The Molecular Epidemiology of Tuberculosis and the Impact of HIV Infection and Antiretroviral Therapy

Rein Maria Geert Jan Houben

London School of Hygiene and Tropical Medicine

Submitted for the degree of Doctor of Philosophy

May 2010

I, Rein Maria Geert Jan Houben, 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.

Signature: .... Date: ....

Abstract

This thesis investigates the epidemiology of tuberculosis (TB) disease in populations. It applies molecular epidemiological methods to elucidate the relative effects of HIV on TB disease following recent first infection, reactivation and recent reinfection with Mycobacterium tuberculosis (Mtb). Finally it aims to explore the impact of Antiretroviral Therapy (ART) on the incidence of TB in a population in rural Sub Saharan Africa.

The data in chapters 2 and 3 is taken from a systematic literature review I performed of population based studies that reported TB molecular epidemiological data. In the subsequent chapters I analyse existing and newly collected data from the Karonga Prevention Study, set in Northern Malawi, to address the research questions.

The results strongly suggest that HIV-infection increases an individual's risk of TB disease due to recent Mtb (re)infection more than through reactivation in populations with generalised HIV epidemics. The last chapter suggests that patients on ART experience a high risk of TB compared to HIV positive/ART naive patients, especially in the first months after initiating ART. Also, it appears that after the introduction of ART in 2005 TB incidence in Karonga District plateaued after declining in the previous 10 years.

These findings strongly suggest that TB programmes in areas with generalised HIV epidemics should focus more of their efforts on reducing Mtb transmission. Improved collaboration between TB and ART programmes may help to reduce TB rates in the highly vulnerable ART receiving population and subsequently in the general population.

Acknowledgements

For nearing the end of a long road that was never lonesome I have many people to thank. Given the space limitations I must unfortunately be brief.

Goodenough College and its community gave me a warm welcome and place to find friends and a partner for life.

To KPS, all its staff in London and Malawi, and especially my colleagues Mia and Lumbani, thank you for your invaluable insights, advice and opportunities that drove me to make the most out of a great year of adventure.

In Dutch we say that success has many fathers. I have 2 parents who always supported me and continue to show by example that hard work and perseverance always pay off in the end.

The Dutch also say that shared suffering is half the suffering. My PhD colleagues in London and the Netherlands, thank you for the inspiration and motivation

I am forever indebted to my supervisor Judith, who retained a steady 'Don't Panic' mode all through the times when I was fully prepared to do so and relentlessly set an example of best practice in PhD supervision. I look forward to our future collaboration to take these results further

That leaves me with you, Fiona. To you I dedicate life, the universe and everything in it, which roughly equates to our shared joy of there being no more theses to write. I look forward to fulfilling the promise of an exciting and PhD-free life with you.

P.S. To the (at least) 42 people I forgot to mention, my mentors, colleagues, family, friends and kind people who offered their hospitality during my travels in the past years, I apologise for the omission. I hope you, as one Douglas Adams once did, will find consolation or joy in the random fact that some young sloths are apparently so inept that they frequently grab their own arms and legs instead of tree limbs, and fall out of trees. I look forward to seeing you again soon.

List of abbreviations

ARI	Annual Risk of Infection
C+ / C-	Culture positive / negative
HIV	Human Immunodeficiency Virus
KPS	Karonga Prevention Study
MICE	Multiple Imputation using Chained Equations
MIRU-VNTR	Mycobacterial Interspersed Repetitive Units-Variable Number of Tandem Repeats
Mtb	*Mycobacterium tuberculosis*
PGRS	Polymorphic GC Rich Sequencing
RFLP	Restriction Fragment Length Polymorphism
SS+ / SS-	Sputum Smear positive / negative
TB	Tuberculosis

Table of Contents

Table of Tables

Table of Figures

1 Introduction

1.1 Chapter Outline

In this chapter I will introduce the epidemiology of Mycobacterium tuberculosis and tuberculosis disease. The molecular epidemiology of TB is introduced, as well as its applications so far. In section 1.6 I will outline the main questions this thesis aims to answer and where I do so. Finally I will give an overview of the chapters and a brief summary of their content.

1.2 The burden of Tuberculosis

1.2.1 Overview

Tuberculosis disease (TB) was declared a “global emergency” by the WHO in 1993. [1] Nevertheless, fuelled by the HIV epidemic TB disease remains one of the major health threats, still causing about 9 million cases and 2 million deaths a year worldwide. [2] A large proportion of these cases occurs in areas with generalised HIV epidemics (defined as a HIV epidemic established in the general population and sustainable without the contribution of high risk groups and where HIV prevalence is consistently measured as > 1% in pregnant women [3]), which includes many of the countries of East and Central Sub Saharan Africa.

1.3 The epidemiology of Tuberculosis

1.3.1 Definition of TB disease

The epidemiology of tuberculosis disease (TB) is complex. The symptoms of TB disease are caused by the immune system’s reaction to an active infection with a member of the *Mycobacterium* genus, predominantly *Mycobacterium tuberculosis* (Mtb). TB can involve other members of the *Mycobacterium* species such as *Mycobacterium bovis* and *Mycobacterium avium*. However, Mtb underlies the majority of TB episodes. For the purpose of this thesis TB is defined as a disease episode caused by an active infection with a strain of Mtb.

1.3.1.1 Measurement of Mtb infection

Infection with Mtb is difficult to measure. The most commonly used test to determine Mtb infection has been the same since the beginning of the 19th century; The Tuberculin Skin Test (TST), also referred to as the Mantoux test. This test determines the body’s immune response after an intradermal injection of ~2ml non species specific purified protein derivative from mycobacteria into the volar surface of the forearm. [4] After 48 to 72 hours after the induration (palpable raised and hardened area) at the location of the injection is measured. If the individual has experienced an Mtb infection in the past an induration is expected to be seen. The required induration for a ‘positive’ TST has been under intense discussion. I use results from TST in Chapter 6 and will discuss the limitations in more detail there.

The TST can be false positive due to cross contamination with other mycobacteria, from a BCG vaccination or due to boosting following multiple tests. The TST can be false negative in immunocompromised individuals, such as HIV positive patients with advanced immunosuppression. Advantages include that the TST is cheap, and its methods and interpretation are reasonably well standardised.

In recent years new blood based tests for measuring Mtb infection have been developed and evaluated. Despite a huge body of research [5], their use and interpretation is still not clear. [6] This thesis does not use any results from these new tests and I will therefore not discuss them further.

1.3.1.2 Presentations of TB disease

I will not go into detail on the diagnosis of TB disease. In short, diagnosis is usually through sputum examination, confirmed by a culture growth positive for Mtb. For the epidemiology the following concepts are important:

1.3.1.2.1 Pulmonary versus Extra Pulmonary TB

A TB episode can have foci in multiple places, although the main focus is often the lung which is referred to as pulmonary TB (PTB). If there is no pulmonary involvement, and TB is present in other organs the TB is referred to as Extra Pulmonary TB (EPTB). Epidemiologically three points are important to mention:

1. Patients can have both PTB and EPTB in the same episode [7, 8]
2. Pulmonary (or laryngeal) involvement in the TB is required for Mtb transmission, as transmission occurs when sputum with live Mtb bacilli is coughed up from the lung and spread. Patients without pulmonary involvement will, in principle, not be the source of Mtb transmission. [8]
3. Patients with an Mtb and HIV co-infection are more likely to have extra pulmonary foci when they develop TB disease. [7]

1.3.1.2.2 Smear positive versus Smear negative TB disease

A TB episode can be smear positive (SS+), which refers to Mtb bacilli seen under a microscope after staining fixed smear of any sputum samples. If no smear can be done, or no Mtb bacilli are seen in any of the smears, the TB episode is referred to as smear negative (SS-). For the epidemiology the following 2 points are important to note:

1. SS- patients are less infectious than SS+, although transmission is still possible [9, 10]
2. Patients with an Mtb and HIV co-infection are more likely to be SS- when they develop TB disease. [11-14]

1.3.1.2.3 Culture Positive versus Culture Negative TB disease

The gold standard of the diagnosis of TB disease is a positive culture of Mtb. If a patient is SS- and culture negative (C-), the diagnosis of TB disease is not confirmed and treatment initiation is only based on clinical symptoms and X-ray if available. For the epidemiology the following points are important:

1. Patients with PTB that are SS- but culture positive (C+) are the subgroup that are likely to still transmit Mtb, albeit at a lower intensity [9, 10]
2. Patients with an Mtb and HIV co-infection are more likely to be culture negative (C-) [8]

1.3.2 Pathways from Mtb infection to a first episode of TB disease

The first step in the pathway to TB disease is an effective contact with an infectious TB case, defined as a contact between an infectious person and someone who has not been infected that leads to the uninfected person to become infected with Mtb. After an Mtb infection has been established, 3 pathways are distinguished in the epidemiology of TB that can lead from Mtb infection to a first episode of TB (see figure 1—1).

1.3.2.1 Rapid progression to TB disease

Following a first Mtb infection an individual can progress 'rapidly' to TB disease. By convention, this period for rapid progression is defined as within the first 5 years after the Mtb infection event. [15, 16]. This 5 year cut off is based on the relatively higher risk of developing TB in the first years after Mtb infection (see figure 1—2). [15, 17] It is arbitrary to an extent, and can be varied. I will discuss this issue in more detail in section 1.5: The molecular epidemiology of TB In this thesis TB following a rapid progression from a first Mtb infection is referred to as TB following recent infection or primary TB. Please note that in clinical discussions primary TB can refer to the first pulmonary foci of Mtb infection before haematogenous spread occurs. However, this definition is not used in this thesis.

1.3.2.2 Reinfection and exogenous disease

After a first infection with Mtb an individual can be re-infected with Mtb before their first TB episode. After a reinfection event the risk of developing TB follows a similar pattern to that seen after a first Mtb infection. [15, 17] If the person develops TB within a set period after reinfection, again conventionally defined as 5 years, this TB episode is referred to as TB following recent re-infection TB or exogenous TB.

Note that it is difficult to distinguish between these two types of TB that follow a recent Mtb infection event. Although the importance of re-infection disease is unclear, its relevance as part of TB pathogenesis in areas with moderate to high rates of Mtb transmission has been demonstrated in modelling studies [15, 17] and elegantly illustrated through population data by Styblo et al.[8] In Chapter 5 I will explore this issue in more detail.

1.3.2.3 Reactivation disease

If an Mtb infection does not progress to TB disease within a set period, usually defined as 5 years, it is assumed to have become established in the body as a latent infection. In this process Mtb bacteria are encapsulated in granulatomous foci in the lungs or other parts of the body. This latent Mtb infection can reactivate in later life to cause an episode TB disease, which is referred to as reactivation or endogenous TB.

Although the risk of reactivation TB from a latent infection is low compared to TB following a recent infection, it is assumed to last life–long and increase with age. [15] It therefore plays an important part of TB epidemiology, especially in populations with low levels of ongoing Mtb transmission such as the non-migrant populations in developed countries.

Please note that under the current TB paradigm an Mtb infection will last lifelong. However, results from recent studies on long term Mtb infection suggest that this process is less clear and that some individuals appear to clear their Mtb infection without ever developing TB or taking TB treatment. *Crampin et al. in preparation*

1.3.3 Recurrent TB

After completing TB treatment for a first TB episode an individual can experience a recurrence. This subsequent TB episode can either be a relapse or re-infection. A relapse is defined as a new episode with the same Mtb strain, which was not cleared by the treatment for the preceding TB episode. A recurrence due to re-infection is defined as a new TB episode caused by a different Mtb strain than the Mtb strain involved in the previous TB episode.

Although I will refer to studies on recurrent TB, the main focus of this thesis in on first TB episodes.

Figure 1—1: Pathways from first Mtb infection to first episode TB

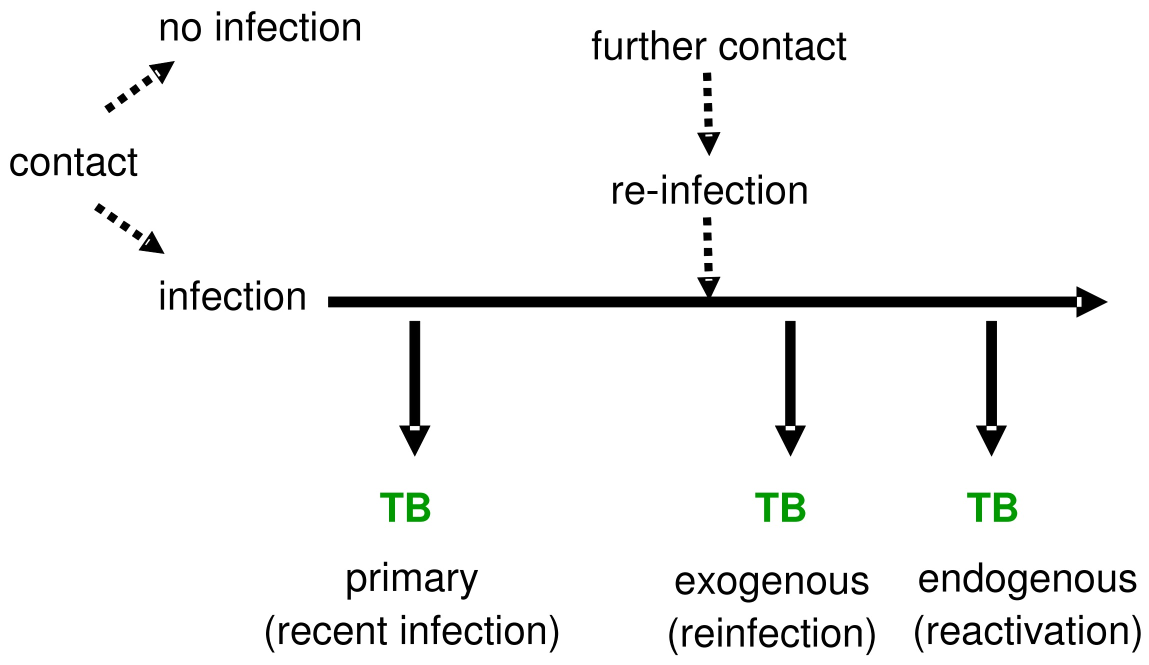
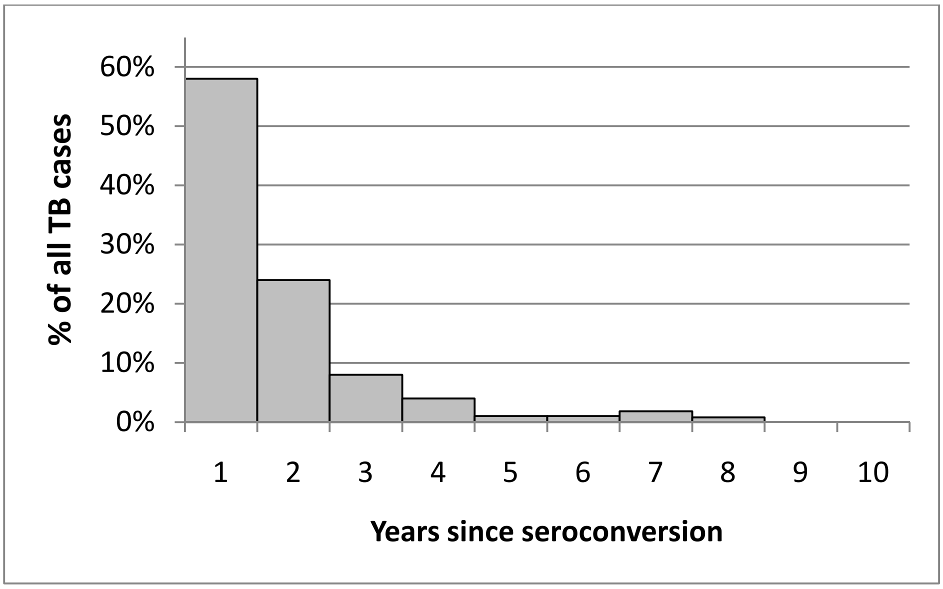


Figure courtesy of Judith Glynn and Paul Fine

Figure 1—2: Percentage of cases occurring by year since first Mtb infection

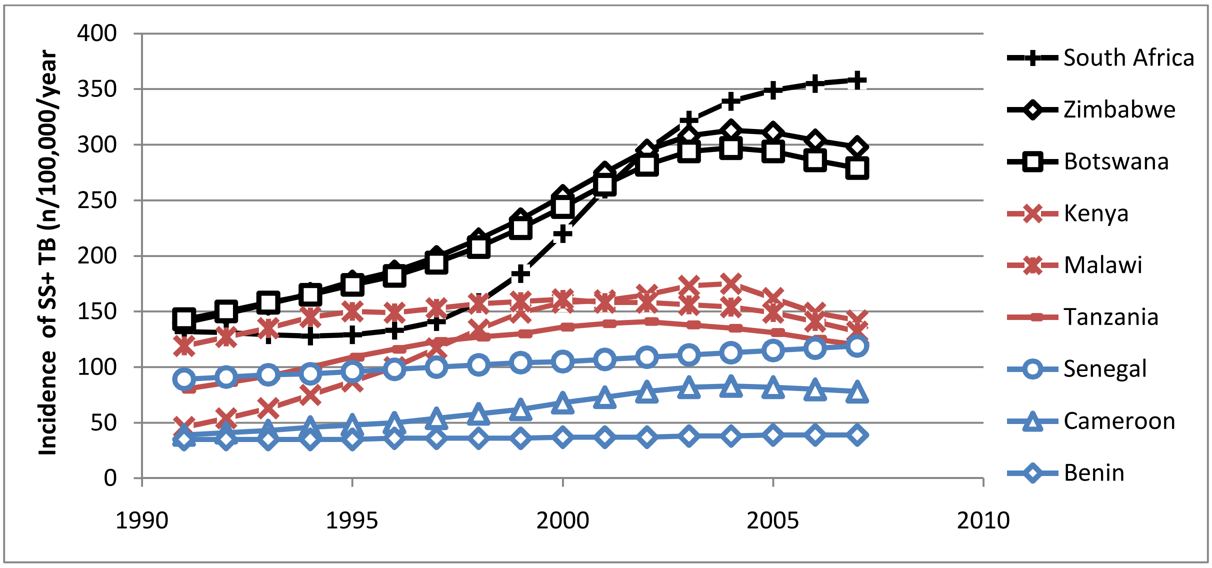


Data in figure reproduced from Medical Research Council vaccine trial [18]

1.4 TB and HIV

The link between HIV and tuberculosis (TB) has been clear since the beginning of the HIV-epidemic [19]. In countries with high prevalences of HIV TB incidence rates increased dramatically, as shown in figure 1—3. [20] Tuberculosis is the main cause of death in HIV-positive individuals from low resource settings [21]. The effect of HIV on TB rates is often separated in a direct effect on the risk of TB in infected with HIV and an indirect effect for the risk of TB in the general population. [22, 23]

Figure 1—3: Annual incidence of SS+ TB by country (1991 – 2007)



Data from Global TB database http://www.who.int/tb/country/global_tb_database/en/index.html (accessed at 13-Feb-2010). Showing data from selected countries.

1.4.1 Direct effect of HIV infection on TB

If an Mtb infected person remains HIV-negative their lifetime risk of TB is estimated at 10%. [17] If they become co-infected with HIV their risk of active TB disease increases to 10% annually [22].

In the beginning of the HIV epidemic it was assumed that the risk of TB was high immediately after HIV infection, then decreased to close to null and then start rising. [24-27] Recent studies have shown that the relative risk of TB actually starts to increase immediately after HIV infection and continues to do so for at least 10 years afterwards [28].

Although it is clear that HIV infection dramatically increases the risk of active TB and affects all three pathways to from Mtb infection to TB disease [22] , the relative effects on the risk for primary, exogenous and endogenous TB may vary [29]. This issue of the relative effect of HIV infection on the different pathways from Mtb infection to TB disease is one of the main focuses of this thesis and will be discussed in more detail in section 1.5.6 and chapters 3, 4 and 5.

1.4.2 Indirect effect of HIV on TB incidence

It is assumed that the high number of additional TB cases in the HIV positive population have a knock on effect in the general population, which is referred to as the indirect effect of HIV. [30]

Odhiambo et al. showed that in Kenyan districts where HIV prevalence in TB cases was high (~50%) the annual risk of Mtb infection (ARI) had increased as well [31].

In addition a long term cohort study among South African miners showed that overall TB incidence rates increased in HIV negatives as well as HIV positives, suggesting that active transmission from the additional TB cases in HIV positives was also affecting TB incidence in the general population [23]. However, a similar gold mine study found no such increase [32] and to date the relative contribution of the indirect effect of HIV on the TB incidence remains under discussion [30, 33].

1.4.3 HIV, TB and the effect of ART

The roll out of antiretroviral therapy (ART) will change the interaction between HIV and TB disease dynamics and affect TB incidence. The extent and direction of this effect will depend on various, interrelated factors, including the immune status of patients starting ART, their risk of TB and survival. [21, 34] Several studies have looked at the relative incidence in cohorts of HIV positive or ART receiving patients [35-46], but none looked at the effect on the overall TB incidence in a a complete population, which is what I will do in Chapter 7. I will discuss the existing literature on ART and TB in more detail there.

1.5 The molecular epidemiology of TB

1.5.1 Measuring TB following recent transmission

Historically one of the complications in the study of TB epidemiology has been the difficulty of distinguishing between TB disease following recent (re-)infection or reactivation. [17] For the last 15 years molecular epidemiology of TB explored this distinction through the application of DNA fingerprinting of Mtb strains involved in TB episodes. [47]

1.5.2 General theory of TB molecular epidemiology

Molecular epidemiology works from the assumption that active TB cases are likely to be epidemiologically related if their Mtb strains have identical, or at least similar DNA fingerprints. These cases with similar or identical Mtb strains are referred to as clustered. DNA fingerprinting techniques thus distinguish between TB cases whose episodes involves a unique Mtb strain and TB cases whose episodes involve clustered (i.e. non-unique) Mtb strains.

If the strain causing the TB episode is unique, it is assumed that the patient's TB episode is more likely to have followed the reactivation of a latent Mtb infection. If the strain is clustered it is assumed that the individual was involved in recent transmission. In each group of clustered cases with identical strains (so called cluster) at least one TB case is due to reactivation of a latent infection. This is referred to as the source case. All other cases are assumed to be due to a recent first infection or reinfection with the Mtb strain transmitted by the source case (or by other cases in the cluster). [48] These cases are referred to as secondary cases. This definition of secondary cases includes all subsequent cases in the cluster, regardless whether the source of their Mtb infection was the original source case or another secondary case.

There are several DNA fingerprinting techniques available; IS*6110* Restriction Fragment Length Polymorphism (RFLP), Spoligotyping, Polymorphic GC Rich Sequencing (PGRS), Mycobacterial Interspersed Repetitive Units-Variable Number of Tandem Repeats (MIRU-VNTR), and more recently whole genome sequencing. In this thesis I will mostly use data acquired with IS*6110* RFLP and will discuss this technique in more detail later in this chapter.

1.5.3 The definition of a clustered case

The main application of clustering is to identify cases involved in ongoing Mtb transmission. Straightforward as this may sound, the application of and interpretation molecular epidemiology has varied between studies and settings. In this section I discuss some of the areas of variation and the impact on the interpretation of the results.

1.5.3.1 Identical versus similar

Firstly, studies can vary the level of required similarity between strains. Some studies allowed a small difference between DNA fingerprint patterns. However, usually only strains with identical DNA patterns are classified as clustered. If similar as well as identical strains are considered to be clustered the proportion of clustered cases will increase. This issue is discussed in more detail in Chapter 2.

1.5.3.2 Time between source and secondary case

The definition of what constitutes 'recent' in the expression recent (re)infection differs between studies. Each researcher needs to decide on the cut-off for the maximum amount of time between diagnosis of 2 cases in which the secondary case is still considered to have been recently infected by the source case and then rapidly progressed to TB disease from that Mtb infection. If the two cases are less than 2 or more than 10 years apart there is usually no discussion; in the first cluster the secondary case is assumed to be due to recent infection; in the second cluster both cases are assumed to be due to reactivation where the first case may have been the source of the Mtb infection, but the infection in the second case first went through a latent phase before reactivating. Usually the cut off is set on 4 to 5 years, following the development of the risk of TB after Mtb infection (see figure 1—2) although most studies do not follow cases sufficiently long to fully capture this effect.

1.5.3.3 Excluding the source case in a cluster

The first case, usually based on time of diagnosis, is usually assumed to be the case that was the source of the cluster and that the TB followed a reactivation [10, 49] However, it is important to note that TB has a highly variable period between Mtb infection and TB disease which has an estimated minimum of 2-3 months [50] to a maximum of the individual's lifetime [17]. Also, some TB cases experience low intensity symptoms for a long period (up to 18 months) without being diagnosed [12], which makes it difficult to be sure when Mtb transmission occurred.

Still some studies have opted to exclude the first case in a cluster with the aim of acquiring a better estimate of the TB cases due to, rather than involved in, recent Mtb transmission. [10, 49]

This approach is commonly referred to as the n-1 method. [48] In this method one calculates the number of cases in a cluster, minus the number of clusters, as a proportion of all cases. In theory, this gives an estimate of the proportion of TB due to recent transmission.

1.5.3.4 Retrospective clustering

Another approach is measuring retrospective clustering, which aims to use clustering as a measure of cases due to, rather than involved in, recent transmission [51]. Here a case is defined as clustered if in a set number of years before (the researchers explored cut offs between 1-7 years) another TB case had been found with an identical Mtb strain. This approach deals with the issue of the cut off as well as the source case.

1.5.3.5 Definitions of clustering in this thesis

Throughout the chapters in this thesis I will use different definitions of what constitutes a clustered case depending on the data available and the analysis goal.

1.5.4 IS*6110* RFLP

Until recent years the main DNA fingerprinting technique used worldwide was Restriction Length Fragment Polymorphism typing using the *6110* insertion sequence (IS*6110* RFLP). The majority of existing molecular epidemiological data is based on this technique, and it forms the mainstay of this thesis.

IS*6110* RFLP was standardised in 1993 [52] and has proven to give valid inferences of recent transmission in outbreak and population settings as I will discuss section 1.5.5.

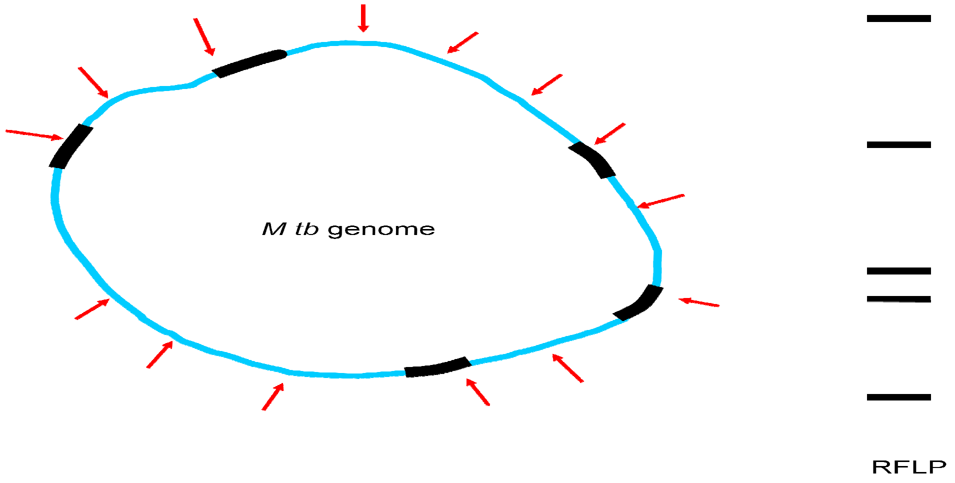
Please note that this thesis focuses on the epidemiological interpretation of DNA fingerprinting results, not on the technique itself. This is why I will keep the description of the technique brief. In short, IS*6110* denotes a repetitive fragment in the circular DNA of Mtb. IS*6110* RFLP uses the variability between Mtb strains in both the number of copies and the chromosomal positions of these fragments by cleaving the DNA at the location of each IS*6110* insertion fragment (see figure 1—4). This results in several fragments of different lengths.

Each fragment is labelled and then put through a Southern gel Blot. Different sized fragments will travel through the blot at different speeds. After photographing a barcode pattern is revealed which is referred to as the 'DNA fingerprint' for that strain (see figure 1—5 for illustration).

The DNA fingerprints are compared to each other using software and visual inspection and grouped clusters of identical DNA patterns.

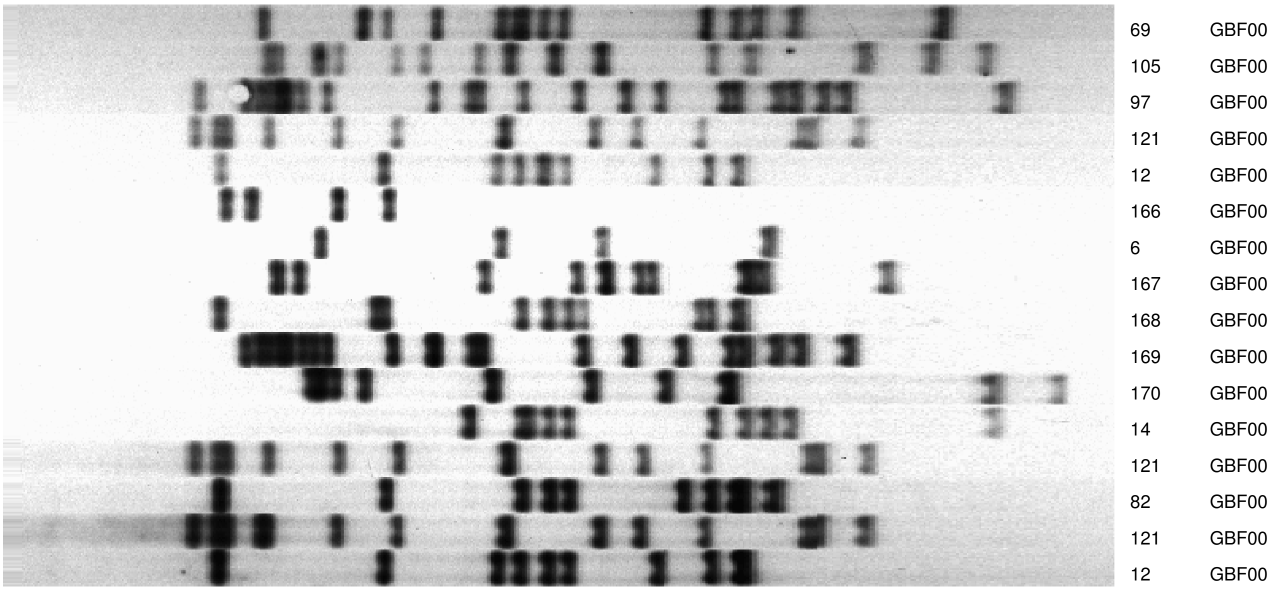
An important determinant of the usefulness of a marker is the expected rate of evolution or half-life of a pattern. This needs to be slow enough to allow identification of epidemiologically related cases, but fast enough to see differences over time. The half life for an IS*6110* RFLP pattern during active disease is roughly estimated at 2-3.2 years [53], which makes it suitable for various applications, which are discussed in section 1.5.5. However, one study observed an IS61110 RFLP pattern that remained unchanged during a latent Mtb infection for 30 years. [54]

Figure 1—4 Graphic representation of the circular Mtb genome



Blsck sections depict IS*6110* sequences, which are the loci where the DNA will be cut. This cutting will result in Mtb DNA fragments of different length. (Figure courtesy of Judith Glynn)

Figure 1—5 Mtb DNA fingerprint patterns from IS*6110* RFLP



Note: each band represents a IS*6110* fragment of a specific length. (Figure courtesy of Judith Glynn).

1.5.4.1 Limitations of molecular epidemiology/RFLP

Unfortunately the molecular epidemiology of TB has some issues that can limit the interpretability of its results. In this section I discuss some of the issues related to the technique/basic theory and how they affect the interpretation of DNA fingerprints. Factors relating to study design, population and setting are discussed in detail in Chapter 2.

Molecular epidemiology relies on the assumption that a TB episode is caused by 1 strain of Mtb. However, recent studies from South Africa extracted multiple strains from the sputum sample of individual patients. [55] In a high incidence setting they detected multiple strains in 19% of their samples, although other studies found percentages between 0.4 and 6% [56, 57]. The frequency and clinical significance of mixed Mtb infections underlying the same TB episodes is still under investigation. [58]

The extent to which mixed infections affect the interpretation of molecular epidemiological results will depend on the proportion of TB episodes in which mixed Mtb infections are shown. This in turn depends on the technique and definition used to detect ‘different’ Mtb strains in one sample. In Karonga District, where the majority of the data for this thesis comes from, the estimated rate of mixed infections is <4% (preliminary and unpublished data). Also, as the next section will show, molecular epidemiology as a method has been validated in different settings and applied successfully for different purposes.

Another issue is that clustering analyses are limited to culture positive TB cases. As the proportion of culture positive TB cases differs by HIV status, analyses that compare HIV positive and HIV negative cases could give biased results if the proportion of cases following recent infection is different for culture negative (e.g. most EPTB cases) and culture positive TB cases. There are few reports that clustering is lower in EPTB (which is hard to detect), and a comparison of 4 year retrospective clustering (which excludes presumptive source cases) between EPTB and PTB cases in the KPS database showed that the proportion clustered was 50% in EPTB and 61% in PTB cases. *Unpublished data* So although there could potentially be some bias, the evidence base is small.

One issue with clustering is stable strains, for which the RFLP pattern appears to remain unchanged over a long period, e.g. 10 years or more. Theoretically, if strains predominate in an area over a long period the interpretation of clustering as a marker for recent transmission becomes problematic as the infection event could have occurred more than 5 years before the disease episode started.

Also, there is a possibility that 2 different strains evolved to an identical pattern simultaneously.

The assumption that only identical strains are part of the same active transmission chain does not always hold; a small deletion or recombination in the mycobacterium's DNA shortly after Mtb transmission can cause DNA fingerprints or RFLP patterns to differ, causing the secondary case to be attributed to reactivation, whereas in reality they are due to recent transmission.

RFLP is quite labour intensive. Up to 2mg of live TB culture is required for DNA extraction, which means a healthy growth is required for each case which takes time to obtain.

Also, despite the computer supported process of comparing strains, IS*6110* RFLP, is vulnerable to technical problems in the preparation which can lead to unclear fingerprints as well as subjective interpretation of the results, both of which can lead to misclassification [47].

In addition, IS*6110* RFLP is insufficiently discriminatory when the number of IS*6110* fragments is low, usually defined as below 6. [59] This makes the technique difficult to interpret in areas where a substantial part of the circulating Mtb strains have less than 6 insertions as in South East Asia. [47]

1.5.5 Applications of Molecular Epidemiology

Despite these potential limitations, DNA fingerprinting data from *IS*6110 RFLP have been extensively validated as a tool to identify TB cases due to recent Mtb infection in various settings. However, caution is required in their interpretation. [59]

1.5.5.1 Outbreaks

IS*6110* RFLP was first validated as a tool in conjunction with epidemiological investigations to map TB outbreaks in industrialised settings. [60] IS*986* RFLP (an alternative marker to *6110*) was used to show that all 11 of the culture positive TB cases involved in the outbreak had a similar Mtb strain to the source case, whereas 2 other cases had a distinctively different RFLP pattern. [60]

In outbreak settings DNA fingerprinting can show which of the suspected cases was actually part of the outbreak, which cases were missed in the epidemiological investigations and which suspected TB cases involved in the outbreak were likely to have developed reactivation TB and not involved in the transmission chain. Molecular epidemiology can thus help to inform interventions aimed to prevent future outbreaks. [60-64]

1.5.5.2 Laboratory quality control

Multiple studies aimed to explore potential cross contaminations in a TB laboratory used IS*6110* RFLP to compare the Mtb strains of both samples in a pair of potential cross contamination. [65-68] Molecular epidemiology was thus able to provide useful information on whether cross contamination had occurred and if so, which of the two samples was the likely source of contamination.

1.5.5.3 Spread of Drug resistance

Molecular epidemiology has been used to study the spread of drug resistance and the relative contribution of the transmission of drug resistant Mtb strains or the development of drug resistance during TB treatment. [61, 64]

1.5.5.4 Population studies

The main area where TB molecular epidemiology has been applied is in population studies. A population can be defined narrowly (e.g. a prison population or hospital workers) or wider as the whole population in hospital catchment area, city or province.

By typing all or a sample of TB cases in such a defined a population it is possible to get an estimate of the proportion of TB cases either involved in (the crude proportion clustered cases) or due to recent Mtb transmission (e.g. through retrospective clustering).

When measured repeatedly or over a longer period, trends in the proportion clustered can give an indication of the progress on TB control in the population. [49, 69] A reduction in the proportion of TB cases that is clustered in an area over time is considered to be a sign of improved TB control [70].

It is important to note however that the observed proportion clustered in a population depends on various factors relating to study design, study population and study setting. In Chapter 2 I describe these factors and how they affect the observed proportion clustered in more detail. In short, an ideal study on TB clustering includes all TB cases from a complete, stable and well defined population. For example all TB cases in a district with low levels of migration over a long period of time, preferably 5 years or more. [59]

1.5.6 Application of TB molecular epidemiology in TB and HIV

Over the years molecular epidemiology has been applied to study the interaction between HIV and TB disease.

1.5.6.1 Relative infectiousness of HIV positive TB cases

Molecular epidemiology of TB has been applied to study the relative infectiousness of HIV positive TB cases within households. DNA fingerprints for TB cases that had identified each other as a contact were compared to see whether the first case could have been the source of the Mtb strain causing the TB episode in the second case.

This study found that if the suspected source case was HIV positive the strains were less likely to match (OR (95%CI) = 0.32 (0.14 – 0.74)). [71] This suggests that HIV positive TB cases are less likely to be the source of infection.

1.5.6.2 HIV and recurrent TB in populations

Two studies of recurrent episodes of TB compared the DNA fingerprints of the strains involved in both episodes and were able to report the rate of recurrent TB due to relapse or recent reinfection, stratified by HIV status. [72, 73] Their analyses showed that the rate of recurrences due to recent re-infection was much higher in HIV positive TB cases. In South Africa the hazard ratio (95% CI) was 18.7 (2.4 – 14.3) [73], in Malawi this was 13.5 (1.8-103.7) [72]. There was no relevant difference by HIV status in the rate of recurrences due to relapses.

1.5.6.3 HIV and the risk of TB following recent infection versus reactivation

In Chapter 3 I report the results of a systematic review on this issue. Please refer to that chapter for a detailed description of the available work on this issue.

1.6 Main questions and thesis structure

1.6.1 Main questions

In this thesis I will address and attempt to answer the following questions

1. Why does the population proportion clustered vary so strongly between studies and settings? (Chapter 2)
2. What is the relative effect of HIV on TB following recent Mtb (re-)infection and TB following reactivation of a latent Mtb infection? (Chapter 2, 3, 4 and 5)
3. What is the effect of the roll out of ART on the incidence of TB in a rural African population? (Chapter 6 and 7)

1.6.2 Structure of the thesis

Chapter by chapter I will discuss the following:

1.6.2.1 Chapter 2

In Chapter 2 I will explore the variation in the observed proportion clustered between studies of general populations using data from a systematic literature review on all population–based studies on TB clustering that used IS*6110* based RFLP as the main DNA fingerprinting technique. I will use the resulting dataset to examine the extent to which this variation in observed clustering can be explained by study design, study setting (the local epidemiology of TB) and study population. In the process I develop a new tool for the interpretation of a locally observed proportion clustered in the context of a study's design and setting.

1.6.2.2 Chapter 3

In this chapter I describe a systematic review and pooled data analysis of studies with information on HIV and TB cluster status of TB cases from populations where HIV is generalized and TB prevalence is high to examine the association between HIV and TB clustering. The overall aim of the chapter is to infer a valid estimate of the relative effect of HIV on TB due to recent (re-)infection and reactivation.

1.6.2.3 Chapter 4

In this chapter I will introduce the general methods and database of the Karonga Prevention Study (KPS) a large, longitudinal research project in Northern Malawi. I will also show the results of an analysis aimed to investigate the stability over time of the associations found in Chapter 3.

1.6.2.4 Chapter 5

In this chapter I combine KPS data from population TST surveys done before 1990 and DNA fingerprint data from Individuals experiencing a first TB episode between 1996 and 2008 in Karonga District to explore the association between recent re-infection with Mtb before an individual's first TB episodes and the effect of HIV infection.

1.6.2.5 Chapter 6

In Chapter 7 I will describe analyses of the incidence of TB in Karonga District over time and by HIV and ART status. This chapter describes the data collection and processing required to obtain valid estimates for the denominators used for those incidence analyses; the sizes of the general population by year, the proportion of that population that was HIV positive and the number of people receiving ART in the district.

1.6.2.6 Chapter 7

This chapter aims to estimate TB incidence over time in the general population as well as by HIV and ART status in Karonga District. Also I explore how ART, which was introduced in mid 2005, has affected these trends.

1.6.2.7 Chapter 8

In chapter 8 I will summarise the results of this thesis, discuss the answers to the questions and how these answer can affect public health and future policy decisions. Finally I will outline remaining research questions and future work that follows from this thesis.

1.6.3 Ethical approval

The majority of the analyses done here were secondary analyses of previously collected and anonymised data. All original studies were approved by the Malawi National Ethics Committee and the ethics committee of the UK home institution (LSHTM). For the individual ethics submission approval numbers, please see the publications of the work. Where I initiated a new study I will describe the ethical approval procedures.

1.6.4 Statistical analysis software

All statistical analyses were done using the Stata statistical software package (version 9 or 10) unless otherwise specified.

1.6.5 Publications from work in this thesis

Since the start of my PhD I developed several pieces of work presented in this thesis into papers. As of July 2010, work described in chapters 2, 3, 4 and 5 have been published, are in print or are accepted for publication. A paper based on Chapter 7 is currently under review with a journal, and I am preparing the ART register work described in Chapter 6 for publication.

1.6.6 Personal contribution to data collection

The data used in Chapter 2 and 3 were collected from literature reviews that I executed. Chapter 4 described the data collection methods from the Karonga Prevention Study, where the data from the remaining chapters comes from. Most of the data was collected before I started my PhD, but I spent 1 year (February 2009 – February 2010) on site in Northern Malawi supervising the ongoing data collection (see Chapter 4). While there I also set up new studies to collected specific data I required for my thesis (see Chapter 6).

2 A systematic review and meta-analysis of molecular epidemiological studies of tuberculosis: development of a new tool to aid interpretation

2.1 Summary

In this chapter I explore the variation in the observed proportion clustered between studies of general populations through a systematic literature review of all population–based studies on TB clustering that used IS*6110* based RFLP as the main DNA fingerprinting technique. I will use the resulting dataset to examine the extent to which this variation in observed clustering can be explained by factors describing study design, study setting (the local epidemiology of TB) and study population. Through the analyses I develop a new tool to aid the interpretation of the locally observed proportion clustered considering the study's design and setting.

The work described in this chapter has been adapted into a paper which has since been published. [74]

2.2 Introduction

2.2.1 Background

It has been observed that proportion clustered varies strongly between populations [51, 75, 76]. However, as mentioned in Chapter 1, the proportion clustered that is observed in a population depends on a variety of factors relating to study design as well as the study population and study setting. [59, 77, 78].

This complicates the comparison of results between studies. It is difficult to know to what extent this variation between studies represents true variation in the proportion of TB that is involved in ongoing Mtb transmission, and how much on other factors.

For example, a low proportion of clustered cases in a study population could be interpreted as that showing there is little TB due to recent transmission and therefore good TB control. However, the proportion could merely be an artefact of study design and be concealing a higher true proportion clustered in the population.

2.2.2 Factors that affect the observed proportion clustered

2.2.2.1 Study design

The observed proportion clustered depends on the design of the study that measures it. Modelling and molecular epidemiological studies have shown that who is included in the study can have a major influence on measured clustering. Clustering is underestimated as the proportion of all TB cases in the region that are included (the sampling fraction) decreases [78].

Conventionally the period between a new Mtb infection and what is considered to be TB disease following recent infection TB disease is 5 years. [15, 17] A study measuring the proportion clustered in a population therefore needs to have a study duration sufficiently long to be able to record whether a unique case in year 1 may still cluster with a case in year 5. [51, 59] A number of long term population based studies looked at how the proportion clustered changed if they artificially reduced the follow up time. They showed that the observed proportion clustered increases with study duration up to a plateau after about 4 years [51, 79, 80].

A clearly defined study area is also important, as high levels of migration in and out of a poorly defined area will artificially decrease the proportion of cases found to be clustered, as cases due to recent infection are misclassified as reactivation. This problem is reduced when an area approximates to a complete and relatively isolated population such as a country or district [59, 78].

In general, any feature which means that the study population is not a complete sample of a closed population will tend to underestimate the true proportion clustered.

Over–estimation of the proportion clustered is less likely, although theoretically possible with biased sampling and contact tracing. In areas where there is insufficient variation in strains, or predominance of strains with few bands, identical strains cannot be assumed always to reflect recent transmission. There is also variation between studies in the laboratory techniques used, and in the rigour of the definition of clustered strains [59, 81].

In practice no study can be designed that satisfies all criteria perfectly; some bias will always remain.

2.2.2.2 Study setting

The proportion of TB due to recent transmission in a population depends on the annual risk of infection (ARI), both currently and in the past. A higher ARI represent higher levels of ongoing transmission which should result in a higher proportion of clustered cases. Modelling studies have shown that if the ARI changed over time the age patterns in the proportion clustered get more pronounced. [82] The decrease in ARI seen in most industrialised countries [83, 84] has been associated with a pronounced pattern of a relatively high proportion of clustered cases in the younger population and low proportion of clustered cases in the old in the non-migrant populations in those areas. [49]

2.2.2.3 Study population

Population based studies have confirmed that with age the risk of reactivation TB increases and in older age groups the proportion of all TB cases that is clustered is lower. This effect was both seen in low and high resource settings [69, 85]

Also, as discussed in Chapter 1, HIV prevalence in the TB case population could also be important as the relative effect on the risk of disease following recent or past infection could be different.

For studies in industrialised settings the proportion of TB cases that are foreign born is also relevant. [59]. In these regions immigrants usually account for a substantial part of the TB case population. They are likely to have been infected with Mtb in their country of origin where the annual risk of infection is relatively high [70, 79, 86], and therefore often have unique (non-clustered) strains, adding to the diversity of Mtb strains in the population [48, 87]. However, through socio–demographic factors they could also be at a higher risk of being in a cluster, thus increasing the proportion clustered.

TB in industrialised countries is often focussed in high risk groups such as people who are homeless, abuse alcohol or use Intravenous drugs. Ongoing Mtb transmission in those pockets of society may be higher than in the general population, resulting in higher rates of clustering.

2.2.3 Other reviews on this subject

During the final phases of the work described in this chapter 2 other systematic literature reviews on the subject of IS*6110* RFLP were published. [88, 89] This piece of work was prepared separately.

My approach distinguishes itself from these other publications by examining the variability of clustering on the population level. Rather than collating results from varying study designs and populations on the risk of clustering for individual TB cases, this chapter sets out to take a more public health approach by investigating the proportion clustered in populations while explicitly taking into account the differences in study design and setting. Furthermore, the methods used for the literature search and statistical analysis of one of these publications gave serious reasons for concern, which I formulated in a letter to the editor. [90]

2.3 Methods

2.3.1 Inclusion and Exclusion criteria systematic literature review

Studies on TB clustering that used IS*6110* RFLP as the main DNA fingerprinting technique were eligible for inclusion in the review if they were population–based and reported clustering results (number of strains in RFLP analysis and proportion clustered) for more than 100 individuals. Studies were excluded if the sample population was not representative of the general population as this could bias the proportion clustered. For example, I excluded studies on prison population, drug resistant patients only and outbreak studies.

In addition, the study area had to be suitable for making a valid estimate of local TB clustering. For example, inclusion of patients from a single hospital would be acceptable if all TB patients in the area are likely to go to that hospital, but not otherwise. In practice study areas had to minimally consist of a geographically defined urban or rural district. Overall, these criteria were necessary to reduce bias (which cannot be corrected for) in the dataset, allowing a valid examination of and correction for factors that influence the estimated proportion clustered.

There was no language restriction, and papers were translated by a fluent speaker as required.

2.3.2 Paper collection

Pubmed and Embase databases were searched between 1990 and November 2006. After maximizing the sensitivity of the search by using a private collection of relevant TB clustering papers, the following Pubmed search query was used: ‘*IS*6110 *OR RFLP OR fingerprint* OR* *cluster* OR genotyp* OR "Epidemiology, Molecular"(MeSH) OR "molecular typing" OR transmission OR "molecular epidemiological" OR "molecular epidemiology") AND (tuberculosis OR tb) AND ("1990" : "3000"(PDAT))'*. A similar search query was used for Embase. The TB publications archive from the Centers for Disease Control and Prevention was accessed but yielded no additional papers [91]. After excluding duplicates all titles were scanned twice for possible relevance. The remaining abstracts were read and eligible full length papers were retrieved if possible.

2.3.3 Data extraction

From each paper I collected information on study methods (secondary DNA typing methods, the cut–off points (if present) for Mtb strains with few IS*6110* bands, matching of DNA fingerprints), the number of patients eligible and included in the final DNA fingerprint analysis, TB clustering results, characteristics of the study population (age, proportion males, HIV positive and foreign born) and TB disease (proportion of smear positive, extra pulmonary, or drug resistant TB cases). In addition, I recorded TB incidence in the region. Where more than one paper provided data on the same population and study period only the one with the longest study duration was included. Queries regarding the extracted data were discussed with Professor Judith Glynn.

2.3.3.1 Measures of TB disease due to recent Mtb transmission

The main study outcome was the proportion of clustered TB cases, i.e. number of cases in clusters / total number of cases. The reported proportion clustered was used, although the strictness of the definition differed between papers; most requiring identical fingerprints, but some allowing one band difference. IS*6110* RFLP is not sufficiently discriminatory in strains with low band numbers (usually 5 or less) [59], and if sufficient information was provided we either excluded these strains or recorded the overall proportion clustered when a secondary DNA typing technique (Spoligotyping, Polymorphic GC Sequencing (PGRS), Direct Repeats) was used for these low band number strains. The proportion of TB due to recent transmission was estimated as (number of clustered cases – number of clusters)/number of cases, the "n-1" method [48]. This assumes that each cluster consists of one source case, due to reactivation disease, the rest being due to recent infection.

2.3.3.2 Study Design

Study duration was recorded in months. Cross sectional surveys were assigned a duration of 0 months. If studies only allowed cases to cluster in a certain period after the source case, this "clustering interval" was used as study duration. I recorded the fraction of all culture positive (C+) TB cases (of all types) in the catchment area that had RFLP results available. I recorded whether the RFLP analysis included Mtb strains with low band numbers, and if so, whether a secondary DNA typing technique (e.g. spoligotyping or polymorphic GC sequencing) was applied for these [92, 93].

2.3.3.3 *Study setting*

When possible the TB incidence in the study region as reported in the paper was used. Otherwise the sampling fraction, reported TB incidence, study duration and study population size were combined to estimate the regional TB rate. If that was not possible, the scientific literature was searched for region specific estimates or we used the WHO website for a country wide estimate [20]. Studies were classified into low, middle and high burden TB areas (TB incidence <10, 11–50 and 50+ TB cases /100,000 /year respectively).

2.3.3.4 *Study population*

Studies from industrialized areas (e.g. Western Europe, North America) were grouped together. For these areas I recorded the proportion of TB cases that were foreign born. Where available I also collected data on the proportion of cases with (a history of) homelessness or drug and alcohol abuse.

Age was recorded as the average for the TB case population; either as the reported mean or median age or, if only age strata were reported, through estimation of the median age within the age stratum that held the median observation. Gender was recorded as the proportion of males in the study population. As historical data on the annual risk of infection were not available for the majority of studies, the mean age of the TB case population can be used a proxy, with a declining annual risk of infection shown by a high mean age of TB disease [82].

If at least 50% of all TB cases were systematically tested for HIV, we recorded the proportion HIV positive of those with test results. If the paper itself did not report the variable, an estimate was extracted from papers that reported on the same population.

2.3.4 Statistical analyses

Non–parametric tests (Wilkoxon rank sum) were used to compare studies from industrialized countries with those from other countries.

To assess the extent to which the variation in clustering seen can be explained by study design, study setting (the local epidemiology of TB) and study population I applied meta–regression (metareg command; Stata v10) [94, 95]. This technique allows multivariate regression analysis to ascertain how well individual or a combination of variables can explain the between study variation in the proportion clustered (the 'tau^2') [96, 97].

Meta–regression assumes a linear association between the outcome (proportion clustered) and the independent variable as well as an approximately normal distribution of the residuals. The latter was checked through visual inspection of the residuals (scatterplots and histograms) and statistical tests (sktest in Stata v10). The choice between entering a variable as categorical or linear was dependent on its distribution, known epidemiological associations and the impact on the between study variation. A variable describing the standard error of the outcome (proportion clustered) is required for each study. I calculated this by taking the square root of 'p*(1-p)/N', where 'p' stands for the proportion clustered in a study, and N the total number of study participants.

The effect of the study design and recorded variables was first examined through univariate meta–regression. To allow for potential negative and positive confounding due to high heterogeneity of variable values between studies, all variables were considered for the multivariate models. Final inclusion of a variable in a multivariate model was based on whether adding the variable had a significant impact (>2.5% reduction) on the between study variation.

Four models were created. The main model was limited to variables describing study design and epidemiological setting (the local TB incidence), to ascertain to what extent these could reduce the variation in the proportion clustered.

The second model included variables describing the study population as well. A third model was limited to studies from industrialized countries, and included the proportion of TB cases that were foreign born. These models were designed to test to what extent the observed variation could be explained by known factors, and how much residual, unexplained variation would remain. A fourth model excluded local TB incidence from the main model.

All models were repeated with the proportion of cases due to recent transmission (n-1 method) as the outcome measure to test the robustness of the findings.

2.3.5 Tool to interpret local proportions clustered

The coefficients from the main model were applied to each study to acquire a predicted value for the proportion clustered based solely on the study's design and local TB incidence. These estimates were compared to the observed values ((observed – expected) / expected *100%) to provide a measure of how much the observed proportion clustered differed from its expected value. This correction for study design and setting provides a new perspective on the proportion clustered, and allows for better comparison between studies. Confidence intervals for the relative differences were estimated using the standard error of the predicted values (stdp option in Stata v10).

2.4 Results

2.4.1 Systematic literature review

The primary literature search yielded 11654 records, and after selection (figure 2—1) 46 papers were included for analysis.

2.4.2 Descriptive analyses

The majority of studies (36) were performed in industrialized country settings (North America, Western Europe, Japan or Hong Kong); the others were done in sub–Saharan Africa (n=3), South America (n=4) and in South East Asia, Eastern Europe, and the Middle East (one each).

Figure 2—2a shows a Forest plot for the proportion clustered in all 46 included studies arranged by study location. It shows a large variation in the proportion clustered between studies (Q–test for heterogeneity chi square=6622 (df=46), p <0.0001), from 6% in Northern Bangladesh to 86% in Greenland. Included studies and their recorded variables are listed in table 2.4 at the end of this chapter.

Table 2.1 summarizes the main characteristics of all studies by the region in which they were conducted. It shows that studies were diverse in their design (e.g. study duration between 0 months (cross sectional study) and >10 years) as well as the setting (recorded local TB incidence between 1.7 and 304 cases/100,000/year). TB incidence differed strongly between industrialized settings and other settings (p <0.001), whereas the proportion of HIV positive TB cases did not (p=0.88). Insufficient data were available on homelessness and drug or alcohol abuse so these variables were not included in the analyses.

Figure 2—1 Flow diagram of systematic literature review process

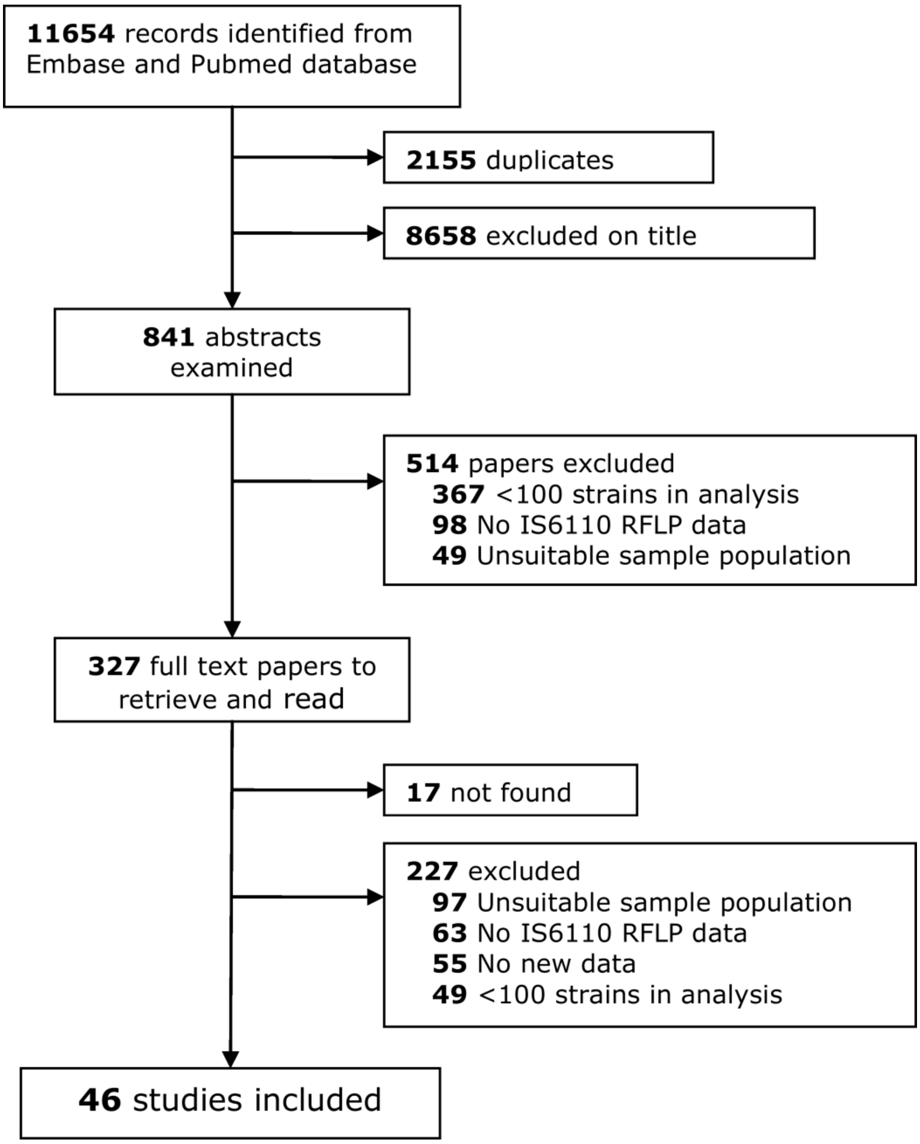
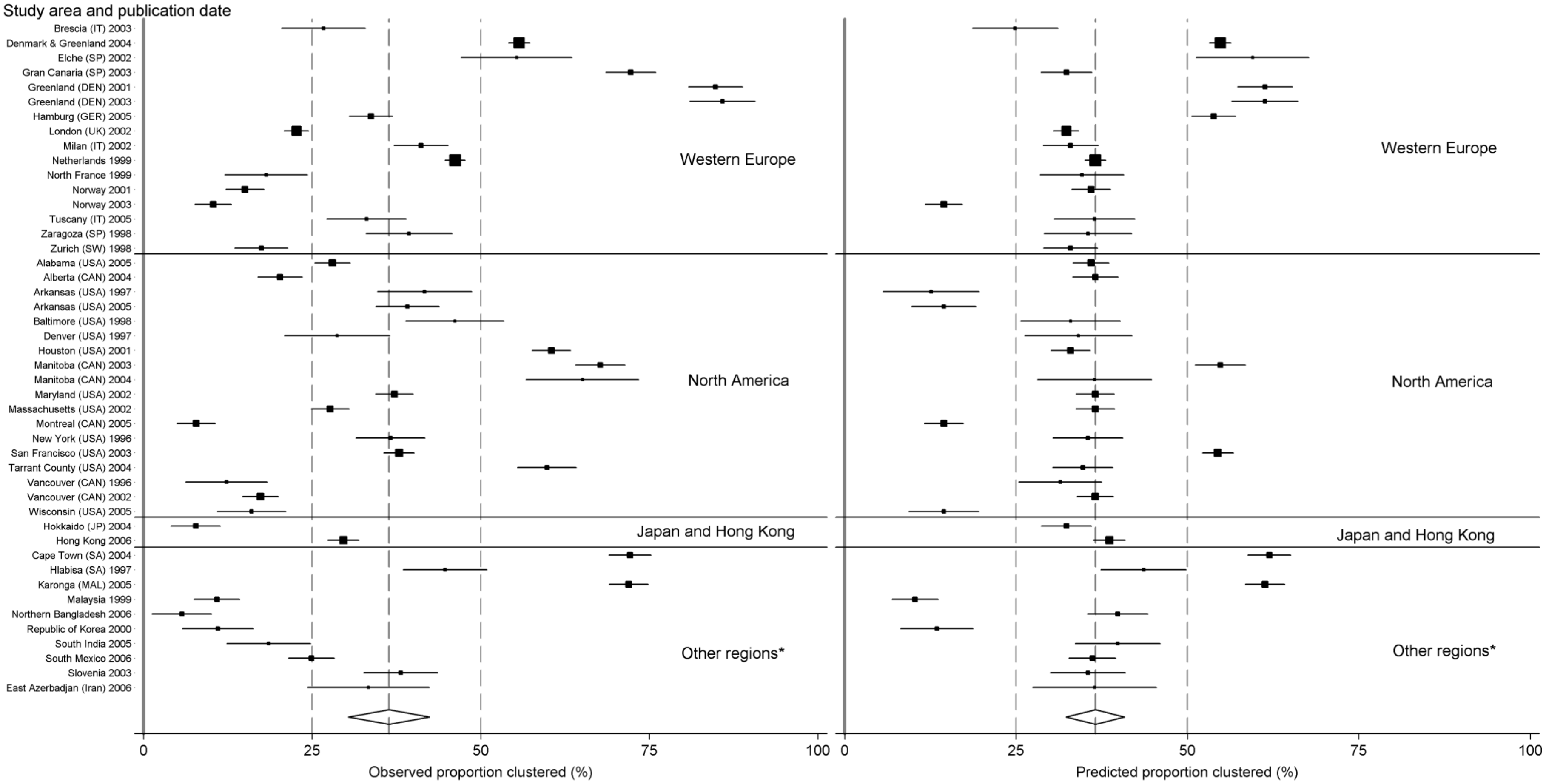

11654 records identified from Embase and Pubmed database
2155 duplicates
8658 excluded on title
841 abstracts examined
514 papers excluded
367 <100 strains in analysis
98 No IS6110 RFLP data
49 Unsuitable sample population
327 full text papers to retrieve and read
17 not found
227 excluded
97 Unsuitable sample population
63 No IS6110 RFLP data
55 No new data
49 <100 strains in analysis
46 studies included

Figure 2—2 Observed (left) and predicted (right) proportions clustered of 46 included studies.



The predicted values for each study were acquired using the coefficients from the meta-regression model 1 (see table 2.2, column 2) that included study duration, sampling fraction, handling of Mtb strains with low band numbers and local TB rate. Table 2.4 (see end of chapter) holds the values for each study, see Table 2.3 for further illustration of the calculations. Box size and error bars indicate number of patients included in the study.
* 'Other regions' refers to sub Saharan Africa, South East Asia, South America, Eastern Europe and Middle East.

Table 2.1: Summary of included studies by study region

Variables[a]	Study region Industrialized countries N=36	 Other countries[b] N=10
Study design		
Study Duration (months)	48 (3 to 120)	13 (0 to 87)
Number of patients	520 (114 to 4266)	372 (105 to 1029)
Sampling fraction[c]	0.90 (0.30 to 1.00) n=36	0.82 (0.002 to 1.00) n=10
Low band numbers included n/N (% studies)	13/36 (36)	6/10 (60)
Secondary DNA typing method used for low band numbers n/N (% studies)	13/36 (36)	3/10 (27)
Study setting		
TB rate in the region (n/100,000/year)	9 (2 to 185)	90 (8 to 761)
Study population		
Average age (years)	45 (30 to 69) n=30	45 (33 to 55) n=9
Sex (% male)	65 (44 to 74) n=28	59 (47 to 76) n=8
% Foreign born	44 (3 to 83) n=28	Not recorded
% HIV positive	16 (1 to 57) n=21	16 (1 to 65) n=4
% Resistant to ≥ 1 drug	10 (0 to 31) n=17	10 (6 to 23) n=5
TB Transmission		
% clustered	35 (8 to 86)	29 (6 to 72)
% recent transmission	25 (4 to 78) n=34	20 (3 to 59)

'N' = total number of studies in TB burden category; 'n' = number of studies used in cell; '%' = Proportion of all study participants.
a Median and range given unless otherwise indicated.
b Studies performed in sub Saharan Africa (n=3) South America (n=4), South East Asia (n=1), Eastern Europe (1), and the Middle East (n=1).
c Sampling fraction is fraction of all culture positive TB patients in the study area with RFLP results available.

2.4.3 Meta–regression analyses

Figure 2—3 shows that the proportion clustered increased with increasing study duration, sampling fraction and TB incidence, decreased with increasing age and proportion foreign born (in industrialized countries) and changed little with increasing study size. In the univariate meta–regression analyses these associations were confirmed (Table 2.2). Some variables reduced the tau^2 by more than 10% (duration of study, handling of low band Mtb strains, age and country of birth of TB case population), whereas others had less or no effect.

Study duration was entered as a categorical variable so its association with clustering could take any shape, including the one shown within populations where clustering increases with study duration, but reaches a plateau after 4 years [51, 78-80].

In the multivariate meta–regression model (table 2.2, model 1) 28% of between study variation was explained by study duration, sampling fraction, handling of strains with low band numbers and local TB incidence. Most coefficients of the included variables were statistically significant at the 0.05 level, and the residuals were approximately normally distributed (p value sktest=0.46). In this model, the proportion clustered increased with study duration and sampling fraction. The model also showed that including strains with a low number of IS*6110* bands increased the proportion clustered, unless secondary typing methods were applied. Additionally, study settings with high TB incidence reported higher proportions clustered.

Incorporating variables describing the study population further reduced the tau^2, explaining up to 60% of the between study variance (table 2.2, model 3). When studies for all countries were considered (table 2.2, model 2), the average age of the TB case population showed a strong negative association with the proportion clustered. In studies from industrialized settings, the proportion clustered decreased as the proportion foreign born increased (p value coefficient <0.001). This negative association was found in 18 out of the 21 studies from industrialized settings that reported the proportion of foreign born TB cases by cluster status, whereas only one study, from Italy, found a positive association [98].

Excluding local TB incidence from the main model (model 4) reduced the explained variation from 28 to 10% as well as the precision of the coefficients. However, the direction and size of the coefficients remained similar.

None of the other variables I recorded had a relevant effect on the tau^2. Similar results were found using the proportion of cases due to recent Mtb transmission (estimated using the n-1 method) as the study outcome, rather than the total proportion clustered (results not shown).

Table 2.2: Meta–regression models: % explained between study variation and coefficients for change in the proportion clustered for variables describing study design, setting and population

	Univariate models[a]	Model 1: Study design and setting	Model 2: Overall	Model 3: Industrialized countries[b]
		(n=46) 28% of variation explained	(n=39) 36% of variation explained	(n=25) 60% of variation explained
Study design				
Study duration (months)	**11.7**	**18.3**	**5.6**	**24.9**
0–12	ref	ref	ref	ref
13–48	-7.3 (-23 to 9)	-3.2 (-20 to 13)	-2.6 (-17 to 12)	21.1 (4 to 38)
>48	11.5 (-5 to 28)	18.3 (2 to 35)	11.1 (-4 to 26)	28.1 (10 – 46)
Sampling fraction (proportion of culture positive cases with RFLP)	**2.2**	**8.9**	**8.7**	**5.6**
0–0.50	ref	ref	ref	ref
0.50–0.75	22.2 (-4 to 48)	27.7 (0 to 55)	30.1 (5 to 55)	29.4 (-2 to 61)
0.75–1	18.9 (-11 to 49)	29.6 (6 to 53)	22.7 (0 to 45)	24.1 (-3 to 51)
Low band strains	**10.8**	**3.9**	–	**16.1**
Excluded	ref	ref		ref
Included with secondary typing	9.1 (-5 to 23)	0.6 (-12 to 13)		3.3 (-9 to 16)
Included, no secondary typing	21.8 (5 to 38)	18.8 (-1 to 39)		21.7 (5 to 38)
Number of patients included	**0**	–	–	–
100–200	ref			
201–500	-0.1 (-17 to 17)			
>500	8.0 (-7 to 24)			
Matching of fingerprints	**0**	–	–	–
Identical	ref			
1 band difference	1.6 (-21 to 18)			
Study setting				
TB incidence in study area	**0.8**	**17.9**	**3.8**	**3.1**
Low (<=10/100,000/year)	ref	ref	ref	ref
Medium (11–50/100,000/year)	3.7 (-11 to 19)	17.9 (2 to 34)	9.8 (-4 to 23)	8.3 (-6 to 22)
High (> 50/100,000/year)	12.2 (4 to 28)	25.4 (9 to 41)	13.6 (-2.5 to 30)	16.4 (-14 to 47)
Study population				
Average age	**27.6** -1.28 (-1.9 to -0.6)	NA	**7.9** -0.83 (-1.6 to 0.04)	**9.3** -0.9 (-1.9 to -0.02)
% foreign born	**17.7** -0.36 (-0.6 to -0.1)	NA	NA	**63.7** -0.54 (-0.8 to -0.3)
Sex (% male)	**0** -18.9 (-102 to 65)	NA	–	–
% HIV positive	**9.5**	NA	–	–
0–10	ref			
10–25	14.8 (-1 to 31)			
>25	12.7 (-3 to 28)			

% Resistant to ≥ 1 drug	**0**	NA	–	–
0–10	ref			
10–20	-10.0 (-34 to 14)			
> 20	-5.9 (-35 to 23)			
Constant (baseline value + (95% CI))		-12 (-40 to 17)	42 (-9 to 92)	46 (-16 to 109)

The bold numbers show the proportion of the heterogeneity explained by each variable, calculated as the absolute reduction in explained variation when the variable is removed from the meta–regression model. Only variables that increased the overall explained variation by at least 2.5% were included in the multivariate models, otherwise a "–" is shown .The coefficients (95% CI) in the multivariate analysis show the difference in proportion clustered between categories (e.g. low versus medium TB burden) or per unit increase (e.g. 1 year of average age) in the variable. The interpretation of the coefficients is further illustrated in Table 2.3

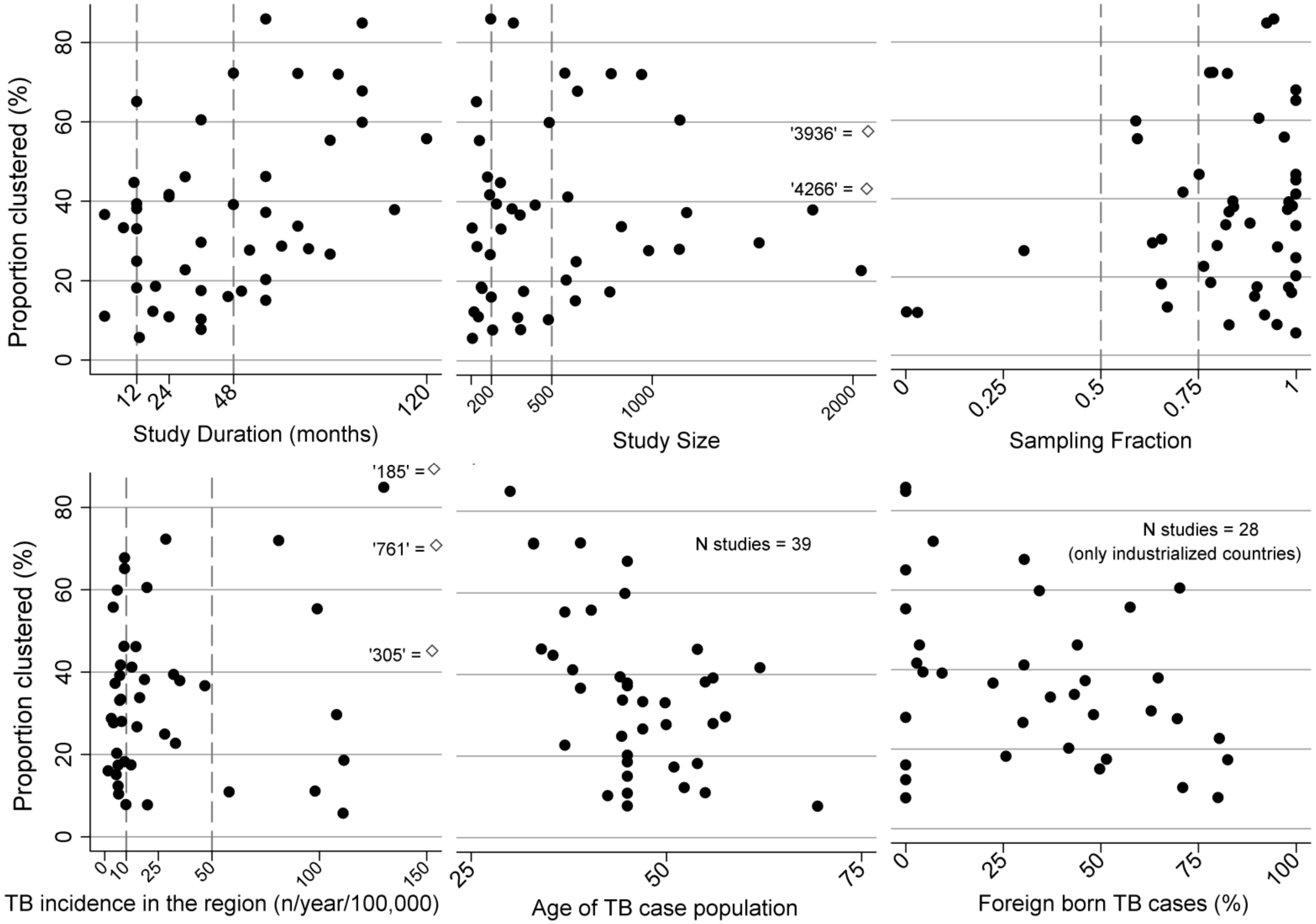
Note: the individual % explained variation per variable do not sum up to the overall % explained variation in the model. This is because of high levels of correlation between some explanatory variables.

NA – not applicable.

a Proportion of explained variation in the univariate analysis.

b Includes studies from Western Europe, North America, Hong Kong and Japan.

Figure 2—3: The association between the proportion clustered and recorded variables

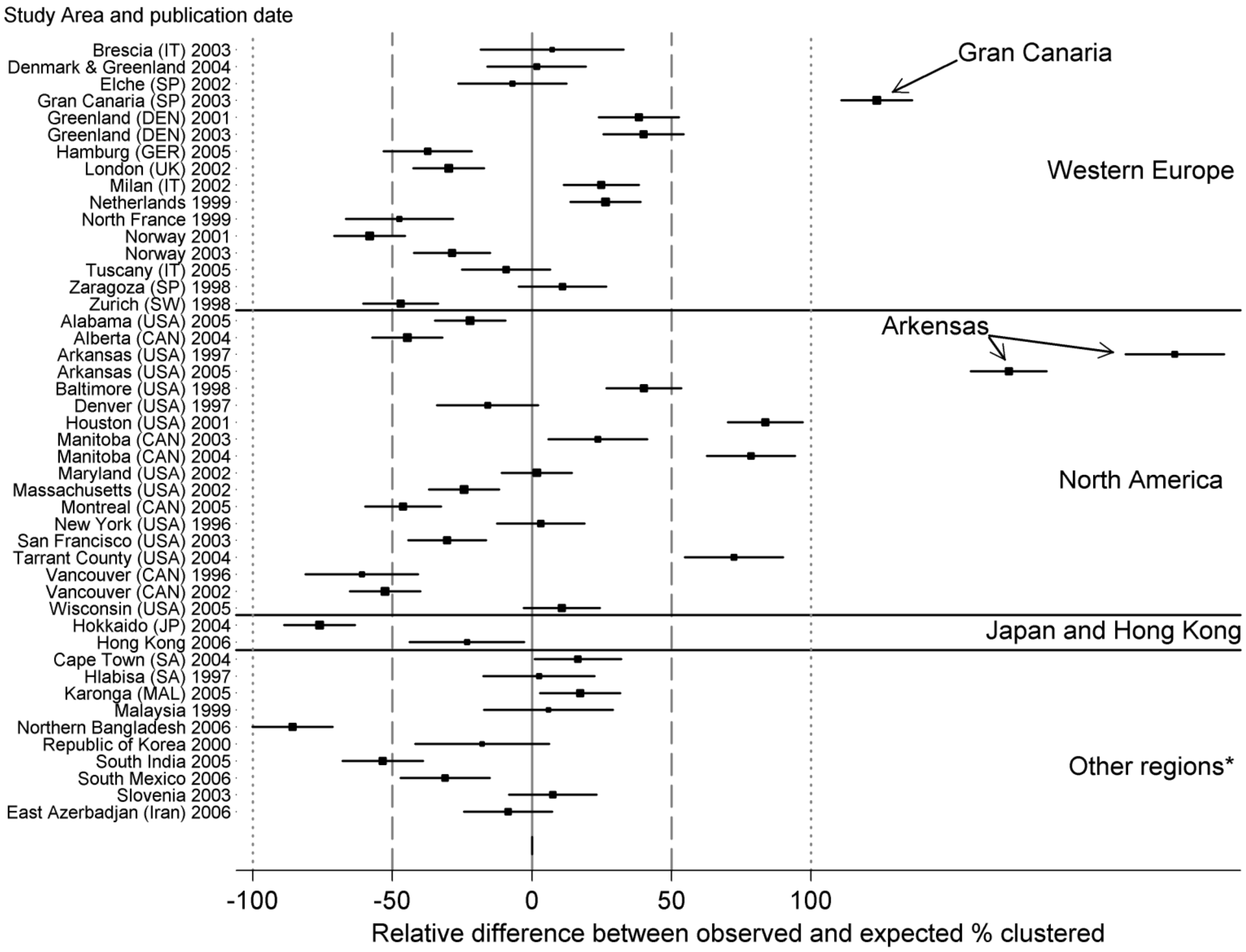


Scatter plots show univariate associations between the proportion clustered and selected recorded variables. Vertical lines show the categories used in the meta-regression. Open diamonds signal outlier values for the recorded variable.

2.4.4 Observed versus expected proportion clustered

For each study the expected proportion clustered could be estimated from the coefficients of the main model. The results are shown in figure 2—2b and the relative difference between the expected and observed values is shown in figure 2—4. Table 2.3 illustrates these calculations. Figure 2—4 illustrates that high observed proportions clustered often lie close to their expected values (e.g. studies from Elche (Spain), Cape Town (South Africa) and Karonga (Malawi) [51, 99, 100]). However, for some studies the levels of reported clustering were twice as high as expected based on study design and setting alone. This applied to regions with apparent moderate as well as higher levels of clustering, for example Arkansas (~40% clustered) and Gran Canaria (72% clustered) [101-103]. On the other hand, studies from Vancouver, Japan and Bangladesh reported proportions clustered half as high as that expected [87, 104, 105].

Figure 2—4: Relative difference (in %) between observed and expected proportion clustered.



*Other regions' refers to sub Saharan Africa, South East Asia, South America, Eastern Europe, Middle East. Error lines indicate 95% confidence intervals of predicted values, calculated using standard error of predictions.

Table 2.3 Calculation of relative difference in proportion clustered

		Study location		
Variable	Coefficient[a]	Hokkaido [106]	Cape Town [100]	Arkensas [101]
Constant[b]	-12	X	X	X
Study duration (months)				
0–12	0			
13–48	-3.2	X		X
>48	18.3		X	
Sampling fraction proportion of culture positive cases included				
0–0.50	0			
0.50–0.75	27.7			
0.75–1	29.6	X	X	X
Low band strains				
Excluded	0	X		X
Included with secondary typing	0.6		X	
Included, no secondary typing	25.4			
TB burden in study area				
Low (<=10/100,000/year)	0			X
Medium (11–50/100,000/year)	17.9	X		
High (> 50/100,000/year)	25.4		X	
Expected proportion clustered		32.3	61.9	14.4
Observed proportion clustered[c]		8	72	42
Relative difference[d]		-75.2%	+16.3%	+191.6%

a Coefficients gives the change in the expected proportion clustered for each category.
b The baseline value of the proportion clustered.
c As reported by the study. See table 2.4 for details.
d Relative difference is calculated as (observed - expected / expected) * 100%.

2.5 Discussion

2.5.1 Systematic review and meta-regression

In this chapter I show that although the proportion clustered varies widely between population based studies, 28% of this variation can be explained by just four variables describing study design and setting. Models including the average age and immigrant status of the TB case population explained up to 60% of the between study variation in industrialized countries.

The residual variation can be due to imprecision of included variables, unmeasured factors, and interactions of the included variables (for which there were insufficient studies to test). With the current level and detail of reporting of clustered studies it seems likely that any explanatory model will have a lot of unknowns.

Although most coefficients of the main model were statistically significant at the 0.05 level, the confidence intervals were wide. This is in part due to the low number of included studies (46–25 depending on the model) and possibly the high heterogeneity in variable values.

The latter is also suggested in model 4. The removal of one explanatory variable (local TB incidence) has a big impact on the unexplained variation between studies as well the precision of the coefficients, without affecting their overall patterns.

This high heterogeneity is one of the main reasons for the exclusion of studies that applied DNA fingerprinting techniques other than IS*6110* RFLP as their main method of strain typing. The number of studies per technique is relatively low, and insufficient to correct for the added variation due to, for example, an unknown level of difference in molecular clocks of the markers used in each technique.

Most associations found in this review are statistically strong and in line with observations made in individual epidemiological and modelling studies, thus giving the meta–regression models additional validity. The importance of study duration in clustering studies has been well documented through modelling [78] and epidemiological studies [51, 79, 80]. The results show that this is also important when comparing between studies. The effect of the sampling fraction has been predicted through modelling [78, 107] and is intuitive, especially when the average cluster size is small: the ‘missing’ part of the population will lead to more clustered cases being classified as unique, thus underestimating the proportion clustered. The expected increases in clustering with local TB incidence and with younger age were also seen [70, 79, 86].

The strong negative association of foreign born TB cases and the proportion clustered, and the fact that 18 out of 21 studies from industrialized countries reported the same statistically significant association on the individual level, both imply that on average foreign born cases have a lower risk of being part of an identified cluster in a study setting.

2.5.2 Observed versus expected clustering

The comparison between observed and expected clustering highlights outliers. More clustering than expected could arise in situations where there is a low number of circulating Mtb strains and little population movement, so that identical strains could reflect transmission many years previously; this could account for the findings from Arkansas [101]. TB outbreaks will increase clustering, which was the case in Gran Canaria where two Mtb strains were involved in 30% of all clustered TB cases [103].

Lower than expected clustering (based on model 1) could reflect an old population with much disease due to reactivation, as in Japan where the average age was 69 years [106]. If we apply model 2, which includes age, to this study the expected clustering is estimated at 18%, which lies much closer to the observed value. The results from the Bangladesh study appear to be due to under sampling; only 111 out of 1264 (9%) notified cases from the region and study period were confirmed by culture and thus potentially included in the clustering analysis [105]. This effect is not included in our model; the sampling fraction was calculated as the fraction of confirmed culture positive cases due to limitations in the reporting of studies.

The comparison of observed and expected clustering also shows the degree to which high observed clustering can be explained: in Malawi, Cape Town and Greenland the high levels of clustering were largely due to the studies' long duration, high sampling fraction and the high local TB rates [51, 100, 108-110].

It would have been preferable to be able to include more studies from high burden countries, but no further studies were available. However, the effect of study design factors is likely to be constant in different regions (as is the case with study duration [51, 79, 80]). A total of 10 included studies were from areas with a high annual TB incidence (>50/100,000O), which should make these results applicable to high burden settings.

2.6 Conclusion

This chapter focussed on the variability of observed proportions clustered in all populations and the extent to which it can be explained by known factors. The methods and associations presented in this chapter can be applied by researchers to acquire a new perspective on the proportion clustered, after adjusting for study design and setting. This will allow a more valid comparison between studies, highlight outliers, and help researchers to assess their local levels of ongoing Mtb transmission.

This chapter aimed to take a broad view on the molecular epidemiology of TB in populations and therefore included studies from all regions. In the next chapter the focus will turn to areas with generalised HIV epidemics [3] and to start examining the association between TB clustering and HIV infection.

Table 2.4 Summary of studies included in meta–regression analysis

Study Location	Study Design						Study Setting & Population					TB trans-mission	
	Study population and DNA fingerprinting methods[a]	Inclusion by IS6110 band number	Secondary DNA typing method (cut–off)[b]	Patients in DNA fingerprint analysis	Duration (months)	Sampling fraction[c]	Local TB incidence (n\year\100,000)	Average age (years)	HIV positives (%)	Sex (% male)	Foreign born (%)	Clustered (%)	Recent transmission (%)
WESTERN EUROPE													
Brescia (Italy) [98]	All c+ cases from Brescia province, '91–'97	All	Spol (<3)	195	84	0.30	15	47	14	67	30	27	16
Denmark & Greenland [111]	Nearly all c+ patients in Denmark & Greenland, '92–'01, 4% of all strains were second or third isolate from same patient	All	.	3936	120	0.97	4	40	.	.	58	56	.
Department Nord (France) [112]	Most (~90%) c+ TB cases in Department Nord, '95[d]	All	.	154	12	0.66	9	54	7	73	26	18	9
Elche (Spain) [99]	Sample of all c+ diagnosed patients in Elche Health district, '93–'99. No information reported about sampling method	>4	.	141	84	0.59	99	37	29	70	.	55	38
Gran Canaria (Spain) [103]	All c+ TB patients in Gran Canaria between '93–'96	>4	.	566	48	0.79	29	39	16	69	7	72	58
Greenland (Denmark) [108]	All identified TB patients from Greenland, '90–'97 15 c+ cases from study region were not notified, and not included in DNA fingerprint analysis	>7[e]	.	310	96	0.93	130	30	.	53	.	85	78
Greenland (Denmark) [109]	All notified patients from Greenland, '98–'02 No data on missed cases, ~60% of notified cases were c+	>7[e]	.	198	60	0.94	185	.	.	.	.	86	
Hamburg (Germany) [113]	All reported c+ TB cases in Hamburg, '97–'02	>4	.	848	72	0.88	16	44	57		43	34	25

London (United Kingdom) [114]	All c+ TB patients in greater London area, July '95–December '96	>4	.	2042	30	0.76	33	37		59	80	23	14
Milan (Italy) [115]	All diagnosed TB cases from Milan metropolitan area residents, '95–'97	All	Spol (<5)	581	24	1.00	13	38	22	63	30	41	28
Netherlands [79]	All reported cases in The Netherlands between '93–'97	All	PGRS (<5)	4266	60	0.75	9	34	4	61	44	46	35
Norway [116]	All diagnosed TB patients in Norway between '1999–'2001	>4	.	485	36	0.92	7	43	0	0	71	10	6
Norway [85]	All diagnosed TB patients in Norway, '94–'98	>4	.	619	60	0.89	5	45	.	54	50	15	11
Tuscany (Italy) [117]	All c+ TB cases in Tuscany, '02[d]	All	.	248	12	1.00	7	50	11	.	37	33	19
Zaragoza (Spain) [118]	All c+ TB patients in Zaragoza in '93 'Nearly all samples' went to 2 participating labs	>4	.	226	12	0.84	32	44	44	69	4	39	27
Zurich (Switzerland) [119]	Patients from the Zurich Canton, '91–'93	All	PGRS (<5)	361	36	.	12	.	10	63	51	17	11
NORTH AMERICA													
Alabama (USA) [120]	All diagnosed TB cases from state of Alabama, '94–'00	>5	.	1136	76	0.80	8	56	6	69	.	28	25
Alberta (Canada) [121]	All c+ TB cases from Alberta province, '94–'98	All	Spol (<6)	573	60	1.00	6	45	.	48	42	20	14
Arkansas (USA) [101]	All c+ cases in Arkensas, '92–'93[d]	>5	.	192	24	0.71	7	62	.	67	3	42	30
Arkansas (USA) [102]	All c+ TB cases in Arkensas, '96–'99	>6	.	419	48	0.98	7	56	.	61	9	39	28
Baltimore (USA) [76]	All c+ cases TB reported in Baltimore City , '94–'96	All	PGRS (<7)	182	30	1.00	15	54	28	69	3	46	32
Denver (USA) [81]	All c+ TB cases from Denver metropolitan area, '88–'94.	>5	.	131	66	0.63	3	.	15	72	48	28	19
Houston (USA) [122]	All reported TB cases in Houston, '95–'98	All	Spol (<5)	1139	36	0.91	20	.	19	70	70	60	53
Manitoba (Canada) [123]	All diagnosed TB cases in Manitoba province, '92–'99[d]	All	.	629	96	1.00	9	45	.	57	30	68	60
Manitoba (Canada)[124]	All diagnosed TB cases Manitoba province, '03	All	.	126	12	1.00	9	.	.	.	.	65	56
Maryland (USA) [125]	All c+ TB cases in Maryland, '96–'00[d]	All	Spol (<7)	1172	60	0.98	5	45	12	44	46	37	28
Massachusetts (USA) [126]	All reported TB cases in Massachusetts July '96–December '00[d]	All	Spol (<7)	983	54	0.95	4	50	27	57	70	28	19
Montreal (Canada) [127]	All reported TB patients in Montreal, '96–'98	>5	.	347	36	0.95	10	40	34	55	80	8	4
New York (USA) [128]	All c+ TB cases in New York in April '91	>4	.	344	0[f]	0.83	47	39	29	74	22	37	28
San Francisco (USA) [129]	All reported TB cases in San Francisco area, '91–'99	All	PGRS (<6)	1800	108	0.84	35	45	20	69	65	38	28

Tarrant County (USA) [130]	All c+ TB patients resident in Tarrant County ‘93–‘00	All	Spol (<7)	488	96	0.59	6	45	.	67	34	60	50
Vancouver (Canada) [104]	All c+ cases in Vancouver area, ‘92–‘94	All	.	114	18	0.67	6	52	.	.	.	12	8
Vancouver (Canada) [87]	All new c+ TB cases in Greater Vancouver area, '95–'99 5 cases were excluded for having experienced a previous TB episode	All	Spol (<6)	791	51	0.98	6	51	5	54	83	17	12
Wisconsin (USA) [131]	All c+ TB cases in Wisconsin, ‘00–‘03	>6	.	200	46	0.99	2	.	.	.	.	16	10
JAPAN AND HONG KONG													
Hokkaido (Japan) [106]	All diagnosed patients in Hokkaido prefecture, '01	>5	.	207	36	0.83	20	69	.	.	.	8	4
Hong Kong [132]	All c+ TB cases with residence on Hong Kong island, May '99– April '02	All	PGRS (<6)	1533	36	0.66	108	58	1	68	63	30	20
OTHER REGIONS													
Cape Town (South Africa) [100]	All patients diagnosed with TB that reported in and are residents of 2 high incidence urban communities of Cape Town. ‘93–‘98	All	Spol (<5)	797	72	0.78	761	33	.	57	.	72	58
Hlabisa (South Africa) [133]	All consecutive SS+ cases in Hlabisa, a rural district in South Africa, May '93 – 'March 94	All	PGRS (<5)	246	11	1.00	305	36	30	62	.	45	29
Karonga (Malawi) [51]	All c+ TB cases in Karonga district, '95–'03	>4	.	948	87	0.82	81	33	65	47	.	72	59
Malaysia [134]	Nationwide random sample of c+TB cases in Malaysia, ‘93–‘94	>4	.	331	24	0.03	58	45	.	.	.	11	6
Republic of Korea [135]	Multistage stratified cluster sample of Korean population	>4	.	136	0[f]	0.002	98	55	.	69	.	11	7
Sunamganj district (Bangladesh) [105]	All SS+ TB cases come from 4 sub districts in Sunamganj district (Northern Bangladesh), November '03 – December–04	>4	.	106	13	0.95	111	.	.	.	.	6	3
Tirruvalur District (India) [136]	All DOTS notified TB cases in Tirruvallur district, ‘99–‘00	>5	.	151	19	0.78	111	45	.	76	.	19	10
Veracruz (Mexico) [137]	All c+ patients in Veracruz state, March ‘95–April ‘03	All	Spol (<6)	623	12[g]	1.00	28	44	2	52	.	25	18
Slovenia [110]	Nearly all (99.7%) of C+ patients in Slovenia, '01	>4	.	304	12	0.99	19	55	1	61	24	38	26
East Azerbadjan (Iran) [138]	All c+ TB cases in East Azerbadjan September '02 – March '03	All	.	105	7	0.82	8	47	.	57	.	33	23

‘c+’ = culture positive; ‘.’ = data not available or found; ‘%’ = percentage of all study participants; ‘Spol’ = Spoligotyping; ‘PGRS’ = polymorphic GC–repetitive sequence typing; ‘no.’ = number; ‘~’ = approximately.

a Sample collection as reported by the authors.
b Secondary DNA fingerprinting method used for clustering analysis, if number of IS6110 RFLP bands is below cut off (shown in parentheses).
c Sampling fraction was based on proportion of all c+ cases that had RFLP results available.
d One band difference in IS6110 RFLP pattern was allowed between clustered strains.
e All typed Mtb strains had >7 IS6110 bands in their RFLP pattern.
f cross sectional survey, duration set to 0 months.
g time difference between first and secondary patients in cluster limited to 12 months.

3 A systematic review and pooled data analysis of population—based studies on HIV infection and TB clustering

3.1 Summary

In this chapter I describe a systematic review and pooled data analysis of studies with information on HIV and TB cluster status of TB cases from populations where HIV is generalized and TB prevalence is high to examine the association between HIV and TB clustering. The overall aim of the chapter is to infer a valid estimate of the relative effect of HIV on TB due to recent (re-)infection and reactivation.

The work described in this chapter has been presented as a poster at the 'Research in Progress' of the Royal Society of Tropical Medicine and Hygiene in 2008, where it won a prize. [139] A paper is currently under review with a peer reviewed journal. [140]

3.2 Introduction

3.2.1 Literature on TB in HIV positive cases

Early studies of TB in HIV positive individuals demonstrated high rates of, presumably, reactivation tuberculosis in HIV–infected patients with latent *M. tuberculosis* (Mtb) infection [141] However, as discussed in Chapter 1 there is evidence from individuals with recurrent episodes of TB, and from age patterns, to suggest that HIV-infection might have a greater effect on (re)infection disease than on reactivation.

3.2.1.1 Studies on recurrent TB

In two studies that measured incidence of relapse (recurrence due to same Mtb strain) and reinfection disease (recurrence due to new Mtb infection) after a first episode of TB in patients with known HIV status, HIV-infection was shown to greatly increase the risk of reinfection disease, but not of relapse [72, 73].

3.2.1.2 Relative risk of TB by HIV and age

After HIV-infection the risk of TB increases over time due to progressive immunosuppression [28]. On average, older individuals will have been HIV-infected for longer so one would expect the effect of HIV on TB risk to increase with age. However, in practice the opposite is found, with lower relative risks in older age groups [142-146]. As younger individuals are more likely to be experiencing their first Mtb infection, and may also be more exposed to Mtb (re)infection, this contradictory age pattern could be explained by HIV mainly increasing the risk of TB disease following first or recent infection.

3.2.2 Relative infectiousness of HIV positive TB cases and clustering

HIV-positive TB cases are, on average, less infectious than HIV-negative cases: their TB episode is more often extra pulmonary or smear negative; and even among smear positive cases they are less likely to be the source of transmission, perhaps because they become ill and seek treatment, or die, earlier [11, 12, 22, 29, 71, 147-150]. This in turn means that HIV-positive TB cases are less likely than HIV-negative cases to be a source case within a cluster.

The next step in the argument is that in each cluster it is assumed there is at least one source case that followed reactivation of a latent Mtb infection and that the other cases presumably followed recent transmission.

Therefore, even if an equal proportion of HIV-positive and HIV-negative cases are clustered, since HIV positive cases are less likely to be source cases, a higher proportion of these cases are likely to be due to recent transmission compared to HIV-negative cases.

So when taking the relative infectiousness, i.e. relative likelihood of being a source case within a cluster of HIV-positive TB cases, into account an equal or higher proportion clustered in the HIV-positive population would suggest that HIV mainly increases the risk of TB disease due to recent Mtb infection.

3.2.3 Study requirements

Considering these arguments molecular epidemiological studies can provide an estimate of the relative effect of HIV on the risk of TB following recent (re-)infection or reactivation. Such studies would require information on HIV and TB cluster status of TB cases from populations where HIV is generalized and TB prevalence is high, so that both HIV-positives and negatives are likely to experience similar exposure to Mtb. Unfortunately, few suitable studies have been done to date.

In this chapter I describe the process of finding those eligible studies, where available collecting the individual patient data from these studies and using appropriate statistical analyses to examine the association between HIV and TB clustering and its consistency across studies. The overall aim of the chapter is to infer a valid estimate of the relative effect of HIV on TB due to recent infection and reactivation.

3.3 Methods

3.3.1 Systematic review

The same literature search as described in chapter 2 was used to find all population-based studies of TB clustering that used IS*6110* Restriction Fragment Length Polymorphism (RFLP) as their main DNA fingerprinting technique. However, for the purpose of this chapter studies had to describe TB clustering by HIV status for a population in which HIV has become generalized (taken as a prevalence consistently over 1% in regular surveillance [3]). As in chapter 2 studies were excluded if the sample was small (<100) or would present a too skewed picture of the population.

Studies only including patients with drug resistance or from a single outbreak were excluded. However, as the goal was to estimate relative clustering between HIV-positive and and HIV-negative TB cases, rather than the absolute proportion clustered in the general population, study populations did not have to be representative of the general population to the same extent as studies included in the analyses of Chapter 2. For example, a study of a gold miner population would not be representative of the proportion clustered in the general population due to the high risk of TB disease. But as long as HIV was generalised so that the assumption that HIV positive and HIV negative cases mix homogeneously is sufficiently reasonable, the study can give a valid indication of the relative clustering between HIV positive and HIV negative TB cases.

See sections 3.4.1 and 3.4.2 for a detailed description of the paper selection process.

Authors of eligible papers were contacted and asked to participate in a collaborative analysis. For each study the individual patient data was requested describing age, sex, HIV status, number of bands in the RFLP fingerprint and whether the TB case was part of a cluster or not.

3.3.2 Variable definition

Within each study, all TB cases whose strain matched that from another TB case within a 5 year period were considered clustered [59]. TB cases were categorized according to their age in years as young adults (15–25), middle aged (26–50) or older (50+ years of age) or in 5 strata (15–25 years, 26–35, 36–45 46–55 and 55+) for a more detailed age-specific analysis. Children <15 years of age were excluded from the analyses as data on this population are often incomplete due to difficulties in the diagnosis.[151] Also, childhood TB cases are much less infectious than adult TB cases and thus have little part in a population's Mtb transmission [151].

As IS*6110* RFLP is insufficiently discriminatory if the number of bands is low (e.g. <5) [59] included studies performed secondary DNA typing (e.g. Spoligotyping or Polymorphic GC sequencing [92, 93]) on strains with <5 bands for further differentiation, or excluded these strains from the clustering analysis.

3.3.3 Statistical analyses

Fixed-effects meta–analysis ('metan' command, Stata v10) was applied to investigate study specific associations between HIV and cluster status. A fixed effect meta–analysis assumes a priori that there is one 'true' or fixed association present in each study. This is opposed to a random effects analysis where the association of interest is expected to vary between studies. More technically speaking, a fixed effect analysis assumes that any variation in the association between studies is due to sampling variation and that there is one true effect, whereas a random effects analysis assumes the effect for each study varies and is drawn from a normal distribution of which the mean represents the true effect. In practice a random effects analysis is often more conservative because of the uncertainty that follows from the heterogeneity in the OR that needs to be adjusted for. [152]

Forest plots were created for the overall association and for each age category. Heterogeneity between studies was assessed visually and using I-squared following guidelines from the Cochrane collaboration [153, 154]. No summary odds ratio (OR) was calculated if more than moderate heterogeneity was present. [155]

For the individual patient data analysis I used random-effects logistic regression (xtlogit command, Stata v10). This is similar to standard logistic regression but allows for within study correlation and between study variation. [156] The association between a patient's cluster status and HIV-infection, age or sex was first tested in univariate analysis. Multivariate models included all variables, unless the Likelihood Ratio (LR) test showed interaction. In that case separate multivariate models were run. Estimates for the proportion clustered by HIV status in each age group, adjusted for study and sex, were estimated from the pooled data by converting the xtlogit coefficients into percentages.

3.3.4 Sensitivity analyses

Analyses were repeated after excluding all strains with <5 IS*6110* bands in their DNA fingerprint. The robustness of the outcomes was further tested by excluding data from each study in turn.

As discussed in detail in chapter 2, several factors (e.g. study duration or population sampling) will affect the reported proportion clustered [59, 74]. To assess the impact of study duration on the results, I re-analysed the data after changing the definition of clustering as 'clustering with a previous isolate within a 12 month period' to compensate for differences between studies.

In small clusters the proportion of clustered cases attributable to ongoing transmission is relatively small compared to larger clusters [59]. I therefore checked whether the effect of HIV on the risk of clustered TB changed if only larger clusters (4 or more cases) were included in the 3 studies that had cluster size data available: Karonga, Namibia and Hlablisa.

3.4 Results

3.4.1 Data collection

The literature search yielded 11654 results, from which seven studies fitted the inclusion criteria [51, 133, 157-161]. For one study the original data could not be traced [158] and it was therefore only included in the overall meta–analysis. Analyses for the still ongoing Karonga Prevention Study from Karonga District, Malawi were updated to include data up to 2006, which constitutes four additional years of data (see Chapter 4 for more details) [51, 162]. Individual patient records were available for 2787 TB cases, 95% (2787/2921) of the total eligible population (table 3.1). HIV results were available for 77% (2141/2787) of these TB cases, of which 25 were excluded either because they were under 15 years old (n=22) or had no information on age (n=3). For one study the number of IS*6110* bands was no longer available for any of the strains [133]. In total 2116 cases were included in the individual patient data analysis.

Table 3.1 Summary of population based studies that reported on HIV and TB cluster status

				HIV-positive			HIV-negative		Study design[a]	
Study	Patient selection	N	% clustered (n/N)	Age median (range)	Sex %male (n/N)	% clustered (n/N)	Age median (range)	Sex %male (n/N)	Strains with <5 IS*6110* bands	Cluster definition
Karonga (Malawi) [51]	All c+ TB cases in Karonga district, '95–'06.	1581	79 (560/706)	34 (3 – 72)	42 (295/706)	72 (278/387)	37 (0 – 83)	52 (201/387)	Excluded	Identical
Gold Mines (South Africa) [159]	All c+ TB cases of pulmonary TB in 4 South African communities of male gold miners, '95.	371	64 (113/177)	37 (24 –62)	100 (177/177)	70 (135/193)	39 (24 – 64)	100 (193/193)	Excluded	Identical
Windhoek (Namibia) [160]	All new patients in hospital serving all TB patients in greater urban area, April '95 – March '96. Patients with TB treatment in previous 6 months were excluded.	263	45 (33/73)	34 (20 –64)	68 (50/73)	45 (63/139)	32 (16 – 72)	75 (104/139)	Spoligo	Identical
Hlablisa (South Africa) [133]	All consecutive SS+ cases in Hlabisa, a rural district in South Africa, May '93 – March '94.	246	49 (32/65)	29 (17–74)	45 (29/65)	44 (68/153)	35 (11 – 80)	70 (107/153)	PGRS	Identical
Addis Ababa (Ethiopia) [157]	Sample of patients with pulmonary TB in Addis Ababa referral hospital (September '96 – August '97). Patients <15 year old and those with TB diagnosis in the last 3 months were excluded.	121	51 (37/72)	30.5 (17 –50)	61 (44/72)	33 (16/49)	25 (15 – 62)	70 (34/49)	Spoligo	Identical
Botswana [161]	Random sample of cases from nationwide surveillance programme (includes 1/10 of all new and 1/5 of all re–treatment cases) in Botswana, January '95 – October '96.	205	44 (28/63)	31 (21 –74)	57 (36/63)	38 (25/65)	36 (14 – 70)	75 (49/65)	Included	1 band difference or identical[b]
Dar es Salaam (Tanzania) [158]	'Randomly selected' sample from TB patients admitted to Muhimbili Medical Centre, a university medical centre in Dar es Salaam, end '92 – mid '93.	134	37 (25/68)	35[c] (18 –61)	N.A.	30 (20/66)	33‡ (14 – 76)	N.A.	Included	Identical
Total		2921	68 (828/1224)	34 (3 – 74)	55 (631/1156)	58 (605/1051)	36 (0 – 83)	70 (688/985)		

'PGRS' = typed with Polymorphic GC Sequencing, 'Spoligo' = Spoligotyped, 'Mtb' = *Mycobacterium tuberculosis*, % = proportion of total, 'n' = number of cases in group with characteristic, 'N' = total number of cases in group, 'N.A.' = not available, 'c+' = culture positive, 'SS+' = sputum smear positive.

a Variables describe how strains with <5 IS*6110* bands were handled and the required similarity of clustered DNA fingerprints.

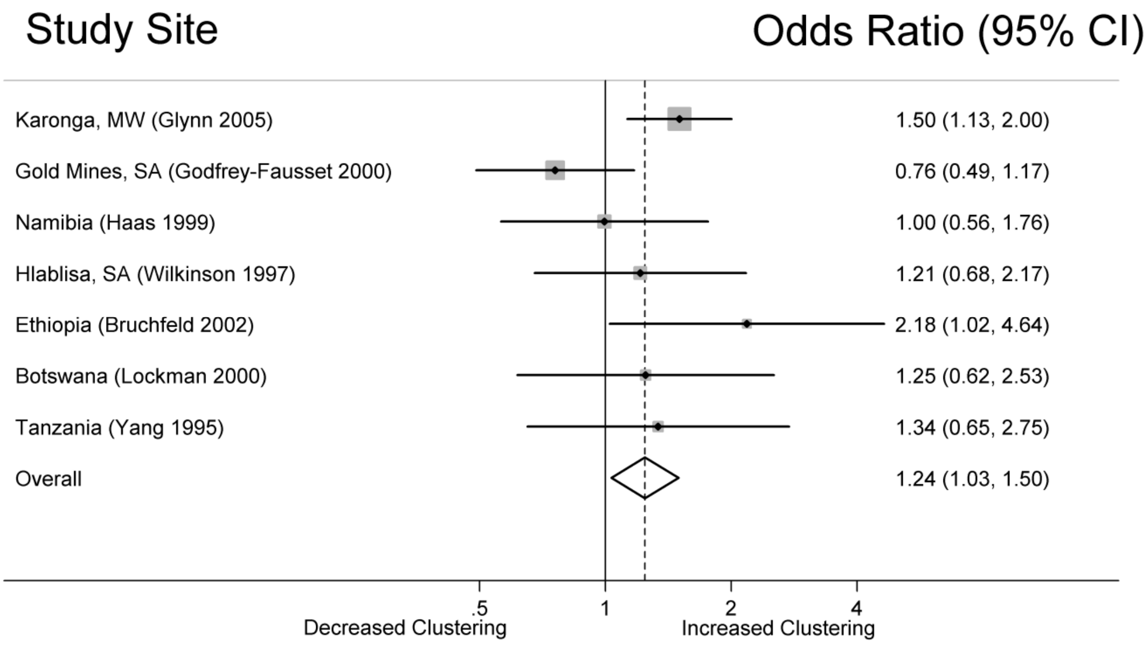
b One band difference allowed in Mtb strains with > 5 IS*6110* bands, other strains had to be identical.

c mean age instead of median age.

3.4.2 Meta–analyses

Five out of 7 included studies found a trend towards a higher proportion clustered in HIV-positives than in HIV-negatives, one found no association and one (the South African Gold Mine study) suggested a trend in the opposite direction (figure 3—1). This is reflected in the combined odds ratio (OR) for HIV and TB clustering (OR = 1.24, 95% CI = 1.03–1.50).

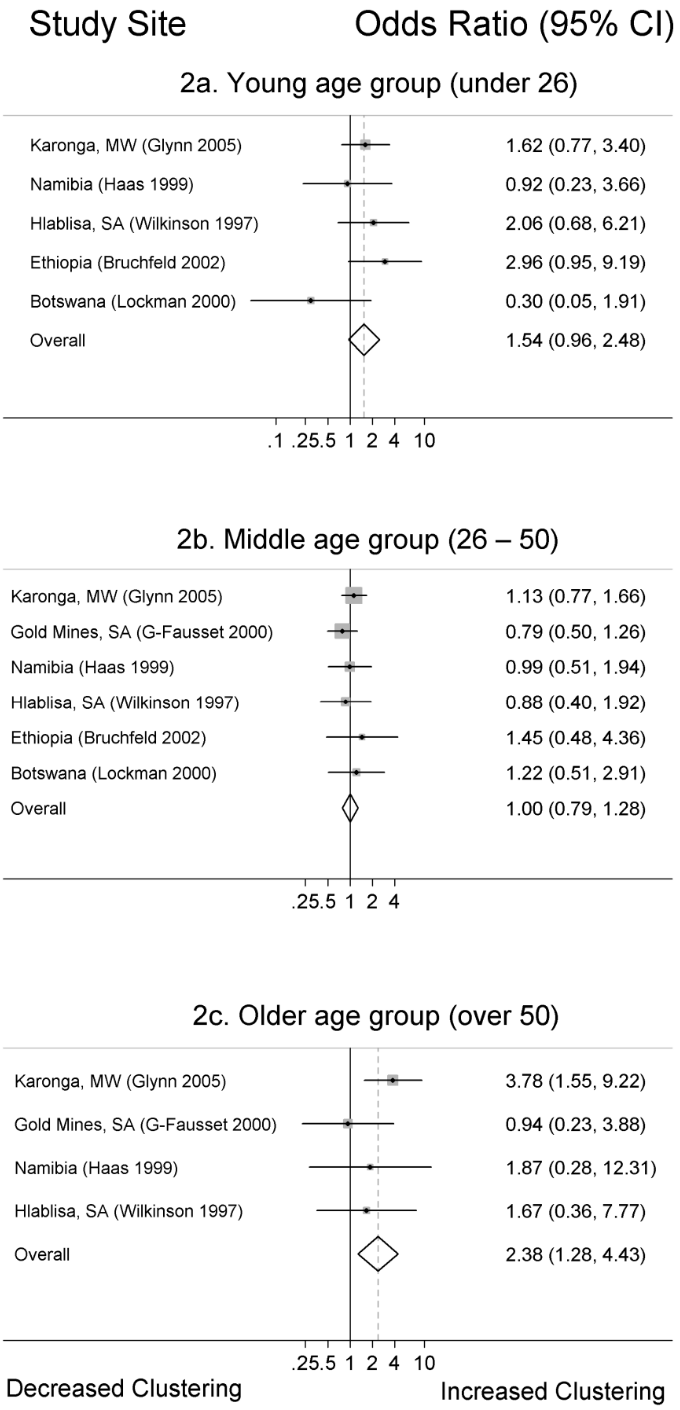
Figure 3—1 Forest plot for HIV and TB cluster status



'MW' = Malawi, 'SA' = South Africa
Note: Heterogeneity between study outcomes was moderate (I^2 = 39%, p–value = 0.13)

The association differed by age. In the young adult population (15–25 years, figure3— 2a) 3 out of 5 studies show a positive association, and only 1 study shows a negative effect (Botswana). This results in an overall positive association (summary OR (95%CI) = 1.62 (1.0–2.6)). In the middle age group (26–50 years, figure 3—2b) HIV has no effect in any of the studies (summary OR (95% CI) = 1.00 (0.8–1.3)) whereas in the older population (>50 years, figure 3—2c) all but 1 of the 4 studies found HIV to be a risk factor for clustering (summary OR (95%CI) = 2.38 (1.3–4.4). In the Botswana study the one older HIV-positive case was clustered, compared to 4 out of 15 (27%) of the HIV-negative TB cases.

Figure 3—2a–c. Forest plots for the association between HIV status and TB clustering, stratified for young adults (15–25 years), middle aged (26–50 years) and older (>50) TB cases



'MW' = Malawi, 'SA' = South Africa
2a: Insufficient data from the gold mines (n=8) to calculate an odds ratio in this age group. In this study 3 out of 5 (60%) HIV-positive cases were part of a cluster, compared to 3 out of 3 of the HIV-negative TB cases. Heterogeneity between study outcomes was low to moderate (I^2 = 22%, p value = 0.3)
2b: No Heterogeneity between study outcomes (I^2 = 0%, p-value = 0.8)
2c: Both the Botswana (n=16) and Ethiopia (n=3) study had insufficient patients in this age category to calculate an odds ratio. In the Botswana study the 1 HIV-positive case was part of a cluster, compared to 4 out of 15 (27%) of the HIV-negative TB cases. Heterogeneity between study outcomes was low (I^2 = 0%, p–value = 0.4).

3.4.3 Individual Patient Data analyses

The univariate random-effects logistic regression (table 3.2) confirmed the results of the meta–analysis. Age and HIV-infection showed a strong interaction (p value LR test = 0.01, DF = 2) so separate multivariate models were run for each age group (table 3.2 column 3–5). These show that clustering is as or more likely in each age group; HIV is a clear risk factor for TB clustering in the older population (OR (95% CI) = 2.57 (1.4–4.7), is positively associated with clustering in the young adult population, albeit less strongly (OR (95%CI) = 1.50 (0.9–2.3)) and in the middle aged population clustering is equally likely in HIV-positive and HIV-negative TB cases (OR (95%CI) = 1.01 (0.8–1.3)).

Within each age group sex had little effect. Additional adjustment for age as a linear variable within each age stratum hardly changed the results in the different age groups (OR (95%CI) = 1.50 (0.9–2.3), 1.00 (0.8–1.3) and 2.57 (1.4–5.7) for the young adult, middle and older age groups respectively).

To investigate the effect of HIV on clustering by age in more detail, the proportions clustered by HIV status were estimated in 5 age categories using coefficients from random-effects logistic regression (figure 3—3a). This shows that in the HIV-negative population the proportion clustered decreases in the older age groups, whereas it increases in the HIV-positive population. The resulting ORs for the 5 age groups are shown in figure 3—3b.

3.4.4 Sensitivity analyses

The exclusion of any individual study did not change the associations found. Figure 3—4a and b show the trends in clustering after either one of the two largest studies (Karonga and Gold Mines) were excluded. Although the difference in the older age groups was less pronounced without the Karonga data (figure 3—4a), clustering remained as or more likely for HIV-positive TB cases in each age group.

Excluding the 341 cases whose Mtb strain had <5 IS*6110* bands or no such information did not affect the outcome. After compensating for study duration (only cases with an identical strain within a 1 year period are considered clustered) the same pattern of equal or higher clustering for HIV- positives persisted, with the OR (95%CI) = 2.10 (1.2–3.7) in the older age category.

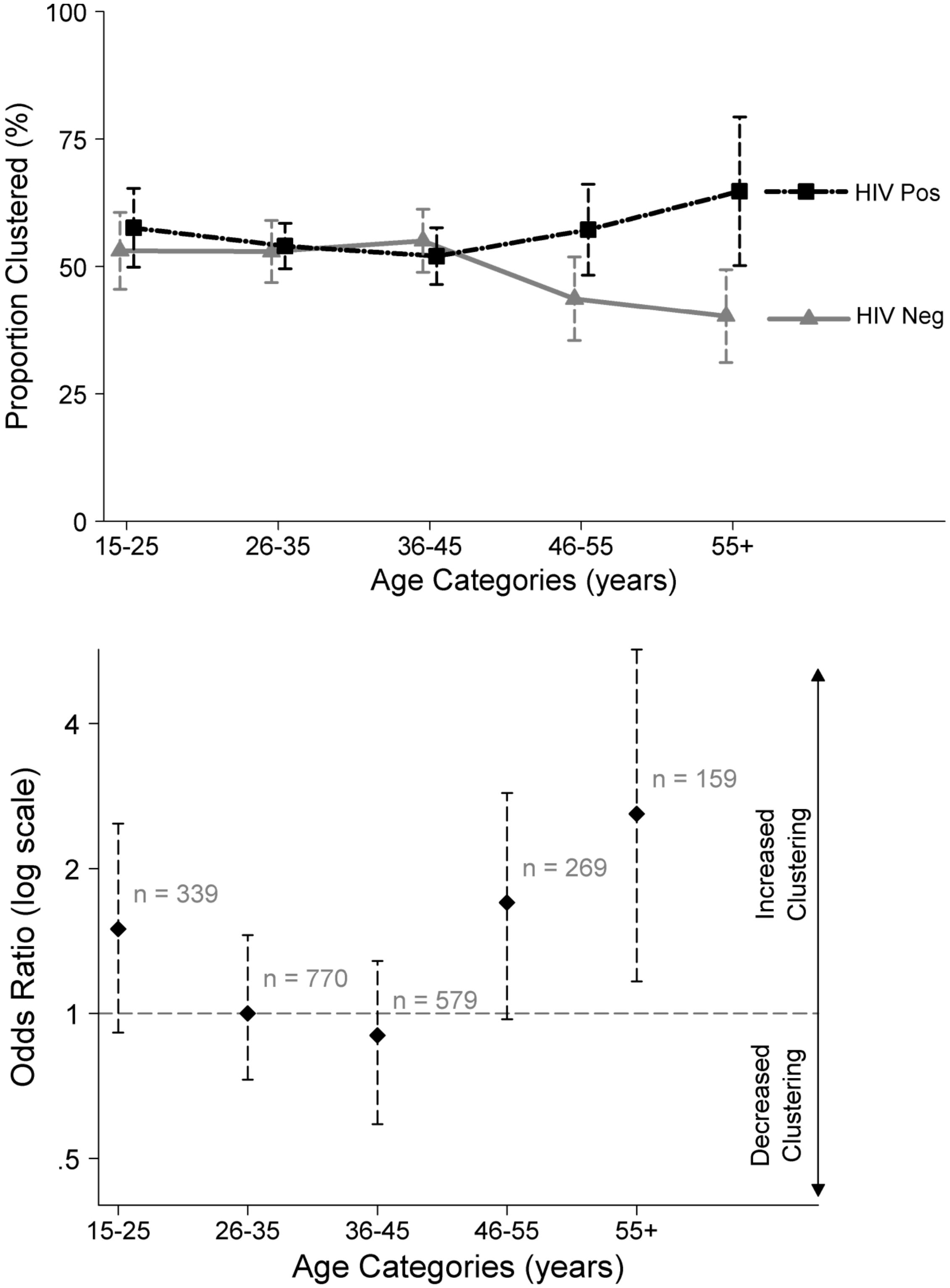
When small clusters of 2–3 cases were excluded from a dataset of the 3 studies with data on cluster size the age-specific effect became more pronounced; the OR in the >50 year old age group for these three studies increased from 2.79 (95%CI = 1.4–5.6) to 3.79 (95%CI = 1.7–8.6), whereas the OR's in the other age groups remained similar.

Table 3.2 Odds Ratio's for being part of cluster

Variable	Univariate OR (95% CI)		Multivariate OR[a] (95% CI)	
	Overall	Young (15–25) n = 339	Middle (26–50) n = 1506	Old (>50) n = 271
HIV status				
Negative	ref	ref	ref	ref
Positive	1.26 (1.0 – 1.5)	1.50 (0.9 – 2.3)	1.00 (0.8 – 1.3)	2.57 (1.4 – 5.7)
Sex				
Female	ref	ref	ref	ref
Male	1.34 (1.1 – 1.6)	1.16 (0.7 – 2.0)	1.19 (0.9 – 1.5)	1.11 (0.6 – 2.1)
Age				
Young	ref			
Middle aged	0.85 (0.7 – 1.1)			
Older	0.68 (0.5 – 0.9)			

Note: Random effects logistic regression was used to adjust for study effect
a Multivariate models included sex and HIV status

Figure 3—3a-b. Study corrected estimates of proportion clustered and OR's in 5 age categories



'n = …' = number of TB cases in each age category

3a. Lines show estimates for the proportion clustered based on coefficients from random-effects logistic regression (converted from log–odds), the error bars represent 2 times the standard error for the proportion clustered: 2 * sqrt(p * (1 - p) / n).

3b. All studies contributed cases to each age category. Estimates are corrected for study effect (acquired using xtlogit command, stata v10) and sex.

Figure 3—4a-b: Study corrected estimates of proportion clustered in 5 age categories after excluding data from Karonga (a) and Gold Mines (b)

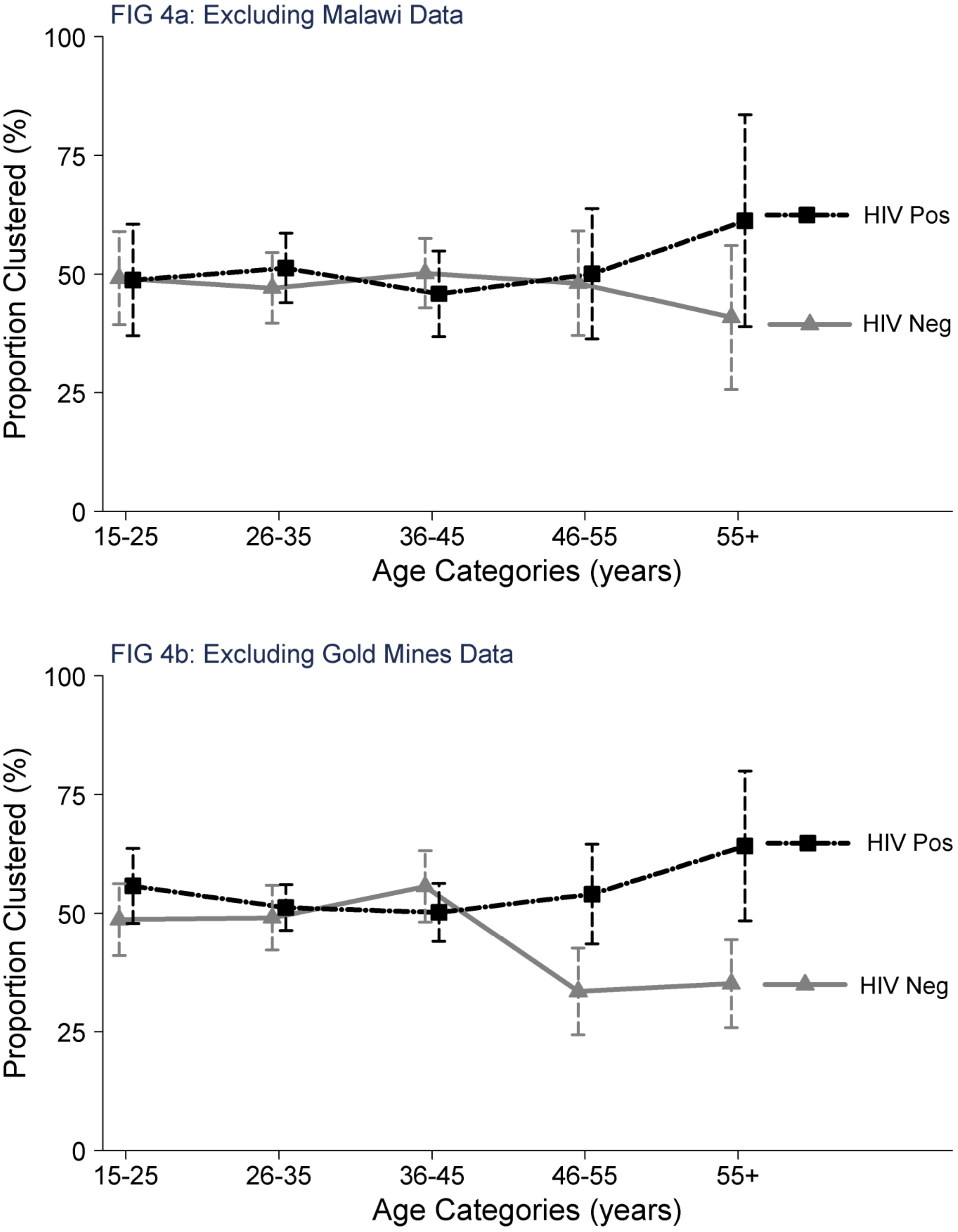


FIG 4a: Excluding Malawi Data

FIG 4b: Excluding Gold Mines Data

4a and 4b. Lines show estimates for the proportion clustered are based on coefficients from random-effects logistic regression (converted from log–odds to proportion) and are corrected for sex. Error bars represent 2 times the standard error for the proportion clustered: 2 * sqrt(p * (1 - p) / n).

3.5 Discussion

3.5.1 Overview

The analyses in this chapter pooled individual patient data for over 2000 TB cases from 6 different populations. The results show that although the association between HIV infection and TB clustering varies across age categories, HIV-positive individuals are consistently as or more likely to be part of a TB cluster as the HIV-negative population. Since HIV-positive TB cases are less likely to be a source of infection than are HIV-negative cases [11, 12, 22, 29, 71, 147-150] this suggests that HIV-infection increases an individual's risk of TB disease due to recent Mtb (re)infection more than through reactivation.

3.5.2 Age specific pattern of HIV infection and TB clustering

The age specific pattern seen here may be the combined effect of two mechanisms. First, in the HIV-negative population, TB clustering decreases with age (see figure 3—3a), which reflects the increasing proportion of the population who have latent tuberculosis in older age groups. This pattern has been shown in multiple studies [51, 69]. Second, in the HIV-positive population it is likely that, on average, older individuals will have been HIV-infected for longer and more immunosuppressed than younger individuals. If HIV-infection has a greater effect on disease following recent infection, this may become more marked with increasing immunosuppression and would thus result in the pattern seen here: increased, rather than decreased, clustering with age in the HIV positive (see figure 3—3a). When combined these two mechanisms would result in the rough J-shape seen in the OR's; little difference in clustering in the young adult and middle aged population and a sharp increase in older individuals (figure 3—3b).

Additional explanations for these patterns could include HIV and age-specific mixing patterns, leading to increased transmission of Mtb strains and clustering within these groups, or other environmental and behavioural factors. For example, HIV-positive and older individuals are likely to spend more time in a hospital setting than their HIV-negative and/or younger counterparts, which could also increase their risk of clustered TB through nosocomial infections.

These results suggest that the emphasis on the influence of HIV on TB following reactivation in early studies may have been misleading [141]. The relative homogeneity of the observed trends across studies in different settings but all with generalized HIV epidemics gives further support to the interpretation conclusions, as does the robustness of the results to the exclusion of the Karonga data, which contributed 51% (1076/2116) of the cases for the individual patient data analysis.

3.5.3 Potential sources of bias

3.5.3.1 Study specific

Some risk of bias remains. The Gold Mine populations experience an extremely high risk of TB disease, partly due to silica dust exposure [163], although this effect is independent of HIV. All the miners were male, but no association with sex was found in the other studies. The Ethiopian population came from a TB referral hospital. It is possible that this attracts patients with complicated TB disease, which is associated with HIV-infection, from a larger population than non–complicated TB cases, which would reduce clustering in the HIV-positive population. However, excluding the data from this study did not change the observed associations.

Anti-retroviral therapy (ART) was rolled out for clinically eligible HIV-positive patients in the Malawi study area from June 2005 onwards [164]. Through subsequent restoration of the immune system it is possible that these patients have a lower risk of TB than HIV-positives not on ART, thus reducing the effect of HIV. However, there was only a 6 month overlap between the study period and the availability of ART in the study area. This makes it unlikely that ART affected these results.

3.5.3.2 Study design issues

Misclassification of clustering is possible, although underestimation is most likely. [59] As discussed in chapter 2, duration, sampling and other factors related to study design and setting can differ between studies and act as an important source of variability of the proportion clustered [59, 74]. Any misclassification is however likely to be non-differential between HIV-infected and non-infected cases and will tend to bring odds ratios closer to one.

However, when misclassification was reduced by removing small clusters, the observed trends became more pronounced. Conversely, when more bias was introduced by limiting cluster duration to 1 year, the associations persisted, albeit less pronounced. Both observations would suggest that our results approach the true association, and probably underestimate it.

3.5.3.3 Biased for culture positive samples

Any clustering analysis is limited to culture positive TB cases. The proportion of culture positive TB cases differs by HIV status which could bias the results. This however requires that culture negative patients are more likely to follow reactivation, i.e. not be part of cluster, compared to culture positive patients. Although the absence of Mtb DNA excludes molecular epidemiological results to support or refute that hypothesis, EPTB is more often culture negative, and more common in HIV positive patients. Within the dataset from the Karonga Prevention Study (see Chapter 4 for a detailed description) I compared 4-year retrospective clustering, which reduces the number of source cases in the analysis, between TB cases with PTB only and EPTB only. The proportion retrospectively clustered was 61% for the PTB cases and 50% in the EPTB cases. So although there could potentially be some bias, the evidence base is small.

3.5.3.4 Assessment of source cases

These analyses extend the well established observation that HIV positive TB cases are on average less likely to be a source of Mtb infection [11, 12, 22, 29, 71, 147-150] to their relative likelihood for being the source case in a cluster compared to HIV negative TB cases. I did not attempt to determine which cases within clusters were likely to be the source case. Most studies were of short duration, so the source cases may not have been included, and the long and variable latent period of TB [17] makes misclassification of source and secondary cases likely, especially in settings with moderate to high rates of ongoing Mtb transmission in the general population.

3.5.4 Implications for public health and patient care

These new insights into the interaction between HIV and TB have implications for policy decisions. Efforts to control TB in settings with generalized HIV epidemics should always be multi-facetted [165], but these results help focus the debate on how control measures should be implemented and prioritised. The higher the proportion of disease attributable to recent transmission, the more important intensified case finding and isoniazid-based prophylaxis of recently (rather than latently) Mtb infected individuals are for TB control.

4 The Karonga Prevention Study

4.1 Summary

After looking at the molecular epidemiology of TB across different populations the remaining chapters of this thesis will use data from the Karonga Prevention Study (KPS). In this chapter I will introduce KPS, its general methods and its data. I will also show the results of an analysis of the KPS data that builds on the results shown in chapter 3. [162]

The majority of the studies described in this chapter were designed, set up and some of them completed before the work on this PhD started in August 2006. My active involvement with the data collection and processing started in February 2009 when I arrived for a year of field work at KPS headquarters in Malawi (see figure 4—1). During my year on site I helped supervise the TB studies described here.

4.2 Introduction

During its 30 years of study activity in Karonga District KPS has performed a wide range of studies, not all of which are relevant for this thesis. I will only describe the methods for the relevant studies most of which were started or even completed before the work on this PhD started in 2006. Where I was actively involved in the data collection of ongoing studies I will outline my contribution.

I also led the data collection in three areas during my time in Malawi; an updated estimate of the population size and distribution in Karonga District, an updated HIV prevalence estimate and a study, which I designed and got funding for, of ART clinic records to quantify the uptake of ART in the district. These pieces of information are described in more detail in Chapter 6.

4.2.1 Setting

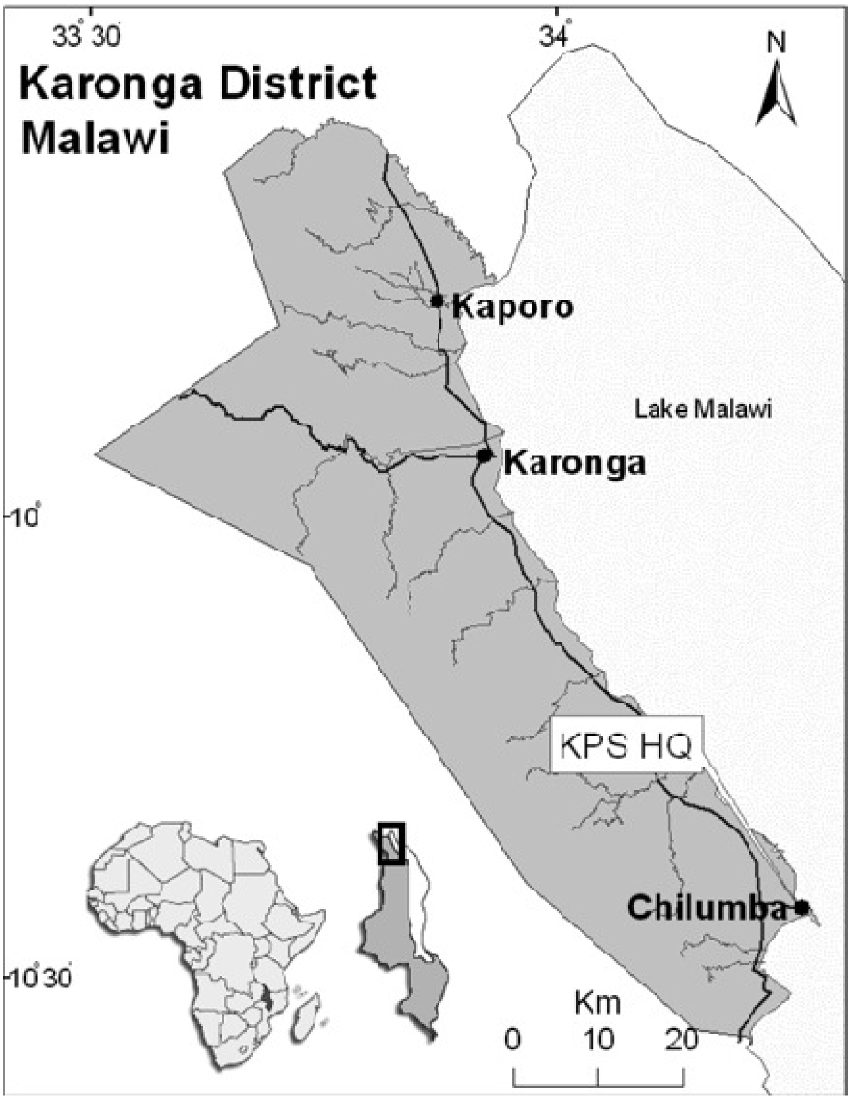
The Karonga Prevention Study is set in Karonga District in Northern Malawi (see figure 4—1). Total area is ~2000 square kilometres. The majority of the population lives along the coastal area, including Karonga Town in the north of the district. KPS headquarters is located in Chilumba, a small harbour town in the south of the District.

4.2.2 Population

The population consists mostly of subsistence farmers, fishermen and small traders. Apart from the main town Karonga (see map) there are few urban areas. Between 1979 and 2009, the population size increased from 112,000 to over 250,000 [166, 167].

Karonga District is relative isolated from the rest of the country and has a low level of population movement. [168] This has facilitated long term follow-up of study participants, ideal for diseases such as leprosy and tuberculosis which can have a long period of latency, incubation or disease. [17]

Figure 4—1 Map of Karonga District



4.2.3 KPS Studies – General

KPS started with large studies on leprosy, with an interest in TB. Currently there are about 15 ongoing studies, ranging from basic sciences such as immunology and to anthropological studies on the uptake of ART.

During two total population surveys carried out between 1979-1984 and 1986-1989 KPS developed a system for identifying individuals who were seen more than once within one study but also across different studies. This system has been used since and the KPS database now holds data on interviews, clinical examinations and biological samples from over 800,000 contacts with ~300,000 individuals. [168] This system allows researchers to link information from the same individual collected across multiple studies and decades.

4.2.4 TB case finding

KPS has strong ties with the TB control programme in Karonga District. From 1985 onwards KPS recorded demographic and clinical information on all diagnosed TB cases in Karonga district, including HIV test results since 1988 [169] Due to the involvement of KPS, Karonga district has had a form of 'enhanced passive case finding' since 1990. [170]

The active involvement of KPS in the District's TB programme has expanded over time. Working in collaboration with the District TB Officer, KPS staff now man all major health facilities where they screen self presenting TB suspects with TB symptoms, including cough and examination of enlarged lymph nodes [171]. Since 1991 participants in all community based KPS studies are asked about chronic cough [171]. Since 2007 KPS has added an annual follow up round in which trained interviewers visit patients who were recorded to have started TB treatment in Karonga District after 1985 and who are still alive and living in the district.

4.2.5 TB diagnostics

From 1985 onwards the KPS laboratory has processed biological samples for TB suspects in Karonga District. Culture isolates from TB cases are sent to the UK for *Mycobacterium* species confirmation, drug susceptibilities and molecular typing.

Sputum smears procedures have been relatively consistent over time. Overall three sputum samples are obtained from TB suspects and individuals starting TB treatment at Karonga District Hospital. For smear positive TB cases review samples are obtained after 2 months and around the end of treatment [5]. Currently, all sputa are examined under a fluorescent microscopy after auramine staining, after which positive samples are confirmed using standard Ziehl Nielsen staining. [172]

Currently all sputa are put on Löwenstein-Jensen medium, regardless of the microscopy result for culture confirmation. Cultures are stored in 37 °C for 10 weeks to allow for slow growing Mtb specimens. Cultures are checked biweekly for growth typical of Mtb.

Since late 1995 at least one sample per patient with a culture growth indicative of Mtb has been sent to the UK for mycobacterium species identification and drug sensitivity testing. Samples are the transported to LSHTM for DNA fingerprinting and strain identification. See section 4.4 for a description of DNA fingerprinting methods and Mtb strain specification.

4.2.6 TB treatment

The TB treatment regimen has followed Malawi government guidelines. Treatment starts with a hospital phase followed by an outpatient phase. During both phases the patients are monitored by KPS staff members.

From 1997 to 2000, individuals with smear-positive TB received 2 months of streptomycin, isoniazid, rifampicin, and pyrazinamide, followed by 6 months of isoniazid and ethambutol. Individuals with smear negative TB received 1 month of streptomycin, isoniazid, and ethambutol, followed by 11 months of isoniazid and ethambutol. The initial phase was in the hospital. Before 1997, thiacetazone was used instead of ethambutol. Currently, both smear-positive and smear-negative patients receive rifampicin, Isoniazid, pyrazinamide and ethambutol during the first 2 months of initial treatment, followed by 6 months continuation phase of rifampicin and Isoniazid. [66, 169]

4.2.7 Interview and HIV test

TB cases are interviewed by a KPS staff member about their disease history, occupation, education, marital status and contacts with known TB cases in the past. Since 1988 all patients starting TB treatment are asked to undergo a HIV test. [173] After informed consent venous blood is taken for HIV testing in the KPS laboratory.

4.3 ART in Karonga District

The national scale up of ART in Malawi started in June 2004 [174] with the first ART clinic in Malawi. The first ART clinic in Karonga District opened in June 2005. [164] Since then over 4500 patients have started on ART in the district (see Chapter 6). In Chapter 6 I will discuss a study I set up and executed with the aim of getting an overview of the uptake of ART in Karonga District over time.

4.4 Mtb DNA fingerprinting and strain identification

4.4.1 Strain identification

Cultures identified as containing Mtb species are selected for DNA fingerprinting using standardised techniques. [52] All DNA fingerprints are scanned and stored electronically for computer assisted visual comparison using the either GelCompar or Bionumerics software (both from Applied Maths).

Strains were considered clustered if they had identical patterns on the RFLP. For the majority of analyses strains with less than 5 bands were not included in the analyses of the molecular epidemiological data.

4.4.2 Strength of KPS molecular typing

The techniques used for DNA fingerprinting of KPS samples have been standardised since 1996. As of December 2009 the database holds 570 different IS*6110* RFLP patterns from 1943 Mtb fingerprinted strains from TB episodes. This database holds a DNA fingerprint for 87% (1723/1991) of culture positive TB cases at KPS between January 1997 and December 2008.

This longevity and relative high coverage of culture positive TB cases from an area with high HIV prevalence make the KPS database unique in the world. Combined with high quality laboratory confirmation and epidemiological description of the (TB case) population stored in the linked KPS database, the Karonga data allow researchers to address questions that no other study can, as I will illustrate in the following chapters.

4.4.3 Missing values

Despite the rigorous and standardised data collection procedures there have been periods where certain parts of the data collection have been less complete, leading to missing data. Missing data is present in all studies, and need not reduce the validity of the results.

In Chapter 7 and Chapter 8 I will address the issue of missing data in more detail and explore the effect of imputing missing values for HIV infection at time of TB episode, cluster status and ART status.

4.5 HIV and the risk of TB due to recent transmission over 12 years in Karonga District, Malawi

4.5.1 Objective

The results from Chapter 3 suggest that HIV mainly increases the risk of TB following recent Mtb infection. In theory this association could be a chance finding, have changed over time or be affected by unmeasured changes in the HIV population (e.g. increase in average time since infection due to a maturing HIV epidemic). In this section I will analyse the KPS data in more detail to investigate the robustness over time of the finding and the stability of the association between HIV and TB clustering found in chapter 3.

The analyses described here were published in the Transactions of the Royal Society of Hygiene and Tropical Medicine. [162]

4.5.2 Methods

I first excluded potential cross contaminations using previously described methods. [66] In this method all TB cases that had only 1 positive sample (either smear or culture) positive were highlighted as Isolated Positive Culture and suspect of cross contamination. The DNA fingerprints of the strains were compared with those from any other positive samples collected or processed (for smear or culture) on the same day. If the strains of an isolated positive culture was identical to that of another sample processed on the same day the likelihood of cross contamination was deemed high, and the case was excluded from further analysis.

For this analysis TB cases were considered clustered if another patient had an identical Mtb strain in the previous 4 years. As discussed in Chapter 1 this retrospective clustering with a fixed time window was used to improve the likelihood that a clustered case was actually due to recent Mtb transmission and allows comparison between time periods which is unbiased by the total duration of the study. [51] The 4 year cut-off was based on a previous KPS study which showed that maximum clustering was reached within 4 years.[51] Cases in the first 4 years were used to determine cluster status of subsequent cases, but were then excluded from the analysis. Cases with less than 5 bands in the RFLP were excluded from the analysis.

For the main statistical analysis two periods were compared; data that were previously reported (October 1999 – March 2003) [51] and new data (April 2003 – October 2007). Cases were stratified into three age groups (15–25, 26–50 and >50 years respectively).

Multivariate logistic regression (Stata v10) was used to calculate age and period stratified odds ratios for the association of HIV with TB clustering, adjusted for sex. Interactions were assessed through likelihood ratio tests.

To test the robustness of the results I explored the impact of restricting clustering to time-windows of either 1 or 2 years [51], or expanding clustering to any case with an identical Mtb strain in the study period, both retrospectively and prospectively. I also used the latter definition of clustering to repeat the analyses including the first 4 years of data, where I compared either 2 6-year periods or 3 4-year periods.

4.5.3 Results

DNA fingerprints were available for 1630/1968 (83%) of all culture positive cases between late 1995 and October 2007. The median age was 35 years (range 17 – 85 years) and 767/1630 (47.1%) were male. Excluding the cases in the first 4 years, 705/1031 (68%) of cases were clustered with a case in the previous 4 years, and 493/746 (66.1%) were HIV-positive.

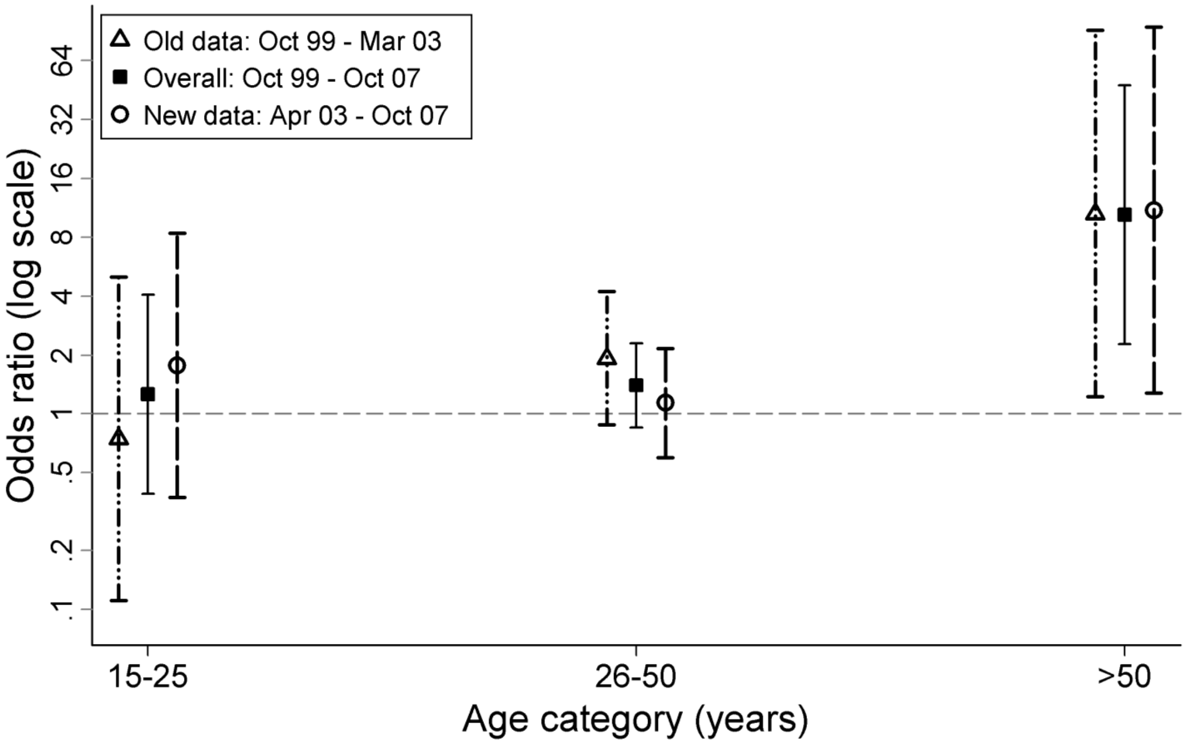
As in chapter 3, an interaction was shown between age and HIV in the overall model (p value LR test =0.05) and both study periods (p value LR tests =0.1). The models (figure 4—2) show that in both time periods and overall the ORs in the young and middle age categories were not statistically different from 1. However, in the older age group HIV infection was associated with increased clustering. This pattern was consistent in all sensitivity analyses.

Overall, in the two periods together, the proportion retrospectively clustered among the HIV-negative was 36/49 (73%), 77/122 (63%) and 41/82 (50%) in age groups 15-25, 26-50 and > 50 respectively. The equivalent numbers for the HIV-positive were 36/45 (80%), 297/411 (72%) and 31/37 (84%).

4.5.4 Discussion

This analysis shows that the association between HIV and TB clustering seen in Chapter 3 is consistent across time periods. What is not yet clear is whether the increase in TB following recent infection applies to patients with and without an established Mtb infection before they become HIV positive.

Figure 4—2: Odds Ratios for TB clustering according to HIV status, by study period



All odds ratios are stratified for age and adjusted for sex. Overall odds ratios are adjusted for study period. Error bars show 95% confidence intervals for the odds ratio

5 The risk of tuberculosis due to recent re-infection in individuals with latent infection and the influence of HIV infection

5.1 Summary

HIV-associated TB can follow reactivation of a latent Mtb infection or recent infection or recent re-infection with Mtb. However little individual patient data is available to show the occurrence of recent re-infection on top of an established Mtb infection as the route to a first episode of TB, or how HIV affects this pathway.

In this chapter I combine data from population TST surveys done before 1990 and DNA fingerprint data from Individuals experiencing a first TB episode between 1996 and 2008 in Karonga District to explore this issue. The results show that HIV-infection strongly increases the risk of TB following recent re-infection in patients with a latent Mtb infection.

The results described in this chapter have been converted into a paper, which is currently in press with the International Journal of Tuberculosis and Lung Disease. [175]

5.2 Introduction

5.2.1 TB following recent reinfection and HIV

As discussed in the previous chapters, the pathogenesis of TB is complex, and further complicated by HIV co-infection. To recapitulate, a first episode of active TB disease can follow a primary infection with *Mycobacterium tuberculosis* (Mtb), reactivation of an established ("latent") Mtb infection, or a subsequent re-infection. [15, 17, 176]

It is known that HIV infection dramatically increases the risk of active TB and that all three pathways to disease are affected. [22] The literature and results from chapters 3 and 4 suggest that HIV-positive individuals are predominantly at risk of TB disease from recently acquired infection. [72, 73, 139, 142-146, 162] Whether this principle applies to the same extent for TB following a recent first Mtb infection and TB following recent re-infection is unknown.

The importance of re-infection disease in TB epidemiology is unclear. Its relevance in areas with moderate to high rates of Mtb transmission has been demonstrated in modelling [15, 17] and elegantly illustrated through population data by Styblo et al [8].

Styblo discussed the case of Eskimo population in north America and Greenland, During the 1950s, these populations experienced an extremely high annual risk of Infection of 25%, which meant everybody was infected in childhood. However, when the ARI fell during the 1960s and 1970s, TB incidence fell simultaneously in all age groups, including those that were already infected with Mtb at least once during their childhood in the 1950s. This showed that recent reinfection with Mtb contributed to the high rates of TB seen in the Eskimo populations during the 1950s

However, to date there are no direct data showing multiple Mtb infections preceding a first TB episode in individuals. It is also unclear whether a prior Mtb infection confers any protection from further Mtb infections or subsequent disease. [8, 177, 178]

5.2.2 Identifying TB following recent reinfection with Mtb

TB research has relied on two main tools to gain insights into Mtb transmission and subsequent TB disease; tuberculin skin tests (TST) to measure Mtb infection [4] and DNA fingerprinting of Mtb strains involved in active TB to distinguish between TB disease due to recent (re-)infection and reactivation. [47]

5.2.2.1 Tuberculin Skin Testing

TST assesses an individual's delayed type hypersensitivity to TB antigens. Although the merits and interpretation of this tool are much debated, a large (i.e. ≥15mm) induration after a standardised Mantoux test is considered strongly indicative of a previous Mtb infection. [4, 179, 180] However, a large TST provides little information on time since infection or the number of (re-)infections.

5.2.2.2 DNA fingerprinting

As discussed in previous chapters, DNA fingerprinting has allowed researchers to distinguish between patients whose TB is likely to have followed a recent infection (or recent reinfection) with a circulating strain of Mtb and those whose TB episode is assumed to have followed reactivation of a prior, latent Mtb infection. For the purposes of this chapter I apply the conventional TB definitions and consider an Mtb infection acquired less than five years before TB disease as being "recent".

5.2.2.3 Combined interpretation of TST and DNA fingerprinting

When the results from DNA fingerprinting and TST are combined for an individual experiencing their first TB episode one can distinguish between the three possible preceding scenarios; a reactivation of a prior, latent Mtb infection as indicated by a unique strain and prior high TST, a progression from recent re-infection as indicated by a clustered strain with prior high TST or a progression from recent primary infection as indicated by a clustered strain with prior low TST.

On average, the DNA fingerprint of the Mtb strain in patients experiencing reactivation disease is less likely to match another in the population compared to those from patients experiencing TB following a recent infection. This should be reflected in a lower proportion of clustered strains in those with a previous Mtb infection compared to those without a previous Mtb infection. By comparing this ratio of clustering in those with and without a prior Mtb infection between HIV-positive and HIV-negative patients, any influence of HIV can be inferred. For example, a weaker association between prior TST result and the proportion of clustered TB in the HIV-positive compared to the HIV-negative population would suggest that HIV particularly increases re-infection disease.

5.2.3 Objective of the chapter

In this chapter I combine TST and DNA fingerprint date from 2 large TST population surveys done by KPS between 1980 and 1989 with the DNA fingerprint database which holds information on TB cases from 1996 onwards. This unique dataset allowed me to study the association of prior Mtb infection status on whether a first TB episode followed recent infection, reinfection or reactivation and examine how HIV affects it.

5.3 Study population and Methods

5.3.1 Study population

Between 1980 and 1989 KPS carried out two large TST population surveys that applied standardised Mantoux testing. Two IU of RT23 (Statens Serum Institut, Copenhagen, Denmark) were injected on the volar surface of the forearm, and indurations were read 2-3 days later. [181] As described before, KPS has performed DNA fingerprinting on culture positive TB cases arising in Karonga district since late 1995 using IS*6110* RFLP. [52, 162]. The last reported TB incidence for Karonga district was ~100 new smear positive cases per 100,000 annually [171].

Patients were included if they had had a DNA fingerprint recorded for their first TB episode and a TST result recorded between 1980 and 1989. Using a case-control approach, I aimed to compare the TST result in cases with TB due to recent infection, and controls with TB due to past infection. IS*6110* RFLP clustering was used to infer likely recent and past infection.

5.3.2 Case definition

Cases were patients with a strain of TB identical to a strain found in another patient in the same population in the previous 5 years. [16] To reduce misclassification, I excluded from analysis the first patient in each cluster of identical strains and all other patients in each cluster diagnosed within the first six months. As IS*6110* RFLP is not sufficiently discriminatory if the number of bands is low [59] strains with <5 bands were excluded from the analysis. Controls were TB patients with unique strains.

5.3.3 Explanatory variables

The risk factor of interest was Mtb infection status before 1990. Restricting the TST results to this period reduced the risk of patients having received Isoniazid Preventive Therapy, which was introduced opportunistically in several later KPS studies for HIV-positive individuals with TST > 5mm. In this analysis Mtb infection is defined as a large TST, i.e. an induration of ≥15mm. Patients with an induration <5mm were considered not to be Mtb infected. Those with an intermediate result (≥5mm and <15mm) were excluded from the main analysis.

Age at first TB episode was grouped in three categories: < 25, 25-49, and 50 or more years old. To correct for the time between an individual's recorded TST and start of first TB episode, this period was divided into three categories of <=10, 11-20 and more than 20 years.

HIV infection status was recorded at time of TST based on anonymised retrospective testing of filter papers [182], and at time of TB episode through testing of serum. [144]

BCG scar status at time of TST was recorded during the TST surveys. [181]

5.3.4 Statistical analysis

I used logistic regression to assess the association between TST history and TB cluster (case-control) status. Age at first TB episode, gap between TST and first TB episode, sex, HIV infection and BCG scar status were investigated as potential confounders. The presence of interaction (e.g. by HIV infection at time of TB episode) was assessed using likelihood ratio (LR) tests. Differences between groups in non-normally distributed continuous variables were analysed using the Wilcoxon rank-sum test. I also compared clustered TB cases with a confirmed prior Mtb infection with all other clustered TB cases from a first TB episode in the KPS database.

5.3.5 Sensitivity analyses

We examined the impact of applying more or less strict criteria to define Mtb infection status based on the TST. Additionally, analyses were repeated including the first cases of each cluster as either unique or as clustered cases. We also explored the effect of excluding all cases diagnosed within the first two years after the index case in each cluster (rather than the first 6 months).

5.4 Results

5.4.1 Patient selection

The selection of cases for analysis is illustrated in figure 5—1. Between October 1995 and September 2008 88% (1707/1946) of all culture positive cases underwent DNA fingerprinting, which resulted in 556 different patterns. After selecting first episodes and cases with a TST recorded between 1980 and 1989, 532 cases were included (see figure 5—2 for the distribution of the induration sizes). Following the strict definitions for clustering and TST result (to ensure distinct groups) left 262 for the main analysis.

Out of these cases, 212 had a TST induration of <5mm and 50 had an induration of ≥15mm, so in total 19% (50/262) of included cases were classified as having an established Mtb infection prior to 1990.

Figure 5—1 Selection of cases and controls for analysis

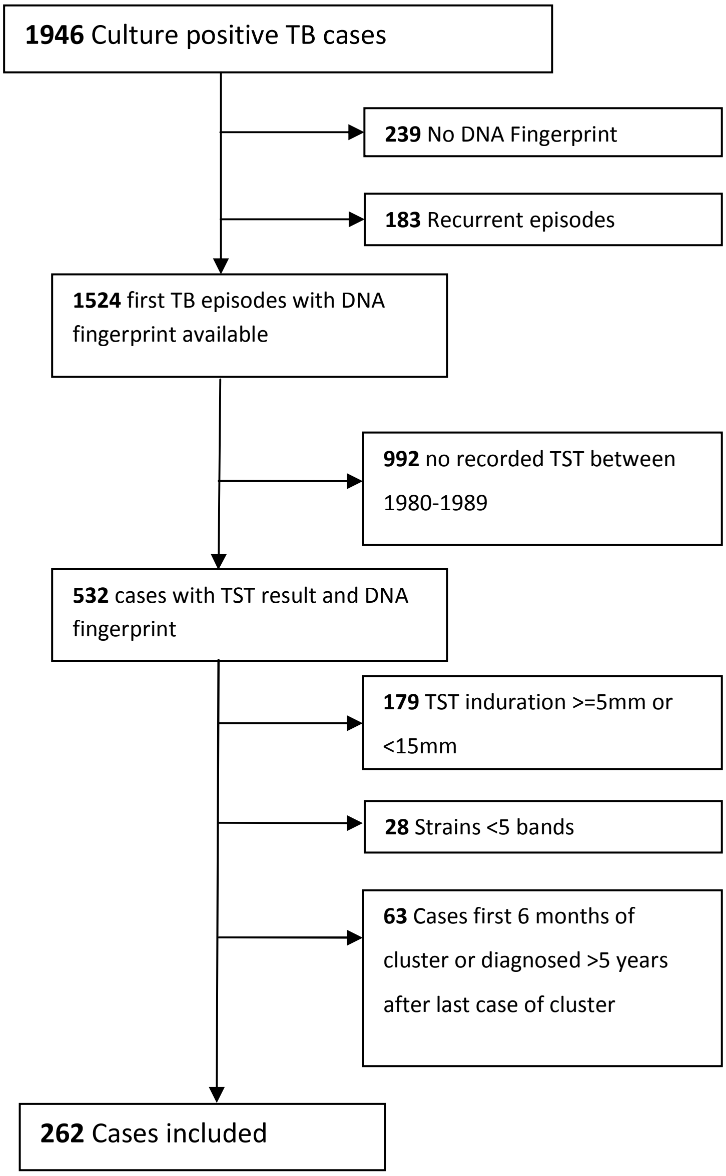
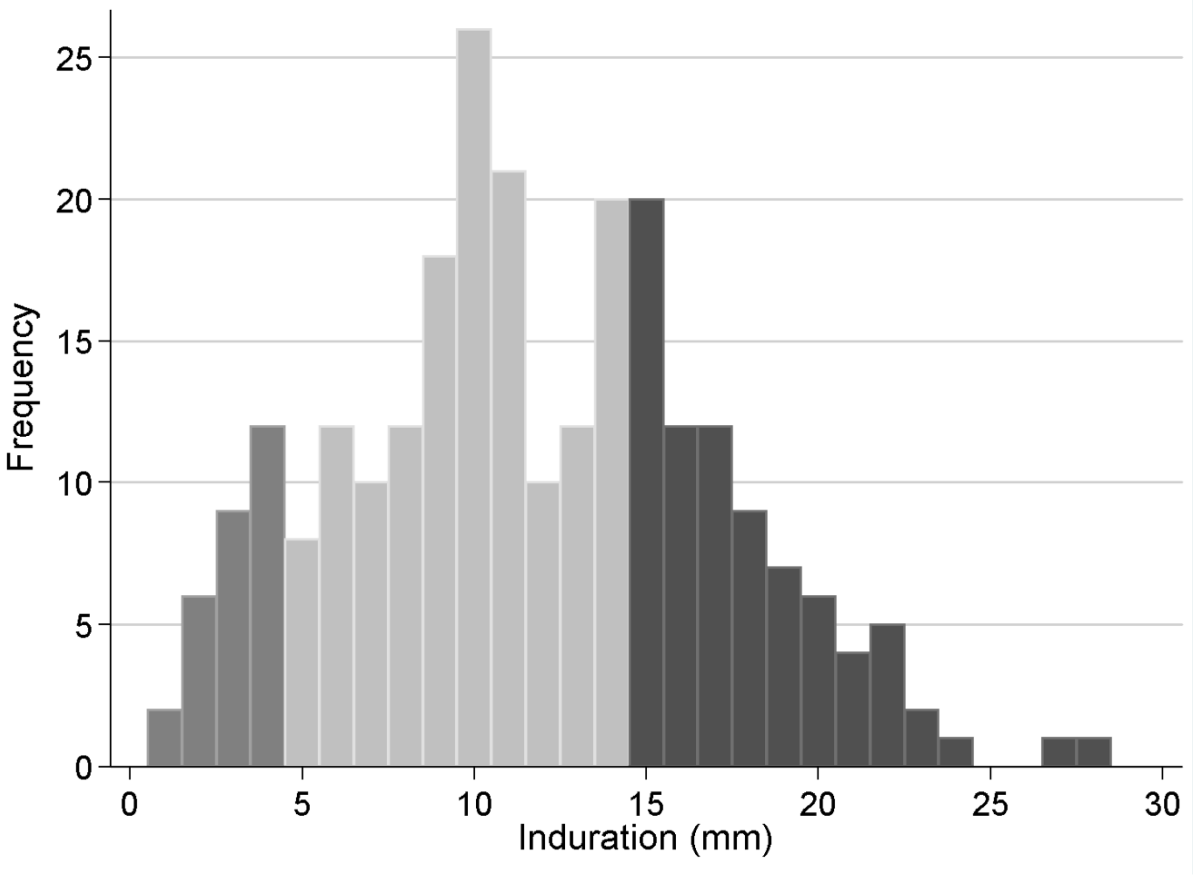

1946 Culture positive TB cases
239 No DNA Fingerprint
183 Recurrent episodes
1524 first TB episodes with DNA fingerprint available
992 no recorded TST between 1980-1989
532 cases with TST result and DNA fingerprint
179 TST induration >=5mm or <15mm
28 Strains <5 bands
63 Cases first 6 months of cluster or diagnosed >5 years after last case of cluster
262 Cases included

Figure 5—2 Histogram of all non-0mm indurations



Note: The 248 zero indurations were included in the statistical analyses. Darker bars on left and right side show TST results that were interpreted as 'no prior Mtb infection' and 'prior Mtb infection' respectively.

HIV test results at the time of TST were negative for all 154 individuals for whom results were available. At the time of the TB episode HIV status was known for 74% (198/262) of individuals of whom 60% were HIV-positive (119/198). Since the overall HIV prevalence in Karonga district was below 1.5% [182] before 1990 and given the long time between the TST and first TB episode (between 8 and 29 years) it is likely that all HIV seroconversions occurred between the TST and the TB episode.

5.4.2 Prior Mtb infection and cluster status

In the overall analysis, the proportion of cases that occurred as part of a cluster was significantly lower among those with a prior Mtb infection (64% (32/50)) compared to those without a prior infection (81% (171/212), OR (95% CI) = 0.43 (0.22–0.83)). After adjustment for key confounding variables the association was weaker (table 5.1, top row).

HIV infection strongly modified the association between a prior Mtb infection and TB clustering (p value LR test for interaction = 0.004 (DF=1)). In individuals who remained HIV-negative the difference between those with and without a prior Mtb infection was pronounced (adjusted OR (95% CI) = 0.15 (0.04–0.59), whereas among those who acquired HIV between their TST and first TB episode the association disappeared (adjusted OR (95%CI)

= 1.85 (0.41–8.29)). For those with unknown HIV status at time of TB episode the OR lay between the OR's of the two other categories, suggesting that there was little bias due to missing HIV status.

Table 5.1 Association between prior Mtb infection status and TB clustering

	Prior TST+ Clustered % (n/N) [b]	Prior TST- Clustered % (n/N)[b]	Crude OR	(95% CI)	Adjusted[a] OR	(95% CI)
All	64 (32/50)	81 (171/212)	0.43	0.22 – 0.83	0.51	0.23 – 1.10
HIV-positive	88 (22/25)	84 (79/94)	1.39	0.37 – 5.24	1.85	0.41 – 8.29
HIV-negative	29 (5/17)	77 (48/62)	0.12	0.04 – 0.40	0.15	0.04 – 0.59
Unknown HIV status	63 (5/8)	79 (44/56)	0.45	0.09 – 2.17	0.24	0.02 – 3.04

Note: HIV was an effect modifier (p value LR test = 0.004 (DF=1))
a adjusted for sex, BCG status at time of TST, time between TST and first TB episode and age at first TB episode
b 'N' is the number of TB cases for that category, 'n' is the number clustered

Among those with no evidence of prior Mtb infection the odds for clustering were similar in the HIV positive and HIV negative (adjusted OR (95%CI) = 1.57 (0.64–3.88), whereas among those with prior Mtb clustering was much more common in the HIV positive (adjusted OR (95%CI) = 74.51 (3.75–1480.95).

5.4.3 TB following recent reinfection

Almost a third of HIV-negative TB cases (5/17) with a prior Mtb infection were part of a cluster despite having an established Mtb infection 11 or more years before their first TB episode. This suggests they experienced a recent re-infection before progressing to active TB disease. In the HIV-positive group, this applied to 88% (22/25) of those with a prior Mtb infection.

Table 5.2 compares the clustered TB cases with a confirmed prior Mtb infection with all remaining clustered TB cases from a first TB episode in the study population. It shows that the number of years a particular strain was found was similar for both groups, regardless of HIV status (all p values for Wilcoxon rank-sum test > 0.12). The 203 clustered cases represented a total of 60 different Mtb strains. There was no evidence that HIV-positive cases with a prior infection were preferentially found in large clusters.

Table 5.2. Comparison of cluster characteristics for all clustered TB cases compared to those with confirmed prior Mtb infection

	N	Time between first and last case of cluster (years) Median (IQR)	p value[b]	Size of Cluster Median (IQR)	p value[b]
All clustered cases[a]	888	10 (7 – 12)	0.98	16 (7 – 27)	0.70
Prior Mtb infection	32	10 (7 – 12)		14 (9 – 27)	
All HIV-negative clustered cases[a]	224	10 (7 – 12)	0.32	15 (6 – 26)	0.35
Prior Mtb infection	5	11 (10 – 12)		17 (13 – 23)	
All HIV-positive clustered cases[a]	433	10 (7 – 12)	0.86	16.5 (8 – 27)	0.33
Prior Mtb infection	22	11 (7 – 12)		17 (10 – 41)	
All clustered cases with unknown HIV status[a]	231	10 (7 – 12)	0.12	16 (7 – 27)	0,06
Prior Mtb infection	5	8 (7 – 9)		7 (5 – 13)	

Note: clustering defined as in main analysis; excluding the first patient and all other patients from the first six months to reduce misclassification.

a Excluding cases with a confirmed prior Mtb infection

b p-value for two-sample Wilcoxon rank-sum test

5.4.4 Sensitivity analyses

Table 5.3 shows the results from the sensitivity analyses. Changing the minimum and maximum values for large and small TST to more or less extreme values, or changing the definitions for clustering did not change the observed trends.

Table 5.3. Association between prior Mtb infection status and clustering in sensitivity analyses

No	Sensitivity analysis	All OR (95% CI)	HIV-negative OR (95% CI)	HIV-positive OR (95% CI)	Unknown HIV status OR (95% CI)
1	Higher minimum value for large TST (=>17mm induration)	0.46 (0.19 – 1.10)	0.09 (0.01 – 0.56)	1.81 (0.31 – 10.6)	0.09 (0.04 – 2.35)
2	Lower maximum value for small TST (0 mm induration).	0.49 (0.22 – 1.11)	0.10 (0.02 – 0.48)	2.15 (0.46 – 10.1)	0.30 (0.02 – 3.87)
1 + 2	More extreme values for large and small TST	0.43 (0.17 – 1.08)	0.07 (0.01 – 0.46)	2.05 (0.33 – 12.6)	0.12 (0.05 – 3.13)
3	TST cut-off at 15mm: =>15mm for large, <15mm for small	0.56 (0.26 – 1.18)	0.11 (0.03 – 0.48)	2.55 (0.48 – 13.49)	0.56 (0.10 – 3.15)
4	All excluded cases classified as clustered	0.68 (0.33 – 1.41)	0.28 (0.09 – 0.89)	1.57 (0.36 – 6.86)	0.83 (0.13 – 5.21)
5	All excluded cases classified as unique	0.63 (0.40 – 0.97)	0.24 (0.09 – 0.66)	1.15 (0.63 – 2.13)	0.43 (0.15 – 1.17)
6	Exclude all cases from first 2 years of cluster	0.51 (0.22 – 1.17)	0.17 (0.04 – 0.75)	1.75 (0.37 – 8.36)	0.29 (0.02 – 3.44)

Note: Table examines the impact of different criteria for Mtb infection status based on the TST induration and different definitions for cluster status. Analyses were conducted including all cases and by HIV infection status at time of first TB episode. All OR's were adjusted for sex, BCG status at time of TST, time between TST and first TB episode and age at first TB episode

5.5 Discussion

5.5.1 Overall interpretation

This chapter describes the association between the long term Mtb infection history of an individual in relation to pathogenesis of their first TB episode. The results showed that HIV-negative TB cases with evidence of Mtb infection many years before their first TB episode were less likely to have currently circulating strains of Mtb than were cases with no evidence of long term prior infection. This was expected, reflecting an increased proportion of re-activation disease. In contrast, in the HIV-positive population the proportion of cases with currently circulating strains was not influenced by prior Mtb infection status.

These results suggest that HIV-associated TB is mainly due to TB following a recent infection or reinfection and give supportive empirical evidence for recent re-infection being important in disease burden for both HIV-negative and HIV-positive groups in a setting with a generalised HIV epidemic and moderate to high levels of ongoing Mtb transmission.

5.5.2 Potential limitations

5.5.2.1 Internal and external validity

These conclusions are based on the combined interpretation of two measurement tools, TST and DNA fingerprinting, and are therefore vulnerable to the assumptions that underlie their interpretation. However, the tools gave expected findings given established understanding of TB pathogenesis in the immunocompetent, i.e. HIV-negative population. [8, 15, 17] As HIV status at time of first TB episode will not affect the result of a TST done 10 or more years before or change the result for a DNA fingerprint, the same credibility should extend to the results in the HIV-positive group, which in itself fit with other results. [72, 73, 139, 142-146, 162] The analyses in the HIV-negatives thus acted as an internal control for the HIV-positive results. The consistency of the results in the sensitivity analyses suggests that they are robust despite the relatively small numbers available for analysis.

The TST can be false positive, due to BCG or infection with environmental non-tuberculous mycobacteria, or false negative due to anergy following HIV infection. [4] We chose conservative values when interpreting TST indurations to reduce the possibility of misclassification and showed that the associations held if more extreme and less extreme values were chosen. The low HIV prevalence and all negative HIV tests at the time of TST make it unlikely that our TST results were influenced by anergy, further strengthening the distinction between those with and those without a prior Mtb infection.

5.5.2.2 Interpretation of TST and DNA fingerprinting results

An important limitation lies in the interpretation of DNA fingerprint results, specifically in interpreting TB disease with a shared Mtb strain as indicating recent infection [59]. The observed proportion clustered using the current definition is ~80%, which is high compared to other settings [128, 132, 133] but consistent with expectations considering study setting, completeness of the sample and the duration of follow up as described in Chapter 2. [74] It is possible that the strains occurring in apparent recent re-infection cases were circulating at the time when these individuals became infected initially, i.e. before 1980 – 1989. Reactivation of these old infections while the strain continued to circulate in the population would lead to them being seen as clustered and hence misclassified as re-infections. [101] While this makes it difficult to draw conclusions about the likelihood of re-infection disease in individual cases, it does not explain the very clear difference seen between the HIV-positive and HIV-negative patients.

Exposure to and (re-) infection with Mtb are relatively rare events. Despite the case-control rather than cohort study approach, these results seem to suggest that in the HIV-positive population the risk of progression following a recent infection is greater than the risk of progression from a prior infection, even in individuals already known to be infected.

5.5.3 Implications and conclusions

These results give further support for WHO's "three I's" policy for TB control in the presence of a generalised HIV epidemic [165]; Intensified case finding in the community to reduce the number and duration of undiagnosed infectious TB disease [183], infection control in hospitals [184] and Isoniazid prophylaxis for all HIV-positive individuals at risk of Mtb infection, not just those with latent infections. [185]

These analyses confirm and extend results from previous studies on recurrent TB and TB in populations and the work described in chapter 3 and 4. Together they point towards an apparent dominance of recent Mtb infection, regardless of previous Mtb infection status, as the cause of TB disease in HIV-positive individuals in settings with moderate to high rates of ongoing Mtb transmission. [72, 73, 139, 162]

6 Estimation of total population size, HIV prevalence and ART uptake in Karonga District

6.1 Summary

This chapter describes the data collection and processing of estimates for the total population size, HIV prevalence and ART uptake in Karonga District, which are needed in order to estimate TB incidence by HIV and ART status in Chapter 7.

6.2 Introduction

6.2.1 Background

The work so far has applied a case control approach to the question of TB molecular epidemiology in settings with generalised HIV epidemics and moderate to high rates of Mtb transmission. Also, thus far I have also not explored the effect of ART on the epidemiology of TB in these settings.

In the coming chapters I will focus on the incidence of TB in different HIV/ART subpopulations (Chapter 7).

This chapter describes the data collection and processing of estimates for the denominators used in the coming incidence chapters; the sizes of the general population, the population HIV positive and HIV negative and finally, the population receiving ART in the district. For this purpose I collected the following pieces of information while working at KPS headquarters in Chilumba between February 2009 and February 2010:

1. An up to date estimate of the population by age and sex as well as their distribution within Karonga district
2. An estimate of the proportion of this population that was HIV positive at certain periods and comparing those to existing estimates.
3. The number of people receiving ART and the distribution of time they spent on ART.

These will allow me to estimate the person years at risk in the different age, sex, HIV and ART categories for different periods.

6.3 Population size

I required estimates of the Karonga population for two purposes. Firstly, to generate a standard population for age, sex and area standardisation of the HIV prevalence estimations found in studies that recorded HIV status from a random sample of the population (see table 6.4 for details of the studies) and secondly, to estimate the person years denominator for the incidence analyses.

6.3.1 Standard population

The purpose of the standard population is to adjust a crude HIV prevalence from study that collected HIV results but potentially oversampled high risk populations, e.g. randomly selected controls from the general population that were age, sex and area matched to TB cases (see table 6.4). Using one standard population for each study I adjusted for potential imbalances from three factors: age group, sex and area of district. That the risk of being HIV positive is different by age and sex is well known and has been shown in the Karonga population. [186, 187] The standardisation by area aims to compensate for the differences in HIV risk and prevalence patterns between the more urban areas (e.g. Karonga Town) and the other areas in the district. [187]

The National Statistics Office (NSO) of Malawi performs a full population census every ten years, most recently in 2008. KPS has used information from the 1988 (area distribution) and 1998 (age and sex distribution) to standardise HIV prevalences. [186] To update these standardisation files I acquired the individual level record for Karonga from the 1998 census to get the area distribution by age and sex. Unfortunately the 2008 census results were not yet publicly available in required detail at the time of these analyses (March 2010).

To convert the 1998 census data into proportions of the population by area I used the expertise available from local KPS field staff, who converted the census enumeration areas into KPS village codes. These were subsequently then divided into one of the six areas described in [187] and section 6.3.1.2 (below).

6.3.1.1 Age structure

Age groups were coded as 15-24, 25-34, 35-44, 45-54, 55+ to allow comparison with models of HIV prevalence and incidence from White et al [186].

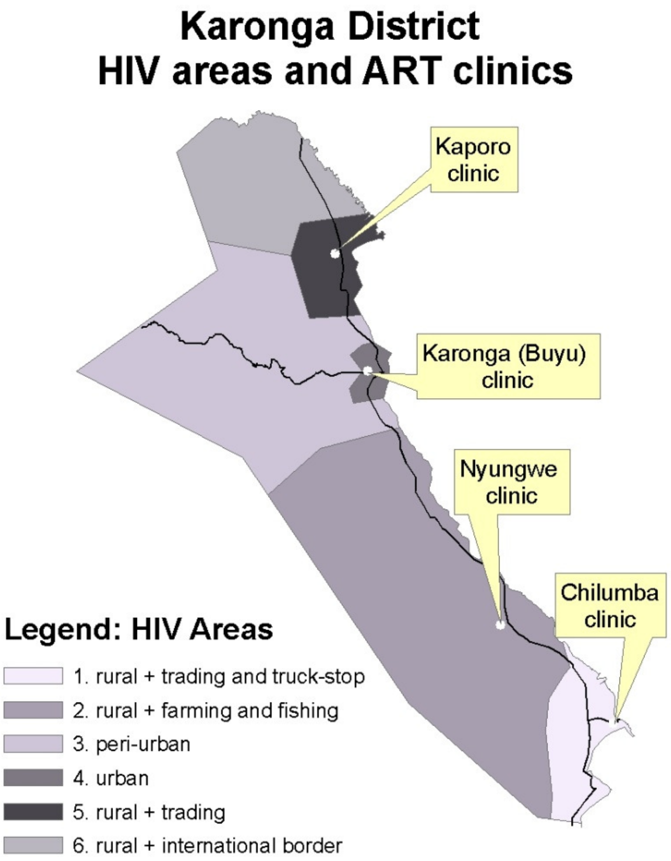
6.3.1.2 Areas in the district

The KPS village codes were grouped to represent the 6 areas within KPS district as defined previously by Crampin et al. [187] Briefly, the district was divided into the following areas (see figure 6—1 on next page) :

1. rural with a trading and truck-stop area
2. rural with farming and fishing
3. peri-urban
4. urban
5. rural with a trading area
6. rural with an international border.

These areas capture part of the heterogeneity of HIV risk in the district and are thus useful to standardise if the population sample that was tested for HIV was likely to be biased to the areas with higher HIV prevalences.

Figure 6—1: KPS defined HIV areas and ART clinics in Karonga District



Map shows Karonga District and the 6 areas defined as the

6.3.1.3 Standard population

Table 6.1 shows the area distribution by age group and sex of the standard population, expressed as the percentage of all individuals in that age and sex group, for males (left side) and females (right side). These were used to acquire area standardised HIV prevalences by age and sex group.

Table 6.2 shows the area and age distribution by sex of the standard population, expressed as the percentage of all males (left) or females (right). These were used to acquire area and age standardised HIV prevalences by sex.

Table 6.1: Standard population distribution within age groups by area and sex (males on left and females on right)

Males						
	15 – 24	25 – 34	35 – 44	45 – 54	55+	All ages
Area 1	17.4	16.2	14.8	14.6	15.1	16.0
Area 2	20.9	18.9	20.6	21.8	21.5	20.6
Area 3	18.2	17.9	18.9	19.5	20.3	18.6
Area 4	12.4	13.6	12.6	10.8	10.5	12.3
Area 5	16.9	16.3	15.4	16.4	16.7	16.4
Area 6	14.6	17.1	17.8	16.9	15.9	16.1
Total %	100	100	100	100	100	100
(N)	(19,206)	(13,226)	(7,162)	(4,813)	(6,734)	(49,229)

Females						
	15 – 24	25 – 34	35 – 44	45 – 54	55+	All ages
Area 1	16.5	16.5	15.3	15.3	19.2	16.5
Area 2	20.0	20.1	21.0	22.8	19.7	20.4
Area 3	178	17.8	19.4	20,3	19.3	18.5
Area 4	12.5	12.2	11.0	11.2	10.0	11.8
Area 5	17.0	17.1	16.7	16.5	16.1	16.8
Area 6	16.2	16.3	16.6	14.0	15.7	16.0
Total %	100	100	100	100	100	100
(N)	(22,394)	(13,226)	(8,475)	(5,670)	(6,574)	(56,339)

Totals represent the total number of males or females in each age group

Table 6.2: Standard population distribution within age and sex groups by area (males on left and females on right)

Males						
	15 – 24	25 – 34	35 – 44	45 – 54	55+	All ages
Area 1	6.7	3.7	2.2	1.4	2.1	16.0
Area 2	8.2	4.4	3.0	2.1	3.0	20.6
Area 3	7.1	4.1	2.8	1.9	2.8	18.6
Area 4	4.8	3.1	1.8	1.1	1.4	12.3
Area 5	6.6	3.7	2.2	1.6	2.3	16.4
Area 6	5.7	3.9	2.6	1.7	2.2	16.1
Karonga	39.0	23.0	14.6	9.8	13.7	100

Females						
	15 – 24	25 – 34	35 – 44	45 – 54	55+	All ages
Area 1	6.6	3.9	2.3	1.5	2.2	16.5
Area 2	7.9	4.7	3.2	2.3	2.3	20.4
Area 3	7.1	4.2	2.9	2.0	2.3	18.5
Area 4	5.0	2.9	1.7	1.1	1.2	11.8
Area 5	6.8	4.0	2.5	1.7	1.9	16.8
Area 6	6.4	3.8	2.5	1.4	1.8	16.0
Karonga	39.8	23.5	15.0	10.1	11.7	100

6.3.2 Population size estimates

6.3.2.1 *Population size between 1986 - 1998*

To estimate the population sized for each year between the 1988 and 1998 census I assumed a standard exponential population growth within each age and sex category:

$$'Population\ year\ x' =' population\ year\ (x-1)' \\ * annual\ growth\ rate$$

The annual growth between 1988 and 1998 was calculated for each age and sex group using the following formula:

$$Annual\ growth\ rate = \left(\frac{population\ in\ 1998}{population\ in\ 1988}\right)^{1/10}$$

For 1986 and 1987 the same growth rate was used for a retrospective estimation of the population decreased in those years.

6.3.2.2 *Population between 1999 - 2009*

For the period between 1999 and 2009 I used a similar approach. I obtained the official 2008 census results by age and sex for Karonga District and assumed a standard exponential growth rate between the 1998 and 2008 census. The population size for 2009 was estimated by extending the 1998 – 2008 population growth 1 more year.

6.3.2.3 *Population size estimates*

Table 6.3 shows the Karonga population estimates by age, sex and time period. These numbers are used as the overall denominator, which for the later chapters will be further subdivided by HIV and ART status. In the next section I will describe the estimates for the HIV prevalence in the district by age and sex between 1986 and 2010.

Table 6.3 Population sizes for Karonga District (1986 – 2009)

		Period				
Sex	Age	1986 - 1989	1990 - 1994	1995 - 1999	2000 - 2004	2005 - 2009
Male	15-24	14,068	16,415	19,415	21,815	24,132
	25-34	8,003	9,490	11,479	14,189	17,655
	35-44	5,611	6,349	7,292	8,539	10,063
	45-54	4,898	4,908	4,945	5,373	5,984
	55+	6,633	6,750	6,906	7,498	8,300
Female	15-24	14,980	18,012	21,977	24,593	26,794
	25-34	10,301	11,674	13,585	16,119	19,206
	35-44	6,707	7,475	8,534	9,576	10,662
	45-54	6,654	6,265	5,916	6,180	6,671
	55+	6,556	6,671	6,847	7,938	9,570
Total		84,412	94,010	106,898	121,820	139,037

Note: population sizes are averages for the specified 4 or 5 year periods

6.4 HIV prevalence in Karonga District

This section builds on previous work by Richard White et al. [186] In their paper they estimated the HIV prevalence in Karonga District using a mathematical model that incorporated the incidence of HIV and used the time since HIV seroconversion to estimate the need for ART in Karonga district. The data guiding the model was based on KPS studies (described below).

One of the main shortcomings of the model is that it does not take into account the roll out of ART (which started in July 2005 (see section 6.5). The prolonged survival of HIV cases starting on ART (who otherwise would be expected to have a very short survival) would probably increase the proportion HIV positive. I explore the effect of this through sensitivity analyses in Chapter 7.

6.4.1 Studies on HIV prevalence in general population

Since its start in 1981 various KPS studies have collected HIV status from random populations which can yield a study HIV prevalence representative for the HIV prevalence in the general population. These include retrospective testing of blood spots collected on filter papers from the 2 surveys done in the 1980s [182] and the various TB case control studies done since. [188-190]

6.4.1.1 Retrospective testing of filter papers

During the 2 district wide house to house surveys filter papers with dried blood spots were collected from all individuals in area 1 and area 6 of Karonga District. With permission from the Malawi Health Sciences Research Committee, these were retrospectively tested for HIV [182].

To estimate the data for the whole district, the area (only for area 1 and 6) and age standardised HIV prevalences were adjusted assuming that the ratio between the HIV prevalence in area 1 and area 6 compared to the whole district was identical to that found in the first study of community controls for TB cases (see next section) which were tested between 1988 and 1990.

6.4.1.2 Random controls for the TB cases

From 1988 onwards KPS has almost continuously conducted case control studies to explore risk factors for TB disease, with the exception of the period between 1993 and 1998. For these studies controls are randomly selected from the population which are matched for age, sex and area of residence in Karonga district.

As HIV is a strong risk factor for TB [171] the population of controls is not directly comparable to the general population as age, sex and areas with a higher HIV prevalence will be oversampled. HIV prevalences for these studies were therefore standardised using the standard populations described in section 6.3

Table 6.4.Crude and standardised HIV prevalences in adults (15 years or older) in Karonga District (by sex)

				Overall	Males		Females	
Study	Period	Description	Consent % (n/N)	Crude	Crude	Stand[a]	Crude	Stand[a]
LEP 1[c]	81 – 84	Retrospectively tested filter paper collected from all individuals in area 1 and 6 of Karonga District	100 (10183/10183)	0.1	0.1	0.1	0.1	0.1
LEP 2[c]	86 – 89	Retrospectively tested filter paper collected from all individuals in area 1 and 6 of Karonga District	100 (11308/11308)	1.7	1.4	1.4	1.9	3.3
TB CC 1	88 – 90	Community controls, matched TB cases [b] on age, sex and area of residence in district	99.6 (1710/1717)	2.8	2.7	2.9	2.9	3.8
TB CC 2	91 – 93	Community controls, matched TB cases [b] on age, sex and area of residence in district	99.4 (722/726)	7.3	6.8	8.5	7.9	11.4
TB CC 2	97 – 01	Community controls, matched TB cases [b] on age, sex and area of residence in district	87 (911/1046)	13.6	13.0	11.0	14.1	11.9
TB CC 3	02 – 05	Community controls, matched TB cases [b] on age, sex and area of residence in district	73 (591/811)	15.5	13.0	11.6	17.8	14.0
TB CC 4	07 – 09	Community controls, matched TB cases [b] on age, sex and area of residence in district	80 (379/475)	17.4	13.6	8.0	20.3	13.5

Note This overall data was combined with the age specific HIV prevalences to inform the mathematical model [186] that estimated HIV prevalence (by age and sex)

a standardisation was done for age and area of residence in Karonga district using standard populations described in table 6.2.

b Controls selection based on TB cases year before. [144, 168]

c Standardised HIV prevalence for LEP1 and LEP2 were adjusted using ratio HIV prevalence area 1+6 / HIV prevalence whole district in TB CC 1.

6.4.2 Mathematical model for HIV prevalence

As said in the previous section, this thesis applies the estimates from the published mathematical model. [186] I will therefore describe the overall workings and main assumptions regarding HIV prevalence, but not go into too much detail.

6.4.2.1 Main assumptions in model

The model extended methods to estimate the incidence of HIV during the initial phase of the epidemic, assuming an exponential increase in HIV incidence at the start of the epidemic. Survival time after HIV infection was assumed to decrease with age of HIV infection, from 12.5 years if infection occurred between the ages of 5-14 to 3.7 years for those infected after 65 years of age.

The combination of incidence and survival (i.e. the duration of infection) allowed the model to estimate the prevalence of HIV at a given time point using the formula

$$Prevalence = {}^{incidence}/_{duration\ of\ infection}$$

The model explored 2 scenarios of the HIV epidemic. HIV incidence was assumed to a stable phase after an initial peak, or HIV incidence was allowed to decrease (e.g. following a change in risk behaviour).

As stated before, the model did not account for prolonged survival following ART initiation.

6.4.2.2 Model input and output

The model was fitted to the age and area standardised HIV prevalences from the LEP and TB CC studies including data up to 2004. Figures 6—2 and 6—3 show the data points used to calibrate the model and added the new data points to assess how well the previous output from a models allowing a decline allowed (figure 6—2) or not allowing a decline (figure 6—3) fit with the new data.

6.4.2.3 Interpretation

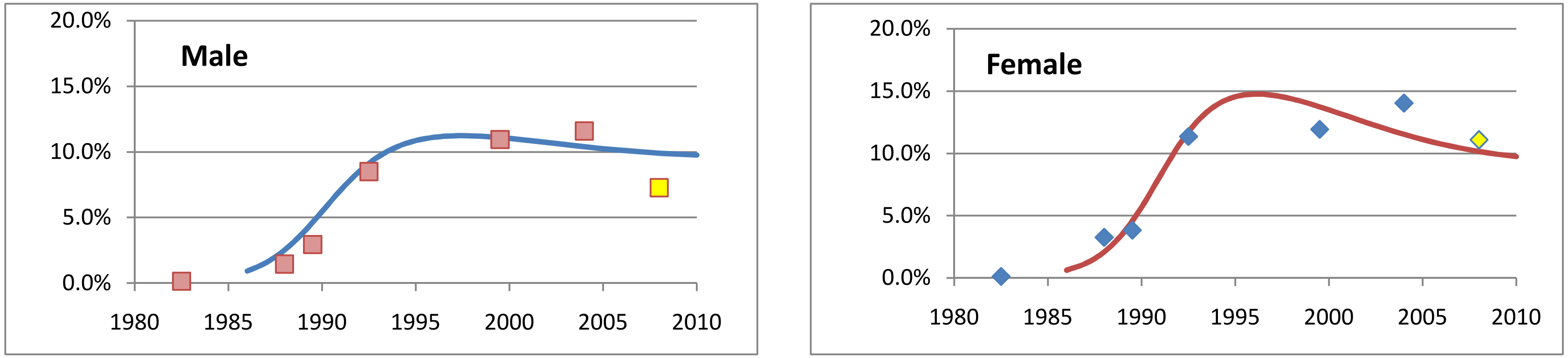
The model fits best when a decline in HIV incidence is allowed (figure 6—2). Considering the relative low numbers in the later data points (see table 6.4) the distance between the last data points in the males and the curve is relatively acceptable. We explored updating the HIV estimates with the newer data, but this proved impossible to do within the time span of this PhD. Considering the weight and location of the additionally available data it is unlikely the prevalence curves would make a substantial difference.

If the model is rerun it would have to be extended to account for improved survival because of ART.

6.4.3 Population estimates

The model output included a year by year HIV prevalence, by sex (see figure 6—2) or by sex and age group. The age specific HIV prevalences were applied to the total population size estimates discussed in section 6.3 to separate the population into HIV negative and HIV positive.

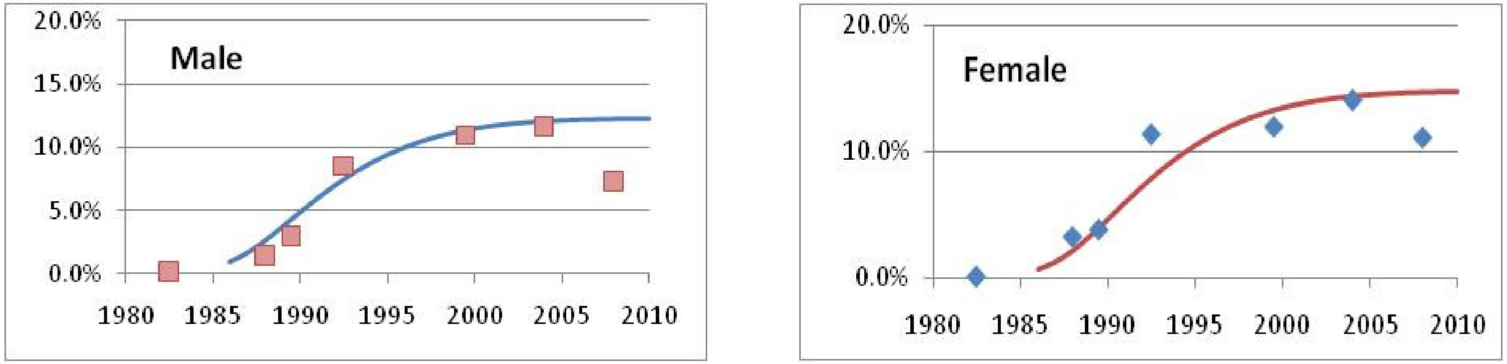
Figure 6—2 Age and area standardised HIV prevalences by sex (markers) and estimated HIV prevalence (line)



Note: Lines show the model estimates of HIV prevalence (in % on Y-axis) over time (in years on X-axis) in the model that allows a decline in HIV incidence. Markers show the age and area standardised HIV prevalences found in KPS studies. The yellow marker shows the data point that was not yet used to fit the model

The figures show that the overall trends fit reasonably well with the last data point as well. For the males the model may have overestimated the HIV prevalence in later years.

Figure 6—3 Age and area standardised HIV prevalences by sex (markers) and estimated HIV prevalence (line)



Note: Lines show the model estimates of HIV prevalence in the model that does not allow a decline in HIV incidence. The figures show that model fit is less good than above overall fit is less good than when a decline is allowed.

6.5 Patients receiving ART

6.5.1 ART uptake in Karonga district.

The first ART clinic, named Buyu (=baobob tree) in Karonga opened up in 2005 in Karonga Town, followed by Mwabi (= chance or luck) clinic in Chilumba which opened in 2006. In January 2008 two other clinics opened, Kaporo clinic in the North and Moonlight clinic in Nyungwe in the middle of the district (figure 6—1).

6.5.1.1 Patient ART mastercards

Patients who start on ART in Malawi are issued a patient mastercard by their ART clinic as part of the government system. On this mastercard several basic pieces of demographic information (age, sex, current village) are recorded, as well as their reason for starting ART, their ART regimen and a record of every time the patient has visited the clinic to receive medication. When a patient transfers between clinics they are supposed to take their mastercard with them so staff at the new ART clinic can copy the data onto a new mastercard.

6.5.1.2 ART patient registers

Each ART clinic also keeps a handwritten patient register. A line is filled for each patient when they register for ART. Apart from the name of the patient and their guardian the register holds the following information:

1. Demographic: age, sex and current village
2. ART history: date registered at this clinic (which is a close proxy for the date first received ART at this clinic) and date first started on ART at any clinic
3. Outcome: alive and receiving ART at this clinic, Transferred to a different clinic, Died, Defaulted or Stopped and the date of outcome if the outcome was not ‘Alive’.

6.5.2 ART patient register study

For the purpose of the analyses in this thesis I required the number of patients receiving ART in Karonga District by age and sex. I also required the time spent on ART to distinguish between patients whose immune system was still recovering and patients for whom ART could have been expected to have had a substantial effect on their immune system. In this chapter I set that cut-off on 6 months. In the coming chapters I will explore the effect of applying different cut-offs on the results of the analyses.

This information was available in the patient registers, which I aimed to record.

Note: this work was in part funded through the Gordon Smith Travelling Scholarship.

6.5.2.1 Objective and data collection

In the ART patient register study I aimed to collect the demographic, ART history and outcome information for each patient record in the patient registers. After obtaining ethical approval from the National Health and Sciences Research Committee in Malawi (reference number 424) and the LSHTM Ethics Committee (reference number 5067) anonymised digital pictures were taken during August 2009 from each page of all patient registers at the 4 clinics (see figure 6—4 for an example of an anonymised picture).

Data were double entered and discrepancies checked and decided on by a third reviewer based on the original picture.

A total of 5575 records were double entered in the KPS database. Village names as recorded by the ART staff were recoded into KPS village codes and further into the 6 HIV areas mentioned described in section 6.2.

6.5.2.2 Transferred patients

Of these records, 917 indicated that the patient in question had either transferred into a Karonga District ART clinic from outside of the district or transferred between ART clinics within the district, probably because an ART clinic opened closer to their home.

Although this migration of patients is understandable, it poses a complication for the interpretation of the data as the same person could potentially be counted twice if one wanted to record the total number of people started on ART in Karonga district.

Ideally, I would be able to follow the course of each individual who started receiving ART as they moved between clinics within Karonga District. I tried an approach to solve this by accessing the individual patient mastercards for those patients that transferred into a clinic. In theory, these mastercards would hold information on a patient's previous clinic and patient number. This would allow us to link records of the same patient as they moved between different clinics. In practice, this system turned out to be unworkable; missing mastercards as well as missing or incorrect numbers caused the required workload per individual record to be very high, and even with the extra effort a high proportion of transferred patients could not be linked.

The exercise turned out to be useful however, as the inspection of mastercards from patients that had transferred into a clinic showed that the date this patient first started ART as recorded in the register was often incorrect. In those cases the date this person was first tested HIV positive was copied into the register as the date first started ART. For all the transferred patients the mastercards were re-checked and if the date first started ART was recorded wrongly in the register the data was amended in the KPS database.

As for the problem of time on ART, the main purpose of these data was to provide a denominator for the incidence of TB in patients receiving ART, by sex, age group and time spent on ART. For that purpose the number of person years in each category is more important than the absolute number of patients starting on ART.

This number of person years at risk of TB in a set period, by age, sex and time spent on ART can be constructed using the available data, as I describe in the next section.

6.5.3 Data processing

6.5.3.1 Survival analysis approach

What was required was a was the number of people receiving ART in a set period (e.g.2005-2006) and the average time these patients had been on ART before they entered that time period. In more technical terms, for set periods I needed to know the total person years alive on ART (and thus at risk of TB) in Karonga district, by age, sex and time on ART.

In essence this is a straightforward survival analysis question in which the total observation time, rather than the number of individuals in the cohort, is split between the required analysis groups.

To extract and split the observation time 3 pieces of information are required, known in Stata as 'origin' (start of clock), 'enter' (start of follow up/observation time) and 'exit'(end of follow up/observation time). See figure 6—5 and table 6.5 for an illustration of the calculation.

Figure 6—4: Calculating observation time from ART patient registers records

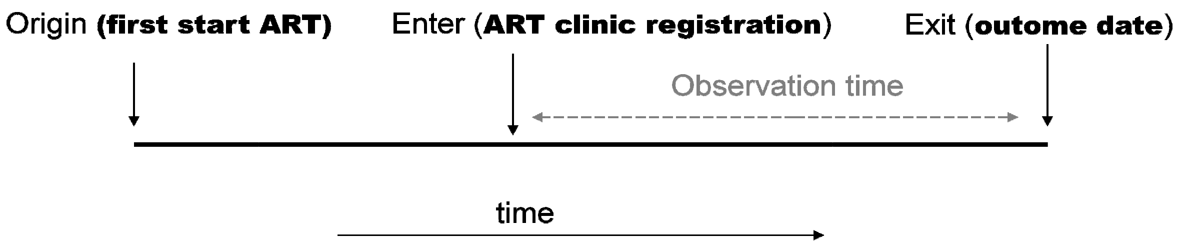
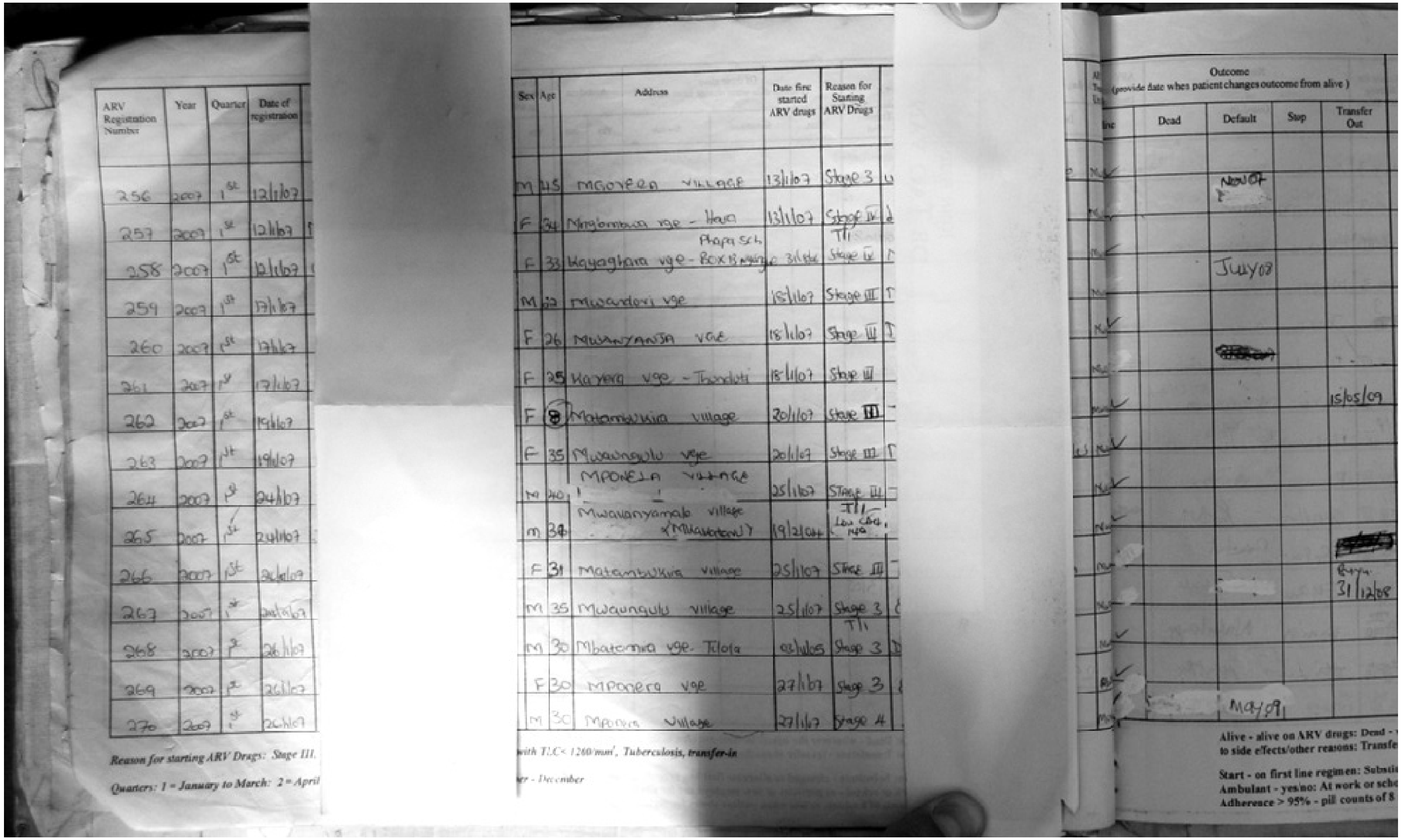


Table 6.5: Variables in observation time calculation

Variable	Description	In this study
Origin	Time when the clock starts	Date first started ART
Enter	Start of follow up time	Date started ART in this clinic
Exit	Exit from study, censor event	If alive: end of study period. If different outcome: date of outcome

This method builds on the idea that patients transferring between clinics may have multiple records, but their observation times are unlikely to overlap, which solves the problem of double counting individuals.

Figure 6—5: Example picture of ART patient register study



ARV Registration Number	Year	Quarter	Date of registration	Sex	Age	Address	Date first started ARV drugs	Reason for Starting ARV Drugs	Alive	Dead	Default	Stop	Transfer Out
256	2007	1st	12/1/07	M	45	MGOYERA VILLAGE	13/1/07	Stage 3	✓				
257	2007	1st	12/1/07	F	34	Mngómbwa vge - Hara Phapa Sch	13/1/07	Stage IV TI	✓		Nov07		
258	2007	1st	12/1/07	F	33	Kayaghara vge - BOX 13 [illegible]	31/1/06	Stage IV	✓				
259	2007	1st	17/1/07	M	22	Mwandovi vge	15/1/07	Stage III	✓		July08		
260	2007	1st	17/1/07	F	26	MWANYANJA VGE	18/1/07	Stage III	✓				
261	2007	1st	17/1/07	F	35	Kayera vge - Thunduti	18/1/07	Stage III	✓		[redacted]		
262	2007	1st	19/1/07	F	8	Matambukira village	20/1/07	Stage III	✓				15/05/09
263	2007	1st	19/1/07	F	35	Mwaungulu vge	20/1/07	Stage III	✓				
264	2007	1st	24/1/07	M	40	MPONELA VILLAGE	25/1/07	STAGE III	✓				
265	2007	1st	24/1/07	m	34	Mwanyamale village (Mkawotcw)	19/2/04	TI Low CD4, 149	✓				
266	2007	1st	26/1/07	F	31	Matambukira village	25/1/07	Stage III	✓				[redacted]
267	2007	1st	26/1/07	m	35	Mwaungulu village	25/1/07	Stage 3	✓				Karonga 31/12/08
268	2007	1st	26/1/07	m	30	Mbatemera vge - Tilola	03/1/05	TI Stage 3	✓				
269	2007	1st	26/1/07	F	30	Mponera vge	27/1/07	Stage 3	✓				
270	2007	1st	26/1/07	M	30	Mponera Village	27/1/07	Stage 4	✓		May09		

Outcome (provide date when patient changes outcome from alive)

Reason for starting ARV Drugs: Stage III, *with T.C< 1200/mm³, Tuberculosis, transfer-in*

Quarters: 1 = January to March: 2 = April *er - December*

Alive - alive on ARV drugs: Dead -
to side effects/other reasons: Transfe

Start - on first line regimen: Substi
Ambulant - yes/no: At work or scho
Adherence > 95% - pill counts of 8

Note: first column gives the clinic registration number.

6.5.3.2 Person years on ART in Karonga District

To calculate the person years of observation I used the stsplit command in Stata. Table 6.6 shows the total number of person years recorded for each calendar year, the age and sex distribution and the proportion of person years where the patient had been on ART more than 6months

Table 6.6: Adult person years receiving ART in Karonga District

	Person years	<=6m on ART	>6m on ART	Age (med+IQR)	Sex (%male)
Aug 05 – Dec 05	137	82	55	37 (32–45)	0.39
Jan 06 – Dec 06	762	289	473	37 (31–45)	0.39
Jan 07 – Dec 07	1517	424	1093	37 (31–44)	0.39
Jan 08 – Dec 08	2225	530	1695	36 (30–45)	0.39
Jan 09 – Aug 09	1615	309	1306	36 (30–45)	0.39

This table shows the number of person years on ART

6.5.4 ART in Karonga District

The data in table show that the delivery of ART in Karonga District expanded quickly in the first years. The records suggest that is a combined effect of people first starting ART in a Karonga clinic because they can now access their clinic, and people who were previously receiving their ART from clinics outside of the district who now have transferred into a Karonga clinic. The pattern of patients repeatedly moving between clinics suggests that ease of access is an important factor for these patients. The ART programme in Karonga District is scheduled to start serving patients at 2 additional ART clinics in the rural South West area and far North area of the District, thus proceeding with the drive to reduce the barrier for patients in need of ART.

6.6 From denominator to incidence

This chapter describes the data collection and processing for the denominators of TB incidence in Chapters 7. In the following chapters I will describe the data collection for the numerator of the incidence equation and analyse the resulting trends.

7 TB incidence in Karonga District after the roll out of ART

7.1 Summary

The roll out of ART is hoped to reduce TB related morbidity and mortality in HIV positive patients. [21] However, mathematical modelling and cohort studies have suggested that this effect may be limited, especially if HIV positive patients are severely immunocompromised when they start on ART. [34, 36]

In this chapter I estimate TB incidence over time by HIV and ART status in Karonga District. I specifically focus on how the introduction of ART in July 2005 has affected incidence trends.

7.2 Introduction

The roll out of ART has further diversified the population of HIV positive patients in sub Saharan settings. Patients with severe immunosuppression and a short life expectancy before ARTs were available are now eligible for, and often have access to ART.

As a part of this highly vulnerable patient group is started on ART a new sub population is created of previously severely immunocompromised patients now supposedly en route to immunological recovery. Although the first months after ART initiation are marked by a highly elevated risk of opportunistic illnesses, including TB [191], after a certain period taking ARTs a patient's immune system is expected to have started recovering, usually measured by the CD4 counts returning to normal levels. [36]

7.2.1 ART and the relative incidence of TB in HIV positive individuals

There are several studies describing TB incidence in patients receiving ART, [35-46, 192] which mostly show that TB incidence is high in the first 3-6 months after ART initiation and then decreases with time spent on ART. Also, when CD4 data were available, studies showed that patients with lower CD4 counts have a higher risk of developing TB. [35, 36, 38, 44, 45, 192]

Several of these studies were able to compare the relative incidence of TB between HIV positive/ART naive patients with patients on ART. [43-46] Overall, they suggested that TB incidence in patients was lower in patients on ART, especially after correction for CD4 count.

However, patients in these studies are often closely monitored and have access to high level laboratory services, which is not representative of the regular setting of ART clinics in Sub Saharan Africa. Also 3 out of 4 of these studies based their results on less than 10 TB episodes in patients on ART. [43, 45, 46] Finally, most did not exclude TB cases with recurrent TB or patients without a laboratory diagnosis.

These studies leave the question of the relative incidence of TB in HIV positive patients started late on ART (WHO stage 3/4 or <200 CD4 count) in complete population settings where the available clinical and laboratory support is minimal. As the majority of HIV positive patients live and will access ART in such settings it is important to get an estimate of what to expect after ART arrives in these populations.

Also, none of the studies directly compared the relative incidence in patients receiving ART with that of a HIV negative cohort from the same population.

7.2.2 ART and TB incidence in the population

Only one study from a European cohort explored the impact of more widespread uptake of ART (between 1995-1997) on the TB incidence in a cohort of HIV positive cases. [44] It showed that TB incidence decreased with the uptake of ART and associated increase in median CD4 count of the population.

No study has been able to look at the effect of ART on the TB incidence rates in a whole population, including HIV negatives, or in a rural Sub Saharan population. It was hoped that the introduction of ART would not only improve survival for HIV positive patients but that ART would have a knock on effect on TB incidence in populations through the improved immune status of HIV positive patients. [21] However, in a mathematical model Williams and Dye warned that because of the prolonged survival in patients receiving ART the effect on TB incidence would be limited unless patients were started early (e.g. from a CD4 = 500 cells/µL) and coverage and compliance were both over 90% [34]. In a study from South Africa Lawn et al. confirmed that patients that started ART with severe immunosuppression (CD4 count <100 cells/µL) remained at higher risk of TB compared to patients starting ART with CD4 count ≥100 cells/µL throughout 5 years of follow up. [36]

7.2.3 TB incidence in Karonga District

In this chapter I explore the effect of ART on the incidence of TB in the context of a well functioning TB control programme and a health centre based ART programme which is unsupported by CD4 or viral load count facilities.

KPS has a complete overview of all diagnosed TB cases in Karonga District since 1986, including reliable data on smear and/or culture confirmed TB. The database also holds information on the HIV status and ART history for most TB cases. The level and quality of health care other than TB in Karonga District is probably representative for other Sub Saharan settings. As described in Chapter 6, ART arrived in mid 2005, and is distributed through local health clinics, with little clinical or laboratory support for the monitoring of patients. This is typical for Malawi and similar to other rural SSA settings. This provides a good opportunity to study the impact of ART roll out on TB rates in a rural population.

The first line ART regimen has been the same since the roll out started in Malawi. The standard first line therapy includes the 2 nucleoside reverse transcriptase inhibitors (NRTI) Stavudine and Lamivudine and 1 non-nucleoside reverse transcriptase inhibitor (NNRTI) Nevirapine. [193] For second line therapy an NRTI backbone of Zidovudine, Lamivudine and Tenofovir is combined with a protease inhibitor combination of Lopinavir/Ritonavir. [193]

Previous KPS studies have shown that the incidence of TB at KPS peaked in the mid 90s and seemed to be decreasing up to 2001[171]. This chapter adds information on TB trends since then and explores the impact of ART in 2005. Based on the assumption that most HIV positive patients starting ART in Karonga will be highly immunosuppressed, the hypothesis is that they will have a high risk of TB. Whether this will lead to a change in TB incidence trends in other groups e.g. the HIV negative population, i.e. whether there is an indirect effect of ART, will be explored as well.

7.3 Methods

7.3.1 Population denominators

These data are described in detail in chapter 6; I will briefly summarise them here.

7.3.1.1 General population

In short, I used the results from 3 full censuses of the population in Karonga District in 1988, 1998 and 2008 as data points. I then assumed an exponential growth between these data points and directly before (1986-1987) and directly after (2009) to get estimates by age and sex group in each year.

7.3.1.2 HIV prevalence

I applied the results from a mathematical model run by Richard White et al, which was published in 2007. [186] This model incorporated an estimate of the incidence of HIV and used the time since HIV seroconversion to estimate the need for ART in Karonga district. The data guiding the model were based on KPS studies that tested randomly selected controls for TB cases or complete populations during whole population surveys (LEP1 and LEP2). [182, 188-190] The model that fitted the data best allowed the incidence of HIV to decline in Karonga after a peak in the early 1990s.[186]

One of the main shortcomings of the model is that it does not take into account the roll out of ART (which started in July 2005 (see section 6.5)). The prolonged survival of HIV cases starting on ART (who otherwise would be expected to have a very short survival) would probably increase the proportion HIV positive. I explore the effect of this through sensitivity analyses. The original model also did not incorporate data beyond 2005, although the new data points fit reasonably well with the estimated curve as is shown in figure 6—2.

7.3.1.3 Person years spent on ART

This was recorded through the ART patient register study, as described in Chapter 6. We recorded demographic, ART history and outcome information for each patient starting ART in Karonga district. This information was converted into person years on ART in Karonga district by age and sex and by time spent on ART. To calculate the incidence of TB in the HIV positive population not receiving ART the total HIV positive person years were reduced by the observed person years on ART.

7.3.2 TB case population

7.3.2.1 Inclusion and exclusion criteria

I included adult (15 years or older) TB cases with a TB diagnosis between Jan 1986 and the start of Aug 2009. The end time was determined by the data from the ART patient register study (see Chapter 6).

Only new TB cases were included. Patients who reported a previous TB episode in the interview or recorded in the KPS database were excluded.

For the main analysis I included patients diagnosed with pulmonary TB who had at least 1 clear positive sputum smear (SS+) result from a sample examined at the KPS laboratory, thus excluding cases with 1 marginally positive smear (fewer than 10 acid-fast bacilli per 100 fields). This case algorithm is reproducible and has been applied with similar rigour throughout the history of KPS. [144, 171] Also, SS+ pulmonary TB cases represent the main source of new Mtb infections in a population, thus drive TB epidemic and thus are a highly relevant group for the study of the TB epidemic in a population.

7.3.3 Variable definitions

7.3.3.1 HIV status of TB cases

HIV status of TB cases at the time of the TB episode was extracted from the KPS database, which holds the results of all confirmed HIV test results for an individual. Tests on urine or saliva were excluded.

A TB case was considered HIV negative at time of their episode if their negative HIV test was recorded after or within 1 year before their TB was diagnosed. Conversely, a TB case was considered to be HIV positive at time of their TB episode if their first recorded positive HIV test was done before or within one year after the end of their TB episode.

7.3.3.2 ART status of TB cases

The first months after an immunocompromised HIV patient is started on ART are usually marked by a high risk of a variety of diseases, including TB. [194] The complex of these disease episodes is sometimes referred to as immune reconstitution syndrome (IRS), although it is unclear which diseases can be referred to as IRS and which are considered as expressions of the regular burden of opportunistic infections in severely immunocompromised patients. [195]

This period marked by increased rates of opportunistic infections precedes the time period where a patient is expected to enter a phase of presumed immune recovery and subsequent lower risk of TB disease. Some studies quote 1 month, others 3. [191]

For the purpose of this analysis I will use a 6 months cut-off for distinguishing between patients recently started on ART and those receiving ART for longer. In patients receiving ART for over 6 months we can expect ART to have had an effect on the risk of TB. This gives 4 groups in the analysis of TB by HIV and ART status:

1. HIV negative
2. HIV positive not receiving ART
3. HIV positive and recently started on ART (i.e. within 6 months)
4. HIV positive and receiving ART for a longer period.

7.3.3.3 Period analysis

Overall and HIV specific TB incidences were calculated by year. TB incidence by ART status was calculated for the complete August 2005-July 2009 period.

In the main analyses I explored TB incidence by HIV and ART status by the following periods:

Period	Start	End	Mid point
1	1997	1999	June 1998
2	1000	2002	June 2001
3	2003	2005	June 2004
4	2006	Aug 2009	October 2007

7.3.3.4 Age

Age was grouped into 5 categories, 15-24, 25-34, 35-44, 45-54, 55 and older.

7.3.4 Incidence estimates

7.3.4.1 Overall Incidence: January 1986 – Aug 2009

The overall TB incidence of new pulmonary SS+ TB cases was estimated by dividing the total number of cases (by age and sex category) by the total number of observed person years, taken as the population size that year. All incidence estimates are expressed as the number of cases per 100,000 person years:

7.3.4.2 Incidence by HIV: January 1996 – August 2009

HIV test results are available for a large proportion of TB patients, and have been routinely offered to TB cases since 1988. From 1997 onwards TB incidence is stratified by HIV status. I report the observed incidence by HIV status, in which TB cases with unknown HIV status at time of their TB episode are excluded. I used Multiple Imputation using Chained Equations (MICE) to fill in and explore the potential bias that followed from excluding them through an analysis of the imputed dataset (see section 7.3.5)

7.3.4.3 Incidence by HIV and ART status: August 2005 – August 2009

As said in Chapter 6 ART became available from a clinic in Karonga District in July 2005 and more widely from 2006. TB incidence was estimated as the total number of TB cases on ART for different lengths of time, between August 2005 and July 2009 divided by the total number of observed person years of patients receiving ART in that age, sex and ART category during that period. TB cases with missing values for ART status were excluded from this analysis. MICE was used to explore the potential bias from those exclusions.

7.3.5 Imputation of missing values

7.3.5.1 Background

As in any study some observations will have missing values. One variable often missing is the HIV status, in this case the HIV status at time of an individual's TB episode. This may be due to refusal to undergo testing, or because the patient died (or defaulted) soon after diagnosis, before a HIV test could be done.

These missing values pose a problem in the analysis. Often patients with missing values for core variables such as HIV are excluded from the analyses. These 'completer' analyses only include cases for which information on all variables in the analysis are available.

Such analyses are inefficient as cases can be lost because of 1 missing value and likely to introduce some bias in the results if the missing data do not satisfy the missing completely at random assumption (MCAR, see below for explanation). One can try to assess the extent of this bias by analysing patients with missing HIV status as a separate group [72], or run sensitivity analyses as well in which all patients with unknown HIV status are assumed to be HIV positive or HIV negative.

However, most of these techniques have the potential to introduce new bias of their own or underestimate the inherent uncertainty of filling in missing data (see http://missingdata.lshtm.ac.uk/start.html, accessed 10-2-2010). In incidence analyses the effect is even more severe, as missing data will result in absolute underestimation of the incidence in each group, e.g. HIV negative and HIV positive group, as well as potentially biasing the relative incidence rates between groups.

In recent years imputation of missing values has become increasingly accessible to medical researchers following the development of computing power and statistical software packages. In modern imputation the value of the missing data is filled in, i.e. imputed, based on the observed values for that and other variables.

7.3.5.2 Missing at random assumption

Imputation of missing values relies on the core assumption that data are either missing completely at random (MCAR) or that any systematic differences in the observed and missing data can be explained by differences in the observed data, which is referred to as data Missing At Random (MAR). [196] If this is not possible the data are considered to be Missing Not At Random (MNAR), and imputation is not possible. In other words, if the reason for missingness is a direct consequence of an unmeasured factor e.g. a prior test result, the MAR assumption is violated. Imputation will only compound the bias introduced by the missing value and its results will not be useful or interpretable.

Unfortunately there is no hard and fast way to determine whether data are MAR or NMAR. In practice part of the data will always be NMAR, i.e. some of the systematic differences will not be explained by covariates. This is similar to residual variation in regression models, with the difference that it cannot be quantified. One has to look carefully at the patterns of missing data and the imputed values across other variables, consider the theoretical framework of why the variable may be missing and then make an informed judgement call on whether the MAR assumption is sufficiently valid and the imputed data can be presented.

7.3.5.2.1 MCAR, MAR and NMAR in the context of HIV

HIV is a good example of how data can be MCAR, MAR and NMAR at the same time. Part of the missing values for HIV status will be MCAR, e.g. a HIV test that was not done because the tube of venous blood was accidently dropped in the lab. These values can be imputed based on information on the person’s age, sex and perhaps area of the district (e.g. urban or rural) they were living in. In the KPS data each of these are shown to be strongly associated with an individual’s HIV status. [182, 186, 187]

Another part of the missing values for HIV will be MAR, for example because a missing HIV test result is more likely if the patient died before he/she could be tested. As death during treatment is also more likely if the patient is HIV positive, the recorded outcome can help explain some of the systematic differences between missing and observed data. In this case, patients with missing HIV values are more likely to be HIV positive than patients with an observed HIV status.

However, for another group of patients with missing HIV status the systematic difference cannot be captured by other observed variables. For example, it is known that patients who have tested positive for HIV are less likely to consent to a future test. [197, 198] In the context of KPS data this issue comes up if the patient was tested positive outside of KPS. Their previous test result will not be recorded in the KPS database and this information cannot be used to correct for the missing data. In Karonga District the number of HIV testing facilities has increased following the roll-out of ART. This has been shown to be a problem in an annual sero-survey. [197]

7.3.5.3 Multiple Imputation using Chained Equations (MICE)

There are various approaches to multiple imputation, most of which lead to similar results. In this chapter I will use the MICE command, which is available in the Stata statistical software package.

7.3.5.3.1 MICE

MICE stands for Multiple imputation using Chained Equations. Multiple imputation refers to the fact that missing values for all specified (i.e. multiple) variables are imputed in the same process. Chained equations refers to the system of regression equations that works cumulatively; after missing values in the first variable are imputed these values are in turn used in the imputation of missing values in the next variable. The imputation of each variable thus builds on previous imputations, in which each imputation is linked through a ‘chain’ of regression equations.

7.3.5.3.2 Basic principle

I will not describe MICE in great technical detail, but give an overview of its basic workings. For a detailed discussion please see [199]

7.3.5.3.2.1 Variables in the imputation

In the MICE command one has to specify the variables to be imputed and the regression model that MICE should apply for each of them.

MICE will always start the imputation cycle with the variable that has the least missing values and build from there. Once the imputation for that variable is complete, it will move to the next variable with the least missing data and thus progress through the variables in order of missingness. As a result the variable with the most values missing will be imputed last, using all observed and imputed values of the other variables.

7.3.5.3.2.2 Imputation variables

First the missing values for the variable are randomly filled in with values that lie within the range of the observed values. Then an appropriate regression model is run in which the imputed variable is the outcome (e.g. a logistic regression to impute values for the binomial variable HIV status) and the variables specified in the MICE command as covariates.

This regression model results in a probability distribution of values for the missing value, given the values, observed and randomly chosen, for the other variables of that observation. The imputed value is drawn at random from that distribution.

Please note that observations for which all covariates are missing are ignored in the MICE command. No imputed value is generated for the simple reason that there are no data to inform the model.

This process is repeated for every variable with missing values in order of the proportion of missing observations. One such sequence of regression models in which all the variables with missing data are imputed once is called a cycle.

7.3.5.3.2.3 Cycles within an imputation

Please note that in the first cycle all missing values were randomly filled in, and then replaced by imputed values. Therefore, in the first cycle the regression models that generate the distributions for these imputed values will be guided to an undesirable extent by the randomly entered values. To get away from those randomly filled in values, the cycles described in the previous section are repeated a specified number of times.

The second cycle will use the imputed values from the first cycle as substitutes for the originally missing values, rather than the random values used in the first cycle, and the third cycle will use the imputed values from the second cycle. Through this mechanism of repeatedly cycling through the chained regression equations the imputed values will move increasingly away from the randomly chosen starting values in the first cycle and have a stronger base in the observed data.

The required number of cycles is not clear. A minimum of 10 cycles is usually recommended. [200] In this thesis each imputation is run with 20 cycles.

7.3.5.3.2.4 Number of imputation runs

The complete imputation process is repeated a number of times, to assess the uncertainty of the imputed values. The recommended number of repeats is under discussion, but a minimum of 10 imputations is usually advised. In this thesis I repeat each imputation 20 times.

7.3.5.4 Covariates in the imputation

For multiple imputation it is generally recommended to use as many and diverse variables as possible, provided that they do not have too many observations missing, i.e. introduce too much additional uncertainty in the imputation model.[196] There is no general rule for what constitutes a high level of missing, one has to make a judgement call. In the imputations for chapter 7 all the covariates I use in MICE have less than 5 percent of observations missing.

7.3.5.4.1 Sex

Imputations were done stratified by sex as various associations between HIV and other variables, for example HIV patterns by age, differ between the sexes. Rather than introduce interaction terms in the MICE command I chose to run separate imputation models for males and females.

7.3.5.4.2 Age

Age is strongly associated with HIV status. Age was not missing for any of the observations, so no regression model needed to be specified.

7.3.5.4.3 Outcome

Outcome of the TB episode was included in the imputation, coded as ‘Completed’, ‘Died before treatment completion’, ‘Left before end of treatment’, ‘Transferred out’. Multinomial logistic regression was used in the imputation model to impute missing values.

7.3.5.4.4 Area in district

The HIV prevalence is higher in some areas of the district, as described in Chapter 6. We included which area the TB case was from in the imputation as a categorical variable with 6 values. Multinomial logistic regression was used to impute missing values for this variable.

7.3.5.4.5 Year of episode

The HIV prevalence among TB cases may well vary over time. Year of diagnosis of the TB episode was included in the imputation. This variable had no missing values.

7.3.5.4.6 Type of TB

TB type was coded as PTB only, or PTB and either lymph node or other EPTB. This variable had no missing values.

7.3.5.5 Imputation of HIV status

7.3.5.5.1 Time span

HIV status for TB cases was imputed for the years 1997 – 2009. This cut off is based on a too high percentage of HIV status missing in the earlier period (HIV status was unknown for over

85% of cases between 1994 and 1996 following concerns of the reliability of HIV tests done during that period).

7.3.5.5.2 MAR assumption -outcome

As illustrated in section 7.3.5.2.1, one of the main variables that could account for HIV status MAR is outcome. Table 7.1 (first 3 columns) shows the distribution of outcome for HIV negative, HIV positive and cases with unknown HIV status.

Table 7.1: HIV status by outcome – missing and imputed

Outcome	HIV Neg % (n/N)	HIV Pos % (n/N)	HIV unknown % (n/N)	Observed HIV prevalence	% of missing imputed as HIV positive[a]
Cured	81.8 (356/435)	68.1 (425/624)	30.2 (90/298)	54.4 (425/781)	54.2
Died	4.6 (20/435)	19.6 (122/624)	44.6 (133/298)	85.9 (122/142)	80.5
Lost	2.3 (10/435)	3.5 (22/624)	13.4 (40/298)	68.8 (22/32)	65.9
Transferred	3.0 (13/435)	3.7 (23/624)	8.7 (26/298)	63.9 (23/36)	60.0
Unknown or missing[b]	8.3 (36/435)	5.1 (32/624)	3.0 (9/298)	47.1 (32/68)	58.3

Note: table includes new SS+ pulmonary TB cases with start of episode before August 2009. Results show that the proportion of outcome = death in the HIV unknown is high compared to those with known HIV status. This fits with a MAR assumption. Imputed values mostly follow the observed proportions in each group, as can be expected with MICE.
a percentage calculated as average across 20 imputations.
b includes 3 failure cases that were merged with unknown category to prevent overfitting in the imputation.

As with any regression analysis, categories with too few data cause problems. For the purpose of the imputation I therefore merged category Failure with the Unknown category. This conservative approach exchanges a bit of additional missingness in the imputation for the benefit of increased stability in the regression models.

The last 2 columns of table 7.1 show that the imputed values for HIV follow the observed HIV prevalence in each outcome category, e.g. a high HIV prevalence in the cases who die before completing treatment.

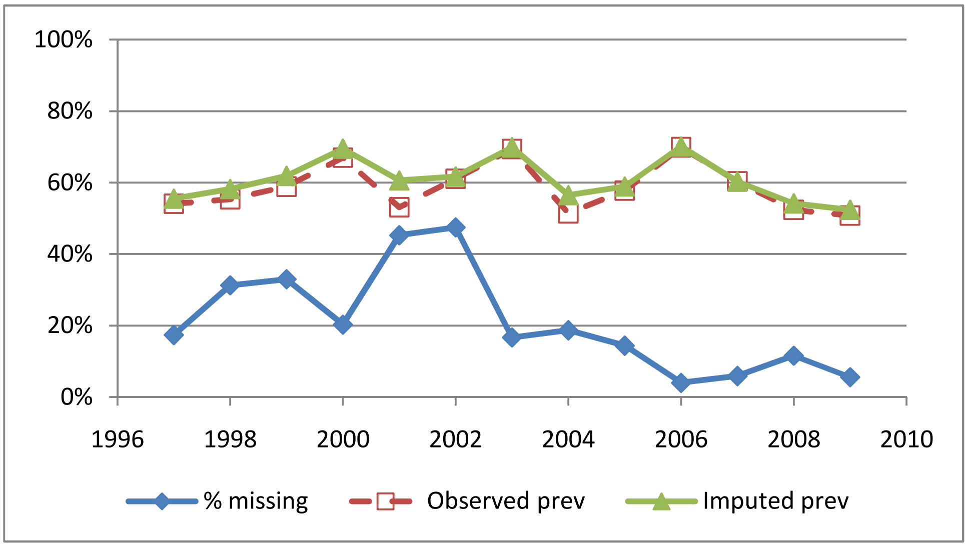
Table 7.2 summarises the results of the HIV imputation based on all variables described in section 7.3.5.4. Figure 7—1 gives a graphical illustration of the imputation. As the majority of TB cases with missing HIV status were imputed as HIV positive the imputed HIV prevalence in TB cases was higher than the observed in most years.

Table 7.2: Observed and imputed HIV prevalence among TB cases by year

Year	missing % (n/N)	Observed prevalence %(n/N)	Imputed prevalence %[a]
1997	17.4 (21/121)	54.0 (54/111)	55.5
1998	31.3 (42/134)	55.4 (51/97)	58.2
1999	33.0 (36/109)	58.9 (43/74)	61.9
2000	20.3 (24/118)	67.0 (63/99)	69.5
2001	45.3 (53/117)	53.1 (34/67)	60.6
2002	47.5 (49/103)	61.1 (33/55)	61.7
2003	16.7 (17/102)	69.4 (59/88)	69.9
2004	18.7 (17/91)	51.4 (38/81)	56.5
2005	14.4 (14/97)	57.8 (48/81)	58.9
2006	4.0 (4/97)	69.9 (65/97)	70.1
2007	5.9 (6/102)	60.4 (58/97)	60.3
2008	11.6 (11/95)	52.4 (44/87)	54.2
2009	5.6 (4/71)	50.8 (34/67)	52.4
Total	22.0 (298/1357)	59.2 (624/1059)	61.0

Note: 2009 cases only included cases with diagnosis before 1st of August 2009
a percentage calculated as average across 20 imputations.

Figure 7—1: Observed and imputed HIV prevalence (1997 – 2009)



This graph shows that the main effect of the imputation was to even out some of the artificial drop in 2001 and 2002. It suggests a decrease in the proportion of TB cases that is HIV positive since 2006.

7.3.5.6 Imputation of ART status

ART status was missing for 17% (70/415) of SS+ pulmonary TB cases with diagnosis between 31st of July 2005 and 1st Aug 2009 (table 7.3)

Broadly the same imputation model was used as in section 7.3.5.5, with the same explanatory variables. The main difference was that the HIV variable was replaced with a 4-category variable that described an individual's HIV and ART status. These 4 categories followed those described in section 7.3.3.2:

1. HIV negative (and therefore not receiving ART)
2. HIV positive not receiving ART
3. HIV positive and recently started on ART (i.e. 6 months or less)
4. HIV positive and receiving ART for a longer period (i.e. more than 6 months).

This joint variable has the advantage that it automatically informs the imputation model of the restriction that HIV negative patients should not be receiving ART. Also, there is a statistical advantage in that HIV and ART status now share a joint probability distribution, which is advantageous in the MICE statistical framework.

Table 7.3 shows the proportion of new SS+ pulmonary TB cases between August 2005 and July 2009 that had a missing value for ART status. Table 7.4 shows the missing, observed and imputed values for ART. The missing values were for the most part imputed as not on ART, although the proportion of cases imputed as being on ART is slightly higher than in the groups.

Table 7.3: Proportion of cases with missing ART status (August 2005 – July 2009)

Year	missing % (n/N
2005	20.0 (10/50)
2006	13.4 (13/97)
2007	15.7 (16/102)
2008	17.7 (17/95)
2009	19.7 (14/71)
Total	16.9 (70/415)

Table 7.4: Observed and imputed values for ART status (August 2005 – July 2009)

ART category	Observed % (n/N)	Imputed %[b]	Total %
Not on ART[a]	82.3 (284/345)	77.1	81.4
Recently started ART (<=6 months before start TB episode)	8.4 (29/345)	9.4	8.6
On ART > 6 months	9.3 (32/345)	13.6	10.0
Total cases	345	1400[a]	8300[a]

a combined HIV negative and HIV positive patients not on ART
a percentage calculated as average across 20 imputations.

7.3.6 Statistical analyses

7.3.6.1 Analyses with imputed datasets

All analyses with imputed datasets were done using the 'mim' prefix in Stata v10. This prefix informs Stata that the dataset is the result of multiple imputations and ensures that the uncertainty from the imputed values (i.e. variation between imputations in imputed value for each observation) is reflected in the estimated accuracy of the effect estimates. The main effect is that the confidence intervals are wider than in a regular analysis.

7.3.6.2 Relative incidence by HIV and ART status

The relative incidence of TB by HIV and ART status was expressed as the incidence rate ratio (IRR). IRR's were calculated using Poisson regression in Stata.

7.3.6.3 Trends in TB incidence in the general, HIV negative and HIV positive population

Trends in the incidence of new SS+ pulmonary TB were inspected visually and through regression analysis. Poisson regression and the lincom command were used to estimate TB incidence by period.

Piecewise Poisson regression was used to explore a change (or the absence thereof) in trends in the annual TB incidence between the period up to the introduction of ART in 2005 (1997 – 2005) and the period from the inclusion of ART (2005 – 2009). Piecewise regression fits 2 variables to describe the annual trend in TB incidence. The first variable described the trend in annual TB incidence for the period 1997 up to and including 2005. The second described the same trend but starting in 2005 running up to Aug 2009. These 2 variables describe 2 linear curves with a shared point in 2005 and thus allow for a 'dent' in the curve.

The exponentiated coefficient (i.e. IRR) for these piecewise regression variables represent the annual increase or decrease in TB incidence for that period, in which the first year is taken as the baseline. The contribution of a second variable to describe the annual trend was assessed using the Wald test.

7.3.6.4 Direct effect of ART on TB incidence trend in the overall and HIV positive population

The additional ART related cases are expected to affect the TB incidence in the population both directly as well as indirectly. There are the additional cases in the ART receiving population who are directly attributable to their treatment with anti-retroviral drugs. For example, without ART these patients would not have survived to develop their first TB episode. However, there will also be an indirect effect as these additional cases will increase Mtb transmission in the overall population, which in turn can lead to additional cases.

As a rough exploration of the direct effect of ART on TB trends I excluded all TB cases that had been on ART for more than 6 months. The reasoning is that these cases were severely immunocompromised at the time of starting ART, otherwise they would not have been eligible under the government ART eligibility criteria [193] which have followed WHO recommendations. It is likely that a proportion of these cases would have died soon afterwards if they had been started on ART, and so would not have developed active TB more than 6 months later.

Please note that some of these cases will actually be due to Mtb transmission from fellow ART receiving TB cases. The line between a direct and indirect becomes blurry in these cases.

7.3.7 Sensitivity analyses

7.3.7.1 SS+ versus All TB

I compared incidence trends in new pulmonary SS+ cases with trends in the incidence of all new TB. Imputations were run identically for the all new TB cases population, apart from the addition of a variable to indicate whether the patient was smear positive or not.

7.3.7.2 No decline in population HIV prevalence

The current model for population HIV prevalence allows for a decline in HIV incidence. Compared to a model that does not allow a decline the relative size of the HIV positive population will be higher in the later years compared to the HIV negative population. The relative reduction in population size also reduces the denominator of the incidence, thus increasing the incidence estimate.

I will explore the impact of not allowing a decrease in HIV incidence to the observed trends in TB incidence.

7.3.7.3 Effect of new studies

From 2007 onwards KPS started intensive screening for TB in admitted patients in the district hospital. This added a level of active case finding, and some of the cases found through this study might otherwise not have been diagnosed with TB and recorded in the KPS database. I explored the impact of cases identified this way on the proportion of TB that was SS+ and the effect on overall trends if I excluded these cases from the analysis.

7.3.7.4 Different methods of imputation

There is no hard and fast rule for setting up imputation models. I therefore explored the effect of imputing HIV alone for the period 1997-2004 and imputing HIV and ART separately for the 2005 – 2009 period.

7.4 Results

7.4.1 TB incidence by ART status (Aug 05 – July 09)

Table 7.5 shows the relative incidence of TB by HIV status (top half) in Karonga District for the 2005-2009 period. TB was much more likely in the HIV positive population compared to the HIV negative population (adjusted IRR (95% CI) = 10.89 (8.87 – 13.36)).

Patients recently started on ART experienced a high incidence of TB (observed and imputed incidence >1600/100,000/year) which was about 6 times the risk experienced by the HIV positive population not receiving ART (observed adjusted IRR (95% CI) = 6.41 (4.25 – 9.67)).

The incidence is lower in patients receiving ART for >6m, but still about 2.5 times higher than in the ART naive population (adjusted IRR (95% CI) = 2.46 (1.66 – 3.67)).

After imputation of HIV and ART status the estimated incidences for this period were increased, most notably in the HIV positive population (observed incidence = 377, versus imputed incidence = 416 cases/100,000/year) although not dramatically different. However, observed and imputed data gave similar results when assessing relative incidences between groups.

Table 7.5: Incidence of SS+ pulmonary TB in Karonga District by HIV/ART status (2005-2009)

	Observed					Imputed				
Period and group				IRR (95% CI)					IRR (95% CI)	
	n	py	inc	Crude	Adjusted[a]	n[c]	py[c]	Inc	Crude	Adjusted[a]
Jan 05- Aug 09										
Overall	462	633686	73	n.a.	n.a.	-	-	-	-	-
HIV negative	174	567555	31	1		186.7	567555	33	1	
HIV positive	249	66130	377	12.28	10.89	275.3	66130	416	12.66	11.44
				(10.12 – 14.91)	(8.87 – 13.36)				(10.45–15.33)	(9.33 – 14.03)
Jul 05 - Aug 09[b]										
HIV+ no ART	124	51602	240	1	1	149.4	51602	323	1	1
HIV+<=6m ART	29	1635	1774	7.38	6.41	35.6	1635	1810	7.48	6.51
				(4.93 – 11.06)	(4.25 – 9.67)				(4.73 – 11.81)	(4.08 – 10.39)
HIV+ >6m ART	32	4624	692	2.88	2.46	41.5	4624	940	3.10	2.66
				(1.95 – 4.25)	(1.66 – 3.67)				(2.12 – 4.53)	(1.79 – 3.94)

n = number of cases; py = person years at risk; inc = Incidence expressed as cases/100,000/year; IRR = Incidence Rate Ratio; n.a. not applicable as no comparison is made

Note: no imputations were done for the overall category. so no separate imputation results are given

a IRR's adjusted for age group (15-24, 25-34, 35-44, 45-54, >=55) and sex

b Period adjusted to opening of first ART clinic in Karonga District

c Numbers represent the average across all 20 imputations

7.4.2 Trends in TB incidence

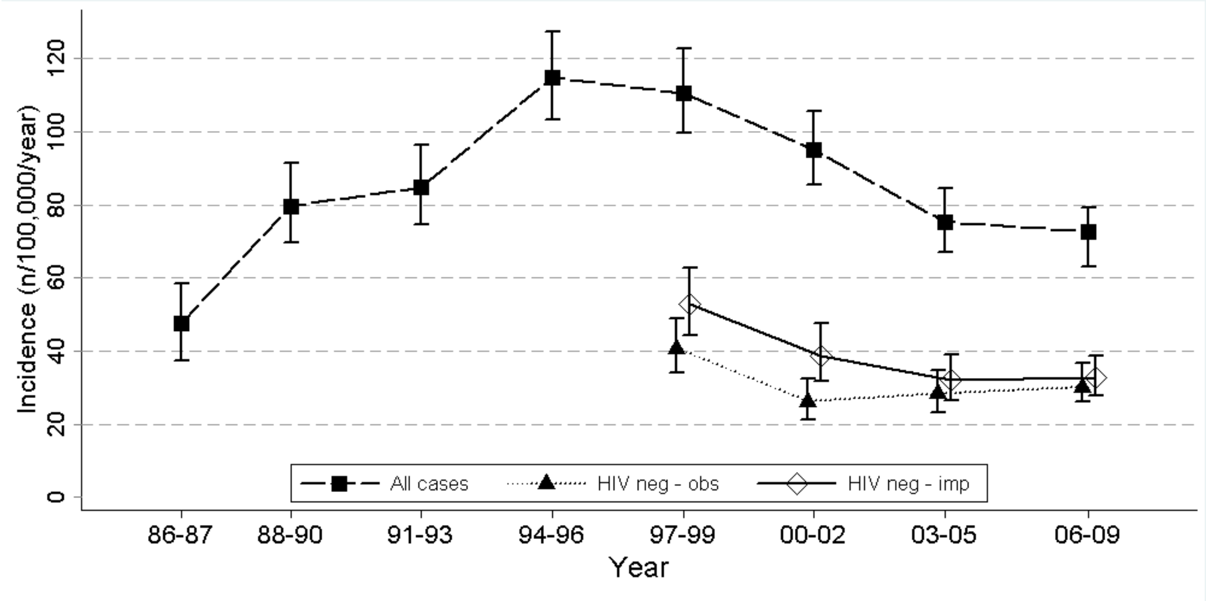
7.4.2.1 TB incidence in general population

In figure 7—2 the top line shows the incidence of all new SS+ cases of pulmonary TB in Karonga district since 1986. It clearly shows the previously reported steep increase through the late 80s and early 90s following the spread of HIV in the district. [171] Incidence peaked around 1995 at 115 cases /100,000 /year. Since then the incidence decreased strongly until the 2003 – 2005 period, after which the incidence appeared to reach a plateau with an average annual incidence between 75 /100,000 in the 2003 – 2005 period and 72/100,000 in the 2006 – 2009 period.

7.4.2.2 TB incidence in HIV negative population (1997 – 2009)

The bottom two curves in figure 7—2 show the incidence in the HIV negative population from 1997 onwards. The dotted line shows the incidence in patients with a known negative HIV status. The solid line shows the incidence after imputation of missing values for HIV status. The observed curve suggests that TB incidence in the HIV negative reached a low on 23 cases / 100,000 around 2002 and has slowly increased since then to 31/100,000 in the 2006 – 2009 period. However, the imputed data suggest that the incidence decreased until the 2003-2005 period, where incidence reached the lowest point 32 /100,000 and suggests stayed low at 33 / 100,000 in the 2006 – 2009 period.

Figure 7—2 Incidence of new SS+ pulmonary TB cases in Karonga District (all and HIV negative only)



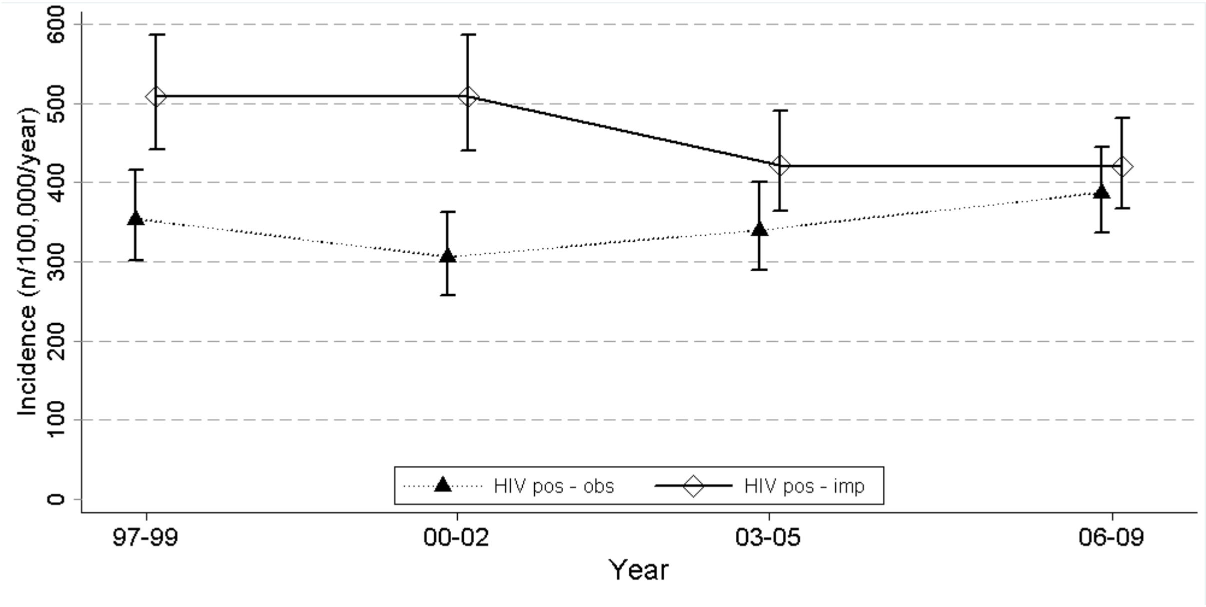
Graph shows observed values and after imputation of missing HIV status. Period markers show mid-point of time periods. Note: ART was introduced in Karonga district in mid 2005.

7.4.2.3 TB incidence in HIV positive population

In figure 7—3 I show the observed (dotted line) and imputed (solid line) incidence of new SS+ pulmonary TB in the HIV positive populaton. It shows how the variation of missing values for HIV status over the years (see table 7.2) skew the observed incidence trends.

The imputed trend suggests that the annual incidence in the HIV positive decreased strongly between the 2000 – 2002 and 2003 – 2005 periods from 509 to 421 cases /100,000. This decline appeared to become plateaued in the period after the introduction of ART.

Figure 7—3 Incidence of new SS+ pulmonary TB in HIV positive patients



Graph shows observed values and after imputation of missing HIV status. Period markers show mid-point of time periods. Note: ART was introduced in Karonga district in mid 2005.

Table 7.6 shows the incidence for the periods used in figures 7—2 and 7—3 for reference.

Table 7.6: TB incidence by period and HIV status

		Overall	HIV negative		HIV positive	
period			Observed	Imputed	Observed	Imputed
1986	1987	47 (37 – 58)				
1988	1990	79 (70 – 91)				
1991	1993	85 (75 – 96)				
1994	1996	115 (103 – 127)				
1997	1999	111 (100 – 123)	41 (34 – 49)	53 (44 – 63)	354 (301 – 416)	508 (441 – 586)
2000	2002	95 (85 – 106)	26 (21 – 33)	39 (32 – 47)	306 (257 – 363)	509 (440 – 589)
2003	2005	75 (67 – 85)	28 (23 – 35)	32 (27 – 39)	340 (289 – 400)	421 (364 – 490)
2006	2009	73 (66 – 81)	31 (26 – 36)	33 (28 – 39)	387 (337 – 445)	421 (367 – 481)

Incidence expressed as n/100,000/year.

7.4.3 Regression analyses

Table 7.7 shows an analysis of the linear trend in annual TB incidence allowing for a break in trend in 2005, the year that ART was introduced in Karonga District.

These analyses confirm the trends seen in figure 7—2 and 7—3. In the overall population TB incidence appeared to decrease roughly 7% annually in the 1997 – 2005 period (IRR (95% CI) = 0.93 (0.91 – 0.96)), after which the trend changed (p value Wald test = 0.03) and incidence plateaued. In the HIV negative population TB incidence declined between 1997 and 2005 (adjusted IRR (95% CI) = 0.91 (0.88 – 0.95)). In the period afterwards the data suggest that incidence was increasing (IRR (95% CI) = 1.09 (1.00 – 1.19)).For the HIV positive population the decline in the pre ART phase is shallower, around 3-4% annually, but in the post ART phase TB incidence plateaued (IRR (95% CI) = 1.00 (0.92 – 1.10)). Adjusting the estimates for age and sex had little effect the results.

The p values for a change in trend suggest that in both the Overall and HIV negative population the change in trend was statistically significant (p-value Wald test < 0.05).

Table 7.7 Trends in annual incidence of TB pre and post the introduction of ART

	Jan 1997 – Dec 2005 IRR (95% CI)	Jan 2005 - Aug 2009 IRR (95% CI)	P values[a]
Overall			
Crude	0.94 (0.92 – 0.96)	1.01 (0.96 – 1.07)	0.03
Adjusted[b]	0.93 (0.91 – 0.95)	1.01 (0.96 – 1.07)	0.03
HIV negative			
Crude	0.91 (0.88 – 0.95)	1.09 (1.00 – 1.19)	0.004
Adjusted[b]	0.91 (0.88 – 0.95)	1.09 (0.99 – 1.19)	0.005
HIV positive			
Crude	0.97 (0.94 – 1.00)	0.98 (0.91 – 1.06)	0.77
Adjusted[b]	0.96 (0.93 – 0.99)	0.99 (0.92 – 1.07)	0.58

Note: Coefficients express linear annual increase or decrease in TB incidence in that period, with the first year of the period (1997 or 2005) taken as baseline. Imputed datasets were used.
IRR = 'Incidence Rate Ratio'
a p-value for change in trend in 2005.
b IRR's adjusted for age group (15-24, 25-34, 35-44, 45-54, >=55) and sex

7.4.3.1 Direct effect of ART

Figure 7—4 shows the TB incidence trends if cases on ART for >6 months (32 in observed data, on average 41 in the 20 imputed datasets) were excluded from the trends analysis.

The curves show that without the TB cases in patients receiving ART for over 6 months the incidence in the HIV positive population would have continued to decrease strongly. As a consequence the overall incidence of TB also seems to continue to decrease.

Table 7.8 shows the incidence for the periods used in figure 7—4.

Table 7.8: TB incidence by period and HIV status after excluding TB cases on ART > 6 months

period		Overall		HIV negative		HIV positive	
1997	1999	111	(100 – 123)	48	(41 – 56)	526	(466 – 594)
2000	2002	95	(85 – 106)	36	(29 – 44)	483	(416 – 561)
2003	2005	75	(67 – 85)	31	(25 – 37)	413	(356 – 480)
2006	2009	65	(58 – 72)	33	(27 – 41)	326	(280 – 379)

Incidence expressed as n/100,000/year.
Note: HIV negative and HIV positive rates are post imputation for HIV and ART status

7.4.4 Sensitivity analyses

7.4.4.1 SS+ versus All TB

Table 7.9 shows the incidence of all new TB episodes diagnosed in Karonga District, regardless of foci of disease or laboratory confirmation. It shows that trends were similar to those in new SS+ pulmonary cases; a decrease up to 2005, after the decrease becomes more shallow (overall and HIV positive population) or TB incidence appears to increase after 2005 (HIV negative population).

Table 7.9: Incidence of all TB cases by period and HIV status

			HIV negative		HIV positive	
period		Overall	Observed	Imputed	Observed	Imputed
1986	1987	78 (66 – 93)				
1988	1990	191 (176 –209)				
1991	1993	197 (181 – 214)				
1994	1996	198 (183 – 215)				
1997	1999	267 (250 – 285)	66 (34 – 49)	101 (44 – 63)	804 (722 – 894)	1409 (1292 – 1537)
2000	2002	217 (202 – 233)	42 (21 – 33)	73 (32 – 47)	607 (537 – 685)	1275 (1163 – 1399)
2003	2005	157 (145 – 170)	44 (23 – 35)	53 (27 – 39)	734 (657 – 820)	986 (894 – 1089)
2006	2009	149 (139 – 160)	56 (26 – 36)	59 (28 – 39)	829 (754 – 911)	925 (846 – 1013)

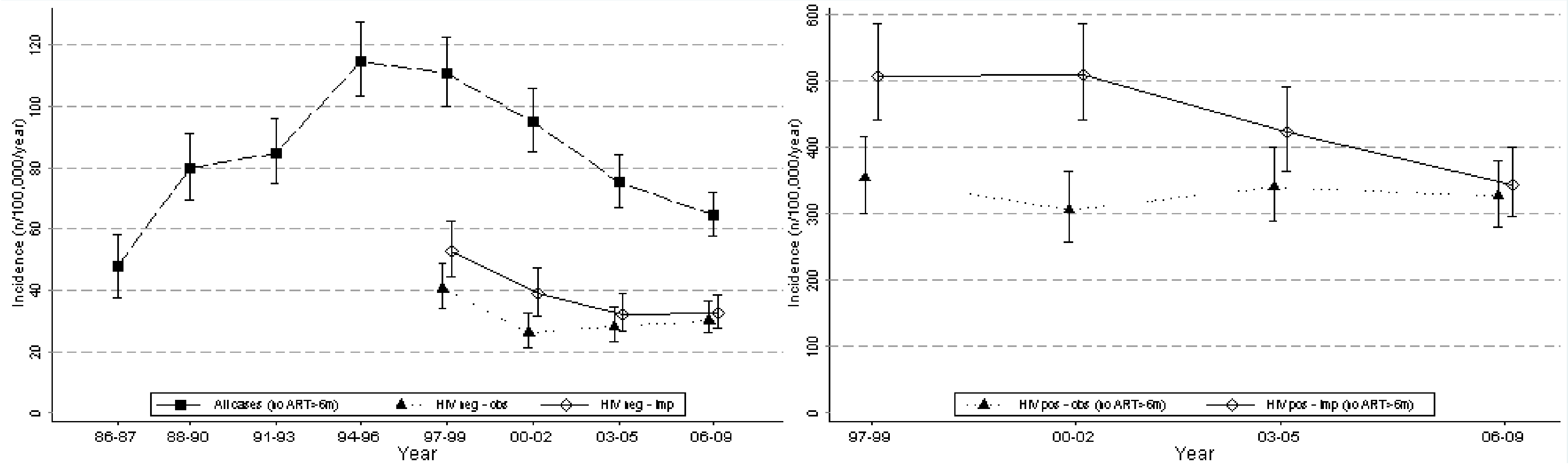
Incidence expressed as n/100,000/year.

7.4.4.2 Model assumptions for HIV prevalence in general population

7.4.4.2.1 No decline in incidence of HIV

Figure 7—5 shows the poor fit of the modelled HIV prevalence to known data points if no decline in HIV incidence was assumed. This suggests that the model with decline was more appropriate.

Figure 7—4: Trends in TB incidence after excluding TB cases with onset over 6 months after starting ART



Incidence expressed as n/100,000/year on y-axis

Figure 7—5: HIV prevalence in Karonga District (by sex) if no decline in HIV incidence is assumed

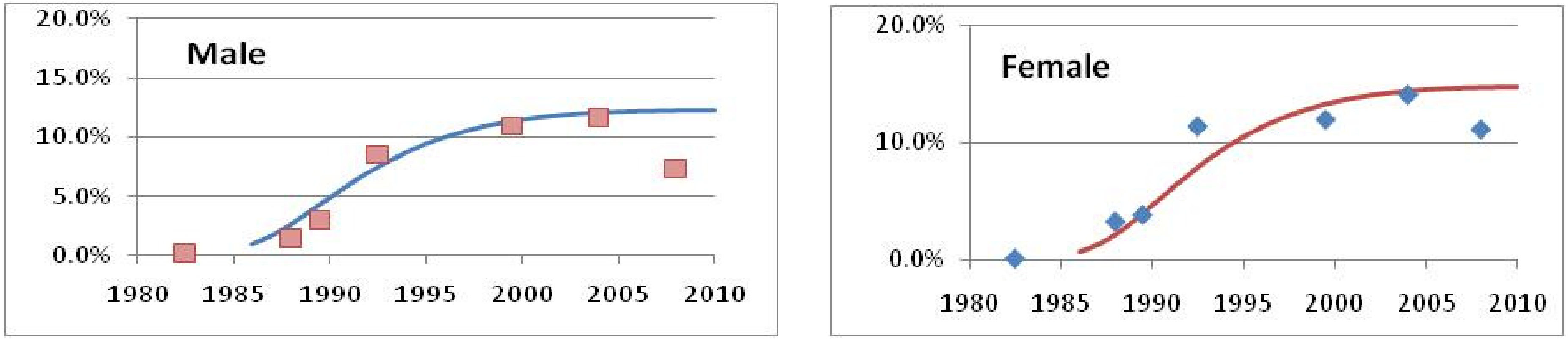


Figure 7—6 shows the incidence of TB by HIV if the prevalence of HIV in the general population is assumed to remain stable. This affects the denominators in the incidence analysis by HIV. For the HIV positive the denominator is increased, which results in lower incidence estimates. Under this assumption, the incidence actually continues to decline lightly, rather than plateau in the last period. For the HIV negative population the denominator is decreased, resulting in a more pronounced increase post ART. In short, changing the model assumption on HIV incidence would only make the change in incidence after 2005 more pronounced in the HIV negative, but less pronounced in the HIV positive.

Table 7.10 shows the results from the annual trend analyses. When these numbers are compared with those in table 7.7 the differences are small, and the trend in TB incidence that was observed in the main analysis with a break in the decline after 2005 is sustained.

Table 7.10 Trends in annual incidence of TB under the no decline in HIV incidence assumption

	Jan 1997 – Dec 2005 IRR (95% CI)	Jan 2005 - Aug 2009 IRR (95% CI)	P values[a]
Overall			
Adjusted[b]	0.94 (0.92 – 0.96)	0.97 (0.92 – 1.03)	0.24
HIV negative			
Adjusted[b]	0.92 (0.88 – 0.96)	1.06 (0.97 – 1.17)	0.02
HIV positive			
Adjusted[b]	0.97 (0.94 – 1.00)	0.95 (0.88 – 1.02)	0.71

Note: Coefficients express linear annual increase or decrease in TB incidence, with the first year of the period (1997 or 2005) taken as baseline. Imputed datasets were used.
IRR = 'Incidence Rate Ratio'
a p-value for 2nd variable describing linear trends in annual TB incidence
b IRR's adjusted for age group (15-24, 25-34, 35-44, 45-54, >=55) and sex

7.4.4.2.2 Prolonged survival in HIV positive patients receiving ART

The model used to estimate the HIV prevalence in the general population does not take into account the prolonged survival of patients starting on ART.

Without rerunning the model one can make some inference about the likely direction and magnitude of including these cases.

The HIV prevalence in the post 2005 period would increase compared to the original estimate. This would have been similar to the scenario explored in the previous section, as HIV prevalence is higher in the last years before increasing the HIV prevalence.

The magnitude of the effect depends on the number of patients receiving ART and their survival. In the extreme case, assuming complete survival, all patients starting ART in Karonga would be added cumulatively to the HIV positive population.

I simulated this scenario which results in a stable HIV prevalence between 2006 and 2009, and the results were similar to the models shown in figures 7—5 and 7—6 and table 7.10.

It is important to note that this assumption of complete survival on ART does not hold in real life. From the patient register data (described in Chapter 6) I recorded that one third of HIV patients starting ART died (or defaulted) in the first 3 years after starting ART (unpublished data). So in reality the effect of including ART is probably less strong.

7.4.4.3 Effect of new studies

Tables 7.11 and 7.12 and figure 7—7 look at the trends in new SS+ pulmonary TB incidence after the exclusion of TB patients who were first picked up from the general wards in Karonga District Hospital after active screening. During 2008 and 2009 a total of 34 cases were excluded for this sensitivity analysis. It shows that the trends were affected slightly and that the change in the annual trend of TB incidence post ART was less strong, but that the overall picture, as seen in the imputed data, was sustained.

Table 7.11: Incidence of new SS+ pulmonary TB cases by period and HIV status after exclusion of potential ICF cases

period		Overall	HIV negative[a]	HIV positive[a]
1997	1999	111 (100 – 123)	53 (44 – 63)	508 (441 – 586)
2000	2002	95 (85 – 106)	39 (32 – 47)	509 (440 – 587)
2003	2005	75 (67 – 85)	32 (27 – 39)	422 (364 – 490)
2006	2009	69 (62 – 77)	31 (27 – 37)	398 (347 – 458)

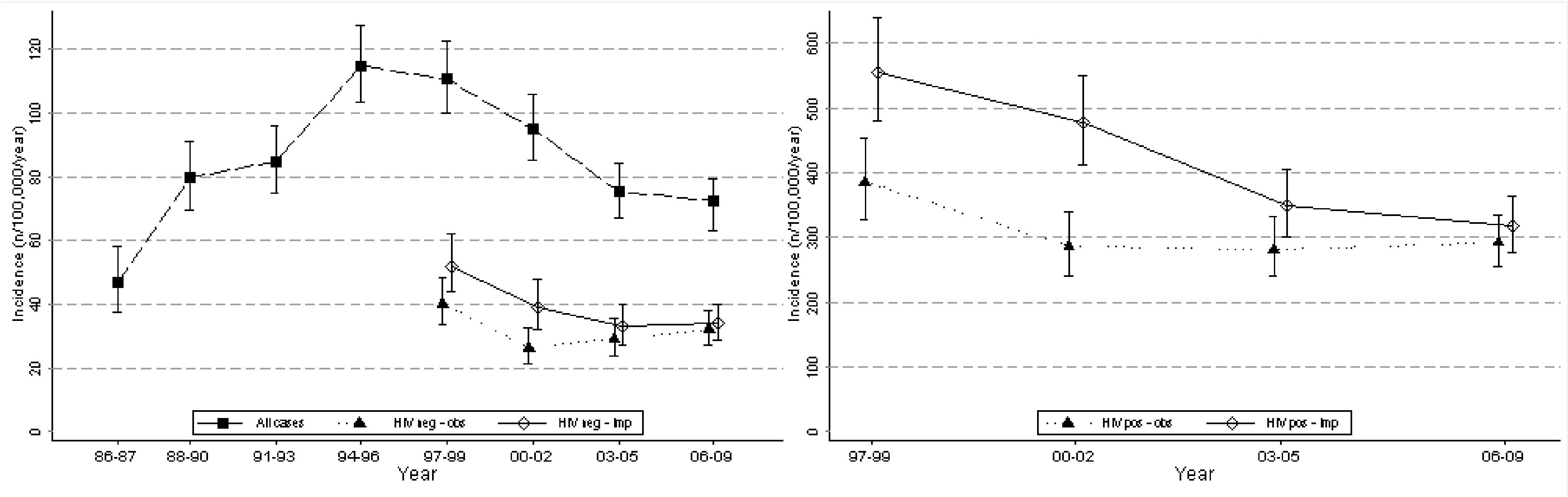
a all results based on imputed datasets

Table 7.12: Trends in annual incidence of new SS+ pulmonary TB pre and post the introduction of ART (excluding potential ICF cases)

	Jan 1997 – Dec 2005	Jan 2005 - Aug 2009	P values[a]
	IRR (95% CI)	IRR (95% CI)	
Overall			
Adjusted[b]	0.94 (0.92 – 0.96)	0.97 (0.92 – 1.03)	0.24
HIV negative			
Adjusted[b]	0.92 (0.88 – 0.96)	1.06 (0.97 – 1.16)	0.005
HIV positive			
Adjusted[b]	0.97 (0.94 – 1.00)	0.95 (0.88 – 1.03)	0.42

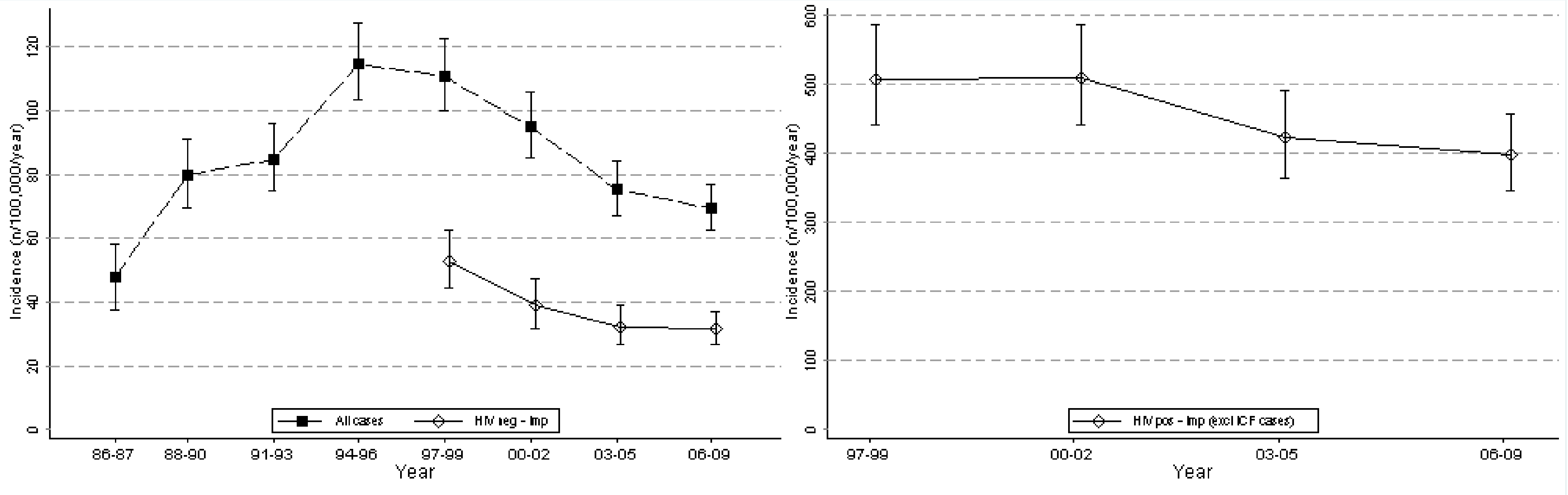
Note: Coefficients express linear annual increase or decrease in TB incidence, with the first year of the period (1997 or 2005) taken as baseline. Imputed datasets were used.
IRR = 'Incidence Rate Ratio'
a p-value for 2nd variable describing linear trends in annual TB incidence
b IRR's adjusted for age group (15-24, 25-34, 35-44, 45-54, >=55) and sex

Figure 7—6: TB incidence in HIV negative (left) and HIV positive (right) under 'no decline in HIV incidence' assumption



Note: Incidence of new SS+ pulmonary cases on y-axis expressed as n/100,000/year.

Figure 7—7: Incidence of smear positive pulmonary TB in HIV negative (left) and HIV positive (right) patients after excluding cases from ICF



Note: Incidence on y-axis expressed as n/100,000/year.

Table 7.13 shows the incidence of all new TB cases over the 1997 – August 2009 period. Again, a similar trend is seen as in the main analysis.

Table 7.13: Incidence of all new TB cases by period and HIV status after exclusion of potential ICF cases

period		Overall	HIV negative	HIV positive
1997	1999	267 (250 – 285)	101 (88 – 115)	1409 (1292 – 1537)
2000	2002	217 (202 – 232)	73 (62 – 85)	1275 (1163 – 1399)
2003	2005	157 (145 – 170)	53 (46 – 62)	986 (894 – 1089)
2006	2009	142 (132 – 153)	58 (51 – 65)	876 (799 – 961)

incidence expressed as n / 100,000 / year

7.4.4.4 Different methods of imputation

Applying different methods of imputation did not affect the results or trends in a relevant way.

7.5 Discussion

7.5.1 Summary of results

This is the first analysis of the effect of ART on the relative incidence of TB using high quality data collected from an uncontrolled ART programme setting. The analyses include data from ~3500 new SS+ TB cases with onset since 1986 of which >60 cases were receiving ART before their TB episode started. The results show that patients starting ART experience a very high risk of TB in the first 6 months. After 6 months the incidence comes down, but remains elevated compared to HIV positive/ART naive patients.

In the first analysis to couple TB surveillance data with population data stratified by HIV and ART status I also show that the incidence in Karonga District appears to have levelled off in the period after ART was rolled out after declining in the previous decade.

On a methodological note, this chapter shows that multiple imputation is useful to highlight and correct for biases due to missing values and is essential for incidence analyses where absolute values, rather than relative rates are the outcome of interest.

7.5.2 TB incidence in patients with ART

TB incidence in Karonga District appears to be about 7 times higher in patients on ART for 6 months or less compared to HIV positive/ART naive patients. This is probably a consequence of the severely compromised immune status of these patients when they are started on ART. Already at a high risk of developing TB disease, their high risk is compounded by the additional ‘TB unmasking’ effect of initiating ART.[45, 191, 195]

Patients that have been on ART for >6 months still suffer from an elevated risk of TB, about 3 times higher than HIV positive patients not receiving ART.

7.5.3 TB incidence in population after the introduction of ART

Between the mid 90s and 2005 the incidence of new SS+ pulmonary TB in Karonga District declined at an average rate of about 7% annually. This decline was seen and statistically significant overall and in both the HIV negative (strongest decline with 9% annually) and HIV positive (3% annual decline) population.

After the introduction of ART the overall incidence appears to have plateaued, which follows from an apparently flat or slightly increasing incidence trend in the HIV negative population and a shallower decrease in the HIV positive population.

The apparent plateau in the HIV positive population seems to be a direct effect of ART, as shown in figure 7—4. The plateau in the HIV negative population could be due to TB following recent transmission from the additional SS+, i.e. infectious, TB cases in the ART receiving group.

7.5.4 Interpretation

7.5.4.1 TB incidence

The data suggest that in the absence of ART a TB control programme following the basic WHO DOTS guidelines can successfully reduce TB incidence in settings with generalised HIV epidemics and moderate rates of TB and ongoing Mtb transmission. For example, HIV negative incidence has fallen below the overall level of 1986, when the effect of HIV was still moderate.

However, the steady decrease of TB was halted after the arrival of ART. A new group of patients at high to very high risk of TB is kept alive longer and will directly (see figure 7—4) and indirectly increase TB incidence in the population. The onward transmission of ART receiving TB cases will be within their own patient group during days that they collect their medication as well as within the general population in which they are again functioning.

It is difficult to quantify the exact size of this indirect effect from ART receiving TB cases. Currently used molecular epidemiological techniques are not sufficiently sensitive and flexible to plot complete transmission chains, i.e. answer the question of ‘who transmitted to whom?’. Whole genome sequencing which is developing fast may be better suited to answer this question, and has been tested in a cluster from the Netherlands. [201]

Also, one needs to consider that not all of the change in incidence trend is due to ART and that some can be due to the new studies (although the trends persisted after excluding those cases) and due to a natural levelling off of the decrease as a standard DOTS programme reaches the limit of its capabilities to control TB in a population like Karonga with moderate Mtb transmission and HIV prevalence.

It is important to note that during the study period there was no active Isoniazid Prophylactic Therapy (IPT) programme running in Karonga District. Several hundred TST positive HIV positive individuals with positive TST have been given IPT in the context of KPS studies, but this is unlikely to have had a relevant impact on the TB incidence trends.

7.5.4.2 *Imputation as a tool to reduce bias*

These analyses explore the effect of imputations. The datasets with imputed values appeared to give more valid values in the incidence trend analyses, and similar IRR's.

7.5.5 Limitations

As any population based study, there are limitations to the level and detail of the available data.

Firstly, KPS did not have the resources to record CD4 count for the complete TB case population (i.e. numerators), let alone from a representative sample of the HIV positive or ART receiving populations (i.e. the denominators). It is likely that the high incidence in the ART receiving population is similar to that seen in a population with the same CD4 cell count distribution. However, from a public health perspective these analyses identify potential high risk populations that can be targeted with intensified TB screening without involving CD4 count machines, which are often not available at the rural clinic level.

The annual population estimates are an approximation and not as precise as one would like. However, the data are anchored in 3 data points (1988, 1998 and 2008) which gives them reasonable strength.

The HIV prevalence follows from a model that fits a likely curve of HIV prevalence to data points. [186] Although the model's approach was based on current best practices, it is still applying assumptions which in turn add uncertainty. The sensitivity analyses showed however that the results were robust.

The data on the ART uptake from clinics in Karonga will be reasonably complete. One risk is that HIV positive patients from Karonga District were receiving ART from an ART clinic outside Karonga. These patients would not add to the denominator for ART, but would add to the numerator of ART associated TB incidence if they were diagnosed with or treated for TB in Karonga District which would lead to an overestimation of the incidence of TB in those categories.

The analyses of imputed datasets were corrected for the uncertainty from imputing missing data.

7.5.6 Setting specificity

These findings will be setting specific to an extent. The direct effect of ART will depend on the immune status of HIV positive patient when they are eligible for and started on ART. [36]

Higher ongoing rates of Mtb transmission in the population will also affect the rates of TB in the ART receiving population and subsequently their contribution to overall Mtb transmission. [202] This depends in part on the relative sensitivity to TB following recent Mtb transmission in the different populations. The ART receiving population is a special subgroup of the HIV positive population. As shown in Chapters 3, 4 and 5, HIV positive patients are more likely to develop TB disease from recent Mtb infection, and it is likely that ART receiving patients have a similar, if not more pronounced if they are started on ART with advanced immunocompromisation, sensitivity to recent Mtb infection.

Karonga District has moderate rates of ongoing Mtb transmission, like the majority of areas in Sub Saharan Africa. The HIV prevalence in the district is estimated to be around 11%, which is also moderate and representative of rural populations in Sub Saharan Africa. Finally, the ART roll out in Malawi and Karonga District is an example of a low tech programme delivering services to a large proportion of the eligible population and thus reflects the likely future situation in most areas in Sub Saharan Africa.

7.5.7 Public health implications and future research

7.5.7.1 Potential Interventions:

This work shows that patients receiving ART are at a high risk of developing TB. Integration of the TB and ART programmes to facilitate intensified case finding in this highly vulnerable population could help reduce the incidence of TB and additional ongoing Mtb transmission in the population.[203] One step forward is that the ventilation considerations recommended for TB clinics should be extended to ART clinics, to reduce the risk of nosomical Mtb infections.[184]

7.5.7.2 Future research

Future studies in Karonga should include measuring CD4 levels in all TB cases and estimate the CD4 count distribution in the population to allow for comparisons between groups in CD4 strata

The main question that remains however is the size and characteristics of the indirect effect of ART (and HIV infection) on TB incidence; to what extent do these extra cases contribute to ongoing Mtb transmission and to whom do they mainly transmit? This can be assessed through modelling studies that incorporate estimates of the relative infectiousness of both cases, or through the application of more flexible molecular epidemiological techniques such as whole genome sequencing, which will hopefully provide detailed transmission chains and provide an estimate of the indirect effect of ART and HIV on the TB rates in the HIV negative population.

8 Conclusion and Discussion

8.1 Overview

In this final chapter I will recap the results from each chapter and integrate them into the main conclusions that result from this thesis. For each conclusion I will summarise the existing knowledge, what this thesis had added and what the implications are. Finally I will describe future research that can build upon the results from this thesis.

8.2 Main results

8.2.1 Chapter 2 -TB clustering in populations

8.2.1.1 What was known

Previous studies had shown that the observed proportion clustered in a population is affected by or dependent on various factors related to study design, setting and population. [59, 77, 78].

8.2.1.2 What does this thesis add

In Chapter 2 I systematically reviewed the literature of population based studies on TB clustering. The results showed that although the observed proportion clustered varies widely between studies, a comparison between the observed and expected proportion clustered (based on factors describing study design and setting) can highlight true outliers.

8.2.1.3 What are the implications

Researchers that explore the proportion of cases in a cluster should make sure they interpret their results in the context of study design and setting, especially when informing policy makers.

8.2.2 Chapter 3-5: HIV and the risk of TB following recent infection

8.2.2.1 What was known?

From early observations it was assumed that TB in HIV positive patients mostly followed reactivation of latent Mtb infections. [141] However, evidence from study on recurrent TB [72, 73] and age patterns [142-146] suggested that TB in HIV positive patients was more likely to follow recent infection with Mtb.

8.2.2.2 What does this thesis add?

Chapters 3 and 4 and 5 focussed on the question of what is the dominant pathway to TB disease in HIV positive individuals. The pooled analysis of all suitable individual patient data in Chapter 3 suggested that TB following recent Mtb infection was the main mechanism, which was further strengthened by the time period analysis of KPS data in Chapter 4. Chapter 5 showed that even in those with a known prior infection a first TB episode in an HIV positive individual may often be due to re-infection. [175]

8.2.2.3 What are the implications?

The policy recommendations following from the results described in chapters 3 through 5 are described in their respective chapters. In short, if the majority of TB disease is the consequence of ongoing Mtb transmission new interventions should focus on interrupting that transmission. The WHO's "three I's" policy emphasize exactly that and should be taken forward, especially in the presence of a generalised HIV epidemic [165]; Intensified case finding in the community and high risk groups (e.g. patients receiving ART) to reduce the number and duration of undiagnosed infectious TB disease [183], infection control in hospitals [184] and Isoniazid prophylaxis for all HIV-positive individuals at risk of Mtb infection, not just those with latent infections. [185]

Each of these efforts should help reduce Mtb transmission, thus reducing TB incidence in the general and HIV positive population.

The results also have implication for vaccine development. If a previous Mtb infection does not confer protection from re-infection before their first episode (as Chapter 5 suggests), vaccines attempting to mimic or boost the body's immune response to an Mtb infection may not be as effective as hoped.

8.2.3 Chapter 7: TB incidence after the introduction of ART

8.2.3.1 What was known?

There are several studies describing TB incidence in patients receiving ART, [35-46, 192] which mostly show that TB incidence is high in the first 3-6 months after ART initiation and then decreases with time spent on ART. Also, when CD4 data were available, studies showed that patients with lower CD4 counts have a higher risk of developing TB. [35, 36, 38, 44, 45, 192] However, most of these studies described ART programmes with reasonable laboratory and clinical support, and none described the overall effect of the introduction of ART on the TB incidence in the overall population, i.e. including the HIV negative population.

8.2.3.2 What does this thesis add?

Chapter 7 described an analysis of the effect of ART on the incidence of TB, by HIV and ART status in the context of an ART programme in rural Africa with limited laboratory and clinical support. The results show that patients starting ART experience a very high risk of TB in the first 6 months, after which the risk decreases but remains elevated compared to HIV positive/ART naive patients. This is in line with observations from other studies.

In the second part of Chapter 7 I described the first analysis of how TB incidence trends in the general population were affected by the introduction of ART. The results suggested that a well supported DOTS programme can control TB incidence in a setting with generalised HIV epidemic and moderate Mtb transmission. However, the burden of additional cases and, probably, additional Mtb transmission, associated with the roll-out of ART has apparently halted this decline in TB incidence in Karonga District.

8.2.3.3 What are the implications?

It seems that starting ART at a severely immunosuppressed state, e.g. at a CD4 level of 200 cells/µL or less, leaves many patients at a prolonged high risk of TB disease.

However, the obvious theoretical solution of starting patients earlier on ART is not necessarily practical. Even though the most recent WHO ART policy recommends that patients are put on ART earlier, at CD4 count of <350 cells/µL [204], some immediate complications arise. Firstly, many of the rural clinics (or some of the more central clinics) do not have CD4 count facilities available, which complicates staging based on this qualification. Also, many ART programmes are already struggling to find the financial or human resources required to meet ART demand for the previously eligible population, and it is unclear where the additional support required to adequately serve an expanded population will come from. Furthermore, if these services become available it requires HIV positive patients to actually come for staging regularly so they are captured before their CD4 count drops (far) below 350. As this level of immunosuppression is less easily characterised by other symptoms such as opportunistic infections, it will probably require that patients come repeatedly for ART staging, which is not easy to achieve.

One approach that warrants discussion in my opinion is putting all patients that test positive on ART immediately, and thus avoid adding further barriers in the pathway from repeated HIV testing to treatment access. Obviously this would compound some of the practical issues above, but from a TB control perspective, the earlier HIV positive patients start on ART the lower the incidence of TB (and subsequently their indirect contribution to Mtb transmission in the general population) will be.

Another clear message from these results is the need for better integration of ART and TB services. [203] The work in this thesis confirms that patients receiving ART are a high risk population for TB, which is relatively easy to reach. Measures could include adding a TB specific checklist or regular sputum collections at every contact, especially during the first months after ART initiation.

One obvious integration (as suggested in section 8.2.2.3) is applying the recommendations on building ventilation for TB clinic buildings [184] to ART clinics. This would reduce the risk of infectious TB cases transmitting their Mtb strain to a vulnerable population of fellow patients receiving ART while both are attending the clinic.

8.3 Future work

8.3.1 Who transmits to whom in an IS*6110* RFLP defined cluster?

One of the main assumptions in this thesis is that HIV positive patients are less likely to be the source of onward Mtb transmission when part of a cluster. Although this assumption is supported by sufficient indirect [11, 12, 22, 29, 147-150] and some direct [71] evidence, it would be useful and interesting to have a more complete overview of who transmitted to whom within a cluster of TB cases defined by IS*6110* RFLP.

Ideally we would be able to directly map transmission chains. In some settings this has been attempted based on timing of the source and secondary case [10], but such data are difficult to interpret considering the highly variable period between Mtb infection and TB disease (estimated between 3-6 months and lifelong), especially in settings with high to moderate rates of ongoing Mtb transmission in the general population where the time of transmission is usually not known.

Whole genome sequencing, a newly developed tool for TB molecular epidemiology detects differences in the Mtb DNA to a much higher extent. It has been successfully used to construct a phylogenetic tree of a single cluster in the Netherlands. [201] When applied to the epidemiologically well described TB clusters in Karonga District, one could construct phylogenetic trees (and therefore likely transmission chains) for all cases and thus obtain a direct estimate of the relative risk of being a source of a secondary TB case by HIV and ART status, and whether/how this changed since 1997.

Genetic studies could also potentially lead to the identification of genetic markers in the Mtb genome that are associated with successful Mtb transmission (i.e. a new Mtb infection that leads to active TB disease), which could be translated into targets for vaccine or drug development.

8.3.2 Incidence by CD4 count strata

As discussed in Chapter 7, it would be interesting to stratify TB incidence estimates in the HIV positive and ART receiving populations by CD4 count. In future KPS studies it would be useful to record CD4 status for all TB cases. For comparison it would be good to have baseline CD4 counts for patients starting ART and an impression of the CD4 count for the general HIV positive population, e.g. by including CD4 count in the HIV test offered to the random TB controls described in chapter 6.

8.3.3 Intervention studies on Mtb transmission in patients with ART

Although it seems obvious that TB and ART services should be integrated better, how this can best be done in the context of small rural ART clinics is not clear. Multiple approaches can be tried, for example improve ventilation in ART clinics as discussed or introduce a more intensive form of TB case finding in this highly vulnerable population, especially during the first six months.

One would like to be able to quantify the effect of such interventions, which would require some thought; mostly of whether there is a comparison group. The problem is that the annual number of TB cases on ART in Karonga District each year is probably too low (less than 10, see chapter 7) to allow a standard trial design with a control group (i.e. control ART clinic). Also analyses in Chapter 7 suggested that TB incidence Karonga is changing, which complicates using it as a historical baseline group. Whole genome sequencing could potentially provide an indirect measure by estimating changes in the proportion of cases in each HIV/ART group that were the source of transmission for another case.

8.4 Conclusion

The work in this thesis has hopefully provided sufficient support for the hypothesis that HIV positive individuals are mainly at risk of TB following recent Mtb infection in settings with generalised HIV and TB epidemics.

The roll out of ART, although an essential lifeline to people living with HIV, has further complicated TB control in settings where the National TB programmes were already struggling to reduce TB incidence. Renewed efforts are required, including a better integration of ART delivery and regular TB services to make sure that the resources spent in one programme are not simply adding work on the other.

Under those conditions I believe that TB can be controlled by a well supported DOTS programme in most settings, as long as TB programmes are sufficiently resourced to consistently diagnose and successfully treat the majority of TB cases in the population. The work in this thesis suggests that DOTS can control TB incidence from exorbitant high to apparently pre-HIV levels (in the HIV negative) and will, with some modifications, probably be able to cope with the roll out of ART as well.

9 References

1. World Health Assembly. WHA 46.36. Tuberculosis Programme. Geneva: WHO, 1993
2. WHO. Global Tuberculosis Control; A short update to the 2009 report. Geneva: World Health Organisation, 2009
3. WHO. Guidelines for Second Generation HIV surveillance. Geneva: WHO/UNAIDS, 2000
4. American Thoracic Society. Targeted tuberculin testing and treatment of latent tuberculosis infection. Am J Respir Crit Care Med **2000**;161:S221-47
5. Richeldi L. An update on the diagnosis of tuberculosis infection. Am J Respir Crit Care Med **2006**;174:736-42
6. van Zyl-Smit RN, Zwerling A, Dheda K and Pai M. Within-subject variability of interferon-g assay results for tuberculosis and boosting effect of tuberculin skin testing: a systematic review. PLoS One **2009**;4:e8517
7. Havlir DV, Barnes PF. Tuberculosis in patients with human immunodeficiency virus infection. N Engl J Med **1999**;340:367-73
8. Styblo K. Epidemiology of Tuberculosis, Selected Papers Vol. 24. The Hague: Royal Netherlands Tuberculosis Association (KNCV), **1991**
9. Behr MA, Warren SA, Salamon H, et al. Transmission of Mycobacterium tuberculosis from patients smear-negative for acid-fast bacilli. Lancet **1999**;353:444-9
10. Tostmann A, Kik SV, Kalisvaart NA, et al. Tuberculosis transmission by patients with smear-negative pulmonary tuberculosis in a large cohort in the Netherlands. Clin Infect Dis **2008**;47:1135-42
11. Johnson JL, Vjecha MJ, Okwera A, et al. Impact of human immunodeficiency virus type-1 infection on the initial bacteriologic and radiographic manifestations of pulmonary tuberculosis in Uganda. Makerere University-Case Western Reserve University Research Collaboration. Int J Tuberc Lung Dis **1998**;2:397-404
12. Corbett EL, Bandason T, Cheung Y-B, et al. Prevalent infectious tuberculosis in Harare, Zimbabwe: burden, risk factors and implications for control. Int J Tuberc Lung Dis **2009**;13:1231-1237
13. Elliott AM, Halwiindi B, Hayes RJ, et al. The impact of human immunodeficiency virus on presentation and diagnosis of tuberculosis in a cohort study in Zambia. J Trop Med Hyg **1993**;96:1-11
14. Githui W, Nunn P, Juma E, et al. Cohort study of HIV-positive and HIV-negative tuberculosis, Nairobi, Kenya: comparison of bacteriological results. Tuber Lung Dis **1992**;73:203-9
15. Vynnycky E, Fine PE. The natural history of tuberculosis: the implications of age-dependent risks of disease and the role of reinfection. Epidemiol Infect **1997**;119:183-201
16. Holm J. Development from Tuberculosis infection to Tuberculosis Disease. TSRU Progress Report. The Hague, The Netherlands: KNCV, 1969
17. Sutherland I, Svandova E and Radhakrishna S. The development of clinical tuberculosis following infection with tubercle bacilli. . Tubercle **1982**;63:255-68
18. BCG and vole bacillus vaccines in the prevention of tuberculosis in adolescence and early adult life. Bull World Health Organ **1972**;46:371-85
19. Slutkin G, Leowski J and Mann J. The effects of the AIDS epidemic on the tuberculosis problem and tuberculosis programmes. Bull Int Union Tuberc Lung Dis **1988**;63:21-4
20. WHO. World Health Organization; Global TB database. Available at: http://www.who.int/tb/country/global_tb_database/en/index.html. Accessed at Feb 12th 2010, 2008
21. Corbett EL, Marston B, Churchyard GJ and De Cock KM. Tuberculosis in sub-Saharan Africa: opportunities, challenges, and change in the era of antiretroviral treatment. Lancet **2006**;367:926-37

22. Corbett EL, Watt CJ, Walker N, et al. The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. Arch Intern Med **2003**;163:1009-21
23. Sonnenberg P, Glynn JR, Fielding K, Murray J, Godfrey-Faussett P and Shearer S. HIV and pulmonary tuberculosis: the impact goes beyond those infected with HIV. Aids **2004**;18:657-62
24. Murray CJ, Salomon JA. Modeling the impact of global tuberculosis control strategies. Proc Natl Acad Sci U S A **1998**;95:13881-6
25. Dye C, Garnett GP, Sleeman K and Williams BG. Prospects for worldwide tuberculosis control under the WHO DOTS strategy. Directly observed short-course therapy. Lancet **1998**;352:1886-91
26. Currie CS, Williams BG, Cheng RC and Dye C. Tuberculosis epidemics driven by HIV: is prevention better than cure? Aids **2003**;17:2501-8
27. Williams BG, Granich R, Chauhan LS, Dharmshaktu NS and Dye C. The impact of HIV/AIDS on the control of tuberculosis in India. Proc Natl Acad Sci U S A **2005**;102:9619-24
28. Glynn JR, Murray J, Bester A, Nelson G, Shearer S and Sonnenberg P. Effects of duration of HIV infection and secondary tuberculosis transmission on tuberculosis incidence in the South African gold mines. Aids **2008**;22:1859-1867
29. Girardi E, Raviglione MC, Antonucci G, Godfrey-Faussett P and Ippolito G. Impact of the HIV epidemic on the spread of other diseases: the case of tuberculosis. Aids **2000**;14 Suppl 3:S47-56
30. Lienhardt C, Rodrigues LC. Estimation of the impact of the human immunodeficiency virus infection on tuberculosis: tuberculosis risks re-visited? Int J Tuberc Lung Dis **1997**;1:196-204
31. Odhiambo JA, Borgdorff MW, Kiambih FM, et al. Tuberculosis and the HIV epidemic: increasing annual risk of tuberculous infection in Kenya, 1986-1996. Am J Public Health **1999**;89:1078-82
32. Corbett EL, Charalambous S, Fielding K, et al. Stable incidence rates of tuberculosis (TB) among human immunodeficiency virus (HIV)-negative South African gold miners during a decade of epidemic HIV-associated TB. J Infect Dis **2003**;188:1156-63
33. Glynn JR. The impact of HIV infection on tuberculosis in Africa, **2007**
34. Williams BG, Dye C. Antiretroviral drugs for tuberculosis control in the era of HIV/AIDS. Science **2003**;301:1535-7
35. Girardi E, Antonucci G, Vanacore P, et al. Impact of combination antiretroviral therapy on the risk of tuberculosis among persons with HIV infection. Aids **2000**;14:1985-91
36. Lawn SD, Badri M and Wood R. Tuberculosis among HIV-infected patients receiving HAART: long term incidence and risk factors in a South African cohort. AIDS **2005**;19:2109-16
37. Bonnet MM, Pinoges LL, Varaine FF, et al. Tuberculosis after HAART initiation in HIV-positive patients from five countries with a high tuberculosis burden. AIDS **2006**;20:1275-9
38. Lawn SD, Myer L, Bekker LG and Wood R. Burden of tuberculosis in an antiretroviral treatment programme in sub-Saharan Africa: impact on treatment outcomes and implications for tuberculosis control. AIDS **2006**;20:1605-12
39. Moh R, Danel C, Messou E, et al. Incidence and determinants of mortality and morbidity following early antiretroviral therapy initiation in HIV-infected adults in West Africa. AIDS **2007**;21:2483-91
40. Moore D, Liechty C, Ekwaru P, et al. Prevalence, incidence and mortality associated with tuberculosis in HIV-infected patients initiating antiretroviral therapy in rural Uganda. AIDS **2007**;21:713-9
41. Dembele M, Saleri N, Carvalho AC, et al. Incidence of tuberculosis after HAART initiation in a cohort of HIV-positive patients in Burkina Faso. Int J Tuberc Lung Dis;14:318-23
42. Ledergerber B, Egger M, Erard V, et al. AIDS-related opportunistic illnesses occurring after initiation of potent antiretroviral therapy: the Swiss HIV Cohort Study. JAMA **1999**;282:2220-6

43. Jones JL, Hanson DL, Dworkin MS and DeCock KM. HIV-associated tuberculosis in the era of highly active antiretroviral therapy. The Adult/Adolescent Spectrum of HIV Disease Group. Int J Tuberc Lung Dis **2000**;4:1026-31
44. Kirk O, Gatell JM, Mocroft A, et al. Infections with Mycobacterium tuberculosis and Mycobacterium avium among HIV-infected patients after the introduction of highly active antiretroviral therapy. EuroSIDA Study Group JD. Am J Respir Crit Care Med **2000**;162:865-72
45. Badri M, Wilson D and Wood R. Effect of highly active antiretroviral therapy on incidence of tuberculosis in South Africa: a cohort study. Lancet **2002**;359:2059-64
46. Santoro-Lopes G, de Pinho AM, Harrison LH and Schechter M. Reduced risk of tuberculosis among Brazilian patients with advanced human immunodeficiency virus infection treated with highly active antiretroviral therapy. Clin Infect Dis **2002**;34:543-6
47. Cohn DL, O'Brien RJ. The use of restriction fragment length polymorphism (RFLP) analysis for epidemiological studies of tuberculosis in developing countries. Int J Tuberc Lung Dis **1998**;2:16-26
48. Small PM, Hopewell PC, Singh SP, et al. The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. N Engl J Med **1994**;330:1703-9
49. Borgdorff MW, van den Hof S, Kremer K, et al. Progress towards tuberculosis elimination: secular trend, immigration and transmission. Eur Respir J **2009**
50. ten Asbroek AH, Borgdorff MW, Nagelkerke NJ, et al. Estimation of serial interval and incubation period of tuberculosis using DNA fingerprinting. Int J Tuberc Lung Dis **1999**;3:414-20
51. Glynn JR, Crampin AC, Yates MD, et al. The importance of recent infection with Mycobacterium tuberculosis in an area with high HIV prevalence: a long-term molecular epidemiological study in Northern Malawi. J Infect Dis **2005**;192:480-7
52. van Embden JD, Cave MD, Crawford JT, et al. Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology. J Clin Microbiol **1993**;31:406-9
53. de Boer AS, Borgdorff MW, de Haas PE, Nagelkerke NJ, van Embden JD and van Soolingen D. Analysis of rate of change of IS6110 RFLP patterns of Mycobacterium tuberculosis based on serial patient isolates. J Infect Dis **1999**;180:1238-44
54. Lillebaek T, Dirksen A, Baess I, Strunge B, Thomsen VO and Andersen AB. Molecular evidence of endogenous reactivation of Mycobacterium tuberculosis after 33 years of latent infection. J Infect Dis **2002**;185:401-4
55. Warren RM, Victor TC, Streicher EM, et al. Patients with active tuberculosis often have different strains in the same sputum specimen. Am J Respir Crit Care Med **2004**;169:610-4
56. Fang R, Li X, Li J, et al. Mixed infections of Mycobacterium tuberculosis in tuberculosis patients in Shanghai, China. Tuberculosis (Edinb) **2008**;88:469-73
57. Das S, Narayanan S, Hari L, et al. Simultaneous infection with multiple strains of Mycobacterium tuberculosis identified by restriction fragment length polymorphism analysis. Int J Tuberc Lung Dis **2004**;8:267-70
58. van Rie A, Victor TC, Richardson M, et al. Reinfection and mixed infection cause changing Mycobacterium tuberculosis drug-resistance patterns. Am J Respir Crit Care Med **2005**;172:636-42
59. Glynn JR, Bauer J, de Boer AS, et al. Interpreting DNA fingerprint clusters of Mycobacterium tuberculosis. European Concerted Action on Molecular Epidemiology and Control of Tuberculosis. Int J Tuberc Lung Dis **1999**;3:1055-60
60. Daley CL, Small PM, Schecter GF, et al. An outbreak of tuberculosis with accelerated progression among persons infected with the human immunodeficiency virus. An analysis using restriction-fragment-length polymorphisms. N Engl J Med **1992**;326:231-5

61. Ikeda RM, Birkhead GS, DiFerdinando GT, Jr., et al. Nosocomial tuberculosis: an outbreak of a strain resistant to seven drugs. Infect Control Hosp Epidemiol **1995**;16:152-9
62. Lemaitre N, Sougakoff W, Truffot-Pernot C, et al. Use of DNA fingerprinting for primary surveillance of nosocomial tuberculosis in a large urban hospital: Detection of outbreaks in homeless people and migrant workers. International Journal of Tuberculosis & Lung Disease **1998**;2:390-396
63. van Deutekom H, Hoijng SP, de Haas PE, et al. Clustered tuberculosis cases: do they represent recent transmission and can they be detected earlier? Am J Respir Crit Care Med **2004**;169:806-10
64. Edlin BR, Tokars JI, Grieco MH, et al. An outbreak of multidrug-resistant tuberculosis among hospitalized patients with the acquired immunodeficiency syndrome. N Engl J Med **1992**;326:1514-21
65. Jasmer RM, Roemer M, Hamilton J, et al. A prospective, multicenter study of laboratory cross-contamination of Mycobacterium tuberculosis cultures. Emerg Infect Dis **2002**;8:1260-3
66. Glynn JR, Yates MD, Crampin AC, et al. DNA fingerprint changes in tuberculosis: reinfection, evolution, or laboratory error? J Infect Dis **2004**;190:1158-66
67. Martin A, Inigo J, Chaves F, et al. Re-analysis of epidemiologically linked tuberculosis cases not supported by IS6110-RFLP-based genotyping. Clin Microbiol Infect **2009**;15:763-9
68. Cook VJ, Stark G, Roscoe DL, Kwong A and Elwood RK. Investigation of suspected laboratory cross-contamination: interpretation of single smear-negative, positive cultures for Mycobacterium tuberculosis. Clin Microbiol Infect **2006**;12:1042-5
69. Borgdorff MW, van der Werf MJ, de Haas PE, Kremer K and van Soolingen D. Tuberculosis elimination in the Netherlands. Emerg Infect Dis **2005**;11:597-602
70. Cattamanchi A, Hopewell PC, Gonzalez LC, et al. A 13-year molecular epidemiological analysis of tuberculosis in San Francisco. Int J Tuberc Lung Dis **2006**;10:297-304
71. Crampin AC, Glynn JR, Traore H, et al. Tuberculosis transmission attributable to close contacts and HIV status, Malawi. Emerg Infect Dis **2006**;12:729-35
72. Crampin AC, Mwaungulu JN, mwaungulu FD, et al. Recurrent TB: relapse of reinfection? The effect of HIV in general population cohort in Malawi. AIDS **2010**;24:417-26
73. Sonnenberg P, Murray J, Glynn JR, Shearer S, Kambashi B and Godfrey-Faussett P. HIV-1 and recurrence, relapse, and reinfection of tuberculosis after cure: a cohort study in South African mineworkers. Lancet **2001**;358:1687-93
74. Houben RMGJ, Glynn JR. A systematic review and meta-analysis of molecular epidemiological studies of tuberculosis: development of a new tool to aid interpretation. Tropical Medicine & International Health **2009**;14:892-909
75. Hermans PW, Messadi F, Guebrexabher H, et al. Analysis of the population structure of Mycobacterium tuberculosis in Ethiopia, Tunisia, and The Netherlands: usefulness of DNA typing for global tuberculosis epidemiology. J Infect Dis **1995**;171:1504-13
76. Bishai WR, Graham NMH, Harrington S, et al. Molecular and geographic patterns of tuberculosis transmission after 15 years of directly observed therapy. Jama **1998**;280:1679-1684
77. Murray M. Determinants of cluster distribution in the molecular epidemiology of tuberculosis. Proc Natl Acad Sci U S A **2002**;99:1538-43
78. Glynn JR, Vynnycky E and Fine PE. Influence of sampling on estimates of clustering and recent transmission of Mycobacterium tuberculosis derived from DNA fingerprinting techniques. Am J Epidemiol **1999**;149:366-71
79. van Soolingen D, Borgdorff MW, de Haas PE, et al. Molecular epidemiology of tuberculosis in the Netherlands: a nationwide study from 1993 through 1997. J Infect Dis **1999**;180:726-36
80. Jasmer RM, Hahn JA, Small PM, et al. A molecular epidemiologic analysis of tuberculosis trends in San Francisco, 1991-1997. Ann Intern Med **1999**;130:971-8

81. Burman WJ, Reves RR, Hawkes AP, et al. DNA fingerprinting with two probes decreases clustering of Mycobacterium tuberculosis. American Journal of Respiratory & Critical Care Medicine **1997**;155:1140-1146
82. Vynnycky E, Borgdorff MW, van Soolingen D and Fine PE. Annual Mycobacterium tuberculosis infection risk and interpretation of clustering statistics. Emerg Infect Dis **2003**;9:176-83
83. Vynnycky E, Fine PE. The annual risk of infection with Mycobacterium tuberculosis in England and Wales since 1901. Int J Tuberc Lung Dis **1997**;1:389-96
84. Vynnycky E, Nagelkerke N, Borgdorff MW, van Soolingen D, van Embden JD and Fine PE. The effect of age and study duration on the relationship between 'clustering' of DNA fingerprint patterns and the proportion of tuberculosis disease attributable to recent transmission. Epidemiol Infect **2001**;126:43-62
85. Dahle UR, Sandven P, Heldal E and Caugant DA. Continued low rates of transmission of Mycobacterium tuberculosis in Norway. Journal of Clinical Microbiology **2003**;41:2968-2973
86. Dahle UR, Eldholm V, Winje BA, Mannsaker T and Heldal E. Impact of immigration on the molecular epidemiology of Mycobacterium tuberculosis in a low-incidence country. Am J Respir Crit Care Med **2007**;176:930-5
87. Hernandez-Gardun~o E, Kunimoto D, Wang L, et al. Predictors of clustering of tuberculosis in Greater Vancouver: A molecular epidemiologic study. CMAJ: Canadian Medical Association Journal **2002**;167:349-352
88. Fok A, Numata Y, Schulzer M and Fitzgerald MJ. Risk factors for clustering of tuberculosis cases: a systematic review of population-based molecular epidemiology studies. Int J Tuberc Lung Dis **2008**;12:480-92
89. Nava-Aguilera E, Andersson N, Harris E, et al. Risk factors associated with recent transmission of tuberculosis: systematic review and meta-analysis. Int J Tuberc Lung Dis **2009**;13:17-26
90. Houben RM, Glynn JR, Fok A, Numata Y, Schulzer M and Fitzgerald JM. Systematic review and analysis of population-based molecular epidemiological studies. Int J Tuberc Lung Dis **2009**;13:275-6
91. CDC. Centers for Disease Control and Prevention; tb-update list. Available at http://listmanager.aspensys.com/read/?forum=tb-update. Accessed at November 2006. **2006**
92. Kamerbeek J, Schouls L, Kolk A, et al. Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. J Clin Microbiol **1997**;35:907-14
93. van Soolingen D, de Haas PE, Hermans PW, Groenen PM and van Embden JD. Comparison of various repetitive DNA elements as genetic markers for strain differentiation and epidemiology of Mycobacterium tuberculosis. J Clin Microbiol **1993**;31:1987-95
94. Harbord RS, Thomas. METAREG: Stata module to perform meta-analysis regression. Statistical Software Components S446201 (Available at http://ideas.repec.org/c/boc/bocode/s446201.html>, or type "ssc install metareg" to install from within Stata). Boston College: Department of Economics, 2004
95. Harbord RM, Higgins JP. Meta–regression in Stata. The Stata Journal **2008**;8:493-519
96. Berkey CS, Hoaglin DC, Mosteller F and Colditz GA. A random-effects regression model for meta-analysis. Stat Med **1995**;14:395-411
97. Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? Stat Med **2002**;21:1559-73
98. Matteelli A, Gori A, Pinsi G, et al. Clustering of tuberculosis among Senegalese immigrants in Italy. International Journal of Tuberculosis & Lung Disease **2003**;7:967-972
99. Ruiz Garcia M, Rodriguez JC, Navarro JF, Samper S, Martin C and Royo G. Molecular epidemiology of tuberculosis in Elche, Spain: A 7-year study. Journal of Medical Microbiology **2002**;51:273-277

100. Verver S, Warren RM, Munch Z, et al. Transmission of tuberculosis in a high incidence urban community in South Africa. International Journal of Epidemiology **2004**;33:351-357
101. Braden CR, Templeton GL, Cave MD, et al. Interpretation of restriction fragment length polymorphism analysis of Mycobacterium tuberculosis isolates from a state with a large rural population. Journal of Infectious Diseases **1997**;175:1446-1452
102. Cave MD, Yang ZH, Stefanova R, et al. Epidemiologic import of tuberculosis cases whose isolates have similar but not identical IS6110 restriction fragment length polymorphism patterns. Journal of Clinical Microbiology **2005**;43:1228-1233
103. Pena MJ, Caminero JA, Campos-Herrero MI, et al. Epidemiology of tuberculosis on Gran Canaria: a 4 year population study using traditional and molecular approaches. Thorax **2003**;58:618-22
104. Blenkush M, Kunimoto D, Black W, Elwood RK and FitzGerald JM. Evidence for TB clustering in Vancouver: results from pilot study using RFLP fingerprinting. Can Commun Dis Rep **1996**;22:49-51
105. Storla DG, Rahim Z, Islam MA, et al. Heterogeneity of Mycobacterium tuberculosis isolates in Sunamganj District, Bangladesh. Scand J Infect Dis **2006**;38:593-6
106. Fujikane T, Fujiuchi S, Yamazaki Y, et al. Molecular epidemiology of tuberculosis in the north Hokkaido district of Japan. International Journal of Tuberculosis & Lung Disease **2004**;8:39-44
107. Murray M. Sampling bias in the molecular epidemiology of tuberculosis. Emerg Infect Dis **2002**;8:363-9
108. Soborg C, Soborg B, Pouelsen S, Pallisgaard G, Thybo S and Bauer J. Doubling of the tuberculosis incidence in Greenland over an 8-year period (1990-1997). International Journal of Tuberculosis & Lung Disease **2001**;5:257-265
109. Thomsen VO, Lillebaek T and Stenz F. Tuberculosis in Greenland--current situation and future challenges. Int J Circumpolar Health **2004**;63 Suppl 2:225-9
110. Zolnir-Dovc M, Poljak M, Erzen D and Sorli J. Molecular epidemiology of tuberculosis in Slovenia: Results of a one-year (2001) nation-wide study. Scandinavian Journal of Infectious Diseases **2003**;35:863-868
111. Lillebaek T, Dirksen A, Kok-Jensen A and Andersen AB. A dominant Mycobacterium tuberculosis strain emerging in Denmark. International Journal of Tuberculosis & Lung Disease **2004**;8:1001-1006
112. Vachee A, Vincent P, Savage C, et al. Molecular epidemiology of tuberculosis in the Nord department of France during 1995. Tubercle & Lung Disease **1999**;79:361-366
113. Diel R, Seidler A, Nienhaus A, Rusch-Gerdes S and Niemann S. Occupational risk of tuberculosis transmission in a low incidence area. Respir Res **2005**;6:35
114. Maguire H, Dale JW, McHugh TD, et al. Molecular epidemiology of tuberculosis in London 1995-7 showing low rate of active transmission. Thorax **2002**;57:617-622
115. Moro ML, Salamina G, Gori A, et al. Two-year population-based molecular epidemiological study of tuberculosis transmission in the Metropolitan area of Milan, Italy. European Journal of Clinical Microbiology & Infectious Diseases **2002**;21:114-122
116. Dahle UR, Sandven P, Heldal E and Caugant DA. Molecular epidemiology of Mycobacterium tuberculosis in Norway. Journal of Clinical Microbiology **2001**;39:1802-1807
117. Lari N, Rindi L, Sola C, et al. Genetic diversity, determined on the basis of katG463 and gyrA95 polymorphisms, spoligotyping, and IS6110 typing, of Mycobacterium tuberculosis complex isolates from Italy. Journal of Clinical Microbiology **2005**;43:1617-1624
118. Samper S, Iglesias MJ, Rabanaque MJ, et al. The molecular epidemiology of tuberculosis in Zaragoza, Spain: A retrospective epidemiological study in 1993. International Journal of Tuberculosis & Lung Disease **1998**;2:281-287

119. Pfyffer GE, Strassle A, Rose N, Wirth R, Brandli O and Shang H. Transmission of tuberculosis in the metropolitan area of Zurich: A 3 year survey based on DNA fingerprinting. European Respiratory Journal **1998**;11:804-808
120. Kempf MC, Dunlap NE, Lok KH, Benjamin WH, Jr., Keenan NB and Kimerling ME. Long-term molecular analysis of tuberculosis strains in alabama, a state characterized by a largely indigenous, low-risk population. J Clin Microbiol **2005**;43:870-8
121. Kunimoto D, Sutherland K, Wooldrage K, et al. Transmission characteristics of tuberculosis in the foreign-born and the Canadian-born populations of Alberta, Canada. International Journal of Tuberculosis & Lung Disease **2004**;8:1213-1220
122. De Bruyn G, Adams GJ, Teeter LD, Soini H, Musser JM and Graviss EA. The contribution of ethnicity to Mycobacterium tuberculosis strain clustering. International Journal of Tuberculosis & Lung Disease **2001**;5:633-641
123. Blackwood KS, Al-Azem A, Elliott LJ, Hershfield ES and Kabani AM. Conventional and molecular epidemiology of Tuberculosis in Manitoba. BMC Infectious DIseases **2003**;3:11
124. Blackwood KS, Wolfe JN and Kabani AM. Application of mycobacterial interspersed repetitive unit typing to Manitoba tuberculosis cases: Can restriction fragment length polymorphism be forgotten? Journal of Clinical Microbiology **2004**;42:5001-5006
125. Cronin WA, Golub JE, Magder LS, et al. Epidemiologic usefulness of spoligotyping for secondary typing of Mycobacterium tuberculosis isolates with low copy numbers of IS6110. Journal of Clinical Microbiology **2001**;39:3709-3711
126. Sharnprapai S, Miller AC, Suruki R, et al. Genotyping analyses of tuberculosis cases in U.S.-and foreign-born Massachusetts residents. Emerging Infectious Diseases **2002**;8:1239-1245
127. Scott AN, Menzies D, Tannenbaum TN, et al. Sensitivities and specificities of spoligotyping and mycobacterial interspersed repetitive unit-variable-number tandem repeat typing methods for studying molecular epidemiology of tuberculosis. Journal of Clinical Microbiology **2005**;43:89-94
128. Frieden TR, Woodley CL, Crawford JT, Lew D and Dooley SM. The molecular epidemiology of tuberculosis in New York City: The importance of nosocomial transmission and laboratory error. Tubercle & Lung Disease **1996**;77:407-413
129. Burgos M, DeRiemer K, Small PM, Hopewell PC and Daley CL. Effect of Drug Resistance on the Generation of Secondary Cases of Tuberculosis. Journal of Infectious Diseases **2003**;188:1878-1884
130. Weis SE, Pogoda JM, Yang Z, et al. Transmission dynamics of tuberculosis in Tarrant county, Texas. Am J Respir Crit Care Med **2002**;166:36-42
131. Cowan LS, Diem L, Monson T, et al. Evaluation of a two-step approach for large-scale, prospective genotyping of Mycobacterium tuberculosis isolates in the United States. Journal of Clinical Microbiology **2005**;43:688-695
132. Chan-Yeung M, Kam KM, Leung CC, et al. Population-based prospective molecular and conventional epidemiological study of tuberculosis in Hong Kong. Respirology **2006**;11:442-448
133. Wilkinson D, Pillay M, Crump J, Lombard C, Davies GR and Sturm AW. Molecular epidemiology and transmission dynamics of Mycobacterium tuberculosis in rural Africa. Tropical Medicine & International Health **1997**;2:747-753
134. Dale JW, Nor RM, Ramayah S, Tang TH and Zainuddin ZF. Molecular epidemiology of tuberculosis in Malaysia. Journal of Clinical Microbiology **1999**;37:1265-1268
135. Park YK, Bai GH and Kim SJ. Restriction fragment length polymorphism analysis of Mycobacterium tuberculosis isolated from countries in the western pacific region. Journal of Clinical Microbiology **2000**;38:191-197

136. Das SD, Narayanan S, Hari L, et al. Differentiation of highly prevalent IS6110 single-copy strains of Mycobacterium tuberculosis from a rural community in South India with an ongoing DOTS programme. Infection, Genetics & Evolution **2005**;5:67-77
137. Jimenez-Corona ME, Garcia-Garcia L, DeRiemer K, et al. Gender differentials of pulmonary tuberculosis transmission and reactivation in an endemic area. Thorax **2006**;61:348-353
138. Asgharzadeh M, Shahbabian K, Majidi J, et al. IS6110 restriction fragment length polymorphism typing of Mycobacterium tuberculosis isolates from East Azerbaijan Province of Iran. Mem Inst Oswaldo Cruz **2006**;101:517-21
139. Houben RMGJ, Crampin AC, Ndlovu R, et al. HIV increases the risk of TB due to recent infection, but the effect varies with age. Research in Progress. London: Royal Society of Tropical Medicine and Hygiene, 2008
140. Houben RMGJ, Crampin AC, Ndlovu R, et al. HIV increases the risk of TB due to recent infection more than that due to reactivation of latent infection. ***submitted***
141. Selwyn PA, Hartel D, Lewis VA, et al. A prospective study of the risk of tuberculosis among intravenous drug users with human immunodeficiency virus infection. N Engl J Med **1989**;320:545-50
142. Houston S, Ray S, Mahari M, et al. The association of tuberculosis and HIV infection in Harare, Zimbabwe. Tuber Lung Dis **1994**;75:220-6
143. Long R, Scalcini M, Manfreda J, et al. Impact of human immunodeficiency virus type 1 on tuberculosis in rural Haiti. Am Rev Respir Dis **1991**;143:69-73
144. Crampin AC, Glynn JR, Floyd S, et al. Tuberculosis and gender: exploring the patterns in a case control study in Malawi. Int J Tuberc Lung Dis **2004**;8:194-203
145. Van den Broek J, Borgdorff MW, Pakker NG, et al. HIV-1 infection as a risk factor for the development of tuberculosis: a case-control study in Tanzania. Int J Epidemiol **1993**;22:1159-65
146. Chum HJ, O'Brien RJ, Chonde TM, Graf P and Rieder HL. An epidemiological study of tuberculosis and HIV infection in Tanzania, 1991-1993. AIDS **1996**;10:299-309
147. Espinal MA, Perez EN, Baez J, et al. Infectiousness of Mycobacterium tuberculosis in HIV-1-infected patients with tuberculosis: a prospective study. Lancet **2000**;355:275-80
148. Elliott AM, Luo N, Tembo G, et al. Impact of HIV on tuberculosis in Zambia: a cross sectional study. BMJ **1990**;301:412-5
149. Kenyon TA, Creek T, Laserson K, et al. Risk factors for transmission of Mycobacterium tuberculosis from HIV-infected tuberculosis patients, Botswana. Int J Tuberc Lung Dis **2002**;6:843-50
150. Carvalho AC, DeRiemer K, Nunes ZB, et al. Transmission of Mycobacterium tuberculosis to contacts of HIV-infected tuberculosis patients. Am J Respir Crit Care Med **2001**;164:2166-71
151. Nelson LJ, Wells CD. Global epidemiology of childhood tuberculosis. Int J Tuberc Lung Dis **2004**;8:636-47
152. Kirkwood BR, Sterne JA. Analysis of clustered data. Essential Medical Statistics. 2nd Edition ed. Oxford: Blackwell Publishing Company, **2003**
153. Higgins JP, Green S. Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.0 ed: The Cochrane Collaboration, 2008
154. Higgins JP, Thompson SG, Deeks JJ and Altman DG. Measuring inconsistency in meta-analyses. Bmj **2003**;327:557-60
155. Ioannidis JP, Patsopoulos NA and Rothstein HR. Reasons or excuses for avoiding meta-analysis in forest plots. Bmj **2008**;336:1413-5
156. Simmonds MC, Higgins JP, Stewart LA, Tierney JF, Clarke MJ and Thompson SG. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. Clin Trials **2005**;2:209-17

157. Bruchfeld J, Aderaye G, Palme IB, et al. Molecular epidemiology and drug resistance of Mycobacterium tuberculosis isolates from ethiopian pulmonary tuberculosis patients with and without human immunodeficiency virus infection. Journal of Clinical Microbiology **2002**;40:1636-1643
158. Yang ZH, Mtoni I, Chonde M, et al. DNA fingerprinting and phenotyping of Mycobacterium tuberculosis isolates from human immunodeficiency virus (HIV)-seropositive and HIV-seronegative patients in Tanzania. Journal of Clinical Microbiology **1995**;33:1064-1069
159. Godfrey-Faussett P, Sonnenberg P, Shearer SC, et al. Tuberculosis control and molecular epidemiology in a South African gold-mining community. Lancet **2000**;356:1066-71
160. Haas WH, Engelmann G, Amthor B, et al. Transmission dynamics of tuberculosis in a high-incidence country: Prospective analysis by PCR DNA fingerprinting. Journal of Clinical Microbiology **1999**;37:3975-3979
161. Lockman S, Sheppard JD, Mwasekaga M, et al. DNA fingerprinting of a national sample of Mycobacterium tuberculosis isolates, Botswana, 1995-1996. International Journal of Tuberculosis & Lung Disease **2000**;4:584-587
162. Houben RMGJ, Crampin AC, Mallard K, et al. HIV and the risk of tuberculosis due to recent transmission over 12 years in Karonga District, Malawi. Transactions of the Royal Society of Tropical Medicine & Hygiene **2009**;103:1187-1189
163. Corbett EL, Churchyard GJ, Clayton TC, et al. HIV infection and silicosis: the impact of two potent risk factors on the incidence of mycobacterial disease in South African miners. Aids **2000**;14:2759-68
164. Jahn A, Floyd S, Crampin AC, et al. Population-level effect of HIV an adult mortality and early evidence of reversal after introduction of antiretroviral therapy in Malawi. Lancet **2008**;371:1603-11
165. WHO. Three I's Meeting; Intensified Case Finding (IC), Isoniazid Preventive Therapy (IPT) and TB Infection Control (IC) for people living with HIV. Geneva, Switzerland, 2008
166. Ponnighaus JM, Fine PE, Bliss L, et al. The Karonga Prevention Trial: a leprosy and tuberculosis vaccine trial in northern Malawi. I. Methods of the vaccination phase. Lepr Rev **1993**;64:338-56
167. Preliminary results 2008 Malawi Population and Housing Census. Zomba: National Statistics Office, Malawi, 2009
168. Crampin AC, Glynn JR and Fine PE. What has Karonga taught us? Tuberculosis studied over three decades. International Journal of Tuberculosis & Lung Disease **2009**;13:153-164
169. Warndorff DK, Yates M, Ngwira B, et al. Trends in antituberculosis drug resistance in Karonga District, Malawi, 1986-1998. Int J Tuberc Lung Dis **2000**;4:752-7
170. Glynn JR, Warndorff DK, Fine PE, Munthali MM, Sichone W and Ponnighaus JM. Measurement and determinants of tuberculosis outcome in Karonga District, Malawi. Bull World Health Organ **1998**;76:295-305
171. Glynn JR, Crampin AC, Ngwira BM, et al. Trends in tuberculosis and the influence of HIV infection in northern Malawi, 1988-2001. Aids **2004**;18:1459-63
172. Steingart KR, Henry M, Ng V, et al. Fluorescence versus conventional sputum smear microscopy for tuberculosis: a systematic review. Lancet Infect Dis **2006**;6:570-81
173. Sterne JA, Turner AC, Fine PE, et al. Testing for antibody to human immunodeficiency virus type 1 in a population in which mycobacterial diseases are endemic. J Infect Dis **1995**;172:543-6
174. Harries AD, Zachariah R, Jahn A, Schouten EJ and Kamoto K. Scaling up antiretroviral therapy in Malawi-implications for managing other chronic diseases in resource-limited countries. J Acquir Immune Defic Syndr **2009**;52 Suppl 1:S14-6

175. Houben RMGJ, Glynn JR, Mallard K, et al. HIV increases the risk of tuberculosis due to recent re-infection in individuals with latent infection. International Journal of Tuberculosis & Lung Disease **2010** ***in print***
176. Mack U, Migliori GB, Sester M, et al. LTBI: latent tuberculosis infection or lasting immune responses to M. tuberculosis? A TBNET consensus statement. Eur Respir J **2009**;33:956-73
177. Canetti G, Sutherland I and Svandova E. Endogenous reactivation and exogenous reinfection: their relative importance with regard to the development of non-primary tuberculosis. Bull Int Union Tuberc **1972**;47:116-34
178. Ziegler JE, Edwards ML and Smith DW. Exogenous reinfection in experimental airborne tuberculosis. Tubercle **1985**;66:121-8
179. Fine PE, Bruce J, Ponnighaus JM, Nkhosa P, Harawa A and Vynnycky E. Tuberculin sensitivity: conversions and reversions in a rural African population. Int J Tuberc Lung Dis **1999**;3:962-75
180. Rieder HL. Methodological issues in the estimation of the tuberculosis problem from tuberculin surveys. Tuber Lung Dis **1995**;76:114-21
181. Fine PE, Sterne JA, Ponnighaus JM and Rees RJ. Delayed-type hypersensitivity, mycobacterial vaccines and protective immunity. Lancet **1994**;344:1245-9
182. Glynn JR, Ponnighaus J, Crampin AC, et al. The development of the HIV epidemic in Karonga District, Malawi. AIDS **2001**;15:2025-9
183. Corbett EL, Bandason T, Duong T, et al. Impact of periodic case-finding for symptomatic smear-positive disease on community control of prevalent infectious tuberculosis: a community randomised trial of two delivery strategies in Harare, Zimbabwe (DETECTB: ISRCTN84352452). 40th IUATLD World Conference on Lung Health. Cancun, Mexico, 2009
184. Escombe AR, Oeser CC, Gilman RH, et al. Natural ventilation for the prevention of airborne contagion. PLoS Med **2007**;4:e68
185. Ait-Khaled N, Alarcon E, Bissell K, et al. Isoniazid preventive therapy for people living with HIV: public health challenges and implementation issues. Int J Tuberc Lung Dis **2009**;13:927-35
186. White RG, Vynnycky E, Glynn JR, et al. HIV epidemic trend and antiretroviral treatment need in Karonga District, Malawi. Epidemiol Infect **2007**;135:922-32
187. Crampin AC, Glynn JR, Ngwira BM, et al. Trends and measurement of HIV prevalence in northern Malawi. Aids **2003**;17:1817-25
188. Ponnighaus JM, Mwanjasi LJ, Fine PE, et al. Is HIV infection a risk factor for leprosy? Int J Lepr Other Mycobact Dis **1991**;59:221-8
189. Crampin AC, Mwinuka V, Malema SS, Glynn JR and Fine PE. Field-based random sampling without a sampling frame: control selection for a case-control study in rural Africa. Trans R Soc Trop Med Hyg **2001**;95:481-3
190. Glynn JR, Warndorff DK, Fine PE, et al. The impact of HIV on morbidity and mortality from tuberculosis in sub-Saharan Africa: a study of rural Malawi and review of the literature. Health Transition Review **1997**;7:75-87
191. Lawn SD, Bekker LG and Miller RF. Immune reconstitution disease associated with mycobacterial infections in HIV-infected individuals receiving antiretrovirals. Lancet Infect Dis **2005**;5:361-73
192. Lawn SD, Myer L, Edwards D, Bekker LG and Wood R. Short-term and long-term risk of tuberculosis associated with CD4 cell recovery during antiretroviral therapy in South Africa. AIDS **2009**;23:1717-25
193. Treatment of AIDS: Guidelines for the use of antriretroviral therapy in Malawi. 3rd Edition ed: Ministry of Health, Malawi, 2008
194. Brodt HR, Kamps BS, Gute P, Knupp B, Staszewski S and Helm EB. Changing incidence of AIDS-defining illnesses in the era of antiretroviral combination therapy. AIDS **1997**;11:1731-8

195. Manabe YC, Breen R, Perti T, Girardi E and Sterling TR. Unmasked tuberculosis and tuberculosis immune reconstitution inflammatory disease: a disease spectrum after initiation of antiretroviral therapy. J Infect Dis **2009**;199:437-44
196. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ **2009**;338:b2393
197. Floyd S, Houben RMGJ, Molesworth A, Crampin AC, Glynn JR and French N. Imputation of mising HIV status in sero-surveys: a case study from Malawi. **in preparation**
198. Reniers G, Eaton J. Refusal bias in HIV prevalence estimates from nationally representative seroprevalence surveys. AIDS **2009**;23:621-9
199. Carlin JB, Galari JC and Royston P. A new framework for managing and analyzing multiply imputed data in Stata. The Stata Journal **2008**;8:49 - 67
200. van Buuren S, Boshuizen HC and Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med **1999**;18:681-94
201. Schurch AC, Kremer K, Kiers K, et al. The tempo and mode of molecular evolution of Mycobacterium tuberculosis at patient-to-patient scale. Infection, Genetics & Evolution **2010**;10:108 - 114
202. Middelkoop K. Impact of antiretroviral therapy on tuberculosis risk in different TB-HIV epidemics. Int J Tuberc Lung Dis;14:261
203. De Cock KM, Marston B. The sound of one hand clapping: tuberculosis and antiretroviral therapy in Africa. Am J Respir Crit Care Med **2005**;172:3-4
204. WHO. Rapid advice: antiretroviral therapy for HIV infection in adults and adolescents. Geneva: World Health Organisation, 2009

10 Appendices

This section holds a copy of publications that are based on the work described in this thesis. Three items have appeared in peer reviewed journals do far:

1. From Chapter 2: 'A System A systematic review and meta-analysis of molecular epidemiological studies of tuberculosis: development of a new tool to aid interpretation'. Paper published In: Tropical Medicine & International Health **2009**;volume 14, pages: 892-909
2. From Chapter 2: 'Systematic review and analysis of population-based molecular epidemiological studies.' Letter to the editor and author reply published in: International Journal of Tuberculosis and Lung Disease. **2009**; volume 13, pages 275-6
3. From Chapter 4: 'HIV and the risk of tuberculosis due to recent transmission over 12 years in Karonga District, Malawi'. Invited paper in Transactions of the Royal Society of Tropical Medicine & Hygiene. **2009**;volume 103, pages 1187-1189

Other papers are currently in press (describing work from Chapters 3 and 5, in IJTLD), submitted (describing work from Chapter 7) and in preparation (work from Chapter 6).

Note: To reduce the size of the document (i.e. make it ready for circulation) I removed the copies. I have attached these publications as .pdf where they were available to me. Others should become available through pubmed in due time.

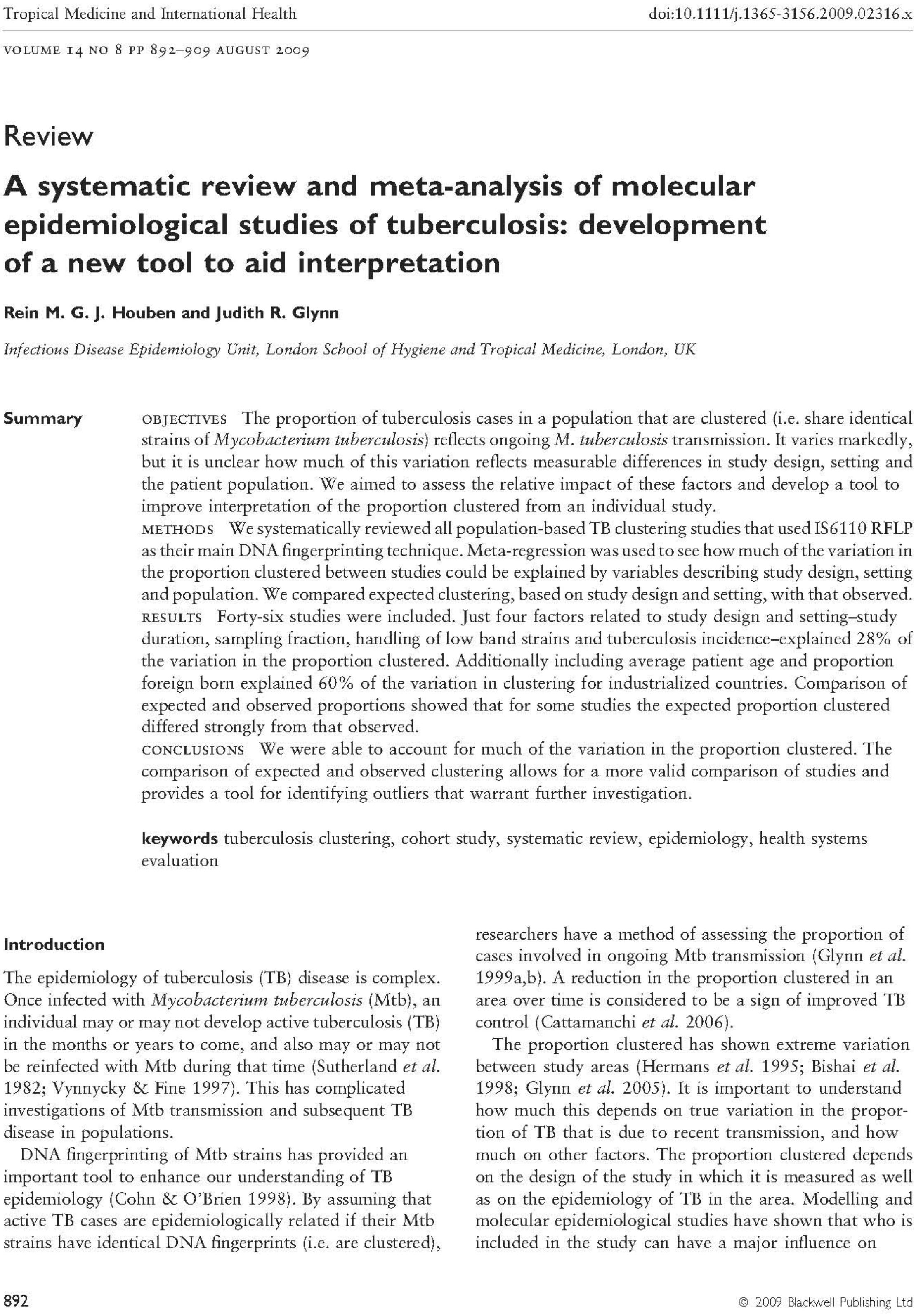
10 Appendices

This section holds a copy of publications that are based on the work described in this thesis. Three items have appeared in peer reviewed journals do far:

1. From Chapter 2: 'A System A systematic review and meta-analysis of molecular epidemiological studies of tuberculosis: development of a new tool to aid interpretation'. Paper published In: Tropical Medicine & International Health **2009**;volume 14, pages: 892-909
2. From Chapter 2: 'Systematic review and analysis of population-based molecular epidemiological studies.' Letter to the editor and author reply published in: International Journal of Tuberculosis and Lung Disease. **2009**; volume 13, pages 275-6
3. From Chapter 4: 'HIV and the risk of tuberculosis due to recent transmission over 12 years in Karonga District, Malawi'. Invited paper in Transactions of the Royal Society of Tropical Medicine & Hygiene. **2009**;volume 103, pages 1187-1189

Other papers are currently in press (describing work from Chapter 5) and submitted (describing work from Chapter 3).

10.1 A systematic review and meta-analysis of molecular epidemiological studies of tuberculosis: development of a new tool to aid interpretation



Tropical Medicine and International Health doi:10.1111/j.1365-3156.2009.02316.x

VOLUME 14 NO 8 PP 892–909 AUGUST 2009

Review

A systematic review and meta-analysis of molecular epidemiological studies of tuberculosis: development of a new tool to aid interpretation

Rein M. G. J. Houben and Judith R. Glynn

Infectious Disease Epidemiology Unit, London School of Hygiene and Tropical Medicine, London, UK

Summary

OBJECTIVES The proportion of tuberculosis cases in a population that are clustered (i.e. share identical strains of *Mycobacterium tuberculosis*) reflects ongoing *M. tuberculosis* transmission. It varies markedly, but it is unclear how much of this variation reflects measurable differences in study design, setting and the patient population. We aimed to assess the relative impact of these factors and develop a tool to improve interpretation of the proportion clustered from an individual study.

METHODS We systematically reviewed all population-based TB clustering studies that used IS6110 RFLP as their main DNA fingerprinting technique. Meta-regression was used to see how much of the variation in the proportion clustered between studies could be explained by variables describing study design, setting and population. We compared expected clustering, based on study design and setting, with that observed.

RESULTS Forty-six studies were included. Just four factors related to study design and setting–study duration, sampling fraction, handling of low band strains and tuberculosis incidence–explained 28% of the variation in the proportion clustered. Additionally including average patient age and proportion foreign born explained 60% of the variation in clustering for industrialized countries. Comparison of expected and observed proportions showed that for some studies the expected proportion clustered differed strongly from that observed.

CONCLUSIONS We were able to account for much of the variation in the proportion clustered. The comparison of expected and observed clustering allows for a more valid comparison of studies and provides a tool for identifying outliers that warrant further investigation.

keywords tuberculosis clustering, cohort study, systematic review, epidemiology, health systems evaluation

Introduction

The epidemiology of tuberculosis (TB) disease is complex. Once infected with *Mycobacterium tuberculosis* (Mtb), an individual may or may not develop active tuberculosis (TB) in the months or years to come, and also may or may not be reinfected with Mtb during that time (Sutherland *et al.* 1982; Vynnycky & Fine 1997). This has complicated investigations of Mtb transmission and subsequent TB disease in populations.

DNA fingerprinting of Mtb strains has provided an important tool to enhance our understanding of TB epidemiology (Cohn & O'Brien 1998). By assuming that active TB cases are epidemiologically related if their Mtb strains have identical DNA fingerprints (i.e. are clustered), researchers have a method of assessing the proportion of cases involved in ongoing Mtb transmission (Glynn *et al.* 1999a,b). A reduction in the proportion clustered in an area over time is considered to be a sign of improved TB control (Cattamanchi *et al.* 2006).

The proportion clustered has shown extreme variation between study areas (Hermans *et al.* 1995; Bishai *et al.* 1998; Glynn *et al.* 2005). It is important to understand how much this depends on true variation in the proportion of TB that is due to recent transmission, and how much on other factors. The proportion clustered depends on the design of the study in which it is measured as well as on the epidemiology of TB in the area. Modelling and molecular epidemiological studies have shown that who is included in the study can have a major influence on measured clustering. Clustering is underestimated as the proportion of all TB cases in the region that are included (the sampling fraction) decreases (Glynn *et al.* 1999a,b). Clustering depends on study duration, increasing with longer duration up to a plateau after about 4 years (Jasmer *et al.* 1999; van Soolingen *et al.* 1999; Glynn *et al.* 2005). A clearly defined study area is important, as high levels of migration in and out of a poorly defined area will artificially decrease the proportion of cases found to be clustered. This problem is reduced when an area approximates to a complete and relatively isolated population such as a country or district (Glynn *et al.* 1999a,b). Immigration and emigration from an area can have a major influence, and will both lead to a failure to identify cases that are due to recent transmission. In general, any feature which means that the study population is not a complete sample of a closed population will tend to underestimate the true proportion clustered.

Overestimation of the proportion clustered is less likely, although theoretically possible with biased sampling and contact tracing. In areas where there is insufficient variation in strains, or predominance of strains with few bands, identical strains cannot be assumed always to reflect recent transmission. There is also variation between studies in the laboratory techniques used, and in the rigour of the definition of clustered strains (Burman *et al.* 1997; Glynn *et al.* 1999a,b).

The proportion of TB due to recent transmission in a population depends on the annual risk of infection (both currently and in the past); the age pattern of TB cases (since older individuals have a higher risk of reactivation disease); and possibly on other factors such as HIV infection, which could have different effects on the risk of disease following recent or past infection (Haas *et al.* 1999; Borgdorff *et al.* 2000; Bruchfeld *et al.* 2002; Murray 2002a,b; Glynn *et al.* 2005).

We performed a systematic literature review on all population-based studies on TB clustering that used IS*6110* based RFLP as the main DNA fingerprinting technique, as this has been widely used as the standard since the early 1990s (van Embden *et al.* 1993). We assess the extent to which the variation in observed clustering can be explained by study design, study setting (the local epidemiology of TB) and study population. Additionally we develop a tool for interpreting a local proportion clustered in the context of a study's design and setting.

This paper distinguishes itself from previous reviews by examining the variability of clustering on the population level, rather than collating results from studies with varying design and settings to attempt to identify individual-level risk factors for being part of a cluster (Fok *et al.* 2008; Nava-Aguilera *et al.* 2009).

Methods

Inclusion and exclusion criteria

Studies on TB clustering that used IS*6110* RFLP as the main DNA fingerprinting technique were eligible for inclusion if they were population-based and reported clustering results (number of strains in RFLP analysis and proportion clustered) for more than 100 individuals. Studies were excluded if the sample population was not representative of the general population as this could bias the proportion clustered (e.g. prison population, drug resistant patients, outbreak studies).

In addition, the study area had to be suitable for making a valid estimate of local TB clustering. For example, inclusion of patients from a single hospital would be acceptable if all TB patients in the area are likely to go to that hospital, but not otherwise. In practice study areas had to minimally consist of a geographically defined urban or rural district. Overall, these criteria were necessary to reduce bias (which cannot be corrected for) in the dataset, allowing a valid examination of and correction for factors that influence the estimated proportion clustered.

There was no language restriction, and papers were translated by a fluent speaker as required.

Literature search

PubMed and Embase databases were searched between 1990 and November 2006. After maximizing the sensitivity of the search by using a private collection of relevant TB clustering papers, the following PubMed search query was used: '*IS6110 OR RFLP OR fingerprint* OR cluster* OR genotyp* OR Epidemiology, Molecular (MeSH) OR molecular typing OR transmission OR molecular epidemiological OR molecular epidemiology) AND (tuberculosis OR tb) AND [1990 : 3000 (PDAT)]*'. A similar search query was used for Embase. The TB publications archive from the Centers for Disease Control and Prevention was accessed but yielded no additional papers (CDC 2006). After excluding duplicates, all titles were scanned twice by RH for possible relevance. The remaining abstracts were read and eligible full length papers were retrieved if possible.

Data extraction

From each paper RH collected information on study methods (secondary DNA typing methods, the cut-off points (if present) for Mtb strains with few IS*6110* bands, matching of DNA fingerprints), the number of patients eligible and included in the final DNA fingerprint analysis, TB clustering results, characteristics of the study population (age, proportion males, HIV positive and foreign born) and TB disease (proportion of smear positive, extra pulmonary or drug resistant TB cases). In addition, we recorded TB incidence in the region. If more than one paper provided data on the same population and study period only the one with the longest study duration was included. Extracted data were checked by JG.

Measures of TB disease due to recent Mtb transmission

The main study outcome was the proportion of clustered TB cases, i.e. number of cases in clusters/total number of cases. The reported proportion clustered was used, although the strictness of the definition differed between papers: most required identical fingerprints, but some allowed one band difference. IS*6110* RFLP is not sufficiently discriminatory in strains with low band numbers (usually five or less) (Glynn *et al.* 1999a,b), and if sufficient information was provided we either excluded these strains or recorded the overall proportion clustered when a secondary DNA typing technique [Spoligotyping, Polymorphic GC Sequencing (PGRS), Direct Repeats] was used for these low band number strains. The proportion of TB due to recent transmission was estimated as (number of clustered cases–number of clusters)/number of cases, the $n - 1$ method (Small *et al.* 1994). This assumes that each cluster consists of one source case, due to reactivation disease, the rest being due to recent infection.

Study design

Study duration was recorded in months, and cross-sectional surveys were assigned a duration of 0 months. If studies only allowed cases to cluster in a certain period after the source case, this clustering interval was used as study duration. We recorded the fraction of all culture positive (c+) TB cases (of all types) in the catchment area that had RFLP results available. We recorded whether the RFLP analysis included Mtb strains with low band numbers, and if so, whether a secondary DNA typing technique (e.g. spoligotyping or polymorphic GC sequencing) was applied for these (van Soolingen *et al.* 1993; Kamerbeek *et al.* 1997).

Study setting

When possible the TB incidence in the study region as reported in the paper was used. Otherwise the sampling fraction, reported TB incidence, study duration and study population size were combined to estimate the regional TB rate. If that was not possible, the scientific literature was searched for region specific estimates or we used the WHO Global Database to acquire a country wide estimate (WHO 2008). Studies were classified into low, middle and high burden TB areas (TB incidence <10, 11–50 and 50 + TB cases per 100 000 per year respectively).

Study population

Studies from industrialized areas (e.g. Western Europe, North America) were grouped together. For these areas we recorded the proportion of TB cases that were foreign born. In these regions immigrants usually account for a substantial part of the TB case population. They are likely to have been infected with Mtb in their country of origin where the annual risk of infection is relatively high (van Soolingen *et al.* 1999; Cattamanchi *et al.* 2006; Dahle *et al.* 2007), and therefore often have unique (non-clustered) strains, adding to the diversity of Mtb strains in the population (Small *et al.* 1994; Hernandez-Garduno *et al.* 2002). However, through socio-demographic factors they could also be at a higher risk of being in a cluster, thus increasing the proportion clustered. Where available we also collected data on the proportion of cases with (a history of) homelessness or drug and alcohol abuse.

Age was recorded as the average for the TB case population; either as the reported mean or median age or, if only age strata were reported, through estimation of the median age within the age stratum that held the median observation. Gender was recorded as the proportion of males in the study population. As historical data on the annual risk of infection were not available for the majority of studies, the mean age of the TB case population can be used a proxy, with a declining annual risk of infection shown by a high mean age of infection (Vynnycky *et al.* 2003). If at least 50% of all TB cases were systematically tested for HIV, we recorded the proportion HIV positive of those with test results. If the paper itself did not report the variable, an estimate was extracted from papers that reported on the same population.

Statistical analyses

Non-parametric tests (Wilkinson rank sum) were used to compare studies from industrialized countries with those from other countries.

To assess the extent to which the variation in clustering seen can be explained by study design, study setting (the local epidemiology of TB) and study population we applied meta-regression (Sterne 2009). This technique allows multivariate regression analysis to ascertain how well individual or a combination of variables can explain the between study variation in the proportion clustered (the tau^2) (Berkey *et al.* 1995; Thompson & Higgins 2002).

Meta-regression assumes a linear association between the outcome (proportion clustered) and the independent variable as well as an approximately normal distribution of the residuals. The latter was checked through visual inspection of the residuals (scatterplots and histograms) and statistical tests (sktest in Stata version 10; Stata Corp LP., College Station, TX, USA). The choice between entering a variable as categorical or linear was dependent on its distribution, known epidemiological associations and the impact on the between study variation. A variable describing the standard error of the outcome (proportion clustered) is required for each study. We calculated this by taking the square root of '$p*(1 - p)/N$', where 'p' stands for the proportion clustered in a study, and N the total number of study participants.

The effect of the study design and recorded variables was first examined through univariate meta-regression. To allow for potential negative and positive confounding due to high heterogeneity of variable values between studies, all variables were considered for the multivariate models. Final inclusion of a variable in a multivariate model was based on whether adding the variable had a significant impact (>2.5% reduction) on the between-study variation.

Four models were created. The main model was limited to variables describing study design and epidemiological setting (the local TB incidence), to ascertain to what extent these could reduce the variation in the proportion clustered.

The second model included variables describing the study population as well. The third model was limited to studies from industrialized countries, and included the proportion of TB cases that were foreign born. These models were designed to test to what extent the observed variation could be explained by known factors, and how much residual, unexplained variation would remain. The fourth model excluded local TB incidence from the main model.

All models were repeated with the proportion of cases due to recent transmission (n–1 method) as the outcome measure to test the robustness of the findings.

Tool to interpret local proportions clustered

The coefficients from the main model were applied to each study to acquire a predicted value for the proportion clustered based solely on the study's design and local TB incidence. These estimates were compared with the observed values [(observed–expected)/expected × 100%] to provide a measure of how much the observed proportion clustered differed from its expected value. This correction for study design and setting provides a new perspective on the proportion clustered, and allows for better comparison between studies. Confidence intervals for the relative differences were estimated using the standard error of the predicted values (stdp option in Stata version 10).

Results

Systematic literature review

Our primary search yielded 11 654 records, and after selection (Figure 1) 46 papers were included for analysis.

Descriptive analyses

The majority of studies (36) were performed in industrialized country settings (North America, Western Europe, Japan or Hong Kong); the others were done in sub-Saharan Africa (n = 3), South America (n = 4) and in South East Asia, Eastern Europe, and the Middle East (one each).

The Forest plot for all 46 included studies arranged by study location (Figure 2a) shows a large variation in the proportion clustered between studies (Q-test for heterogeneity chi square = 6622 (d.f. = 46), P < 0.0001), from 6% in Northern Bangladesh to 86% in Greenland. Included studies and their recorded variables are listed in Appendix 1.

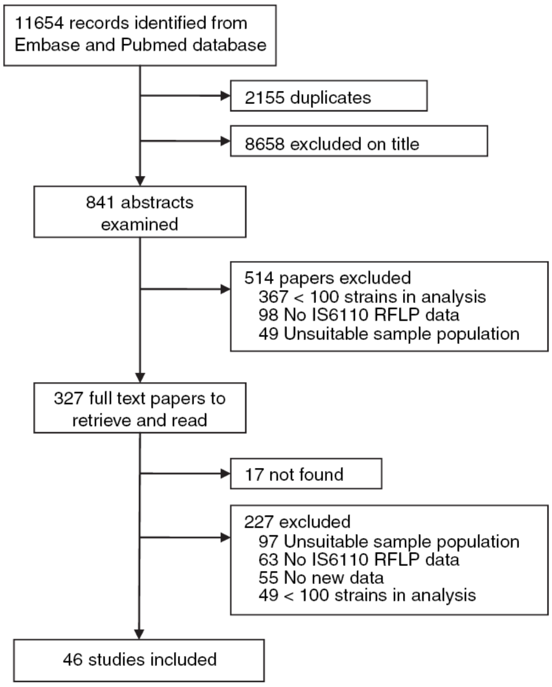

11654 records identified from Embase and Pubmed database
2155 duplicates
8658 excluded on title
841 abstracts examined
514 papers excluded
367 < 100 strains in analysis
98 No IS6110 RFLP data
49 Unsuitable sample population
327 full text papers to retrieve and read
17 not found
227 excluded
97 Unsuitable sample population
63 No IS6110 RFLP data
55 No new data
49 < 100 strains in analysis
46 studies included

Figure 1 Flow diagram of systematic literature review process.

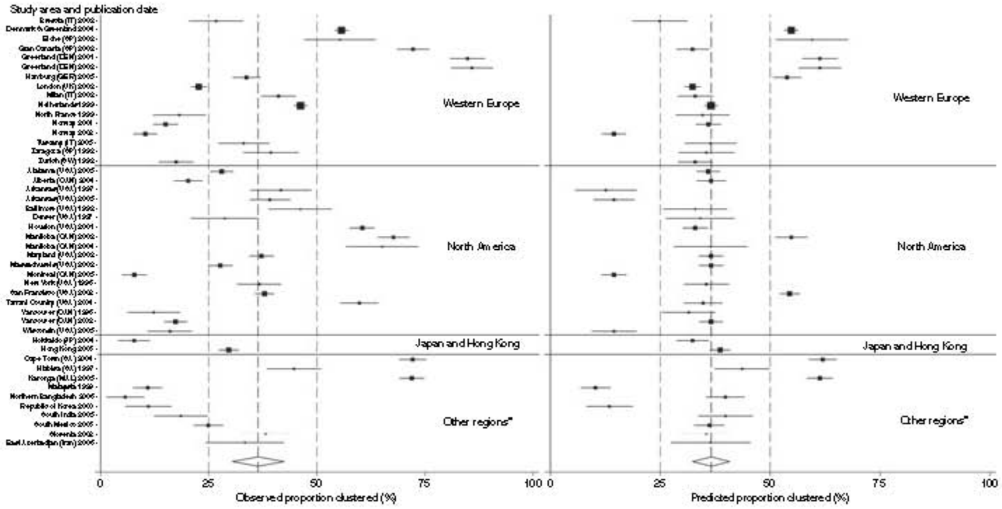


Figure 2 Observed (left) and predicted (right) proportions clustered of 46 included studies. The predicted values for each study were acquired using the coefficients from the meta-regression model **1** (see Table 2, column 2) that included study duration, sampling fraction, handling of Mtb strains with low band numbers and local TB incidence. Appendix **1** holds the values for each study, see appendix 2 for further illustration of the calculations. Box size and error bars indicate number of patients included in the study. *'Other regions' refers to sub Saharan Africa, South East Asia, South America, Eastern Europe and Middle East.

Table 1 shows that studies were diverse in design (e.g. study duration between 0 months (cross-sectional study) and >10 years) as well as setting (recorded local TB incidence between 1.7 and 304 cases/100 000/year). TB incidence differed strongly between industrialized settings and other settings ($P < 0.001$), whereas the proportion of HIV positive TB cases did not ($P = 0.88$). Insufficient data were available on homelessness and drug or alcohol abuse so these variables were not included in the analyses.

Meta-regression analyses

Figure 3 shows that the proportion clustered increased with increasing study duration, sampling fraction and TB incidence, decreased with increasing age and proportion foreign born (in industrialized countries) and changed little with increasing study size. In the univariate meta-regression analyses these associations were confirmed (Table 2). Some variables reduced the tau^2 by more than 10% (duration of study, handling of low band Mtb strains, age and country of birth of TB case population), whereas others had less or no effect.

Study duration was entered as a categorical variable so its association with clustering could take any shape, including the one shown within populations where clustering increases with study duration, but reaches a plateau after 4 years (Glynn *et al.* 1999a,b; van Soolingen *et al.* 1999a,b; Glynn *et al.* 2005; Jasmer *et al.* 1999).

In the multivariate meta-regression model (Table 2, Model 1) 28% of between-study variation was explained by study duration, sampling fraction, handling of strains with low band numbers and local TB incidence. Most coefficients of the included variables were statistically significant at the 0.05 level, and the residuals were approximately normally distributed (P-value sktest = 0.46). In this model, the proportion clustered increased with study duration and sampling fraction. The model also showed that including strains with a low number of IS*6110* bands increased the proportion clustered, unless secondary typing methods were applied. Additionally, study settings with high TB incidence reported higher proportions clustered.

Incorporating variables describing the study population further reduced the tau^2, explaining up to 60% of the between study variance (Table 2, Model 3). When studies for all countries were considered (Table 2, Model 2), the average age of the TB case population showed a strong negative association with the proportion clustered. In studies from industrialized settings, the proportion clustered decreased as the proportion foreign born increased (P-value coefficient <0.001). This negative association was found in 18 of the 21 studies from industrialized settings that reported the proportion of foreign born TB cases by

Table 1 Summary of included studies by study region

Variables*	Study region Industrialized countries N = 36	Other countries† N = 10
Study design		
Study Duration (months)	48 (3–120)	13 (0–87)
Number of patients	520 (114–4266)	372 (105–1029)
Sampling fraction‡	0.90 (0.30–1.00) n = 36	0.82 (0.002–1.00) n = 10
Low band numbers included n/N (% studies)	13/36 (36)	6/10 (60)
Secondary DNA typing method used for low band numbers n/N (% studies)	13/36 (36)	3/10 (27)
Study setting		
TB rate in the region (n/100 000/year)	9 (2–185)	90 (8–761)
Study population		
Average age (years)	45 (30–69) n = 30	45 (33–55) n = 9
Sex (% male)	65 (44–74) n = 28	59 (47–76) n = 8
% Foreign born	44 (3–83) n = 28	Not recorded
% HIV positive	16 (1–57) n = 21	16 (1–65) n = 4
% Resistant to ≥1 drug	10 (0–31) n = 17	10 (6–23) n = 5
TB Transmission		
% clustered	35 (8–86)	29 (6–72)
% recent transmission	25 (4–78) n = 34	20 (3–59)

N = total number of studies in TB burden category; n = number of studies used in cell; '%' = proportion of all study participants.

*Median and range given unless otherwise indicated.

†Studies performed in sub-Saharan Africa (n = 3), South America (n = 4), South East Asia (n = 1), Eastern Europe (1) and the Middle East (n = 1).

‡Sampling fraction is fraction of all culture positive TB patients in the study area with RFLP results available.

cluster status, whereas only one study, from Italy, found a positive association (Matteelli *et al.* 2003).

Excluding local TB incidence from the main model (Model 4) reduced the explained variation from 28 to 10% as well as the precision of the coefficients. However, the direction and size of the coefficients remained similar.

None of the other variables we recorded had a relevant effect on the tau^2. Similar results were found using the proportion of cases due to recent Mtb transmission (estimated using the n–1 method) as the study outcome, rather than the total proportion clustered (results not shown).

Observed *vs.* expected proportion clustered

For each study the expected proportion clustered could be estimated from the coefficients of the main model. The results are shown in Figure 2b and the relative difference between the expected and observed values is shown in Figure 4. Appendix 2 illustrates these calculations. Figure 4 shows that high observed proportions clustered often lie close to their expected values [e.g. studies from Elche (Spain) Cape Town (South Africa) and Karonga (Malawi) (Ruiz Garcia *et al.* 2002; Verver *et al.* 2004; Glynn *et al.* 2005)]. However, for some studies the levels of reported clustering were twice as high as expected based on study design and setting alone. This applied to regions with apparent moderate as well as higher levels of clustering, for example Arkansas (~40% clustered) and Gran Canaria (72% clustered) (Braden *et al.* 1997; Pena *et al.* 2003; Cave *et al.* 2005). On the other hand, studies from Vancouver, Japan and Bangladesh reported proportions clustered half as high as that expected (Blenkush *et al.* 1996; Hernandez-Gardun~o *et al.* 2002; Storla *et al.* 2006).

Discussion

We show that although the proportion clustered varies widely between population-based studies, 28% of this variation can be explained by just four variables describing study design and setting. Models including the average age and immigrant status of the TB case population explained up to 60% of the between-study variation in industrialized countries.

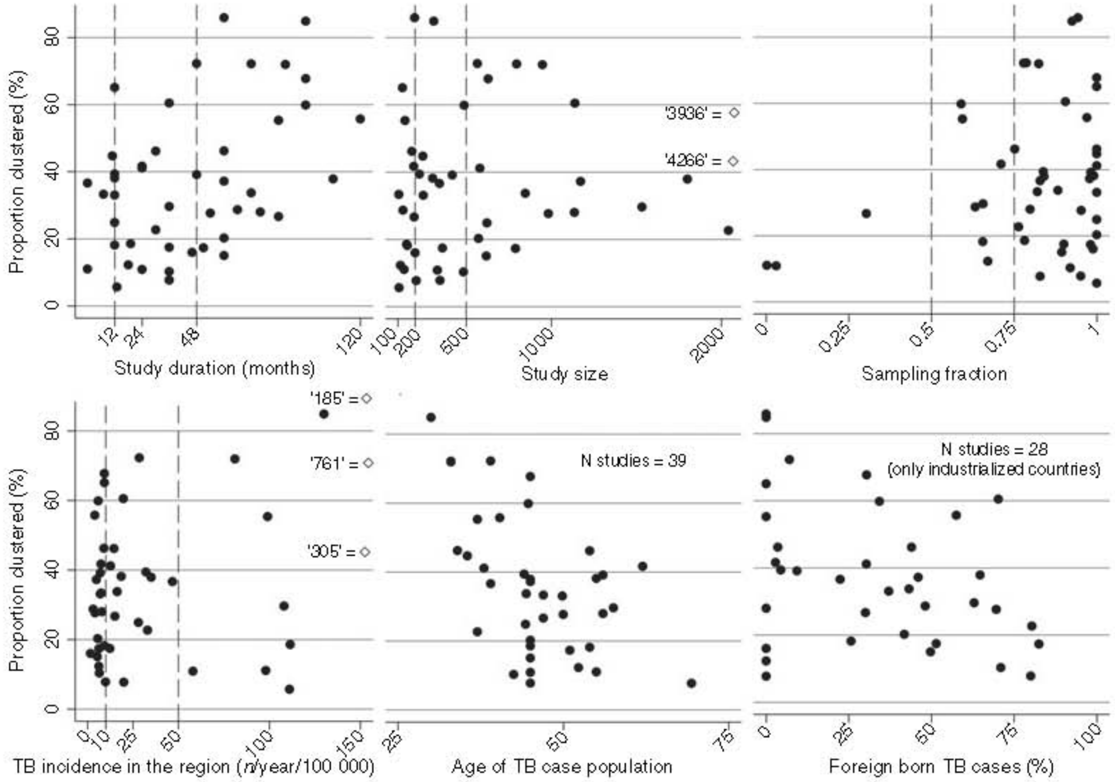


Figure 3 The association between the proportion clustered and recorded variables. Scatter plots show univariate associations between the proportion clustered and selected recorded variables. Vertical lines show the categories used in the meta-regression. Open diamonds signal outlier values for the recorded variable.

The residual variation can be due to imprecision of included variables, unmeasured factors and interactions of the included variables (for which there were insufficient studies to test). With the current level and detail of reporting of clustered studies it seems likely that any explanatory model will have a lot of unknowns.

Although most coefficients of the main model were statistically significant at the 0.05 level, the confidence intervals were wide. This is in part due to the low number of included studies (46–25 depending on the model) and possibly the high heterogeneity in variable values. The latter is also suggested in Model 4. The removal of one explanatory variable (local TB incidence) has a big impact on the unexplained variation between studies as well the precision of the coefficients, without affecting their overall patterns.

This high heterogeneity is one of the main reasons we excluded studies that applied DNA fingerprinting techniques other than IS*6110* RFLP as their main method of strain typing. The number of studies per technique is relatively low, and insufficient to correct for the added variation due to, for example, an unknown level of difference in molecular clocks of the markers used in each technique. However, with the recent standardisation and subsequent more widespread use of PCR based Mycobacterial Interspersed Repetitive Unit-Variable Number of DNA Tandem Repeats (MIRU-VNTR) a valid comparison between RFLP and MIRU-VNTR may in the future become viable (Supply *et al.* 2006).

Most associations found in this review are statistically strong and in line with observations made in individual epidemiological and modelling studies, thus giving our meta-regression models additional validity. The importance of study duration in clustering studies has been well documented through modelling (Glynn *et al.* 1999a,b) and epidemiological studies (Jasmer *et al.* 1999; van Soolingen *et al.* 1999; Glynn *et al.* 2005). Our results show

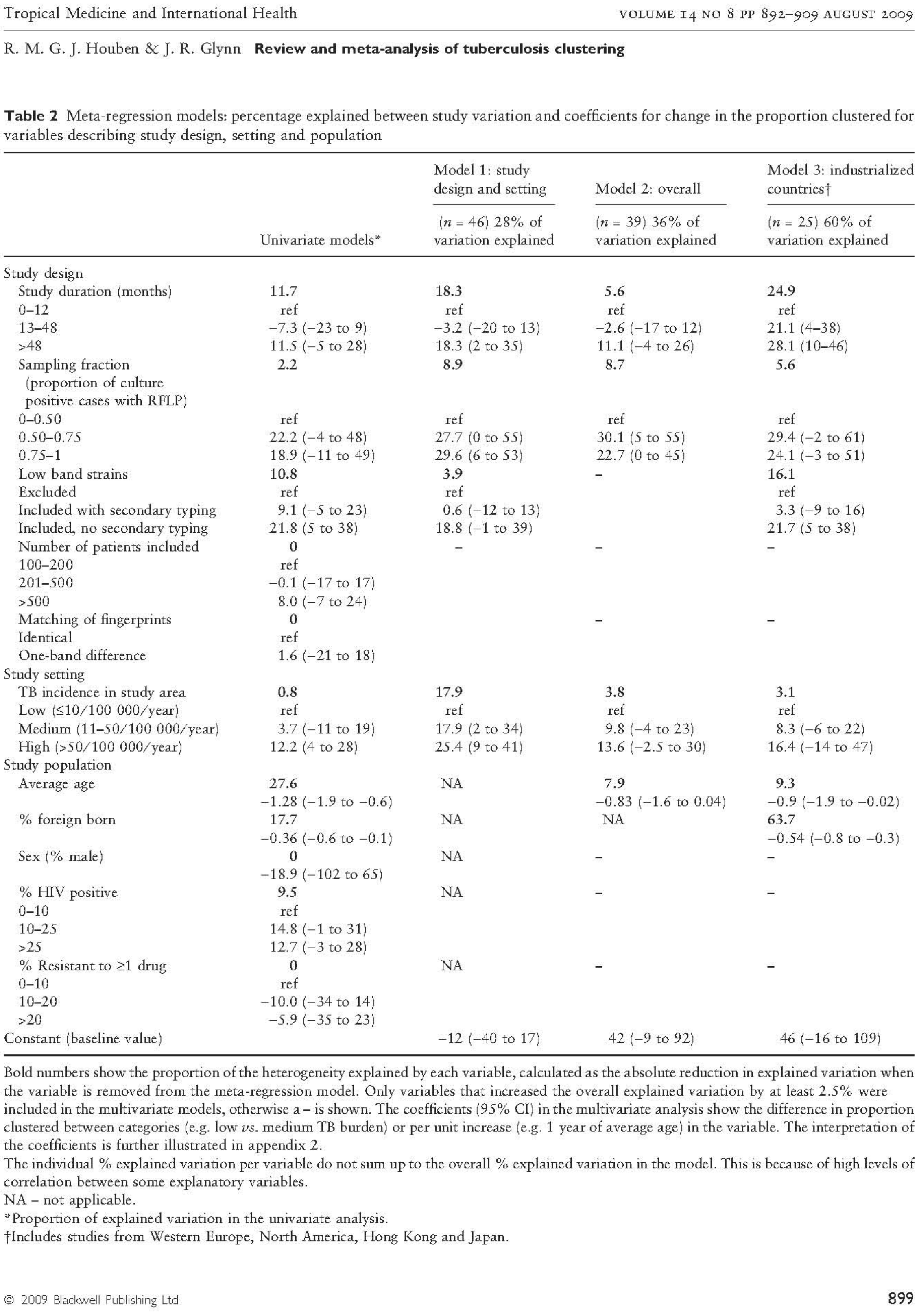


Table 2 Meta-regression models: percentage explained between study variation and coefficients for change in the proportion clustered for variables describing study design, setting and population

	Univariate models*	Model 1: study design and setting (*n* = 46) 28% of variation explained	Model 2: overall (*n* = 39) 36% of variation explained	Model 3: industrialized countries† (*n* = 25) 60% of variation explained
Study design				
Study duration (months)	**11.7**	**18.3**	**5.6**	**24.9**
0–12	ref	ref	ref	ref
13–48	−7.3 (−23 to 9)	−3.2 (−20 to 13)	−2.6 (−17 to 12)	21.1 (4–38)
>48	11.5 (−5 to 28)	18.3 (2 to 35)	11.1 (−4 to 26)	28.1 (10–46)
Sampling fraction (proportion of culture positive cases with RFLP)	**2.2**	**8.9**	**8.7**	**5.6**
0–0.50	ref	ref	ref	ref
0.50–0.75	22.2 (−4 to 48)	27.7 (0 to 55)	30.1 (5 to 55)	29.4 (−2 to 61)
0.75–1	18.9 (−11 to 49)	29.6 (6 to 53)	22.7 (0 to 45)	24.1 (−3 to 51)
Low band strains	**10.8**	**3.9**	–	**16.1**
Excluded	ref	ref		ref
Included with secondary typing	9.1 (−5 to 23)	0.6 (−12 to 13)		3.3 (−9 to 16)
Included, no secondary typing	21.8 (5 to 38)	18.8 (−1 to 39)		21.7 (5 to 38)
Number of patients included	**0**	–	–	–
100–200	ref			
201–500	−0.1 (−17 to 17)			
>500	8.0 (−7 to 24)			
Matching of fingerprints	**0**		–	–
Identical	ref			
One-band difference	1.6 (−21 to 18)			
Study setting				
TB incidence in study area	**0.8**	**17.9**	**3.8**	**3.1**
Low (≤10/100 000/year)	ref	ref	ref	ref
Medium (11–50/100 000/year)	3.7 (−11 to 19)	17.9 (2 to 34)	9.8 (−4 to 23)	8.3 (−6 to 22)
High (>50/100 000/year)	12.2 (4 to 28)	25.4 (9 to 41)	13.6 (−2.5 to 30)	16.4 (−14 to 47)
Study population				
Average age	**27.6** −1.28 (−1.9 to −0.6)	NA	**7.9** −0.83 (−1.6 to 0.04)	**9.3** −0.9 (−1.9 to −0.02)
% foreign born	**17.7** −0.36 (−0.6 to −0.1)	NA	NA	**63.7** −0.54 (−0.8 to −0.3)
Sex (% male)	**0** −18.9 (−102 to 65)	NA	–	–
% HIV positive	**9.5**	NA	–	–
0–10	ref			
10–25	14.8 (−1 to 31)			
>25	12.7 (−3 to 28)			
% Resistant to ≥1 drug	**0**	NA	–	–
0–10	ref			
10–20	−10.0 (−34 to 14)			
>20	−5.9 (−35 to 23)			
Constant (baseline value)		−12 (−40 to 17)	42 (−9 to 92)	46 (−16 to 109)

Bold numbers show the proportion of the heterogeneity explained by each variable, calculated as the absolute reduction in explained variation when the variable is removed from the meta-regression model. Only variables that increased the overall explained variation by at least 2.5% were included in the multivariate models, otherwise a – is shown. The coefficients (95% CI) in the multivariate analysis show the difference in proportion clustered between categories (e.g. low *vs.* medium TB burden) or per unit increase (e.g. 1 year of average age) in the variable. The interpretation of the coefficients is further illustrated in appendix 2.

The individual % explained variation per variable do not sum up to the overall % explained variation in the model. This is because of high levels of correlation between some explanatory variables.

NA – not applicable.

*Proportion of explained variation in the univariate analysis.

†Includes studies from Western Europe, North America, Hong Kong and Japan.

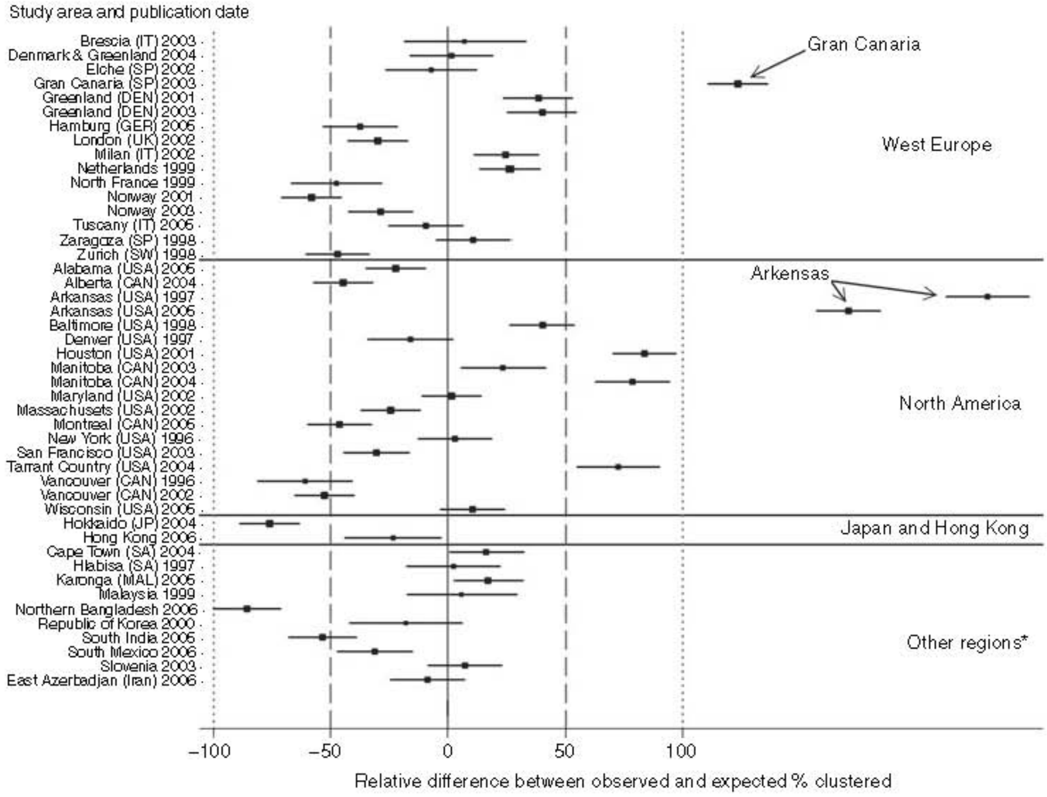


Figure 4 Relative difference (in %) between observed and expected proportion clustered. *Other regions' refers to sub Saharan Africa, South East Asia, South America, Eastern Europe, Middle East. Error lines indicate 95% confidence intervals of predicted values, calculated using standard error of predictions.

that this is also important when comparing between studies. The effect of the sampling fraction has been predicted through modelling (Glynn *et al.* 1999a,b; Murray 2002a,b) and is intuitive, especially when the average cluster size is small: the 'missing' part of the population will lead to more clustered cases being classified as unique, thus underestimating the proportion clustered. The expected increases in clustering with local TB incidence and with younger age were also seen (van Soolingen *et al.* 1999; Cattamanchi *et al.* 2006; Dahle *et al.* 2007).

The strong negative association of foreign born TB cases and the proportion clustered, and the fact that 18 of 21 studies from industrialized countries reported the same statistically significant association on the individual level, both imply that on average foreign born cases have a lower risk of being part of an identified cluster in a study setting.

Our comparison between observed and expected clustering highlights outliers. More clustering than expected could arise in situations where there is a low number of circulating Mtb strains and little population movement, so that identical strains could reflect transmission many years previously; this could account for the findings from Arkansas (Braden *et al.* 1997). TB outbreaks will increase clustering, which was the case in Gran Canaria where two Mtb strains were involved in 30% of all clustered TB cases (Pena *et al.* 2003).

Lower than expected clustering (based on model 1) could reflect an old population with much disease due to reactivation, as in Japan where the average age was 69 years (Fujikane *et al.* 2004). If we apply Model 2, which includes age, to this study the expected clustering is estimated at 18%, which lies much closer to the observed value. The results from the Bangladesh study appear to be due to under sampling; only 111 of 1264 (9%) notified cases from the region and study period were confirmed by culture and thus potentially included in the clustering analysis (Storla *et al.* 2006). This effect is not included in our model; the sampling fraction was calculated as the fraction of confirmed culture positive cases due to limitations in the reporting of studies.

The comparison of observed and expected clustering also shows the degree to which high observed clustering can be explained: in Malawi, Cape Town and Greenland the high levels of clustering were largely due to the studies' long duration, high sampling fraction and the high local TB rates (Soborg *et al.* 2001; Zolnir-Dovc *et al.* 2003; Thomsen *et al.* 2004; Verver *et al.* 2004; Glynn *et al.* 2005a,b).

We would have liked to include more studies from high burden countries, but there were no further studies available. However, the effect of study design factors is likely to be constant in different regions (as is the case with study duration (Jasmer *et al.* 1999; van Soolingen *et al.* 1999a,b; Glynn *et al.* 2005)). We did include 10 studies from areas with a high annual TB incidence (>50/100 000), which should make our results applicable to high burden settings.

Previous reviews have had limitations, either through not including all relevant studies or potential problems in the analysis (Fok *et al.* 2008; Houben *et al.* 2009; Nava-Aguilera *et al.* 2009). Also, rather than collating results from varying study designs and populations on the individuals' risk of clustering, we chose to take a more public health-oriented approach by investigating the proportion clustered in populations while explicitly taking into account the differences in study design and setting.

We have focussed on the extent to which measured clustering can be explained by known factors, and the methods and associations presented here can be applied by researchers to acquire a new perspective on the proportion clustered, after adjusting for study design and setting. This will allow a more valid comparison between studies, highlight outliers and help researchers to assess their local levels of ongoing Mtb transmission.

Acknowledgements

We thank Sara Thomas for fruitful discussions on the systematic literature review and Simon Cousens, Jonathan Sterne and Chris Frost for their useful statistical advice.

References

Asgharzadeh M, Shahbabian K, Majidi J *et al.* (2006) IS6110 restriction fragment length polymorphism typing of Mycobacterium tuberculosis isolates from East Azerbaijan Province of Iran. *Memorias do Instituto Oswaldo Cruz* **101**, 517–521.

Berkey CS, Hoaglin DC, Mosteller F & Colditz GA (1995) A random-effects regression model for meta-analysis. *Statistics in Medicine* **14**, 395–411.

Bishai WR, Graham NMH, Harrington S *et al.* (1998) Molecular and geographic patterns of tuberculosis transmission after 15 years of directly observed therapy. *The Journal of American Medical Association* **280**, 1679–1684.

Blackwood KS, Al-Azem A, Elliott LJ, Hershfield ES & Kabani AM (2003) Conventional and molecular epidemiology of Tuberculosis in Manitoba. *BMC Infectious Diseases* **3**, **11**.

Blackwood KS, Wolfe JN & Kabani AM (2004) Application of mycobacterial interspersed repetitive unit typing to Manitoba tuberculosis cases: can restriction fragment length polymorphism be forgotten? *Journal of Clinical Microbiology* **42**, 5001–5006.

Blenkush M, Kunimoto D, Black W, Elwood RK & FitzGerald JM (1996) Evidence for TB clustering in Vancouver: results from pilot study using RFLP fingerprinting. *Canada Communicable Disease Report* **22**, 49–51.

Borgdorff MW, Behr MA, Nagelkerke NJ, Hopewell PC & Small PM (2000) Transmission of tuberculosis in San Francisco and its association with immigration and ethnicity. *International Journal of Tuberculosis and Lung Disease* **4**, 287–294.

Braden CR, Templeton GL, Cave MD *et al.* (1997) Interpretation of restriction fragment length polymorphism analysis of *Mycobacterium tuberculosis* isolates from a state with a large rural population. *Journal of Infectious Diseases* **175**, 1446–1452.

Bruchfeld J, Aderaye G, Palme IB *et al.* (2002) Molecular epidemiology and drug resistance of Mycobacterium tuberculosis isolates from ethiopian pulmonary tuberculosis patients with and without human immunodeficiency virus infection. *Journal of Clinical Microbiology* **40**, 1636–1643.

Burgos M, DeRiemer K, Small PM, Hopewell PC & Daley CL (2003) Effect of drug resistance on the generation of secondary cases of tuberculosis. *Journal of Infectious Diseases* **188**, 1878–1884.

Burman WJ, Reves RR, Hawkes AP *et al.* (1997) DNA fingerprinting with two probes decreases clustering of *Mycobacterium tuberculosis*. *American Journal of Respiratory & Critical Care Medicine* **155**, 1140–1146.

Cattamanchi A, Hopewell PC, Gonzalez LC *et al.* (2006) A 13-year molecular epidemiological analysis of tuberculosis in San Francisco. *International Journal of Tuberculosis and Lung Disease* **10**, 297–304.

Cave MD, Yang ZH, Stefanova R *et al.* (2005) Epidemiologic import of tuberculosis cases whose isolates have similar but not identical IS6110 restriction fragment length polymorphism patterns. *Journal of Clinical Microbiology* **43**, 1228–1233.

CDC (2006) Centers for Disease Control and Prevention; TB-update list. Available at http://listmanager.aspensys.com/read/?forum=tb-update. Accessed at November 2006.

Chan-Yeung M, Kam KM, Leung CC *et al.* (2006) Population-based prospective molecular and conventional epidemiological study of tuberculosis in Hong Kong. *Respirology* **11**, 442–448.

Cohn DL & O'Brien RJ (1998) The use of restriction fragment length polymorphism (RFLP) analysis for epidemiological studies of tuberculosis in developing countries. *International Journal of Tuberculosis and Lung Disease* **2**, 16–26.

Cowan LS, Diem L, Monson T *et al.* (2005) Evaluation of a two-step approach for large-scale, prospective genotyping of

Mycobacterium tuberculosis isolates in the United States. *Journal of Clinical Microbiology* 43, 688–695.

Cronin WA, Golub JE, Magder LS *et al.* (2001) Epidemiologic usefulness of spoligotyping for secondary typing of *Mycobacterium tuberculosis* isolates with low copy numbers of IS6110. *Journal of Clinical Microbiology* 39, 3709–3711.

Dahle UR, Sandven P, Heldal E & Caugant DA (2001) Molecular epidemiology of *Mycobacterium tuberculosis* in Norway. *Journal of Clinical Microbiology* 39, 1802–1807.

Dahle UR, Sandven P, Heldal E & Caugant DA (2003) Continued low rates of transmission of *Mycobacterium tuberculosis* in Norway. *Journal of Clinical Microbiology* 41, 2968–2973.

Dahle UR, Eldholm V, Winje BA, Mannsaker T & Heldal E (2007) Impact of immigration on the molecular epidemiology of *Mycobacterium tuberculosis* in a low-incidence country. *American Journal of Respiratory and Critical Care Medicine* 176, 930–935.

Dale JW, Nor RM, Ramayah S, Tang TH & Zainuddin ZF (1999) Molecular epidemiology of tuberculosis in Malaysia. *Journal of Clinical Microbiology* 37, 1265–1268.

Das SD, Narayanan S, Hari L *et al.* (2005) Differentiation of highly prevalent IS6110 single-copy strains of *Mycobacterium tuberculosis* from a rural community in South India with an ongoing DOTS programme. *Infection, Genetics & Evolution* 5, 67–77.

De Bruyn G, Adams GJ, Teeter LD, Soini H, Musser JM & Graviss EA (2001) The contribution of ethnicity to *Mycobacterium tuberculosis* strain clustering. *International Journal of Tuberculosis and Lung Disease* 5, 633–641.

Diel R, Seidler A, Nienhaus A, Rusch-Gerdes S & Niemann S (2005) Occupational risk of tuberculosis transmission in a low incidence area. *Respiratory Research* 6, 35.

van Embden JD, Cave MD, Crawford JT *et al.* (1993) Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *Journal of Clinical Microbiology* 31, 406–409.

Fok A, Numata Y, Schulzer M & Fitzgerald MJ (2008) Risk factors for clustering of tuberculosis cases: a systematic review of population-based molecular epidemiology studies. *International Journal of Tuberculosis and Lung Disease* 12, 480–492.

Frieden TR, Woodley CL, Crawford JT, Lew D & Dooley SM (1996) The molecular epidemiology of tuberculosis in New York City: The importance of nosocomial transmission and laboratory error. *Tubercle & Lung Disease* 77, 407–413.

Fujikane T, Fujiuchi S, Yamazaki Y *et al.* (2004) Molecular epidemiology of tuberculosis in the north Hokkaido district of Japan. *International Journal of Tuberculosis and Lung Disease* 8, 39–44.

Glynn JR, Bauer J, de Boer AS *et al.* (1999a) Interpreting DNA fingerprint clusters of *Mycobacterium tuberculosis*. European Concerted Action on Molecular Epidemiology and Control of Tuberculosis. *International Journal of Tuberculosis and Lung Disease* 3, 1055–1060.

Glynn JR, Vynnycky E & Fine PE (1999b) Influence of sampling on estimates of clustering and recent transmission of Mycobacterium tuberculosis derived from DNA fingerprinting techniques. *American Journal of Epidemiology* 149, 366–371.

Glynn JR, Crampin AC, Yates MD *et al.* (2005) The importance of recent infection with Mycobacterium tuberculosis in an area with high HIV prevalence: a long-term molecular epidemiological study in Northern Malawi. *Journal of Infectious Diseases* 192, 480–487.

Haas WH, Engelmann G, Amthor B *et al.* (1999) Transmission dynamics of tuberculosis in a high-incidence country: prospective analysis by PCR DNA fingerprinting. *Journal of Clinical Microbiology* 37, 3975–3979.

Hermans PW, Messadi F, Guebrexabher H *et al.* (1995) Analysis of the population structure of *Mycobacterium tuberculosis* in Ethiopia, Tunisia, and The Netherlands: usefulness of DNA typing for global tuberculosis epidemiology. *Journal of Infectious Diseases* 171, 1504–1513.

Hernandez-Garduno E, Kunimoto D, Wang L *et al.* (2002) Predictors of clustering of tuberculosis in Greater Vancouver: a molecular epidemiologic study. *CMAJ: Canadian Medical Association Journal* 167, 349–352.

Houben RM, Glynn JR, Fok A, Numata Y, Schulzer M & Fitzgerald JM (2009) Systematic review and analysis of population-based molecular epidemiological studies. *International Journal of Tuberculosis and Lung Diseases* 13, 275–276.

Jasmer RM, Hahn JA, Small PM *et al.* (1999) A molecular epidemiologic analysis of tuberculosis trends in San Francisco, 1991-1997. *Annals of Internal Medicine* 130, 971–978.

Jimenez-Corona ME, Garcia-Garcia L, DeRiemer K *et al.* (2006) Gender differentials of pulmonary tuberculosis transmission and reactivation in an endemic area. *Thorax* 61, 348–353.

Kamerbeek J, Schouls L, Kolk A *et al.* (1997) Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *Journal of Clinical Microbiology* 35, 907–914.

Kempf MC, Dunlap NE, Lok KH, Benjamin WH Jr, Keenan NB & Kimerling ME (2005) Long-term molecular analysis of tuberculosis strains in alabama, a state characterized by a largely indigenous, low-risk population. *Journal of Clinical Microbiology* 43, 870–878.

Kunimoto D, Sutherland K, Wooldrage K *et al.* (2004) Transmission characteristics of tuberculosis in the foreign-born and the Canadian-born populations of Alberta, Canada. *International Journal of Tuberculosis and Lung Disease* 8, 1213–1220.

Lari N, Rindi L, Sola C *et al.* (2005) Genetic diversity, determined on the basis of katG463 and gyrA95 polymorphisms, spoligotyping, and IS6110 typing, of *Mycobacterium tuberculosis* complex isolates from Italy. *Journal of Clinical Microbiology* 43, 1617–1624.

Lillebaek T, Dirksen A, Kok-Jensen A & Andersen AB (2004) A dominant *Mycobacterium tuberculosis* strain emerging in Denmark. *International Journal of Tuberculosis and Lung Disease* 8, 1001–1006.

Maguire H, Dale JW, McHugh TD *et al.* (2002) Molecular epidemiology of tuberculosis in London 1995–7 showing low rate of active transmission. *Thorax* 57, 617–622.

Matteelli A, Gori A, Pinsi G *et al.* (2003) Clustering of tuberculosis among Senegalese immigrants in Italy. *International Journal of Tuberculosis & Lung Disease* 7, 967–972.

Moro ML, Salamina G, Gori A *et al.* (2002) Two-year population-based molecular epidemiological study of tuberculosis transmission in the Metropolitan area of Milan, Italy. *European Journal of Clinical Microbiology & Infectious Diseases* **21**, 114–122.

Murray M (2002a) Determinants of cluster distribution in the molecular epidemiology of tuberculosis. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 1538–1543.

Murray M (2002b) Sampling bias in the molecular epidemiology of tuberculosis. *Emerging Infectious Diseases* **8**, 363–369.

Nava-Aguilera E, Andersson N, Harris E *et al.* (2009) Risk factors associated with recent transmission of tuberculosis: systematic review and meta-analysis. *International Journal of Tuberculosis and Lung Disease* **13**, 17–26.

Park YK, Bai GH & Kim SJ (2000) Restriction fragment length polymorphism analysis of *Mycobacterium tuberculosis* isolated from countries in the western pacific region. *Journal of Clinical Microbiology* **38**, 191–197.

Pena MJ, Caminero JA, Campos-Herrero MI *et al.* (2003) Epidemiology of tuberculosis on Gran Canaria: a 4 year population study using traditional and molecular approaches. *Thorax* **58**, 618–622.

Pfyffer GE, Strassle A, Rose N, Wirth R, Brandli O & Shang H (1998) Transmission of tuberculosis in the metropolitan area of Zurich: A 3 year survey based on DNA fingerprinting. *European Respiratory Journal* **11**, 804–808.

Ruiz Garcia M, Rodriguez JC, Navarro JF, Samper S, Martin C & Royo G (2002) Molecular epidemiology of tuberculosis in Elche, Spain: A 7-year study. *Journal of Medical Microbiology* **51**, 273–277.

Samper S, Iglesias MJ, Rabanaque MJ *et al.* (1998) The molecular epidemiology of tuberculosis in Zaragoza, Spain: A retrospective epidemiological study in 1993. *International Journal of Tuberculosis and Lung Disease* **2**, 281–287.

Scott AN, Menzies D, Tannenbaum TN *et al.* (2005) Sensitivities and specificities of spoligotyping and mycobacterial interspersed repetitive unit-variable-number tandem repeat typing methods for studying molecular epidemiology of tuberculosis. *Journal of Clinical Microbiology* **43**, 89–94.

Sharnprapai S, Miller AC, Suruki R *et al.* (2002) Genotyping analyses of tuberculosis cases in U.S.-and foreign-born Massachusetts residents. *Emerging Infectious Diseases* **8**, 1239–1245.

Small PM, Hopewell PC, Singh SP *et al.* (1994) The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. *New England Journal of Medicine* **330**, 1703–1709.

Soborg C, Soborg B, Pouelsen S, Pallisgaard G, Thybo S & Bauer J (2001) Doubling of the tuberculosis incidence in Greenland over an 8-year period (1990–1997). *International Journal of Tuberculosis and Lung Disease* **5**, 257–265.

van Soolingen D, de Haas PE, Hermans PW, Groenen PM & van Embden JD (1993) Comparison of various repetitive DNA elements as genetic markers for strain differentiation and epidemiology of *Mycobacterium tuberculosis*. *Journal of Clinical Microbiology* **31**, 1987–1995.

van Soolingen D, Borgdorff MW, de Haas PE *et al.* (1999) Molecular epidemiology of tuberculosis in the Netherlands: a nationwide study from 1993 through 1997. *Journal of Infectious Diseases* **180**, 726–736.

Sterne J (2009) *Meta-Analysis in Stata: An Updated Collection from the Stata Journal*. Stata Press, College Station, TX, 259 p.

Storla DG, Rahim Z, Islam MA *et al.* (2006) Heterogeneity of *Mycobacterium tuberculosis* isolates in Sunamganj District, Bangladesh. *Scandinavian Journal of Infectious Diseases* **38**, 593–596.

Supply P, Allix C, Lesjean S *et al.* (2006) Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. *Journal of Clinical Microbiology* **44**, 4498–4510.

Sutherland I, Svandova E & Radhakrishna S (1982) The development of clinical tuberculosis following infection with tubercle bacilli. *Tubercle* **63**, 255–268.

Thompson SG & Higgins JP (2002) How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine* **21**, 1559–1573.

Thomsen VO, Lillebaek T & Stenz F (2004) Tuberculosis in Greenland–current situation and future challenges. *International Journal of Circumpolar Health* **63**(Suppl. 2), 225–229.

Vachee A, Vincent P, Savage C *et al.* (1999) Molecular epidemiology of tuberculosis in the Nord department of France during 1995. *Tubercle & Lung Disease* **79**, 361–366.

Verver S, Warren RM, Munch Z *et al.* (2004) Transmission of tuberculosis in a high incidence urban community in South Africa. *International Journal of Epidemiology* **33**, 351–357.

Vynnycky E & Fine PE (1997) The natural history of tuberculosis: the implications of age-dependent risks of disease and the role of reinfection. *Epidemiology and Infection* **119**, 183–201.

Vynnycky E, Borgdorff MW, van Soolingen D & Fine PE (2003) Annual *Mycobacterium tuberculosis* infection risk and interpretation of clustering statistics. *Emerging Infectious Diseases* **9**, 176–183.

Weis SE, Pogoda JM, Yang Z *et al.* (2002) Transmission dynamics of tuberculosis in Tarrant county, Texas. *American Journal of Respiratory and Critical Care Medicine* **166**, 36–42.

WHO. (2008) World Health Organization; Global TB database. Available at: http://www.who.int/tb/country/global_tb_database/en/index.html (accessed on 15 February 2008).

Wilkinson D, Pillay M, Crump J, Lombard C, Davies GR & Sturm AW (1997) Molecular epidemiology and transmission dynamics of *Mycobacterium tuberculosis* in rural Africa. *Tropical Medicine & International Health* **2**, 747–753.

Zolnir-Dovc M, Poljak M, Erzen D & Sorli J (2003) Molecular epidemiology of tuberculosis in Slovenia: Results of a one-year (2001) nation-wide study. *Scandinavian Journal of Infectious Diseases* **35**, 863–868.

Corresponding Author Rein M. G. J. Houben, Infectious Disease Epidemiology Unit, London School of Hygiene and Tropical Medicine, Keppel Street, WC1E 7HT, London, UK. Tel.: +44-20-7958 4752; Fax: +44-20-7637 4314; E-mail: Rein.Houben@lshtm.ac.uk

Appendix 1. Summary of studies included in meta-regression analysis

	Study design						Study Setting & population					TB transmission	
Study Location	Study population and DNA fingerprinting methods*	Inclusion by IS6110 band number	Secondary DNA typing method (cut-off)†	Patients in DNA fingerprint analysis	Duration (months)	Sampling fraction‡	Local TB incidence (*n*\year\ 100 000)	Average age (years)	HIV positives (%)	Sex (% male)	Foreign born (%)	Clustered (%)	Recent transmission (%)
Brescia (Italy) (Matteelli *et al.* 2003)	All c + cases from Brescia province, '91–'97	All	Spol (<3)	195	84	0.30	15	47	14	67	30	27	16
Denmark & Greenland (Lillebaek *et al.* 2004)	Nearly all c + patients in Denmark & Greenland, '92–'01, 4% of all strains were second or third isolate from same patient	All	.	3936	120	0.97	4	40	.	.	58	56	.
Department Nord (France) (Vachee *et al.* 1999)	Most (~90%) c + TB cases in Department Nord, 1995§	All	.	154	12	0.66	9	54	7	73	26	18	9
Elche (Spain) (Ruiz Garcia *et al.* 2002)	Sample of all c + diagnosed patients in Elche Health district, 1993–1999. No information reported about sampling method	>4	.	141	84	0.59	99	37	29	70	.	55	38
Gran Canaria (Spain) (Pena *et al.* 2003)	All c + TB patients in Gran Canaria between 1993–1996	>4	.	566	48	0.79	29	39	16	69	7	72	58
Greenland (Denmark) (Soborg *et al.* 2001)	All identified TB patients from Greenland, 1990–1997, 15 c + cases from study region were not notified, and not included in DNA fingerprint analysis	>7¶	.	310	96	0.93	130	30	.	53	.	85	78
Greenland (Denmark) (Thomsen *et al.* 2004)	All notified patients from Greenland, 1998–2002 No data on missed cases, ~60% of notified cases were c+	>7¶	.	198	60	0.94	185	.	.	.	.	86	
Hamburg (Germany) (Diel *et al.* 2005)	All reported c + TB cases in Hamburg, 1997–2002	>4	.	848	72	0.88	16	44	57		43	34	25

Appendix 1. (*Continued*)

	Study design						Study Setting & population					TB transmission	
Study Location	Study population and DNA fingerprinting methods*	Inclusion by IS6110 band number	Secondary DNA typing method (cut-off)†	Patients in DNA fingerprint analysis	Duration (months)	Sampling fraction‡	Local TB incidence (*n*\year\ 100 000)	Average age (years)	HIV positives (%)	Sex (% male)	Foreign born (%)	Clustered (%)	Recent transmission (%)
London (United Kingdom) (Maguire *et al.* 2002)	All c + TB patients in greater London area, July 1995–December 1996	>4	.	2042	30	0.76	33	37		59	80	23	14
Milan (Italy) (Moro *et al.* 2002)	All diagnosed TB cases from Milan metropolitan area residents, 1995–1997	All	Spol (<5)	581	24	1.00	13	38	22	63	30	41	28
Netherlands (van Soolingen *et al.* 1999)	All reported cases in The Netherlands between 1993–1997	All	PGRS (<5)	4266	60	0.75	9	34	4	61	44	46	35
Norway (Dahle *et al.* 2001)	All diagnosed TB patients in Norway between 1999–2001	>4	.	485	36	0.92	7	43	0	0	71	10	6
Norway (Dahle *et al.* 2003)	All diagnosed TB patients in Norway, 1994–1998	>4	.	619	60	0.89	5	45	.	54	50	15	11
Tuscany (Italy) (Lari *et al.* 2005)	All c + TB cases in Tuscany, 2002§	All	.	248	12	1.00	7	50	11	.	37	33	19
Zaragoza (Spain) (Samper *et al.* 1998)	All c + TB patients in Zaragoza in 1993 'Nearly all samples' went to two participating labs	>4	.	226	12	0.84	32	44	44	69	4	39	27
Zurich (Switzerland) (Pfyffer *et al.* 1998)	Patients from the Zurich Canton, 1991–1993	All	PGRS (<5)	361	36	.	12	.	10	63	51	17	11
Alabama (USA) (Kempf *et al.* 2005)	All diagnosed TB cases from state of Alabama, 1994–2000	>5	.	1136	76	0.80	8	56	6	69	.	28	25
Alberta (Canada) (Kunimoto *et al.* 2004)	All c + TB cases from Alberta province, 1994–1998	All	Spol (<6)	573	60	1.00	6	45	.	48	42	20	14
Arkansas (USA) (Braden *et al.* 1997)	All c + cases in Arkensas, 1992–1993§	>5	.	192	24	0.71	7	62	.	67	3	42	30
Arkansas (USA) (Cave *et al.* 2005)	All c + TB cases in Arkensas, 1996–1999	>6	.	419	48	0.98	7	56	.	61	9	39	28
Baltimore (USA) (Bishai *et al.* 1998)	All c + cases TB reported in Baltimore City, 1994–1996	All	PGRS (<7)	182	30	1.00	15	54	28	69	3	46	32
Denver (USA) (Burman *et al.* 1997)	All c + TB cases from Denver metropolitan area, 1988–1994.	>5	.	131	66	0.63	3	.	15	72	48	28	19

Appendix 1. (*Continued*)

	Study design						Study Setting & population					TB transmission	
Study Location	Study population and DNA fingerprinting methods*	Inclusion by IS6110 band number	Secondary DNA typing method (cut-off)†	Patients in DNA fingerprint analysis	Duration (months)	Sampling fraction‡	Local TB incidence (*n*\year\ 100 000)	Average age (years)	HIV positives (%)	Sex (% male)	Foreign born (%)	Clustered (%)	Recent transmission (%)
Houston (USA) (De Bruyn *et al.* 2001)	All reported TB cases in Houston, 1995–1998	All	Spol (<5)	1139	36	0.91	20	.	19	70	70	60	53
Manitoba (Canada) (Blackwood *et al.* 2004)	All diagnosed TB cases in Manitoba province, 1992–1999§	All	.	629	96	1.00	9	45	.	57	30	68	60
Manitoba (Canada) (Blackwood *et al.* 2004)	All diagnosed TB cases Manitoba province, 2003	All	.	126	12	1.00	9	.	.	.	.	65	56
Maryland (USA) (Cronin *et al.* 2001)	All c + TB cases in Maryland, 1996–2000§	All	Spol (<7)	1172	60	0.98	5	45	12	44	46	37	28
Massachusets (USA) (Sharnprapai *et al.* 2002)	All reported TB cases in Massachusetts July 1996–December 2000§	All	Spol (<7)	983	54	0.95	4	50	27	57	70	28	19
Montreal (Canada) (Scott *et al.* 2005)	All reported TB patients in Montreal, 1996–1998	>5	.	347	36	0.95	10	40	34	55	80	8	4
New York (USA) (Frieden *et al.* 1996)	All c + TB cases in New York in April 1991	>4	.	344	0**	0.83	47	39	29	74	22	37	28
San Francisco (USA) (Burgos *et al.* 2003)	All reported TB cases in San Francisco area, 1991–1999	All	PGRS (<6)	1800	108	0.84	35	45	20	69	65	38	28
Tarrant County (USA) (Weis *et al.* 2002)	All c + TB patients resident in Tarrant County 1993–2000	All	Spol (<7)	488	96	0.59	6	45	.	67	34	60	50
Vancouver (Canada) (Blenkush *et al.* 1996)	All c + cases in Vancouver area, 1992–1994	All	.	114	18	0.67	6	52	.	.	.	12	8

Appendix 1. (*Continued*)

	Study design						Study Setting & population					TB transmission	
Study Location	Study population and DNA fingerprinting methods*	Inclusion by IS6110 band number	Secondary DNA typing method (cut-off)†	Patients in DNA fingerprint analysis	Duration (months)	Sampling fraction‡	Local TB incidence (*n*\year\ 100 000)	Average age (years)	HIV positives (%)	Sex (% male)	Foreign born (%)	Clustered (%)	Recent transmission (%)
Vancouver (Canada) (Hernandez-Garduno *et al.* 2002)	All new c + TB cases in Greater Vancouver area, 1995–1999 five cases were excluded for having experienced a previous TB episode	All	Spol (<6)	791	51	0.98	6	51	5	54	83	17	12
Wisconsin (USA) (Cowan *et al.* 2005)	All c + TB cases in Wisconsin, 2000–2003	>6	.	200	46	0.99	2	.	.	.	.	16	10
Hokkaido (Japan) (Fujikane *et al.* 2004)	All diagnosed patients in Hokkaido prefecture, 2001	>5	.	207	36	0.83	20	69	.	.	.	8	4
Hong Kong (Chan-Yeung *et al.* 2006)	All c + TB cases with residence on Hong Kong island, May 1999–April 2002	All	PGRS (<6)	1533	36	0.66	108	58	1	68	63	30	20
Cape Town (South Africa) (Verver *et al.* 2004)	All patients diagnosed with TB that reported in and are residents of two high incidence urban communities of Cape Town. 1993–1998	All	Spol (<5)	797	72	0.78	761	33	.	57	.	72	58
Hlabisa (South Africa) (Wilkinson *et al.* 1997)	All consecutive SS + cases in Hlabisa, a rural district in South Africa, May 1993–March 1994	All	PGRS (<5)	246	11	1.00	305	36	30	62	.	45	29
Karonga (Malawi) (Glynn *et al.* 2005)	All c + TB cases in Karonga district, 1995–2003	>4	.	948	87	0.82	81	33	65	47	.	72	59
Malaysia (Dale *et al.* 1999)	Nationwide random sample of c + TB cases in Malaysia, 1993–1994	>4	.	331	24	0.03	58	45	.	.	.	11	6
Republic of Korea (Park *et al.* 2000)	Multistage stratified cluster sample of Korean population	>4	.	136	0**	0.002	98	55	.	69	.	11	7

Appendix 1. (*Continued*)

	Study design						Study Setting & population					TB transmission	
Study Location	Study population and DNA fingerprinting methods*	Inclusion by IS6110 band number	Secondary DNA typing method (cut-off)†	Patients in DNA fingerprint analysis	Duration (months)	Sampling fraction‡	Local TB incidence (*n*\year\ 100 000)	Average age (years)	HIV positives (%)	Sex (% male)	Foreign born (%)	Clustered (%)	Recent transmission (%)
Sunamganj district (Bangladesh) (Storla *et al.* 2006)	All SS + TB cases come from four sub-districts in Sunamganj district (Northern Bangladesh), November 2003–December 2004	>4	.	106	13	0.95	111	.	.	.	.	6	3
Tirruvalur District (India) (Das *et al.* 2005)	All DOTS notified TB cases in Tirruvallur district, 1999–2000	>5	.	151	19	0.78	111	45	.	76	.	19	10
Veracruz (Mexico) (Jimenez-Corona *et al.* 2006)	All c + patients in Veracruz state, March 1995–April 2003	All	Spol (<6)	623	12††	1.00	28	44	2	52	.	25	18
Slovenia (Zolnir-Dovc *et al.* 2003)	Nearly all (99.7%) of c + patients in Slovenia, 2001	>4	.	304	12	0.99	19	55	1	61	24	38	26
East Azerbadjan (Iran) (Asgharzadeh *et al.* 2006)	All c + TB cases in East Azerbadjan September 2002–March 2003	All	.	105	7	0.82	8	47	.	57	.	33	23

c = culture positive; '.' = data not available or found; % = percentage of all study participants; Spol, Spoligotyping; PGRS, polymorphic GC–repetitive sequence typing; 'no.' = number; '~' = approximately.

*Sample collection as reported by the authors.

†Secondary DNA fingerprinting method used for clustering analysis, if number of IS6110 RFLP bands is below cut-off (shown in parentheses).

‡Sampling fraction was based on proportion of all c+ cases that had RFLP results available.

§One band difference in IS6110 RFLP pattern was allowed between clustered strains.

¶All typed Mtb strains had >7 IS6110 bands in their RFLP pattern.

**Cross sectional survey, duration set to 0 months.

††Time difference between first and secondary patients in cluster limited to 12 months.

Appendix 2. Calculation of relative difference in proportion clustered

		Study location (year of publication)		
Variable	Coefficient*	Hokkaido (Fujikane *et al.* 2004)	Cape Town (Verver *et al.* 2004)	Arkensas (Braden *et al.* 1997)
Constant†	−12	X	X	X
Study duration (months)				
0–12	0			
13–48	−3.2	X		X
>48	18.3		X	
Sampling fraction proportion of culture positive cases included				
0–0.50	0			
0.50–0.75	27.7			
0.75–1	29.6	X	X	X
Low band strains				
Excluded	0	X		X
Included with secondary typing	0.6		X	
Included, no secondary typing	25.4			
TB burden in study area				
Low (≤10/100 000/year)	0			X
Medium (11–50/100 000/year)	17.9	X		
High (>50/100 000/year)	25.4		X	
Expected proportion clustered		32.3	61.9	14.4
Observed proportion clustered‡		8	72	42
Relative difference§		−75.2%	+16.3%	+191.6%

*Coefficients give the change in the expected proportion clustered for each category.

†The baseline value of the proportion clustered.

‡As reported by the study. See appendix 1 for details.

§Relative difference is calculated as (observed–expected/expected) × 100%.

10.2 Systematic review and analysis of population-based molecular epidemiological studies

INT J TUBERC LUNG DIS 13(2):275–276
© 2009 The Union

Correspondence

Systematic review and analysis of population-based molecular epidemiological studies

Fok et al. recently described the results of a systematic literature review of TB molecular epidemiological studies that used IS*6110* RFLP as the primary genotyping method.[1] A total of 36 studies were included, and factors associated with the proportion clustered were identified. While we welcome this timely attempt to make sense of the variation in reported clustering, we raise some questions regarding their study selection and analysis methods.

The authors state that included studies should describe populations of TB cases that represent (or are a random sample of) all cases in a geographically defined area (their inclusion criterion 1 and reference 2). We agree that this is necessary, but not all included studies appear to comply strictly with this criterion. The USA sentinel study[3] treats cases from seven geographically separate areas as a single population, which is not the same as a random sample of all TB cases in the USA. Similar problems exist for a further seven studies (from Los Angeles, Ile de France, Manhattan, Turkey, Madrid [2] and Equatorial Guinea).

In addition, the authors' use of the search term 'cluster' as a conditional rather than optional term may have led to relevant studies being missed: in a similar systematic review we found not 1413 but close to 12 000 hits, and 47 eligible studies. Twenty-one of these were not included in this review, all of which could have contributed to the meta-regression analyses and at least eight of which appear to meet the stated inclusion criteria for the meta-analysis.

We also have some concerns about the analytical method. Meta-regression makes two important assumptions that are analogous to linear regression, and it is not clear from the paper if these were met. The first assumption is that a linear association exists between the dependent and explanatory variable.[4] The paper does not report on this, even though previous studies have shown that study duration is not linearly associated with the proportion clustered. [5] The second assumption is that the residuals are normally distributed.[4] Again the paper does not report on this, and it seems likely that this was not fulfilled for the incidence rate, which appears to be highly skewed (Table 2: mean = 40.1, SD = 86.1).

Finally, we are concerned about the inclusion of the average cluster size as an explanatory variable, as it is not sufficiently independent of the measured proportion clustered. It would be interesting to repeat the analysis without it.

The number of studies reporting results on TB molecular epidemiology has grown considerably in the past decade, allowing researchers to summarise available evidence and explore differences between populations. However, the methods involved are complicated. Through debate we hope that a consensus can be reached on the most rigorous and unbiased approach, thus furthering an exciting and challenging field.

R. M. G. J. Houben
J. R. Glynn
Department of Epidemiology and Public Health
London School of Hygiene and Tropical Medicine
London, United Kingdom
e-mail: Rein.Houben@lshtm.ac.uk

References

1 Fok A, Numata Y, Schulzer M, FitzGerald J M. Risk factors for clustering of tuberculosis cases: a systematic review of population-based molecular epidemiology studies. Int J Tuberc Lung Dis 2008; 12: 480–492.

2 Glynn J R, Bauer J, de Boer A S, et al. Interpreting DNA fingerprint clusters of *Mycobacterium tuberculosis*. European Concerted Action on Molecular Epidemiology and Control of Tuberculosis. Int J Tuberc Lung Dis 1999; 3: 1055–1060.

3 Ellis B A, Crawford J T, Braden C R, McNabb S J, Moore M, Kammerer S, National Tuberculosis Genotyping and Surveillance Network Work Group. Molecular epidemiology of tuberculosis in a sentinel surveillance population. Emerg Infect Dis 2002; 8: 1197–1209.

4 Thompson S G, Sharp S J. Explaining heterogeneity in meta-analysis: a comparison of methods. Stat Med 1999; 18: 2693–2708.

5 Glynn J R, Crampin A C, Yates M D, et al. The importance of recent infection with *Mycobacterium tuberculosis* in an area with high HIV prevalence: a long-term molecular epidemiological study in Northern Malawi. J Infect Dis 2005; 192: 480–487.

In reply

We would like to thank Drs. Houben and Glynn for their thoughtful comments with regard to the inclusion of studies that did not strictly comply with our selection criteria. As not all articles provided detailed information about study settings, we decided initially to be inclusive and intended to identify a comprehensive collection of studies on risk factors for clustering of TB cases. We conducted subgroup analyses instead, by excluding certain studies during the analysis. For example, whether TB registry or active surveillance data had been utilized was incorporated into our subgroup selection. We had also conducted the analysis without the US sentinel study and reported that our meta-regression results were not greatly influenced by this study. Of the eight studies that Drs. Houben and

Glynn identified as problematic, two were from low-incidence countries and six were from high-incidence countries. We repeated meta-analyses without these two[1,2] for the low-incidence group. The results did not change significantly. For the high-incidence group, however, the number of available studies became too small to obtain reliable estimates. We consider that assessing the impact of particular studies could be more informative than excluding them upfront. We appreciate their input regarding the need for additional subgroup analyses.

Our systematic review was designed to estimate the impact of commonly investigated risk factors for TB clustering. We therefore selected studies with five criteria besides the one mentioned in the letter. To provide a different perspective, we attempted to investigate associations between TB clustering proportions and prevalence of risk factors among those studies selected for our primary objective. We should mention that our selection criteria were not meant to address the associations between TB clustering and study characteristics per se. In fact we excluded many studies that reported TB clustering proportions, because these studies did not report on risk factors. We agree that those who aim to investigate heterogeneity of TB clustering need to undertake a broader search.

Regarding our analytical approach, we confirmed that our dependent variable, the TB clustering proportion, was normally distributed within 36 studies (the Shapiro-Wilk W test $W = 0.97$, P value $= 0.43$) while distributions of explanatory variables for regression analysis need not be normal.[3] We are satisfied with the basic statistical assumption of residual normality in conducting our meta-regression. Further, we tested each model for non-linearity by adding higher order terms. In no case did these added terms contribute significantly to the fit.

Drs. Houben and Glynn expressed their concerns about including the average cluster size in the models. As stated in our discussion, we investigated whether maximum cluster size could serve as a better parameter. The number of studies that reported this value, however, limited us from proceeding. Statistically speaking, nonetheless, the F test comparing the two models, with vs. without the average cluster size, was highly significant ($P < 0.001$) in improving the model. We examined the impact of excluding this variable and noted that the coefficients of other two variables became somewhat larger, while their P values remained significant and comparable.

A. Fok
Y. Numata
M. Schulzer
J. M. FitzGerald
Center for Clinical Epidemiology and Evaluation
Vancouver, British Columbia, Canada
e-mail: markf@interchange.ubc.ca

References

1 Barnes P F, Yang Z, Preston-Martin S, et al. Patterns of tuberculosis transmission in Central Los Angeles. JAMA 1997; 278: 1159–1163.

2 Ellis B A, Crawford J T, Braden C R, et al. Molecular epidemiology of tuberculosis in a sentinel surveillance population. Emerg Infect Dis 2002; 8: 1197–1209.

3 Montgomery D C, Peck E A, Vining D C. Introduction to linear regression analysis. 4th ed. Hoboken, NJ: Wiley-Interscience, 2006.

10.3 HIV and the risk of tuberculosis due to recent transmission over 12 years in Karonga District, Malawi

Transactions of the Royal Society of Tropical Medicine and Hygiene (2009) **103**, 1187—1189

available at www.sciencedirect.com

ScienceDirect

journal homepage: http://www.elsevier.com/locate/trstmh

SOCIETY MEETING PAPER

HIV and the risk of tuberculosis due to recent transmission over 12 years in Karonga District, Malawi☆

Rein M.G.J. Houben[a,b,*], **Amelia C. Crampin**[a,b], **Kim Mallard**[c], **J. Nimrod Mwaungulu**[a], **Malcolm D. Yates**[d], **Frank D. Mwaungulu**[a,✠], **Bagrey M.M. Ngwira**[a,b], **Neil French**[a,b], **Paul E.M. Fine**[a,b], **Judith R. Glynn**[a,b]

[a] *Karonga Prevention Study, Malawi*
[b] *Infectious Disease Epidemiology Unit, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK*
[c] *Pathogen Molecular Biology Unit, London School of Hygiene and Tropical Medicine, London, UK*
[d] *Mycobacterium Reference Unit, Health Protection Agency, London, UK*

Received 6 March 2009; received in revised form 11 March 2009; accepted 12 March 2009
Available online 11 April 2009

KEYWORDS
HIV;
Tuberculosis;
Mycobacterium tuberculosis;
Molecular epidemiology;
Transmission;
Malawi

Summary Tuberculosis (TB) patients with strains common to other recent cases ('clustering') suggest recent transmission. HIV status and age may affect proportions clustered. We investigated TB clustering by HIV and age in a population-based study in Malawi. Among 746 patients, HIV infection increased the proportion clustered. Sex-period-adjusted odds ratios for the association of HIV and clustering were 1.26 (95% CI 0.4—4.1) for ages 15—25 years, 1.40 (0.9—2.3) for 25—50 years and 10.44 (2.3—47.9) for >50 years and remained stable over two periods examined. These results suggest that HIV increases the proportion of TB due to recent transmission in the elderly.

1. Introduction

Tuberculosis (TB) is the main cause of death in HIV-positive patients from low-resource settings, and HIV infection increases an infected person's risk of active TB disease from 10% in their lifetime to 10% annually.[1] However, the relative effect of HIV on TB due to recent or past infection is not known.[2]

Active TB disease follows recent infection or reinfection with *Mycobacterium tuberculosis* (Mtb), or reactivation of a

☆ Based on a presentation to the Royal Society of Tropical Medicine & Hygiene Research in Progress meeting on 18 December 2008. This oral presentation was awarded third prize at the meeting.

* Corresponding author. Tel.: +44 20 7927 2668; fax: +44 20 7637 4314.
E-mail address: rein.houben@lshtm.ac.uk (R.M.G.J. Houben).

✠ Deceased.

0035-9203/$ — see front matter © 2009 Royal Society of Tropical Medicine and Hygiene. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.trstmh.2009.03.013

latent Mtb infection.[3] It is assumed that cases with identical strains are likely to represent recent infection and that each 'cluster' of cases with common strains contains one index case due to reactivation.[4] The number of cases in clusters, minus the number of clusters, as a proportion of all cases, thus gives an estimate of the proportion of TB due to recent transmission (the 'n − 1' method).[4]

We have previously shown that HIV increased the proportion of TB cases clustered in rural Malawi, but only among older adults.[5] Here we repeat the analysis, including four more years of data to investigate the robustness of the finding and the stability of this association between HIV and TB clustering over time.

2. Materials and methods

The Karonga Prevention Study in rural Malawi has been collecting TB molecular epidemiological data since late 1995, with results now available up to October 2007. The methods have been described previously: all patients with TB in Karonga District (population ∼250 000) were included and were asked to undergo HIV testing.[6] Isolates from all culture-confirmed TB cases were typed by *IS*6110 RFLP fingerprinting.[5]

After excluding possible cross-contamination, TB cases were considered clustered if another patient had an identical Mtb strain in the previous 4 years, based on a previous study in Karonga showing that maximum clustering was reached within 4 years.[5] Cases in the first 4 years were used to determine cluster status of subsequent cases, but were then excluded from the analysis. This retrospective clustering[5] with a fixed time window gives an estimate of 'n − 1' clustering, and allows comparison between time periods that is unbiased by the total duration of the study.

For the main statistical analysis two periods were compared: using data that have previously been reported (October 1999—March 2003)[5] and new data (April 2003—October 2007). Cases were stratified into three age groups (15—25, 26—50 and >50 years).

Multivariate logistic regression (Stata v.10; Stata Corp., College Station, TX, USA) was used to calculate age and period-stratified odds ratios (ORs) for the association of HIV with TB clustering, adjusted for sex. Interactions were assessed through likelihood ratio tests.

To test the robustness of the results we repeated the statistical analyses, including cases from the first 4 years, comparing either two 6-year periods or three 4-year periods. We also explored the impact of restricting clustering to time windows of either 1 or 2 years,[5] or expanding clustering to any case with an identical Mtb strain in the study period, both retrospectively and prospectively.

3. Results

DNA fingerprints were available for 1630/1968 (83%) of all culture-positive cases between late 1995 and October 2007. The median age was 35 years (range 17—85 years) and 767/1630 (47.1%) were male. Excluding the cases in the first 4 years, 705/1031 (68.4%) of cases were clustered with a case in the previous 4 years, and 493/746 (66.1%) were HIV-positive.

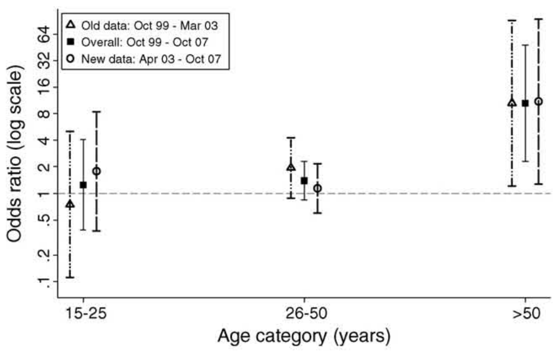


Figure 1 Odds ratios for tuberculosis clustering according to HIV status, by study period. All odds ratios are stratified for age and adjusted for sex. Overall odds ratios are adjusted for study period. Error bars show 95% CIs for the odds ratio.

An interaction was shown between age and HIV in the overall model (*P*-value LR test = 0.05) and both study periods (*P*-value LR tests = 0.1). The models (Figure 1) show that in both time periods and overall the ORs in the young and middle age categories were not statistically different from 1. However, in the older age group HIV infection was associated with increased clustering. This pattern was the same in all sensitivity analyses.

Overall, in the two periods together, the proportion retrospectively clustered among the HIV-negative was 36/49 (73%), 77/122 (63%) and 41/82 (50%) in age groups 15—25, 26—50 and >50 years, respectively. The equivalent figures for the HIV-positive were 36/45 (80%), 297/411 (72%) and 31/37 (84%).

4. Discussion

This paper shows that the association we reported previously between HIV and TB clustering, at least in the elderly, persists and is thus very unlikely to be due to chance. Combined with the observation that HIV-positive TB cases are on average less likely to be the source of Mtb transmission,[1,7], these results strengthen the hypothesis that HIV mainly increases the risk of TB disease due to recent (re)infection. Further work is needed to study how antiretroviral therapy (ART) affects the association between HIV and TB clustering. ART has been available in Karonga since June 2005 and has already been shown to reduce mortality in the population.[8]

Although efforts to control TB in settings with generalized HIV epidemics should always be multifaceted, our results suggest that measures aimed at reducing TB disease attributable to recent transmission could be more effective in reducing TB incidence than concentrating on those with latent infection.

Authors' contributions: JRG, RMGJH and ACC contributed to the conception and design of the study; RMGJH, ACC, KM, JNM, MDY, FDM, BMMN, NF, PEMF and JRG all acquired data or actively participated in data-management activities; RMGJH performed the statistical analysis and drafted the report. All authors commented on the manuscript, suggested revisions

and read and approved the final version. RMGJH and JRG are guarantors of the paper.

Acknowledgements: We thank the Government of the Republic of Malawi for their interest in, and support of, the Karonga Prevention Study.

Funding: The study was funded by The Wellcome Trust UK and the British Leprosy Relief Association.

Conflicts of interest: None declared.

Ethical approval: The study was approved by the National Health Sciences Research Committee of Malawi (reference numbers HSRC-64-96, NHSRC-01-38, NHSRC 424) and the Ethics Committee of the London School of Hygiene and Tropical Medicine, UK (reference numbers 384, 745A, 5067).

References

1. Corbett EL, Watt CJ, Walker N, Maher D, Williams BG, Raviglione MC, et al. The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. *Arch Intern Med* 2003;**163**:1009—21.
2. Girardi E, Raviglione MC, Antonucci G, Godfrey-Faussett P, Ippolito G. Impact of the HIV epidemic on the spread of other diseases: the case of tuberculosis. *AIDS* 2000;**14**(Suppl 3):S47—56.
3. Vynnycky E, Fine PE. The natural history of tuberculosis: the implications of age-dependent risks of disease and the role of reinfection. *Epidemiol Infect* 1997;**119**:183—201.
4. Small PM, Hopewell PC, Singh SP, Paz A, Parsonnet J, Ruston DC, et al. The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. *N Engl J Med* 1994;**330**:1703—9.
5. Glynn JR, Crampin AC, Yates MD, Traore H, Mwaungulu FD, Ngwira BM, et al. The importance of recent infection with *Mycobacterium tuberculosis* in an area with high HIV prevalence: a long-term molecular epidemiological study in Northern Malawi. *J Infect Dis* 2005;**192**:480—7.
6. Crampin AC, Glynn JR, Ngwira BM, Mwaungulu FD, Ponnighaus JM, Warndorff DK, et al. Trends and measurement of HIV prevalence in northern Malawi. *AIDS* 2003;**17**:1817—25.
7. Crampin AC, Glynn JR, Traore H, Yates MD, Mwaungulu L, Mwenebabu M, et al. Tuberculosis transmission attributable to close contacts and HIV status, Malawi. *Emerg Infect Dis* 2006;**12**: 729—35.
8. Jahn A, Floyd S, Crampin AC, Mwaungulu F, Mvula H, Munthali F, et al. Population-level effect of HIV an adult mortality and early evidence of reversal after introduction of antiretroviral therapy in Malawi. *Lancet* 2008;**371**:1603—11.