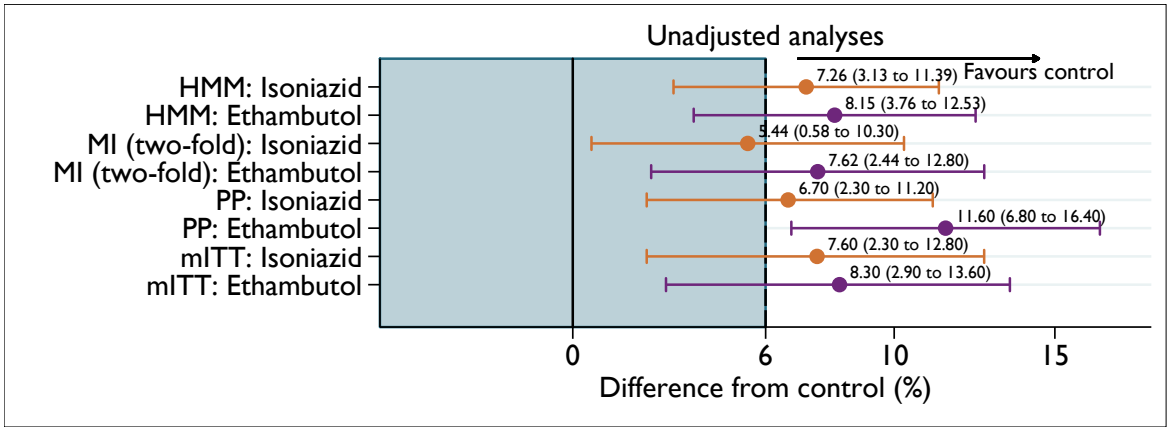


Figure I2: Analysis of REMoxTB using the Viterbi algorithm (unadjusted analysis).

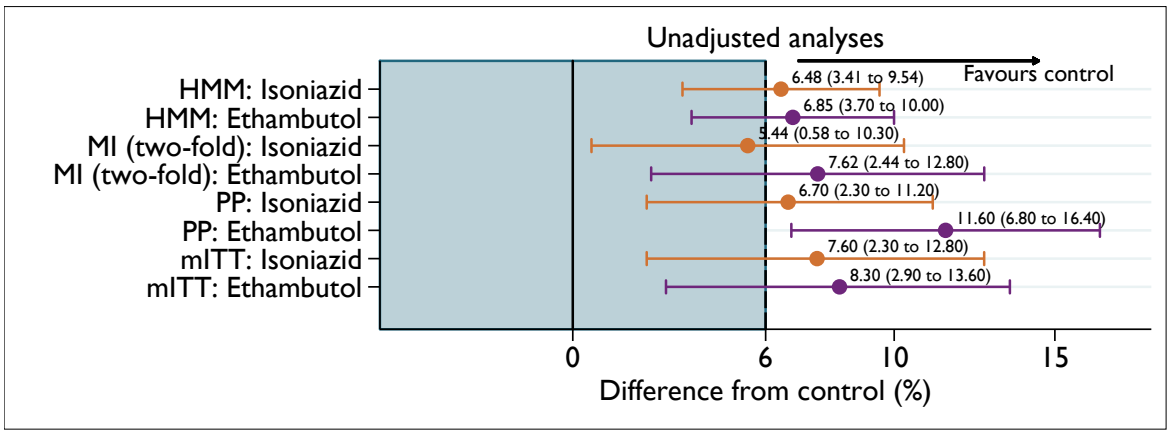


Table I1: Proportion of patients meeting the primary outcome for REMoxTB following imputation using the forwards/backwards algorithm.

		PP (n=237)			**mITT** (n=111)		
		Control	Isoniazid	Ethambutol	Control	Isoniazid	Ethambutol
Excluded from primary analysis		35 (6%)	41 (7%)	35 (6%)	80 (14%)	95 (16%)	62 (11%)
Primary outcome	**Imputed set**						
Failure	1	11 (14%)	18 (19%)	22 (35%)	9 (26%)	13 (32%)	13 (37%)
Success	1	69 (86%)	77 (81%)	40 (65%)	26 (74%)	28 (68%)	22 (63%)
Failure	2	10 (13%)	17 (18%)	22 (35%)	8 (23%)	10 (24%)	13 (37%)
Success	2	70 (88%)	78 (82%)	40 (65%)	27 (77%)	31 (76%)	22 (63%)
Failure	3	14 (18%)	15 (16%)	15 (24%)	9 (26%)	10 (24%)	12 (34%)
Success	3	66 (83%)	80 (84%)	47 (76%)	26 (74%)	31 (76%)	23 (66%)
Failure	4	12 (15%)	14 (15%)	17 (27%)	10 (29%)	10 (24%)	14 (40%)
Success	4	68 (85%)	81 (85%)	45 (73%)	25 (71%)	31 (76%)	21 (60%)
Failure	5	14 (18%)	20 (21%)	14 (23%)	9 (26%)	12 (29%)	11 (31%)
Success	5	66 (83%)	75 (79%)	48 (77%)	26 (74%)	29 (71%)	24 (69%)
Failure	6	14 (18%)	15 (16%)	15 (24%)	8 (23%)	10 (24%)	11 (31%)
Success	6	66 (83%)	80 (84%)	47 (76%)	27 (77%)	31 (76%)	24 (69%)
Failure	7	15 (19%)	19 (20%)	15 (24%)	11 (31%)	12 (29%)	12 (34%)
Success	7	65 (81%)	76 (80%)	47 (76%)	24 (69%)	29 (71%)	23 (66%)
Failure	8	17 (21%)	20 (21%)	18 (29%)	10 (29%)	13 (32%)	13 (37%)
Success	8	63 (79%)	75 (79%)	44 (71%)	25 (71%)	28 (68%)	22 (63%)
Failure	9	18 (23%)	19 (20%)	19 (31%)	12 (34%)	10 (24%)	14 (40%)
Success	9	62 (78%)	76 (80%)	43 (69%)	23 (66%)	31 (76%)	21 (60%)
Failure	10	11 (14%)	16 (17%)	18 (29%)	10 (29%)	10 (24%)	15 (43%)
Success	10	69 (86%)	79 (83%)	44 (71%)	25 (71%)	31 (76%)	20 (57%)
Failure	11	10 (13%)	19 (20%)	19 (31%)	8 (23%)	11 (27%)	12 (34%)
Success	11	70 (88%)	76 (80%)	43 (69%)	27 (77%)	30 (73%)	23 (66%)
Failure	12	13 (16%)	22 (23%)	20 (32%)	9 (26%)	13 (32%)	14 (40%)
Success	12	67 (84%)	73 (77%)	42 (68%)	26 (74%)	28 (68%)	21 (60%)
Failure	13	13 (16%)	19 (20%)	17 (27%)	9 (26%)	12 (29%)	12 (34%)
Success	13	67 (84%)	76 (80%)	45 (73%)	26 (74%)	29 (71%)	23 (66%)
Failure	14	11 (14%)	19 (20%)	18 (29%)	9 (26%)	11 (27%)	13 (37%)
Success	14	69 (86%)	76 (80%)	44 (71%)	26 (74%)	30 (73%)	22 (63%)
Failure	15	12 (15%)	18 (19%)	19 (31%)	9 (26%)	10 (24%)	12 (34%)
Success	15	68 (85%)	77 (81%)	43 (69%)	26 (74%)	31 (76%)	23 (66%)
Failure	16	10 (13%)	18 (19%)	15 (24%)	8 (23%)	11 (27%)	12 (34%)
Success	16	70 (88%)	77 (81%)	47 (76%)	27 (77%)	30 (73%)	23 (66%)
Failure	17	10 (13%)	19 (20%)	15 (24%)	8 (23%)	11 (27%)	11 (31%)

Success	17	70 (88%)	76 (80%)	47 (76%)	27 (77%)	30 (73%)	24 (69%)
Failure	18	10 (13%)	15 (16%)	17 (27%)	8 (23%)	11 (27%)	13 (37%)
Success	18	70 (88%)	80 (84%)	45 (73%)	27 (77%)	30 (73%)	22 (63%)
Failure	19	13 (16%)	18 (19%)	15 (24%)	9 (26%)	11 (27%)	13 (37%)
Success	19	67 (84%)	77 (81%)	47 (76%)	26 (74%)	30 (73%)	22 (63%)
Failure	20	10 (13%)	14 (15%)	14 (23%)	8 (23%)	10 (24%)	11 (31%)
Success	20	70 (88%)	81 (85%)	48 (77%)	27 (77%)	31 (76%)	24 (69%)

J Probability transitions for RIFAQUIN

Figure J1: Positive to negative probability transitions with piecewise constant at 2, 4 and 10 months for RIFAQUIN.

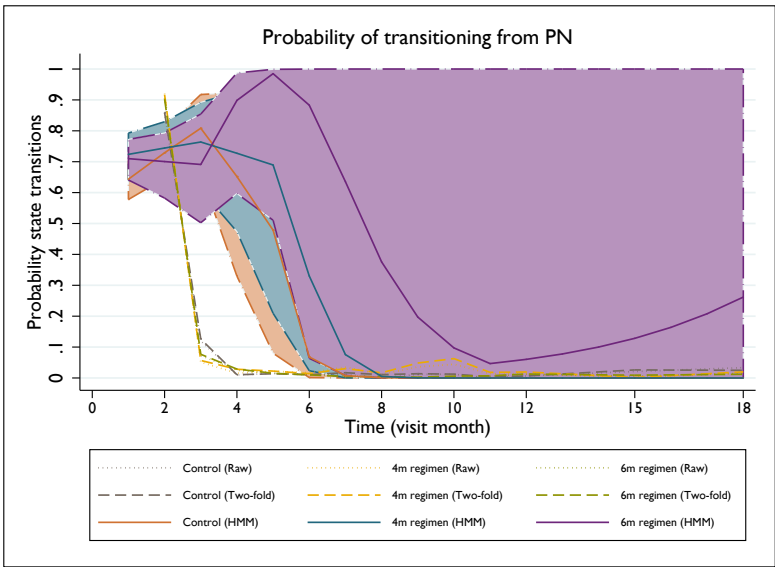
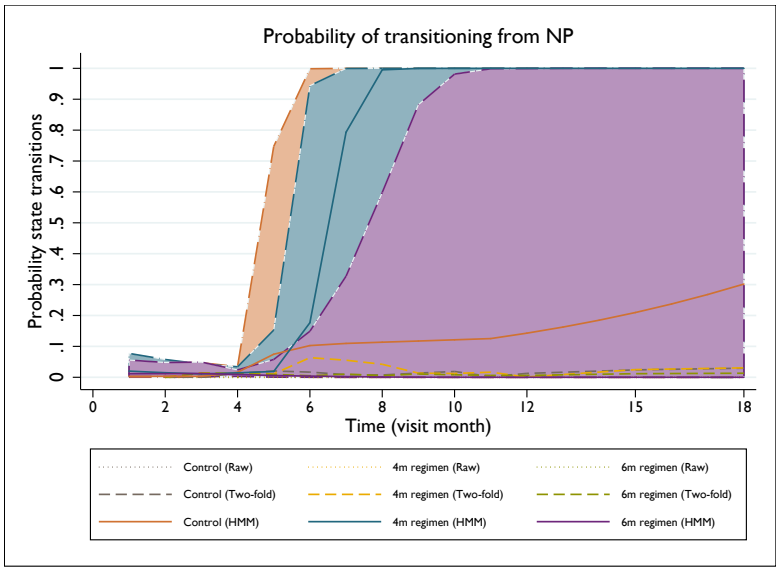


Figure J2: Negative to positive probability transitions with piecewise constant at 2, 4 and 10 months for RIFAQUIN.



K Analyses Viterbi forwards/backwards for RIFAQUIN

Figure K1: Analysis of RIFAQUIN using the forwards/backwards algorithm (unadjusted analysis).

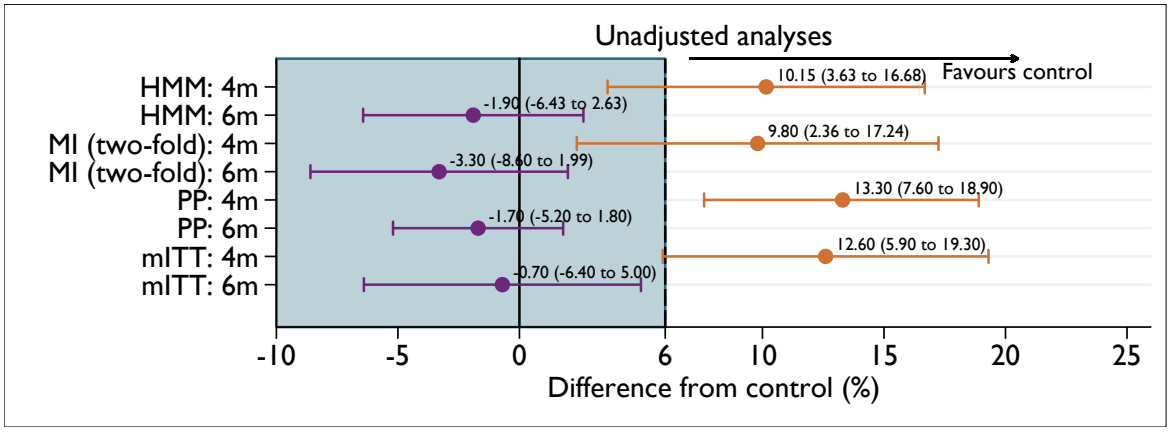
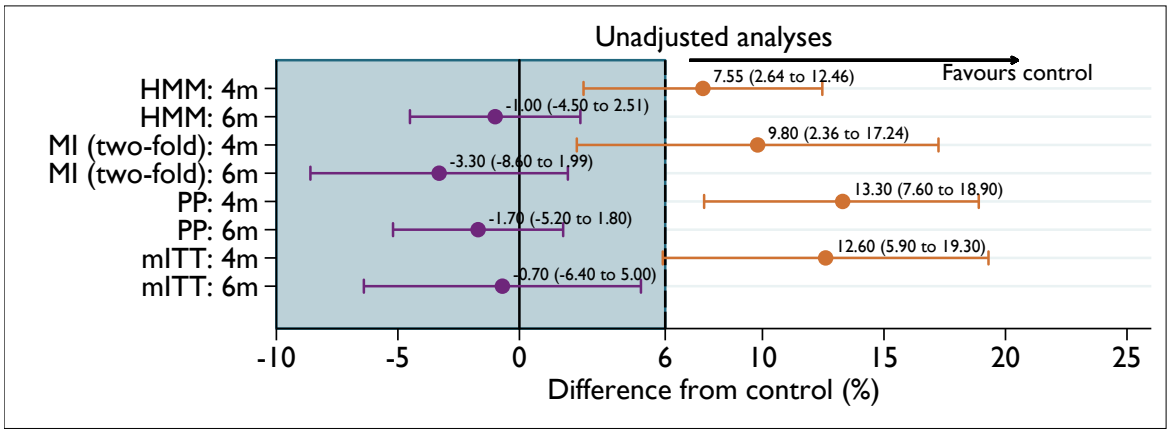


Figure K2: Analysis of RIFAQUIN using the Viterbi algorithm (unadjusted analysis).



L Technical details to construct the joint multivariate normal distribution for each reference-based option.

To construct the joint MVN distribution for a patient's pre-deviation data and post-deviation outcome data, we begin by describing the jump to reference option which leads to how we approach the other options. We use the methods describe by Carpenter and Kenward[79] (pp 251-252) to describe these details. For step 2 of the algorithm for reference-based sensitivity analyses (see step 2 in §6.1.1), a mean vector and variance-covariance matrix is drawn from the posterior distribution for each randomised arm. Let the current draw of the control (reference) arm means be denoted by $\mu_\xi = (\mu_{\xi,0}, \mu_{\xi,0}, ..., \mu_{\xi,d_k})$ and variance-covariance be denoted by Σ_ξ for deviation time d_k from the posterior. Let the current draw of the treatment group means be denoted by $\mu_\tau = \mu_{\tau,0}, \mu_{\tau,1}, ..., \mu_{\tau,d_k}$ and the variance-covariance matrix be denoted by Σ_τ for deviation time d_k from the posterior.

Under jump to reference, the joint distribution for the observed (pre-deviation) and missing data (post-deviation) outcomes is formed as MVN with mean and variance-covariance matrix from a patient's randomised treatment arm for pre-deviation measurements. Post-deviation, we assume the mean and variance-covariance matrix matches the observed mean for those who were randomised to the control arm. The variance-covariance matrix for the treatment arm and the control arm conditions on components of the post-deviation data given the pre-deviation data. For patients randomised to the control arm who deviate, missing data is imputed under MAR as for standard imputation (§3.5.1).

The new variance-covariance matrix from the control arm (ξ) and treatment arm (τ), partitioned at time d_k, for pre-deviation data (1) and post-deviation data (2) can be formed as[79]:

$$\Sigma_\xi = \begin{bmatrix} \xi_{11} & \xi_{12} \\ \xi_{21} & \xi_{22} \end{bmatrix}, \Sigma_\tau = \begin{bmatrix} \tau_{11} & \tau_{12} \\ \tau_{21} & \tau_{22} \end{bmatrix}.$$

From these matrices we form the new variance-covariance matrix as[79]:

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

to match the variance-covariance matrix from the treatment arm for the pre-deviation data and the control arm for the conditional components for the post-deviation data given the pre-deviation data. Then, as shown by Carpenter, Roger and Kenward[107]:

$$\Sigma_{11} = \tau_{11},$$
$$\Sigma_{21} = \xi_{21}\xi_{11}^{-1}\tau_{11},$$
$$\Sigma_{22} = \xi_{22} - \xi_{21}\xi_{11}^{-1}(\xi_{11} - \tau_1 1)\xi_{11}^{-1}\xi_{12},$$

For copy increments in reference, the mean for a patient that deviates from the treatment arm and follows the control arm becomes:
$\mu_k = [\mu_{\tau,0}, \mu_{\tau,1}, ..., \mu_{\tau,d_{k-1}}, \mu_{\tau,d_k} + (\mu_{\xi,d_{k+1}} - \mu_{\xi,d_k}), \mu_{\tau,d_i} + (\mu_{\xi,d_{k+2}} - \mu_{\xi,d_k}), ...]^T$.

For copy reference, the mean and variance covariance matrix comes from the control arm, irrespective of deviation time[79].

Under MAR, post-deviation data for deviating patients is assumed to behave like that of their original randomisation[109].

For last mean carried forward, the mean for a patient that deviates from the control arm and follows the mean observations for patients who were randomised to the treatment arm becomes:
$\mu_k = [\mu_{\tau,0}, \mu_{\tau,1}, ..., \mu_{\tau,d_{k-1}}, \mu_{\tau,d_k}, \mu_{\tau,d_i}, ...]^T$, where $\Sigma = \Sigma_\tau$.

M Unadjusted analyses using reference-based sensitivity analyses for REMoxTB.

Figure M1: Jump to reference sensitivity analyses for the REMoxTB study (unadjusted analysis).

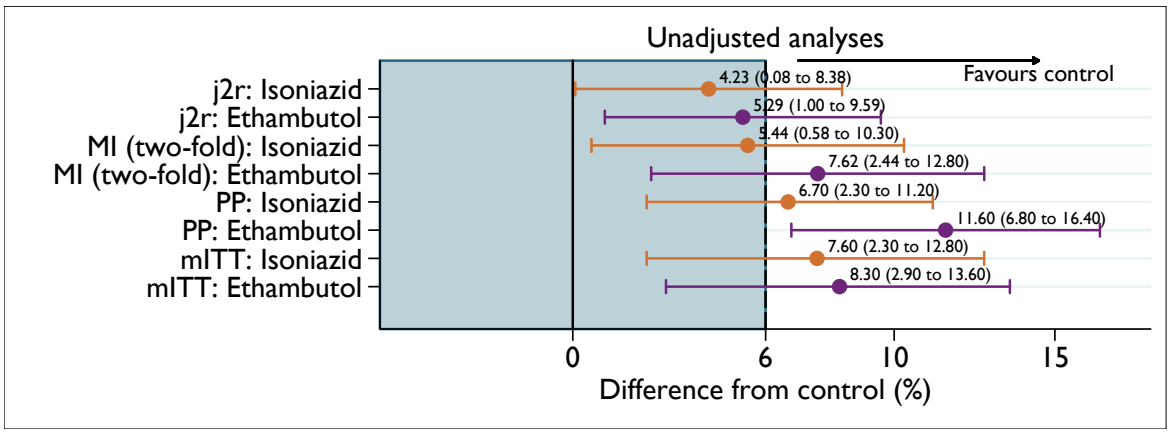


Figure M2: Copy increments in reference sensitivity analysis for the REMoxTB study (unadjusted analyses).

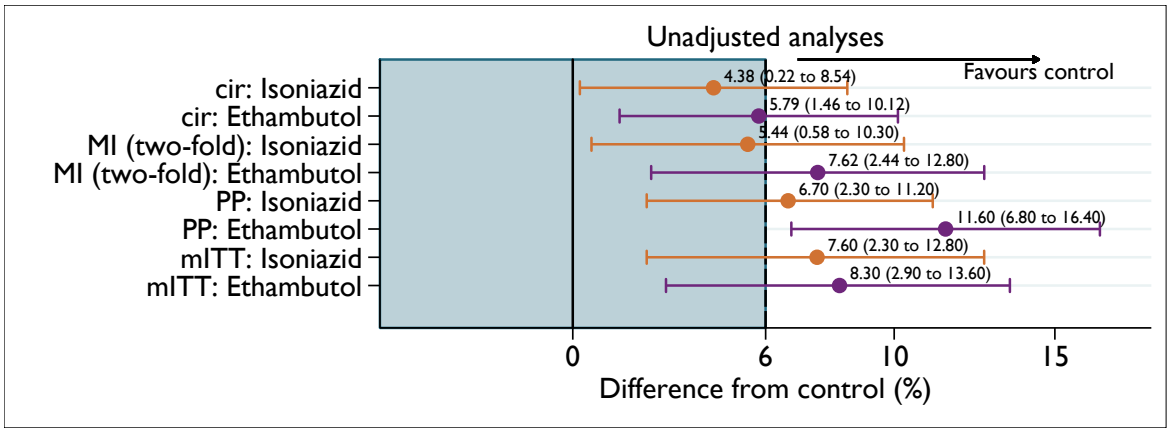


Figure M3: Copy reference sensitivity analysis for the REMoxTB study (unadjusted analyses).

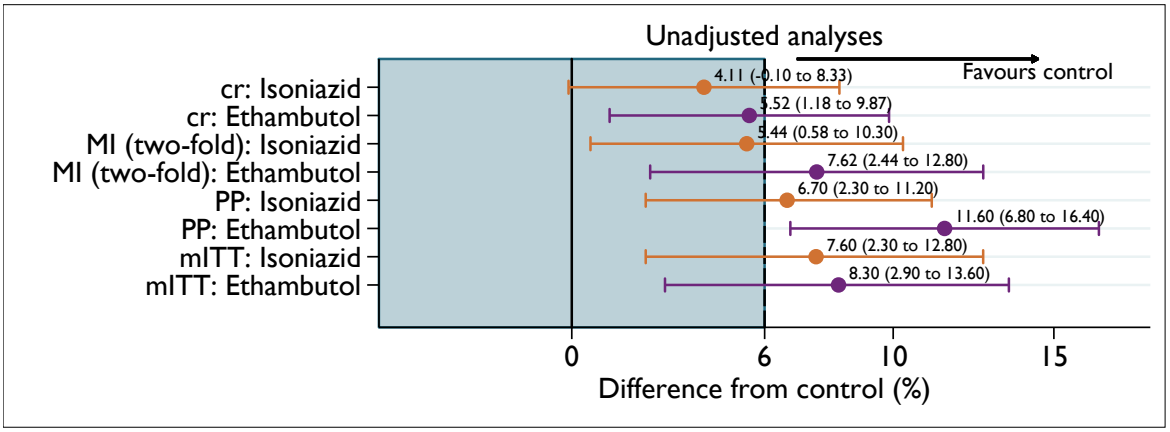


Figure M4: Last mean carried forward sensitivity analysis for the REMoxTB study (unadjusted analyses).

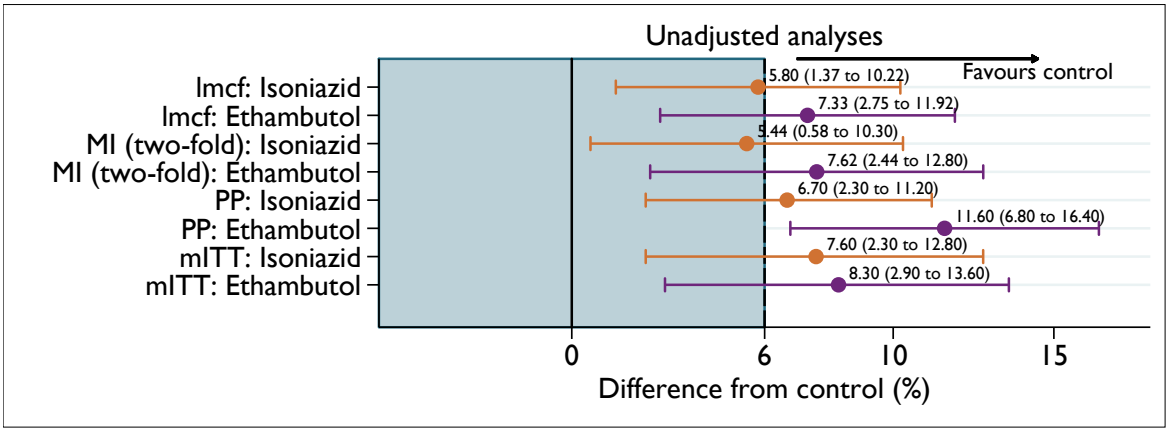
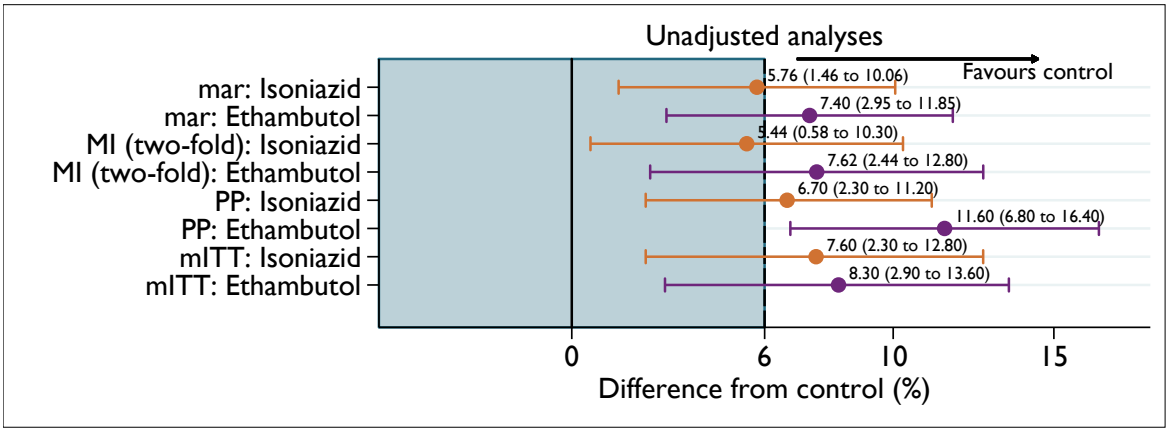


Figure M5: Missing at random sensitivity analysis for the REMoxTB study (unadjusted analyses).



N Unadjusted analyses using reference-based sensitivity analyses for RIFAQUIN.

Figure N1: Jump to reference sensitivity analyses for the RIFAQUIN study (unadjusted analysis).

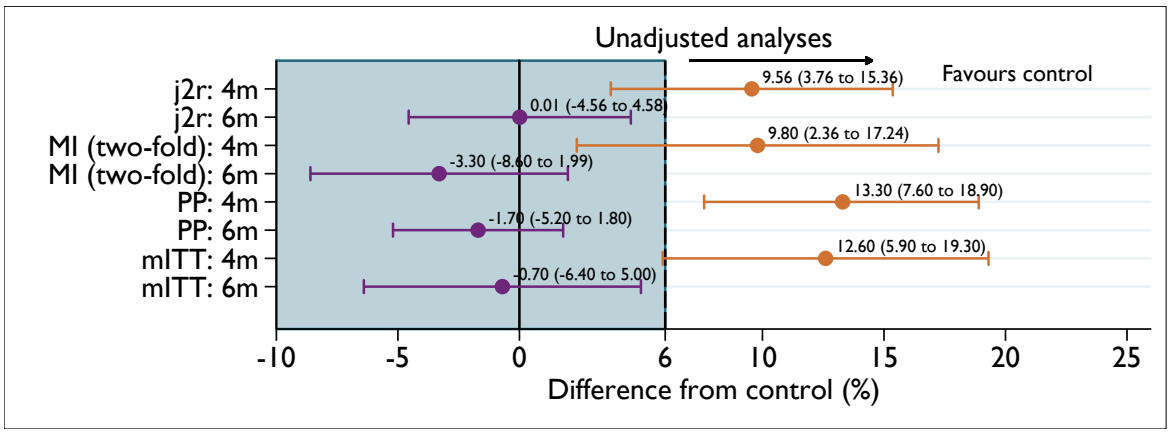


Figure N2: Copy increments in reference sensitivity analyses for the RIFAQUIN study (unadjusted analysis).

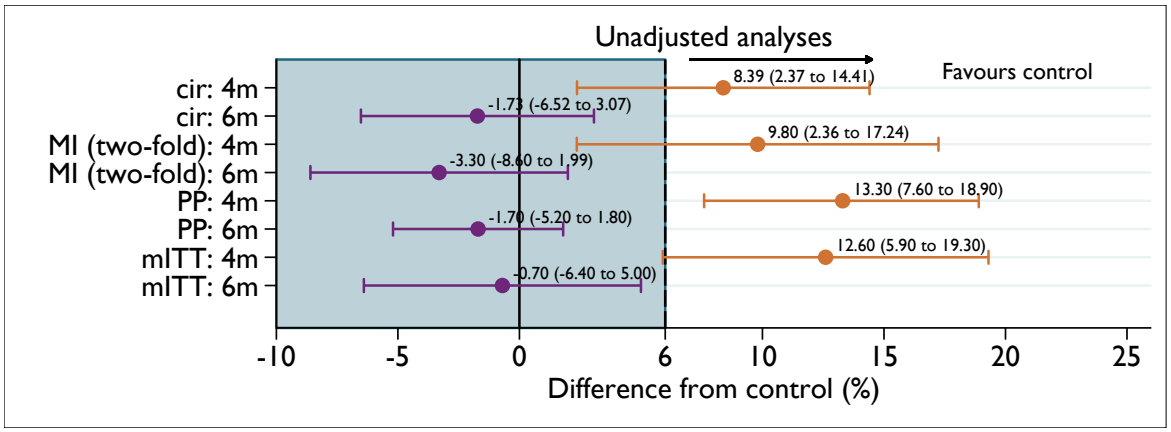


Figure N3: Copy reference sensitivity analyses for the RIFAQUIN study (unadjusted analysis).

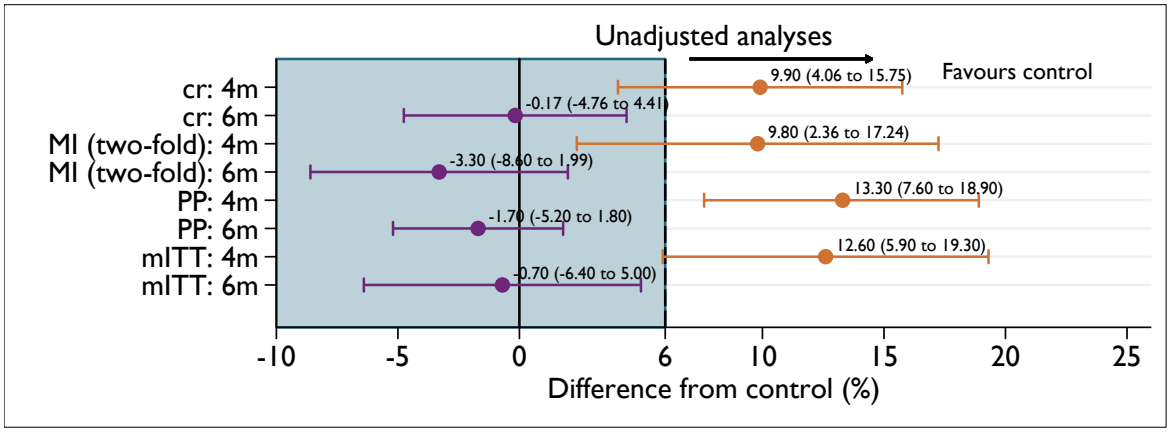


Figure N4: Last mean carried forward sensitivity analyses for the RIFAQUIN study (unadjusted analysis).

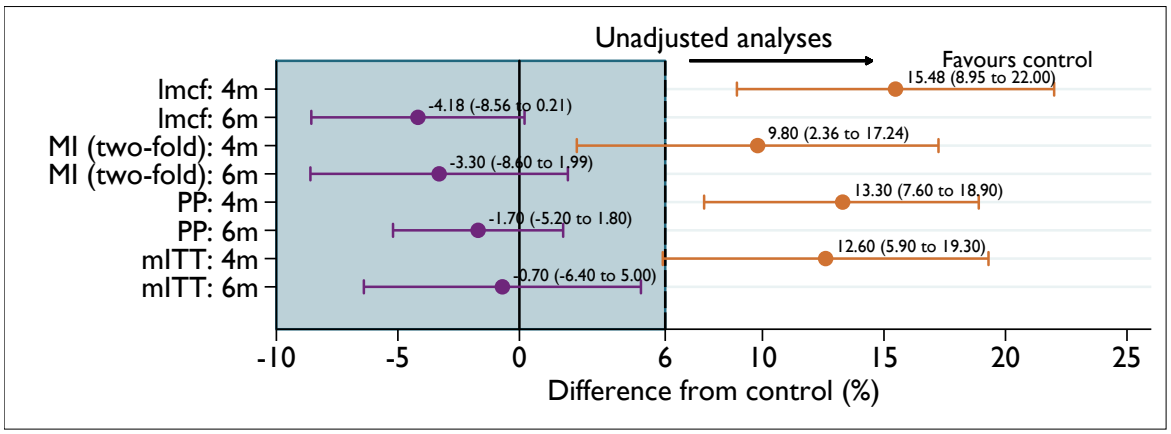
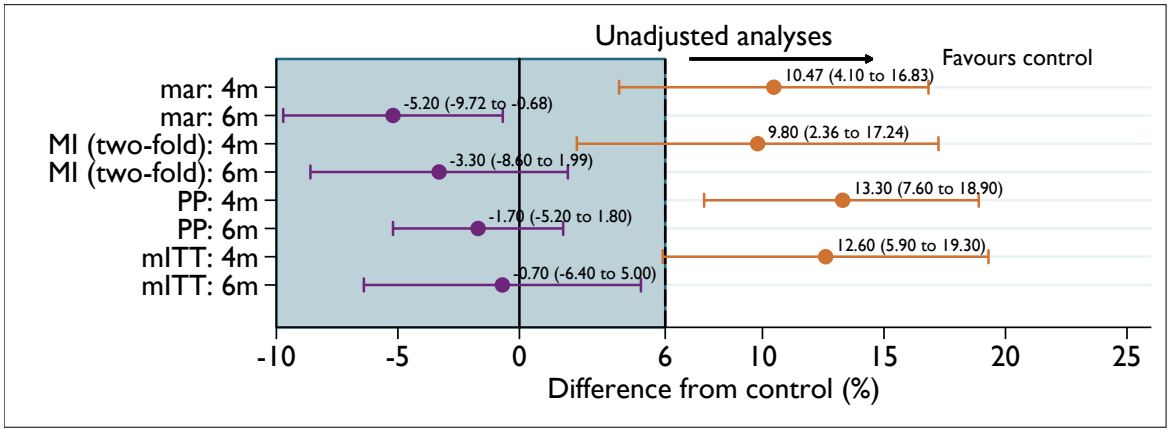


Figure N5: Missing at random sensitivity analyses for the RIFAQUIN study (unadjusted analysis).



Bibliography

[1] G. Piaggio, D. R. Elbourne, D. G. Altman, S. J. Pocock, S. J. Evans, and CONSORT Group. Reporting of noninferiority and equivalence randomized trials: an extension of the consort statement. *JAMA*, 295(10):1152–60, 2006.

[2] European Medicines Agency. [cited 15th november 2017]. http://www.ema.europa.eu/ema/index.jsp?curl=pages/about_us/document_listing/document_listing_000426.jsp&mid=.

[3] U.S. Food and Drug Administrators. [cited 15th november 2017]. https://www.fda.gov/AboutFDA/CentersOffices/default.htm.

[4] Consolidated Standards of Reporting Trials. [cited 15th november 2017]. http://www.consort-statement.org/.

[5] Standard Protocol Items: Recommendations for Interventional Trials. [cited 15th november 2017]. http://www.spirit-statement.org/.

[6] International conference on harmonisation; guidance on statistical principles for clinical trials; availability–fda. notice. *Fed Regist*, 63(179):49583–98, 1998.

[7] Food and H. H. S. Drug Administration. International conference on harmonisation; choice of control group and related issues in clinical trials; availability. notice. *Fed Regist*, 66(93):24390–1, 2001.

[8] Committee for Proprietary Medicinal Products Committee for Medicinal Products for Human Use (CHMP). Points to consider on switching between superiority and non-inferiority, 2000. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003658.pdf.

[9] Use Committee for Medicinal Products for Human, Party Efficacy Working, and Consultation Committee for Release for. Committee for medicinal products for human use (chmp) guideline on the choice of the non-inferiority margin. *Stat Med*, 25(10):1628–38, 2006.

[10] G. Piaggio, D. R. Elbourne, S. J. Pocock, S. J. Evans, D. G. Altman, and Consort Group. Reporting of noninferiority and equivalence randomized trials: extension of the consort 2010 statement. *JAMA*, 308(24):2594–604, 2012.

[11] D. Moher, S. Hopewell, K. F. Schulz, V. Montori, P. C. Gotzsche, P. J. Devereaux, D. Elbourne, M. Egger, and D. G. Altman. Consort 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ*, 340:c869, 2010.

[12] C. Begg, M. Cho, S. Eastwood, and et al. Improving the quality of reporting of randomized controlled trials: The consort statement. *JAMA*, 276(8):637–639, 1996.

[13] H.H.S Food, ; Drug Administration. Non-inferiority clinical trials to establish effectiveness. guidance for industry. 2016. https://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm202140.pdf.

[14] H.H.S Food, ; Drug Administration. Draft guidance for industry non-inferiority clinical trials. 2010. http://www.fdanews.com/ext/resources/files/archives/n/NoninferiorityGuidance.pdf.

[15] A. W. Chan, J. M. Tetzlaff, P. C. Gotzsche, and et al. Spirit 2013 explanation and elaboration: guidance for protocols of clinical trials. *BMJ*, 346:e7586, 2013.

[16] S. Lange and G. Freitag. Choice of delta: requirements and reality–results of a systematic review. *Biom J*, 47(1):12–27, 2006.

[17] T. R. Ten Have, S. L. Normand, S. M. Marcus, C. H. Brown, P. Lavori, and N. Duan. Intent-to-treat vs. non-intent-to-treat analyses under treatment non-adherence in mental health randomized trials. *Psychiatr Ann*, 38:772–783, 2008.

[18] R. J. Little, Q. Long, and X. Lin. A comparison of methods for estimating the causal effect of a treatment in randomized clinical trials subject to noncompliance. *Biometrics*, 65:640–9, 2009.

[19] K. L. Sainani. Making sense of intention-to-treat. *PM R*, 2:209–13, 2010.

[20] Sr. D'Agostino, R. B., J. M. Massaro, and L. M. Sullivan. Non-inferiority trials: design concepts and issues - the encounters of academic consultants in statistics. *Stat Med*, 22(2):169–86, 2003.

[21] Y. Matsuyama. A comparison of the results of intent-to-treat, per-protocol, and g-estimation in the presence of non-random treatment changes in a time-to-event non-inferiority trial. *Stat Med*, 29(20):2107–16, 2010.

[22] A. H. Kaji and R. J. Lewis. Noninferiority trials: Is a new treatment almost as effective as another? *JAMA*, 313(23):2371–2, 2015.

[23] C. H. Lee, T. Steiner, E. O. Petrof, and et al. Frozen vs fresh fecal microbiota transplantation and clinical resolution of diarrhea in patients with recurrent clostridium difficile infection: A randomized clinical trial. *JAMA*, 315(2):142–9, 2016.

[24] N. M. Rahman, J. Pepperell, S. Rehal, and et al. Effect of opioids vs nsaids and larger vs smaller chest tube size on pain control and pleurodesis efficacy among patients with malignant pleural effusion: The time1 randomized clinical trial. *JAMA*, 314(24):2641–53, 2015.

[25] A. R. Stevenson, M. J. Solomon, J. W. Lumley, and et al. Effect of laparoscopic-assisted resection vs open resection on pathological outcomes in rectal cancer: The alacart randomized clinical trial. *JAMA*, 314(13):1356–63, 2015.

[26] J. Fleshman, M. Branda, D. J. Sargent, and et al. Effect of laparoscopic-assisted resection vs open resection of stage ii or iii rectal cancer on pathologic outcomes: The acosog z6051 randomized clinical trial. *JAMA*, 314(13):1346–55, 2015.

[27] I. R. White, J. Carpenter, and N. J. Horton. Including all individuals is not enough: lessons for intention-to-treat analysis. *Clin Trials*, 9(4):396–407, 2012.

[28] R. J. Cook, L. Zeng, and G. Y. Yi. Marginal analysis of incomplete longitudinal binary data: a cautionary note on locf imputation. *Biometrics*, 60(3):820–8, 2004.

[29] G. Molenberghs, H. Thijs, I. Jansen, C. Beunckens, M. G. Kenward, C. Mallinckrodt, and R. J. Carroll. Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, 5(3):445–64, 2004.

[30] Committee for Proprietary Medicinal Products Committee for Medicinal Products for Human Use (CHMP). Points to consider on missing data, 2001. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003641.pdf.

[31] K. Unnebrink and J. Windeler. Sensitivity analysis by worst and best case assessment: Is it really sensitive? *Drug Information Journal*, 33(3):835–839, 1999.

[32] J. R. Carpenter, M. G. Kenward, and S. Vansteelandt. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society Series a-Statistics in Society*, 169:571–584, 2006.

[33] S. R. Seaman and I. R. White. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res*, 22(3):278–95, 2011.

[34] D. B. Rubin. *Multiple Imputation for Nonresponse In Surveys*. John Wiley and Sons: New York, 1987.

[35] J.R. Carpenter and M.G. Kenwood. Missing data in randomised controlled trials a practical guide, 2007. http://missingdata.lshtm.ac.uk/downloads/rm04_jh17_mk.pdf.

[36] P. J. Karanicolas, F. Farrokhyar, and M. Bhandari. Blinding: Who, what, when, why, how? *Can J Surg*, 53(5):345–8, 2010.

[37] Simon J Day and Douglas G Altman. Blinding in clinical trials and other studies. *BMJ*, 321(7259):504, 2000.

[38] M. Rothmann, B. L. Wiens, and I. S. Chan. [taken from google books] design and analysis of non-inferiority trials, 2012. https://books.google.co.uk/books?id=57LNBQAAQBAJ&printsec=frontcover#v=onepage&q&f=false.

[39] S. L. George, X. Wang, and H. Pang. [taken from google books] cancer clinical trials: Current and controversial issues in design and

analysis, 2016. https://books.google.co.uk/books?id=qcLBDAAAQBAJ&printsec=frontcover&printsec=frontcover#v=onepage&q&f=false.

[40] B.L. Wiens and G. K. Rosenkranz. Missing data in noninferiority trials. *Statistics in Biopharmaceutical Research*, 5:4:383–393, 2013.

[41] Steven M. Snapinn. Noninferiority trials. *Current Controlled Trials in Cardiovascular Medicine*, 1(1):19–21, 2000.

[42] World Health Organisation. [cited 22nd november 2017]. http://apps.who.int/iris/bitstream/10665/259366/1/9789241565516-eng.pdf?ua=1.

[43] Larissa Shamseer, David Moher, Mike Clarke, and et al. Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015: elaboration and explanation. *BMJ*, 349, 2015.

[44] S. Rehal, T. P. Morris, K. Fielding, J. R. Carpenter, and P. P. Phillips. Non-inferiority trials: are they inferior? a systematic review of reporting in major medical journals. *BMJ Open*, 6(10):e012594, 2016.

[45] ISI Web of Knowledge. [cited 31st may 2015]. http://admin-apps.webofknowledge.com/JCR/JCR.

[46] T. Hwang, I. K. Morikawa. Design issues in noninferiority equivalence trials. *Drug Information Journal*, 33:1205–1218, 1999.

[47] A. Vickers, N. Goyal, R. Harland, and R. Rees. Do certain countries produce only positive results? a systematic review of controlled trials. *Control Clin Trials*, 19(2):159–66, 1998.

[48] G. Tunes da Silva, B. R. Logan, and J. P. Klein. Methods for equivalence and noninferiority testing. *Biol Blood Marrow Transplant*, 15(1 Suppl):120–7, 2009.

[49] P. Schiller, N. Burchardi, M. Niestroj, and M. Kieser. Quality of reporting of clinical non-inferiority and equivalence randomised trials–update and extension. *Trials*, 13:214, 2012.

[50] T. A. Althunian, A. de Boer, O. H. Klungel, W. N. Insani, and R. H. Groenwold. Methods of defining the non-inferiority margin in randomized, double-blind controlled trials: a systematic review. *Trials*, 18(1):107, 2017.

[51] G. Wangge, O. H. Klungel, K. C. Roes, A. de Boer, A. W. Hoes, and M. J. Knol. Room for improvement in conducting and reporting non-inferiority randomized controlled trials on drugs: a systematic review. *PLoS One*, 5(10):e13550, 2010.

[52] A. Le Henanff, B. Giraudeau, G. Baron, and P. Ravaud. Quality of reporting of noninferiority and equivalence randomized trials. *JAMA*, 295(10):1147–51, 2006.

[53] T. Q. Gallagher, C. Hill, S. Ojha, and et al. Perioperative dexamethasone administration and risk of bleeding following tonsillectomy in children: a randomized controlled trial. *JAMA*, 308(12):1221–6, 2012.

[54] J. Radford, T. Illidge, N. Counsell, and et al. Results of a trial of pet-directed therapy for early-stage hodgkin's lymphoma. *N Engl J Med*, 372(17):1598–607, 2015.

[55] Felicity Hasson, Sinead Keeney, and Hugh McKenna. Research guidelines for the delphi survey technique. *Journal of Advanced Nursing*, 32(4):1008–1015, 2000.

[56] Comet initiative. http://www.comet-initiative.org/.

[57] E. Lesaffre. Superiority, equivalence, and non-inferiority trials. *Bull NYU Hosp Jt Dis*, 66(2):150–4, 2008.

[58] D. F. Postma, C. H. van Werkhoven, L. J. van Elden, and et al. Antibiotic treatment strategies for community-acquired pneumonia in adults. *N Engl J Med*, 372(14):1312–23, 2015.

[59] P. C. Gotzsche. Lessons from and cautions about noninferiority and equivalence randomized trials. *JAMA*, 295(10):1172–4, 2006.

[60] J. A. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, 338:b2393, 2009.

[61] T. P. Morris, B. C. Kahan, and I. R. White. Choosing sensitivity analyses for randomised trials: principles. *BMC Med Res Methodol*, 14:11, 2014.

[62] C. L. Vale, J. F. Tierney, and S. Burdett. Can trial quality be reliably assessed from published reports of cancer trials: evaluation of risk of bias assessments in systematic reviews. *BMJ*, 346:f1798, 2013.

[63] V. N. Chihota, A. D. Grant, K. Fielding, B. Ndibongo, A. van Zyl, D. Muirhead, and G. J. Churchyard. Liquid vs. solid culture for tuberculosis: performance and cost in a resource-constrained setting. *Int J Tuberc Lung Dis*, 14(8):1024–31, 2010.

[64] S.K Gupta. Non-inferiority clinical trials: Practical issues and current regulatory perspective. *Indian J Pharmacol*, 43(4):371–4, 2011.

[65] S. H. Gillespie, A. M. Crook, T. D. McHugh, C. M. Mendel, S. K. Meredith, S. R. Murray, F. Pappas, P. P. Phillips, A. J. Nunn, and R. EMoxTB Consortium. Four-month moxifloxacin-based regimens for drug-sensitive tuberculosis. *N Engl J Med*, 371(17):1577–87, 2014.

[66] A. Jindani, T. S. Harrison, A. J. Nunn, P. P. Phillips, and et al. High-dose rifapentine with moxifloxacin for pulmonary tuberculosis. *N Engl J Med*, 371(17):1599–608, 2014.

[67] C. S. Merle, K. Fielding, O. B. Sow, and et al. A four-month gatifloxacin-containing regimen for treating tuberculosis. *N Engl J Med*, 371(17):1588–98, 2014.

[68] Critical path institute. https://codr.c-path.org/. Accessed: 2016/06/20.

[69] R. J. A Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley Sons, Inc., 2002.

[70] A. M. Wood, I. R. White, and S. G Thompson. Are missing outcome data adequately handled? a review of published randomized controlled trials in major medical journals. *Clin Trials*, 1:368–76, 2004.

[71] Gordon D. Murray and Janet G. Findlay. Correcting for the bias caused by drop-outs in hypertension trials. *Statistics in Medicine*, 7(9):941–946, 1988.

[72] Stata multiple imputation reference manual release 13. https://www.stata.com/manuals13/mi.pdf. Accessed: 2017/10/16.

[73] Sas/stat(r) 9.3 user's guide. https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#mi_toc.htm. Accessed: 2017/10/16.

[74] Missing data imputation and model checking. https://www.rdocumentation.org/packages/mi/versions/1.0. Accessed: 2017/10/16.

[75] R. R. Andridge and R. J. A. Little. A review of hot deck imputation for survey non-response. *Int Stat Rev*, 78(1):40–64, 2010.

[76] J. Nevalainen, M. G. Kenward, and S. M. Virtanen. Missing values in longitudinal dietary data: a multiple imputation approach based on a fully conditional specification. *Stat Med*, 28(29):3657–69, 2009.

[77] S. van Buuren, J. P. L. Brand, C. G. M. Groothuis-Oudshoorn, and D.B. Rubin. Fully conditional specification in multivariate imputation. *Stat Med*, 76(12):1049–1064, 2006.

[78] Catherine A. Welch, Irene Petersen, Jonathan W. Bartlett, and et al. Evaluation of two-fold fully conditional specification multiple imputation for longitudinal electronic health record data. *Statistics in Medicine*, 33(21):3725–3737, 2014.

[79] James R. Carpenter and Michael G. Kenward. *Multiple Imputation and its Application*. Wiley. A John Wiley Sons, Ltd., Publication, New York, 2013.

[80] https://www.tbfacts.org/tb-drugs/. Accessed: 2017/10/16.

[81] Ian R. White, Rhian Daniel, and Patrick Royston. Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. *Computational Statistics Data Analysis*, 54(10):2267 – 2275, 2010.

[82] J. Carpenter, G. Rucker, and G. Schwarzer. Assessing the sensitivity of meta-analysis to selection bias: a multiple imputation approach. *Biometrics*, 67(3):1066–72, 2011.

[83] Geert Molenberghs and Michael G. Kenward. *Missing Data*. John Wiley and Sons: New York, 2007.

[84] R. Faria, M. Gomes, D. Epstein, and I. R. White. A guide to handling missing data in cost-effectiveness analysis conducted within randomised controlled trials. *Pharmacoeconomics*, 32(12):1157–70, 2014.

[85] P. P. Phillips, C. M. Mendel, D. A. Burger, A. M. Crook, A. J. Nunn, R. Dawson, A. H. Diacon, and S. H. Gillespie. Limited role of culture conversion for decision-making in individual patient care and for advancing novel regimens to confirmatory clinical trials. *BMC Med*, 14:19, 2016.

[86] Alan Agresti. *An Introduction to Categorical Data Analysis*. Wiley. A John Wiley Sons, Ltd., Publication, New York, 2007.

[87] A. G. Barnett, N. Koper, A. J. Dobson, F. Schmiegelow, and M. Manseau. Using information criteria to select the correct variance-covariance structure for longitudinal data in ecology. *Methods in Ecology and Evolution*, 1(1):15–24, 2010.

[88] Ware J.H. Fitzmaurice G.M., Laird N.M. *Applied Longitudinal Analysis. Second Edition.* Wiley. A John Wiley & Sons, Ltd., Publication, New York, 2011.

[89] Philip Hougaard. Multi-state models: A review. *Lifetime Data Analysis*, 5(3):239–264, Sep 1999.

[90] Z. Chen, S. Vijayan, R. Barbieri, M. A. Wilson, and E. N. Brown. Discrete- and continuous-time probabilistic models and algorithms for inferring neuronal up and down states. *Neural Comput*, 21(7):1797–862, 2009.

[91] Kulkarni V.G. *Introduction to Modeling and Analysis of Stochastic Systems*. Springer, 2011.

[92] L. R. Rabiner. A tutorial on hidden markov-models and selected applications in speech recognition. *Proceedings of the Ieee*, 77(2):257–286, 1989.

[93] J. D. Kalbfleisch and J. F. Lawless. The analysis of panel data under a markov assumption. *Journal of the American Statistical Association*, 80(392):863–871, 1985.

[94] Richard Durbin and et al. *Biological sequence analysis. Probabilistic models of proteins and nucleic acids.* Cambridge University Press, 1998.

[95] Multi-state modelling with r: the msm package. version 1.6.4. https://cran.ms.unimelb.edu.au/web/packages/msm/vignettes/msm-manual.pdf, 2016. Accessed: 2017/10/16.

[96] Walter Zucchini. Iain L. MacDonald. *Hidden Markov and Other Models for Discrete-valued Time Series*. Chapman and Hall, 1997.

[97] H. Binder, W. Sauerbrei, and P. Royston. Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. *Stat Med*, 32(13):2262–77, 2013.

[98] B. G. Leroux. Maximum-likelihood-estimation for hidden markov-models. *Stochastic Processes and Their Applications*, 40(1):127–143, 1992.

[99] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in statistical analysis of probabilistic functions of markov chains. *Annals of Mathematical Statistics*, 41(1):164–, 1970.

[100] J. Lember and A. Koloydenko. The adjusted viterbi training for hidden markov models. *Bernoulli*, 14(1):180–206, 2008.

[101] A. J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Ieee Transactions on Information Theory*, 13(2):260–+, 1967.

[102] G. Walz. Computing the matrix exponential and other matrix functions. *Journal of Computational and Applied Mathematics*, 21(1):119–123, 1988.

[103] B. C. Kahan, H. Rushton, T. P. Morris, and R. M. Daniel. A comparison of methods to adjust for continuous covariates in the analysis of randomised trials. *Bmc Medical Research Methodology*, 16, 2016.

[104] Chris Jackson. R documentation to calculate prevalence in the msm package. version 1.6.5. https://www.rdocumentation.org/packages/msm/versions/1.6.5/topics/prevalence.msm. Accessed: 2017/12/20.

[105] P. Saint-Pierre, C. Combescure, J. P. Daures, and P. Godard. The analysis of asthma control under a markov assumption with use of covariates. *Statistics in Medicine*, 22(24):3755–3770, 2003.

[106] Guideline on Missing Data in Confirmatory Clinical Trials. Points to consider on missing data, 2010. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/09/WC500096793.pdf.

[107] J. R. Carpenter, J. H. Roger, and M. G. Kenward. Analysis of longitudinal trials with protocol deviation: A framework for relevant, accessible assumptions, and inference via multiple imputation. *Journal of Biopharmaceutical Statistics*, 23(6):1352–1371, 2013.

[108] R. Little and L. Yau. Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics*, 52(4):1324–1333, 1996.

[109] S Cro. Relevant accessible sensitivity analysis for clinical trials with missing data. April 2017.

[110] Kenward M.G. Controlled multiple imputation methods for sensitivity analysis in longitudinal clinical trials with dropout and protocol deviation. *Clinical Investigation*, 5(3):311–320, 2015.

[111] S. Cro, M. Kenward, and J. Carpenter. Variance estimation in reference based sensitivity analysis for longitudinal trials with protocol deviation. *Trials*, 16, 2015.

[112] Nicholas J. Horton, Stuart R. Lipsitz, and Michael Parzen. A potential for bias when rounding in multiple imputation. *The American Statistician*, 57(4):229–232, 2003.

[113] O. N. Keene, J. H. Roger, B. F. Hartley, and M. G. Kenward. Missing data sensitivity analysis for recurrent event data using controlled imputation. *Pharm Stat*, 13(4):258–64, 2014.

[114] Y. Zhao, A. H. Herring, H. Zhou, M. W. Ali, and G. G. Koch. A multiple imputation method for sensitivity analyses of time-to-event data with possibly informative censoring. *J Biopharm Stat*, 24(2):229–53, 2014.

[115] C. A. Bernaards, T. R. Belin, and J. L. Schafer. Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Stat Med*, 26(6):1368–82, 2007.

[116] G. Wangge, O. H. Klungel, K. C. Roes, A. de Boer, A. W. Hoes, and M. J. Knol. Should non-inferiority drug trials be banned altogether? *Drug Discov Today*, 18(11-12):601–4, 2013.

[117] P. Ranganathan, C. S. Pramesh, and R. Aggarwal. Common pitfalls in statistical analysis: Intention-to-treat versus per-protocol analysis. *Perspect Clin Res*, 7(3):144–6, 2016.

[118] G. Casazza, M. Solbiati, and Metodologica Gruppo di Autoformazione. Can we trust equivalence and non-inferiority trials? *Intern Emerg Med*, 8(5):439–42, 2013.

[119] E. Brittain and D. Lin. A comparison of intent-to-treat and per-protocol results in antibiotic non-inferiority trials. *Stat Med*, 24(1):1–10, 2005.

[120] A. D. Garrett. Therapeutic equivalence: fallacies and falsification. *Stat Med*, 22(5):741–62, 2003.

[121] J. Rohmel. Therapeutic equivalence investigations: statistical considerations. *Stat Med*, 17(15-16):1703–14, 1998.

[122] S. K. Aberegg, A. M. Hersh, and M. H. Samore. Empirical consequences of current recommendations for the design and interpretation of noninferiority trials. *J Gen Intern Med*, 2017.

[123] Committee for Human Medicinal Products European Medicines Agency. Ich e9 (r1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials, 2017. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2017/08/WC500233916.pdf.

[124] Maha Hussain, Catherine M. Tangen, Donna L. Berry, and et al. Intermittent versus continuous androgen deprivation in prostate cancer. *New England Journal of Medicine*, 368(14):1314–1325, 2013. PMID: 23550669.

[125] Q. Ruan, Q. Liu, F. Sun, L. Shao, J. Jin, S. Yu, J. Ai, B. Zhang, and W. Zhang. Moxifloxacin and gatifloxacin for initial therapy of tuberculosis: a meta-analysis of randomized clinical trials. *Emerg Microbes Infect*, 5:e12, 2016.

[126] Panteha Hayati Rezvan, Katherine J. Lee, and Julie A. Simpson. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC Medical Research Methodology*, 15(1):30, Apr 2015.