INCIDENCE OF TUBERCULOSIS IN HIGH AND MEDIUM BURDEN COUNTRIES: INSIGHTS FROM THE TRANSMISSION CYCLE

by

MARIA EUGENIA CASTELLANOS REYNOSA

(Under the Direction of Cristopher Whalen)

ABSTRACT

In October 2018, the United Nations General Assembly confirmed its commitment to eliminate tuberculosis by 2030. To achieve this goal, it would be required to urgently reduce the incident cases around the world. This dissertation was designed to further our understanding of what factors contribute to tuberculosis incidence in high and medium burden tuberculosis countries. We founded our work in the model of the cycle of tuberculosis transmission, studying factors that affect both individual and population levels.

At the individual level, exposure leads to infection and then disease. The risk from exposure to infection increases according to the contact rate, but the definition of adequate contact for transmission is still poorly understood. In the first and second aims of the study we examined the nature of the interaction between tuberculosis cases and their social contacts in a high-burden country, Uganda and developed a method to estimate adequate contact for transmission of *Mycobacterium tuberculosis* as the conjunction of two domains-setting and relationship. We proved that these domains affected the likelihood of infection with *M. tuberculosis* for members of a social network of a tuberculosis case, particularly children contacts.

At the population level, clustering of cases might indicate a recent transmission chain that feeds again into new infections. In the third aim of this research we characterized the proportion

of clustered tuberculosis cases based on genotypic matching in Guatemala between 2010 and 2014, providing for the first time an insight in the molecular epidemiology of tuberculosis in this middle-burden country. We found high levels of ongoing transmission of *M. tuberculosis* in Guatemala as indicated by clustering in a convenience sample. Moreover, we detected previously unreported strains of *M. tuberculosis* that contribute to tuberculosis morbidity in the country.

Tuberculosis affects disproportionally to the marginalized population. Continued efforts like ours to increase our knowledge and understanding of the factors that contribute to the burden of tuberculosis in low-income settings are urgently required. This work could provide the basis for innovative measures and evidence-based policies that would effectively halt tuberculosis transmission and occurrence in these areas, leading the pathway for a feasible global elimination.

INDEX WORDS: tuberculosis, incidence, transmission, contact, clustering, genotyping, Uganda, Guatemala

INCIDENCE OF TUBERCULOSIS IN HIGH AND MEDIUM BURDEN COUNTRIES: INSIGHTS FROM THE TRANSMISSION CYCLE

by

MARIA EUGENIA CASTELLANOS REYNOSA

B.Sc., Universidad de San Carlos de Guatemala, Guatemala, 2005

MSc., University of Liverpool, United Kingdom, 2008

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2019

INCIDENCE OF TUBERCULOSIS IN HIGH AND MEDIUM BURDEN COUNTRIES:

INSIGHTS FROM THE TRANSMISSION CYCLE

by

MARIA EUGENIA CASTELLANOS REYNOSA

Major Professor:	Cristopher Whalen
Committee:	Mark Ebell
	Kevin K. Dobbin
	Frederick D. Quinn

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2019

DEDICATION

To God and my Mother of Heaven, today, now and always. In memory of Silvita my beloved cousin, whose faith, grace and strength in the face of adversity continues to be one of the main inspirations of my life.

ACKNOWLEDGEMENTS

Thank you to Dr. Cristopher Whalen for this guidance, patience and teachings over these years. Thank you for being such a kind human being and mentor who teach me about the importance and integrity of our work.

Thank you to Dr. Mark Ebell, Dr. Fred Quinn and Dr. Kevin Dobbin. All of you were great teachers and key instruments in the completion of this thesis

Thank you to all the EIA team for all the great feed-back provided in all my enterprises during my time in Athens.

Thanks to all the friends that have being at my side during these years. I have been extremely fortunate to have you all and you are close to my heart. Dear Mari, thank you for your quiet but constant presence and help in my life and during these years. I could have not achieved this without you.

Finally, my biggest thanks are to my beloved family, back in Guatemala. Thank you for being my strength and my main source of calmness during these times.

Table of Contents

LIST OF TABLES

Page

LIST OF SUPPLEMENTARY TABLES

LIST OF FIGURES

Page

LIST OF SUPPLEMENTARY FIGURES

LIST OF EQUATIONS

Chapter 1
INTRODUCTION

Despite many efforts from public, private and civil sectors, 10.4 million (range, 8.7-12.2 million) new cases of tuberculosis are estimated to have occurred globally in 2015 (World Health Organization., 2016). Roughly, 11% of these cases occurred in HIV-infected individuals, with this co-infection affecting principally Africa (World Health Organization., 2016). Moreover, tuberculosis continues to have high mortality, accounting for an estimated 1.4 million deaths in HIV-negative individuals and 0.4 million deaths in those with HIV infection (World Health Organization., 2016).

The World Health Organization has created a post-2015 global tuberculosis strategy framework called the "End TB strategy". This framework sets ambitious milestones, targeting a 50% reduction in the tuberculosis incidence rate in 2025 and a 90% reduction by 2035 (WHO, official text approved by the 67th World Health Assembly, 2014). However, currently, the annual reduction of TB incidence has been estimated to be approximately 1.5%, lower than the 4-5% per year by 2020 estimated to be needed to accomplish the goals of this global strategy (World Health Organization., 2016).

The risk of acquiring tuberculosis infection after exposure is determined both by exogenous factors, such as crowding, smoking and host characteristics such as HIV and malnutrition (Narasimhan, Wood, Macintyre, & Mathai, 2013). After infection, the risk of progressing to active disease is mainly determined by host related characteristics (Narasimhan et al., 2013).

The goal of this dissertation was to further the understanding of what factors contribute to tuberculosis incidence in low-income settings in order to advise policy measures preventing tuberculosis transmission. We focused our project in two countries. Uganda, in East Africa, a high burden tuberculosis country with over 200 new cases per 100,000 population and is one of the 30 high tuberculosis/HIV

burden countries, globally (World Health Organization., 2016). Another one is Guatemala, a Central American country, with 25 new cases per 100,000 population and with the highest burden of tuberculosis in Central America (Pan American Health Organization, 2013; World Health Organization., 2016).

The present dissertation is founded in the model of the cycle of tuberculosis transmission (Figure 1.1).

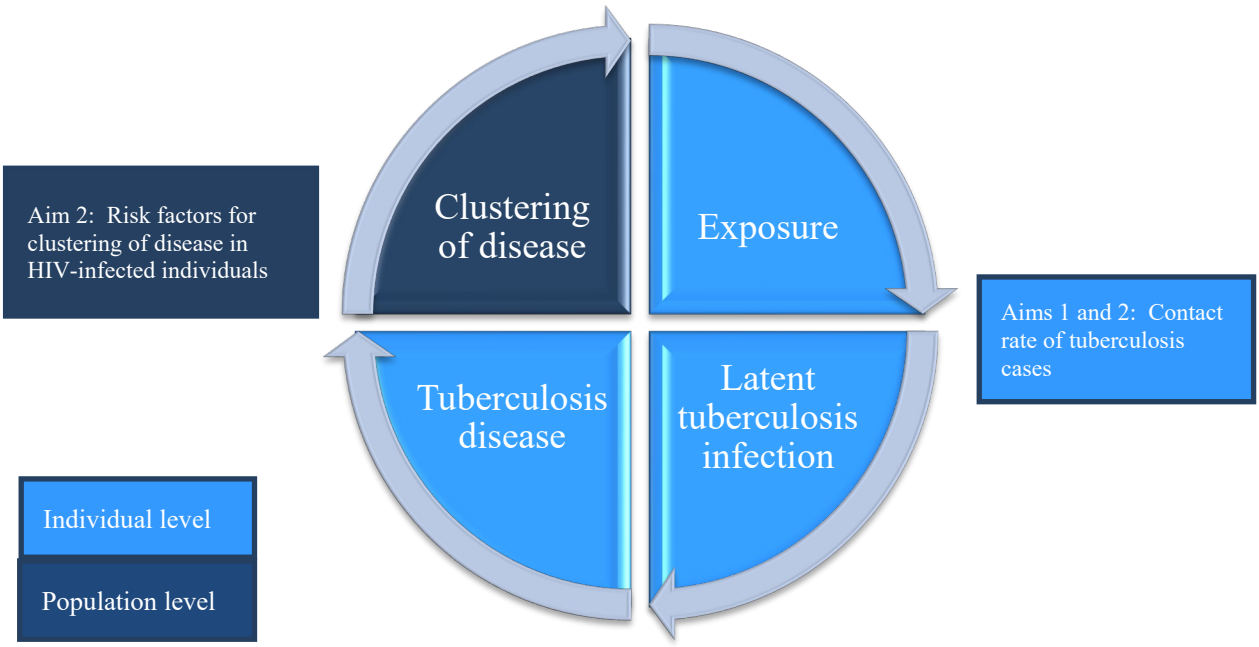


Figure 1.1. Cycle of tuberculosis transmission and aims of this study

Epidemics of tuberculosis are driven by the occurrence of incident cases. The incidence rate of tuberculosis infections can be expressed with three factors: contact rate between individuals, transmission probability and prevalence (Thomas & Weber, 2001). The overarching goal of Aims 1 and 2 was to increase our understanding of what constitutes adequate contact between an infectious case and a susceptible host (i.e., contact). In Aim 1 we defined adequate contact for tuberculosis transmission in an African urban environment by examining the interaction within ego-centric networks and developed a

score that measures the degree of contact. In Aim 2 we determined whether this score covaries with the presence of tuberculosis infection among contacts of tuberculosis cases in an African urban environment.

At the population level, clustering of cases might indicate a recent transmission chain that feeds again into new infections. In Aim 3, we estimated the level of tuberculosis transmission by measuring the proportion of clustered tuberculosis cases based on genotypic matching and identified potential risk factors associated with clustered strains that might indicate a recent transmission chain (Davies, Gordon, & Davies, 2014).

Specific Aims

Aims 1 and 2.

To define adequate contact for transmission of *Mycobacterium tuberculosis* in an African urban environment. This overarching aim will be addressed through two specific aims:

Aim 1. To define adequate contact for tuberculosis transmission in an African urban environment by examining the interaction within ego-centric networks and develop a score that measures the degree of contact.

Rationale. The definition of "contact" when studying the spread of respiratory or close-contact transmitted infectious diseases is not standardized but generally, a contact occurs when, at a minimum, a short face-to-face conversation occurs within a short distance and/or physical contact (Dodd et al., 2016; Edmunds, O'callaghan, & Nokes, 1997; Mossong et al., 2008). Further characterization requires, among others features, knowledge of the order, frequency and duration of the contact (Bansal, Read, Pourbohloul, & Meyers, 2010). Additionally, several factors may modify the frequency and nature of the contact between an infectious case and their contacts, such as age and gender of the individuals, usual place of interaction, and ventilation of the setting (Dodd et al., 2016; Feenstra, Nahar, Pahan, Oskam, & Richardus, 2013; Johnstone-Robertson et al., 2011; Melegaro, Jit, Gay, Zagheni, & Edmunds, 2011; Mossong et al., 2008).

In order to define adequate contact for tuberculosis transmission, all these factors should be included in studies aiming to understand the dynamics of social mixing among population. These variables tend to be highly correlated, so methods such as principal component analyses and exploratory factor analyses (EFA) have been used to detect the interrelationships among observed variables using data reduction (Pett, Lackey, & Sullivan, 2003).

In Aim 1 we carried out an exploratory factor analysis from information collected from a social network survey conducted among tuberculosis cases that assessed their social mixing with contacts within established ego-centric networks in Rugaba, Uganda. Our main aim was to identify underlying factors that would explain the level of contact among them. As a secondary aim we evaluated the construct validity of these factors by evaluating their association with other variables related to social mixing.

Aim 2. To determine whether the contact score covaries with the presence of tuberculosis infection among social contacts of tuberculosis cases.

Rationale. There is variability in the infectiousness of tuberculosis cases, and that variability depends on the contact rate among a tuberculosis case and his/her contact, the probability of transmission and the prevalence of the infection in the population. The prevalence of tuberculosis, although challenging, can be measured using several methods, most notably surveillance data from tuberculosis programs or by population-based surveys (Glaziou, Van der Werf, Onozaki, & Dye, 2008). Case reports and estimation of secondary attack rates in high endemic areas have shown that the probability of transmission given adequate contact seems to cover a wide spectrum. However, the third component that defines incidence, the contact rate is still poorly defined.

Several studies have highlighted the heterogeneity and complexity of the social contact patterns among human populations (Dodd et al., 2016; Mossong et al., 2008; Wallinga, Teunis, & Kretzschmar, 2006). Moreover, we know that tuberculosis occurs in clusters and disproportionately impacts certain high-risk groups (Sulis, Roggi, Matteelli, & Raviglione, 2014). Research is warranted to further understand social mixing dynamics among tuberculosis cases and their contacts. Nevertheless, the

quantification of adequate contact between a tuberculosis case and their social network has not been performed in African settings. In this Aim 2, we assessed the congruence between the contact scores developed in Aim 1 and the presence of tuberculosis infection among the social networks of tuberculosis index cases in Rugaba, Uganda.

Aim 3. To characterize the proportion of clustered tuberculosis cases based on genotypic matching in Guatemala City, Guatemala between 2010 and 2014 and to identify risk factors associated with these clustered cases in HIV-infected subjects.

Rationale. *Mycobacterium tuberculosis* strains with shared genotypes are considered clustered. Clustered strains represent a chain of transmission, and may represent ongoing, or recent, transmission, depending on the sampling interval of cases. Thus, when the sample collection is restricted in time, and in a well-defined geographical area (J. Glynn et al., 1999) we can infer that clustered cases represent recent transmission; the assumption being that clusters are "epidemiologically linked chains of recently transmitted disease" (Murray & Nardell, 2002). Strains with unique genotypes are thought to represent reactivation of an old tuberculosis infection and are considered non-clustered. Clustered strains are affected by several host and population-level characteristics (Fok, Numata, Schulzer, & FitzGerald, 2008; M. Murray, 2002). Age structure, prevalence of latent tuberculosis infection, and HIV prevalence are population-level variables that have been shown to effect cluster distribution. Individual host characteristics that influence levels of tuberculosis clustering include (but are not limited to) place of birth, pulmonary tuberculosis disease (rather than extrapulmonary), and alcohol abuse (Fok et al., 2008; M. Murray, 2002). Several studies have been performed evaluating levels of tuberculosis clustering and potential risk factors for increased cluster risk. Most of these studies, however, were conducted in low-incidence tuberculosis settings and little is known about relevant risk factors for recent transmission in HIV-infected patients in settings with a medium-burden TB burden, such as Guatemala. Few molecular tuberculosis studies have been published from Guatemala, and knowledge of existing circulating genotypes and the identification of risk factors associated with recent transmission will allow an

evidenced-based approach for health policymakers to direct and concentrate targeted tuberculosis control measures to high-risk populations.

Chapter 2
LITERATURE REVIEW

Transmission Dynamics of Infectious Diseases

An infectious disease can be studied by focusing on the dynamics of the disease or the dynamics of the infectiousness (Thomas & Weber, 2001). In the context of public health, the study of the dynamics of infectiousness is essential, as the success of an infectious disease relies on its capacity for transmission to susceptible hosts (J. e. Cohen, Powderly, & Opal, 2017).

Mathematical models have been developed to try to better explain the dynamics of transmission of infectious diseases. One of the simplest ones is the binomial model. In this model, the probability of transmission ("*p*") is the frequency of infection after one exposure. The assumption for this model is that each exposure to an infectious host is discrete and independent. If we want to add to the model the possibility that contact happens in continuous time, then the probability of being infected per unit time if all contacts occur with infectious individuals can be described as *cp,* "*c*" being the contact rate per unit time (Thomas & Weber, 2001).

Halloran developed a modified model to estimate the probability of infection from a contact with unknown infection status, this is $\rho=pP$. P represents the probability that an individual with whom contact is made is infectious and ρ represents an infection probability (Thomas & Weber, 2001). These expressions can help us to study the dynamics of infectious diseases, as epidemics are driven by the occurrence of incidence cases. So, incidence rate can be expressed with three factors: contact rate between individuals; transmission probability "*p*", which is the probability that "a contact between an infectious individual and a susceptible host leads to a successful transmission event"; and "P", which is the probability that an individual contact occurs between a susceptible individual and an infectious

individual (Equation 2.1). Usually, "P" is assumed to be equivalent to the prevalence, fraction of infectious individuals in the total population at a given time (Real & Biek, 2007; Thomas & Weber, 2001).

I (t) = c(t)*p*P(t)

I(t)= Incidence rate
c=contact rate
p=transmission probability
P (t)= Prevalence of infectious persons at time (t)

Equation 2.1. Estimation of the incidence of an infectious disease

We extended Halloran's model to estimate the number of incident active cases by including a latent period between infection and the development of a disease that can be observed. If "D" represents probability of disease in an infected individual, then p*D will express the probability of developing disease from a new infection. In other words, the probability of new active cases resulting from transmission. So, the incidence rate of new disease cases can be expressed as shown in Equation 2.2.

I (t) = c(t)*p*D*P(t)

I(t)= Incidence rate
c(t)=contact rate
p=transmission probability
D= probability of disease
P (t)= Prevalence of infectious persons at time (t)

Equation 2.2. Estimation of the incidence of infections that will progress to disease

Transmission of *Mycobacterium tuberculosis*

Tuberculosis has been present in humans for thousands of years (Levy, 2012). Tuberculosis is caused by the *Mycobacterium tuberculosis* complex and the principal mode of transmission is by the spread of respiratory aerosols (Dheda, Barry, & Maartens, 2016). Transmission occurs primarily when an infectious, diseased individual generates infectious dried droplets via coughing, sneezing, singing, laughing and/or talking and are inhaled by a healthy individual (Aït-Khaled et al., 2010; Nelson & Williams, 2006). In particular, the small droplets can stay in the air for several hours and are responsible for infection (Esmail, Barry, Young, & Wilkinson, 2014). After inhalation by a healthy individual, these droplets will travel to the alveoli or bronchioles and the infection will begin (Nelson & Williams, 2006).

The majority of infected individuals will develop latent tuberculosis infection (LTBI) (Dheda et al., 2016). LTBI is defined as an immunological based-laboratory diagnosis of *Mycobacterium tuberculosis* infection -but without clinical symptoms, or an irregular chest radiography (Lin & Flynn, 2010). Two diagnostic tests currently exist for the diagnosis of latent tuberculosis infection – the tuberculin skin test and interferon gamma release assays.

A small proportion (5-10%) of people infected with tuberculosis will progress to active tuberculosis disease (Nelson & Williams, 2006). Active tuberculosis disease can present as pulmonary or extrapulmonary, however the former is more common (World Health Organization., 2016). In both presentations, a myriad of sign and symptoms resembling other respiratory and systematic diseases appear. Among them, fever, night sweats and weight loss are most likely (Nelson & Williams, 2006). In pulmonary tuberculosis, a persistent cough is described in the majority of the cases (Heemskerk, Caws, Marais, & Farrar, 2015). Extra pulmonary tuberculosis may manifest symptoms according to the involved organs/tissues: Pleura, pericardium, spine, central nervous system and lymph nodes are a few examples (Aït-Khaled et al., 2010; Heemskerk et al., 2015).

Factors that modify the risk of exposure, infection and tuberculosis disease

Understanding the differences between the risk of exposure, risk of infection and risk for tuberculosis disease is critical to understand the epidemiology of tuberculosis and its transmission dynamics (Figure 2.1). Risk of exposure relates to the risk that a susceptible host is in contact with *Mycobacterium tuberculosis* (Nelson & Williams, 2006). In exposed subjects, the risk of infection will depend mainly on external factors related to the index case (infectivity) and environmental factors (duration contact, smoking, ventilation, etc.) (Narasimhan et al., 2013; Nelson & Williams, 2006). If infection has occurred, then the risk of progression to disease is primarily determined by host characteristics and environmental factors.

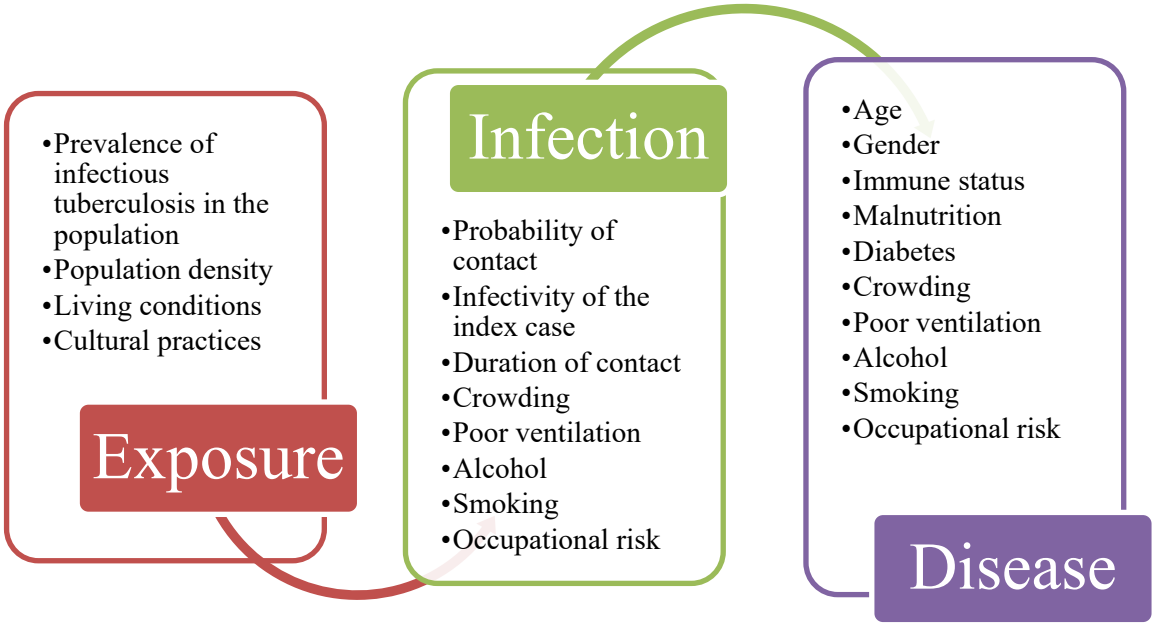


Figure 2.1. Risk factors for tuberculosis exposure, infection and disease

Source: (Nelson & Williams, 2006) and (Narasimhan et al., 2013)

Infection with human immunodeficiency virus (HIV), immunosuppression therapy and chronic renal failure that requires dialysis are important risk factors identified for progression from tuberculosis

infection to active tuberculosis disease, but other comorbidities, therapies, demographic and social factors may be involved in this process (Dheda et al., 2016) (Table 2.1).

Table 2.1. Risk factors for tuberculosis disease

Risk factor	General population Fold risk of developing active tuberculosis compared to subjects with no risk factor	Subjects with Latent Tuberculosis infection Risk of progression to active tuberculosis compared to subjects with no risk factor
Demographics		
Male sex	2 (after adolescence)	Unclear
Age	Higher in under 4 and over 20 years of age	2.2-5 (when infected young)
Genetic polymorphisms	Higher risk	Higher risk of infection, not disease
Social factors		
Smoking	2	2-3
Alcohol abuse	3	1.5
Biomass fuel exposure	2	Unclear
Overcrowding and poverty	Higher risk	Higher risk
Other comorbidities		
HIV	20-40	50-100
Silicosis	3-4	30
Malignancy	4-5	16 for cases of carcinoma of head and neck
Diabetes mellitus	3	2-3.6
Under nutrition or underweight	12	2-3
Chronic Obstructive Pulmonary Disease	2 (subjects using inhaled corticosteroids)	Unclear
Chronic renal failure requiring dialysis	7-50	10-25
Tuberculosis infection in the last two years	No data	15
Apical fibronodular changes in chest radiograph	No data	6-19
Therapies		
TNF-α inhibitors	1.5	1.7-9
Transplantation/ immunosuppressive therapy	15-20	20-74
Treatment with glucocorticoids	2	4.9

Source: Slightly modified from Dheda et al. (2016)

Epidemiology of Tuberculosis

Global epidemiology of tuberculosis in the general population and in HIV-infected subjects

Historically, one in three individuals globally is infected with this *Mycobacterium tuberculosis* infection. Recent modelling data suggest that a quarter of the world is infected worldwide equating to approximately 1.7 billion people were latently infected with *Mycobacterium tuberculosis* in 2014 (R. M. Houben & Dodd, 2016).

The World Health Organization estimated that 10.4 million (range 8.7-12.2 million) new active cases of tuberculosis occurred in 2015 in the world (World Health Organization., 2016). Roughly, 11% of these cases occurred in HIV-infected subjects, primarily impacting sub-Saharan Africa (World Health Organization., 2016). Moreover, tuberculosis continues to have high mortality, accounting for an estimated 1.4 million deaths in HIV-uninfected people and 0.4 million deaths in HIV-infected individuals in 2015 (World Health Organization., 2016).

Epidemiology of Tuberculosis in Guatemala

Currently, Guatemala has the highest burden of tuberculosis (TB) in Central America (Pan American Health Organization, 2013). The TB burden in Guatemala has not significantly declined over the last 10 years, with an estimated incidence of around 25-36 cases per 100,000 population between 2004-2014 (World Health Organization., 2016). The true number of tuberculosis cases is likely higher. A recent study found that the incidence of tuberculosis in 2013 varied greatly among the different provinces in the country (1-52 cases/100, 000 population) (Figure 2.2). The authors of the study discussed that these vast differences among provinces are likely due to programmatic issues, such as underreporting and underdiagnoses, than to actual differences in TB burden among regions (Samayoa-Peláez, Ayala, Yadon, & Heldal, 2016).

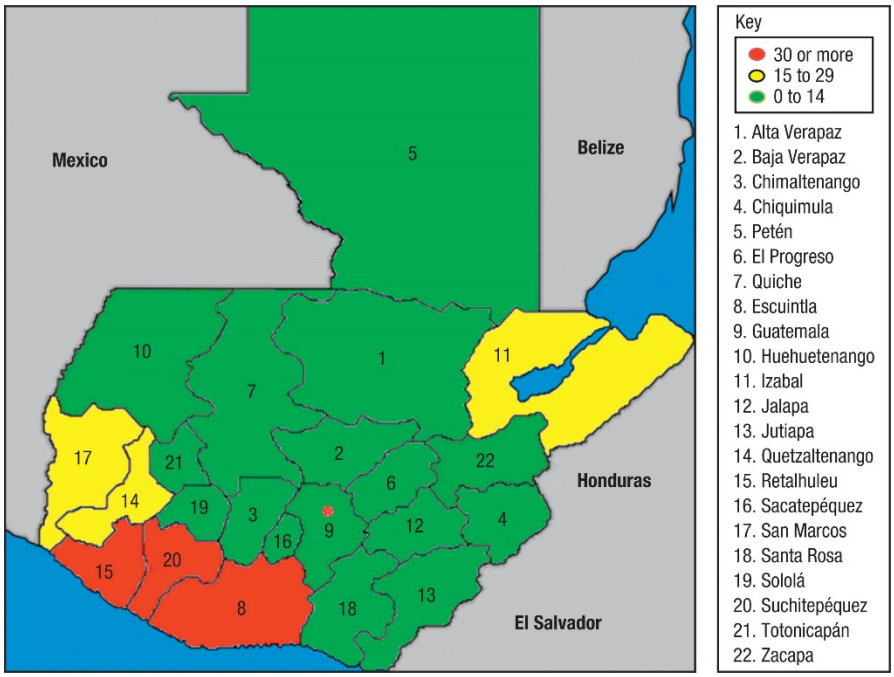


Figure 2.2. Incidence of tuberculosis (per 100,000 population) in the different departments (provinces) of Guatemala.

Source: (Samayoa-Peláez et al., 2016)

Epidemiology of Tuberculosis in Uganda

Uganda is one of the 30 high tuberculosis/HIV burden countries as described by the World Health Organization (World Health Organization., 2016). The estimated incidence for 2015 was 202 new cases (95% CI 120-304) per 100,000 population, totaling 79,000 new cases (95% CI 47,000-119,000 cases). More than a third of these cases occurred in HIV-infected subjects (26,000 cases).

The role of social contact patterns in the incidence of tuberculosis

In this section of the literature review, we will first review the relevance of an adequate definition of contact rate to characterize the incident cases of tuberculosis in a population. Then we will define the concept of contact, describe the evidence regarding determinants of social mixing and then we will describe the impact of different patterns of contact and social mixing on the incidence of tuberculosis. We will finalize with the introduction of the Wells-Riley equation, which has been of use to identify the probability of transmission in closed spaces.

Relevance of contact rate in the incidence of tuberculosis

In Aims 1 and 2, we intend to define adequate contact for transmission of *Mycobacterium tuberculosis* in an African urban environment. If we refer to Equation 2.1 (Incidence=c*p*P), and apply it to tuberculosis infection, we could argue that two of three components that are needed to estimate the incidence of an infection, transmission probability and prevalence, can be measured. The prevalence of tuberculosis, although challenging, can be measured using several methods, most notably surveillance data from tuberculosis programs or by population-based surveys (Glaziou et al., 2008). The probability of transmission given adequate contact seems to cover a wide spectrum. On one end of the spectrum, there has been reports of transmission after flying in planes where a tuberculosis case is also flying or after embalming a tuberculosis culture-positive cadaver (Kenyon, Valway, Ihle, Onorato, & Castro, 1996; Sterling et al., 2000). This suggests that some individuals may be infected after a single exposure. On the other end of the spectrum, and through the calculation of secondary attack rates, it has been estimated that in Uganda, the probability of infection in household contacts of tuberculosis cases is around 47%, meaning that over 50% of subjects that live in the home of a tuberculosis case remain uninfected (C. C. Whalen et al., 2011). However, the definition of the third component, contact rate, is still poorly understood. Mathematical models of respiratory infectious diseases assume homogeneous mixing in the population, but evidence suggest otherwise (Kong, Wang, Han, & Cao, 2016; Mossong et al., 2008; Wallinga et al., 2006). We know that tuberculosis occurs in clusters and affects certain groups, such as

marginalized populations (Sulis et al., 2014). Further research is warranted to better understand social mixing patterns among tuberculosis cases and their contacts and that leads to the rationale for Aims 1 and 2.

Definition of contact

The definition of "contact" when studying the spread of respiratory or close-contact transmitted infectious diseases is not standardized (Table 2.2), but generally a contact has occurred when, at a minimum, a short face-to-face conversation occurs within a short distance and/or physical contact.

Table 2.2. Definition of "contact" among studies focusing on respiratory or close-contact transmitted diseases.

Name	Definition provided by the study	Reference
Close contact	Contact with someone with whom the interviewee had a face-to-face conversation that was longer than a greeting and within an arm's reach.	(Dodd et al., 2016)
Casual contact	Contacts with people who were inside buildings other than the interviewee's home that the interviewee had visited.	(Dodd et al., 2016)
Contact	Two-way conversation (at a distance which did not require raising the voice) in which at least two words were spoken by each party and in which there was no physical barrier between the two parties (such as security screens	(Edmunds et al., 1997)
Contact	Persons in their household and Number of different persons they conversed with during a typical week	(Wallinga et al., 2006)
Contact person	Person sitting or standing within arm's length of the participant for 30 seconds or longer.	(Stein et al., 2014)
Physical contact	Skin-to-skin contact such as a kiss or handshake	(Mossong et al., 2008)

Name	Definition provided by the study	Reference
Nonphysical contact	Two-way conversation with three or more words in the physical presence of another person but no skin-to-skin contact	(Mossong et al., 2008)
Social contact	All contacts among people: Range from a short conversation on the street to close physical contact when sleeping with someone in the same bed	(Feenstra et al., 2013)
Close contact	Those involving physical touch (type I) or those involving a 2-way conversation with 3 or more words in the physical presence of another person without physical touch (type II).	(Johnstone-Robertson et al., 2011)
Casual contact	Those occurring in an indoor location but not satisfying the criteria for a close contact	(Johnstone-Robertson et al., 2011)
Contact	Being in close proximity for more than roughly five minutes"	(Potter, Handcock, Longini Jr, & Halloran, 2012)
Contact	Physical contact or a two-way conversation of at least three words in the physical presence of another person	(Potter, Handcock, Longini Jr, & Halloran, 2011)

Determinants of social mixing

Social patterns among individuals are not homogeneous. Different studies have consistently shown that assuming a homogenous mixing among subjects is incorrect and will yield to inaccurate estimates of the level of contact and transmission of infectious diseases (Mossong et al., 2008; Wallinga et al., 2006).

Several factors have been identified that determine the frequency and nature of the contact within an infectious case and their contacts.

Age. Younger subjects interacts more commonly with people of their own age, whereas older subjects have a wider range of contact of different ages (Edmunds et al., 1997). Similar results have been described by other reports (Dodd et al., 2016; Mossong et al., 2008).

Sex. There is evidence that the contact patterns among men and women are different. In Bangladesh, a focus group-based study, showed that the majority of men had contacts in the household, within the neighborhood and outside the neighborhoods whereas the social interaction of women was confined to the household and to a lesser extent within the neighborhood (Feenstra et al., 2013).

Moreover, subjects tend to mix preferably with subjects of their same sex. In a cross-sectional study conducted in Zambia and South-Africa, Dodd and collaborators found that 63% of the close contacts of women were also women. Similarly, 61% of close contacts of the male interviewees were men (Dodd et al., 2016).

Education level. Another factor that might also influence the social patterns among subjects is education level. A survey using respondent driven sampling (first subjects recruited were asked to recruit more subjects) conducted in two distinct settings, The Netherlands and Thailand, showed that besides age and sex, social mixing was influenced by education, with subjects more prone to contact other subjects of their same education level, particularly in Thailand (Stein et al., 2014).

Incidence of tuberculosis and social mixing patterns

Andrews et al conducted a study in South Africa in which they studied the contribution of five different settings in the transmission of tuberculosis: own household, transit, school, workplace and other households (Andrews, Morrow, Walensky, & Wood, 2014). They discovered that irrespective of age, tuberculosis transmission occurred mainly outside of the household. Among children under 14 years old, own household, transit and school were the most relevant transmission sites, whereas in adults over 18 years of age, workplace and to a lesser extent transit were the ones with the highest contribution of transmission. Overall, just over 15% of tuberculosis transmission occurs within the household.

Adult men seem to be the group that drives tuberculosis transmission. A report has indicated that more than 50% of tuberculosis infection in men, women or children may be attributed to contact with adult men (Dodd et al., 2016).

Wells-Riley Equation or the probability of transmission in *Mycobacterium tuberculosis*

Although in this aim we are interested in furthering our understanding of adequate contact through social mixing on the risk of tuberculosis transmission, it will be essential to capture the probability of transmission (the "p" of Equation 2.1) as both components are intrinsically linked with the risk of infection. To understand the factors that determine the probability of transmission of *Mycobacterium tuberculosis*, we will refer to the Wells-Riley model (Equation 2.3) which estimates the probability of transmission of a respiratory infection in an indoor space (Noakes & Sleigh, 2008; Rudnick & Milton, 2003).

$$P_i = \frac{C}{S} = 1 - \exp\left(-\frac{Iqpt}{Q}\right)$$

P_i = Probability of infection
C= Number of infectious cases
S= Number of susceptible subjects
I= Number of infectious individuals in a space
p= Pulmonary ventilation rate
q = Infectious quanta generation rate
t = Exposure time interval
Q= room ventilation rate with clean air

Equation 2.3. The Wells Riley Model

Source: (Rudnick & Milton, 2003) and (Andrews et al., 2014)

Based on the Wells-Riley model, the probability of transmission after a certain time of exposure relates to several factors, namely the concentration of microorganisms in the contaminated air ("q" or quanta), the exposure time ("t"), the breathing rate ("p") and room ventilation (Q) (Sze To & Chao, 2010). There is a wide range in the quanta production among infectious tuberculosis patients, with some subjects expelling less than one quanta per hour and the most infectious ones producing 60 to 226 quanta per hour (Yates et al., 2016). In our aim, we are focusing on the risk between an index case and a contact, so the "I" will be 1 in this context. Upon knowing the probability of transmission and, subsequently multiplying it by the contact rate and prevalence of tuberculosis infection, we will be able to fully characterize the incidence of tuberculosis/risk of infection in this population using a theory-based model.

The role of recent transmission in the incidence of tuberculosis

In this section of the literature review, we will describe the different genotyping techniques that have been used to characterize *M. tuberculosis*. Then, we will focus in the application of these molecular tools to estimate recent transmission and summarized of the risk factors that have been associated with the risk of clustering. We will finalize with a review of the most relevant methodological considerations when conducting a molecular study focusing in transmission dynamics.

The use of molecular tools to determine recent transmission.

There are different techniques to genotype *Mycobacterium tuberculosis* (Mathema, Kurepina, Bifani, & Kreiswirth, 2006). To describe in depth each of them is beyond the scope of this literature review. We will focus on the three most commonly used techniques: IS-6110 RFLP, spoligotyping and MIRU-VNTR (Table 2.3).

IS6110-RFLP. In this technique, a restriction enzyme is used to cleave the DNA where the IS6110 element is present. The fragments are separated using electrophoresis. An IS6110 probe

hybridizes to these fragments and each strain will produce a particular pattern (Barnes & Cave, 2003) (Van Soolingen, 2001) (Jagielski et al., 2016) (Moström, Gordon, Sola, Ridell, & Rastogi, 2002).

Mycobacterial interspersed repeat units -Variable numbers of tandem repeats (MIRU-VNTR). In this technique, several polymerase chain reaction (PCR) assays are conducted in different loci The size of each PCR product depends on the number of MIRUs present (Moström et al., 2002) (Jagielski et al., 2016) (Barnes & Cave, 2003). This technique can be conducted using 12 loci. However, using 15 loci and 24-loci yield a higher discrimination and it is recommended for standard studies (15 loci) and for evolutionary studies (24-loci) (Philip Supply, 2005; P. Supply et al., 2006)

Spacer oligonucleotide typing (Spoligotyping). Kamerbeek and collaborators develop this technique at the end of the 1990s (Kamerbeek et al., 1997) (Figure 2.3) . *Mycobacterium tuberculosis* contains the direct repeat (DR) region with 36-bp direct repeats that are separated by polymorphic DNA sequences called “spacers”(Mathema et al., 2006). PCR assay is conducted to amplify this region. The presence or absence of 43 spacers will produce a unique pattern, that is used to genotype a strain (Barnes & Cave, 2003; Jagielski et al., 2016; Mathema et al., 2006; Van Soolingen, 2001).

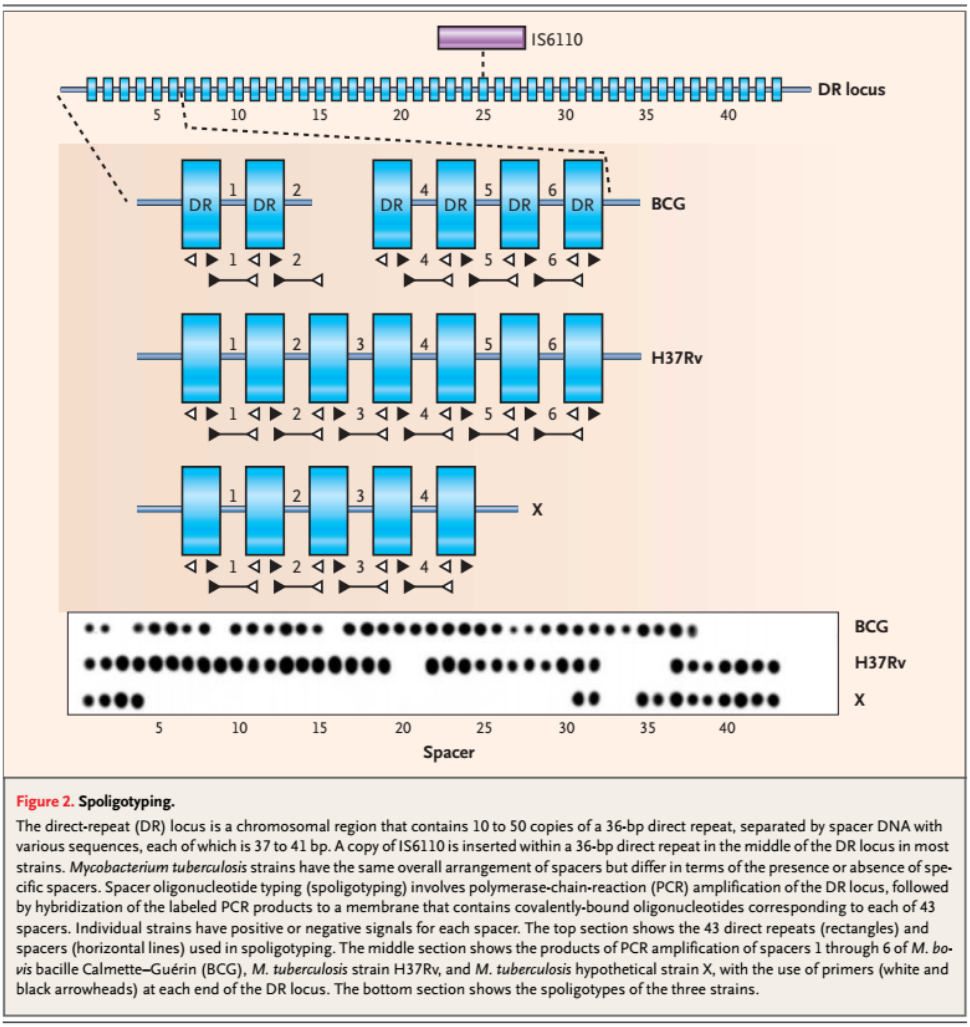


Figure 2. Spoligotyping.

The direct-repeat (DR) locus is a chromosomal region that contains 10 to 50 copies of a 36-bp direct repeat, separated by spacer DNA with various sequences, each of which is 37 to 41 bp. A copy of IS6110 is inserted within a 36-bp direct repeat in the middle of the DR locus in most strains. *Mycobacterium tuberculosis* strains have the same overall arrangement of spacers but differ in terms of the presence or absence of specific spacers. Spacer oligonucleotide typing (spoligotyping) involves polymerase-chain-reaction (PCR) amplification of the DR locus, followed by hybridization of the labeled PCR products to a membrane that contains covalently-bound oligonucleotides corresponding to each of 43 spacers. Individual strains have positive or negative signals for each spacer. The top section shows the 43 direct repeats (rectangles) and spacers (horizontal lines) used in spoligotyping. The middle section shows the products of PCR amplification of spacers 1 through 6 of *M. bovis* bacille Calmette–Guérin (BCG), *M. tuberculosis* strain H37Rv, and *M. tuberculosis* hypothetical strain X, with the use of primers (white and black arrowheads) at each end of the DR locus. The bottom section shows the spoligotypes of the three strains.

Source: (Barnes & Cave, 2003)

Figure 2.3. Description of the spoligotyping technique.

Each technique has its advantages and disadvantages (Table 2.3). The decision on which of this techniques should be used depends on several factors including the specific research question, level of discrimination needed, financial considerations, and amount of DNA available (Barnes & Cave, 2003; Moström et al., 2002; Van Soolingen, 2001).

Table 2.3. Selected genotyping techniques for *Mycobacterium tuberculosis*

Name	IS6110-RFLP	Spoligotyping	MIRU-VNTR
Advantage	High resolution Required significant amount of DNA	Inexpensive Can be performed in clinical samples Easy interpretation	High resolution High sensitivity Relatively fast
Disadvantages	Not useful in strains with low IS-6110 copies Required high amounts of DNA Expensive	Low resolution	Might have instability of mutations Unrelated strains might have same genotype
Useful for transmission studies	++++	+	++++
Stability	±	++	?

Source: Data abstracted from (Barnes & Cave, 2003; Borgdorff & van Soolingen, 2013; Jagielski et al., 2016).

Application of genotyping for tuberculosis transmission studies.

Molecular genotyping has contributed to understand in a better way the levels of transmission tuberculosis in a particular geographical region (Jagielski et al., 2016). Strains with identical or closely related genotypes are called clustered strains, that is belonging to a cluster and are considered to indicate recent transmission (primary infection or reinfection) (Barnes & Cave, 2003; J. Glynn et al., 1999). In molecular studies, the term "recent transmission" usually describes transmission that occurred in the last two to five years (Borgdorff & van Soolingen, 2013). Thus, when the sample collection is restricted in time, and a well-defined geographical area (J. Glynn et al., 1999) we can infer that clustered cases represent recent transmission; the assumption being that clusters are "epidemiologically linked chains of recently transmitted disease" (Murray & Nardell, 2002). Strains with a unique pattern are called non-

clustered or orphans and are considered to have been acquired in the past (reactivation) (Barnes & Cave, 2003).

Tuberculosis clustering. The proportion of clustering in different molecular epidemiology studies varies greatly. A systematic review found proportions as low at 7% and as high as 72% from 36 studies (Figure 2.4) (Fok et al., 2008). Higher tuberculosis incidence rates, larger mean cluster size, and the use of conventional contact tracing were factors associated with higher tuberculosis clustering proportion (Fok et al., 2008). In low TB burden settings, the pooled TB clustering proportion was 40.9% (95% CI 40.3-41.5), whereas in high/medium incidence studies, this value was 44.7% (95% CI 43.9-45.6).

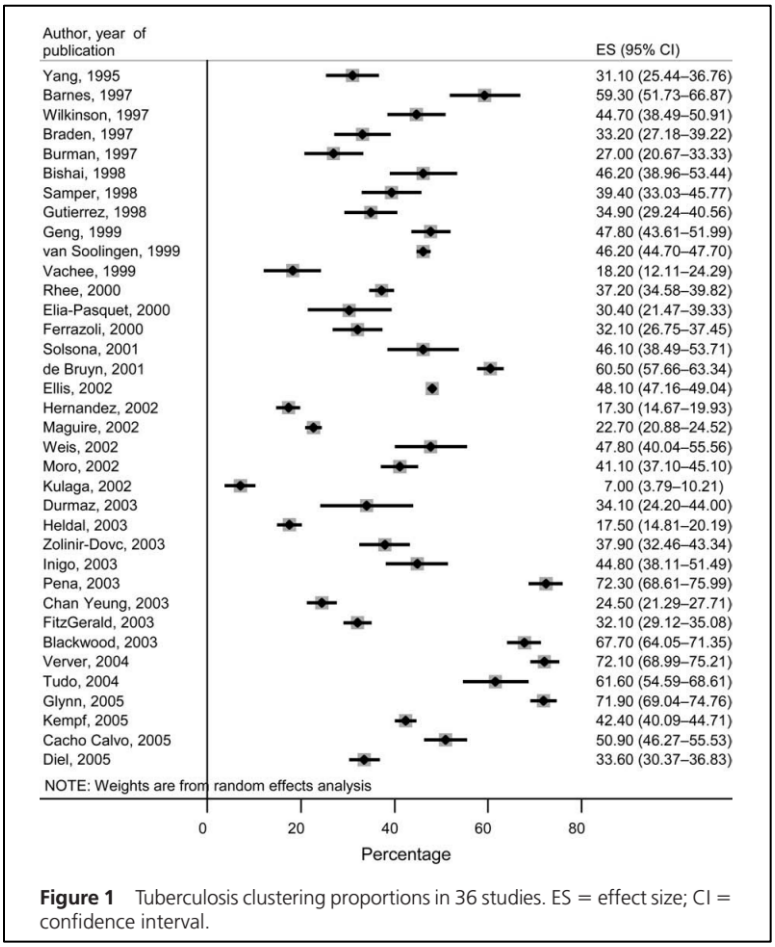


Figure 2.4. Proportion of clustering in molecular epidemiology studies of tuberculosis.

Source: (Fok et al., 2008)

Risk factors for clustering

Risk factors that are associated with clustering seem to be different in low tuberculosis countries as opposed to high/medium tuberculosis countries. A meta-analysis was conducted by Fok and others in molecular epidemiological studies of tuberculosis in which IS6110-RFLP was used as the primary genotyping technique. Eight different risk factors were explored to assess their relationship with the risk of tuberculosis clustering: male sex, local birth, pulmonary tuberculosis, smear positive status, HIV status, alcohol abuse, injection drug use and homelessness (Fok et al., 2008). When focusing only in the studies that reported adjusted odds ratio (OR), in low tuberculosis incidence settings risk factors for clustering included male sex (OR: 1.3, 95% CI 1.2-1.5), local birth (OR 3.5, 95% CI 2.3-5.4), pulmonary tuberculosis (OR: 1.6, 95% CI 1.1-2.4), alcohol abuse (OR: 1.8, 95% CI 1.3-2.6), injection drug use (OR: 2.2, 95% CI 1.5-3.3) and homelessness (OR: 2.1, 95% CI 1.1-3.9). However, in high/medium TB burden settings, only local birth (OR: 1.9, 95% CI 1.3-3.0), pulmonary tuberculosis (OR: 2.0, 95% CI 1.5-2.7) and homelessness (OR: 1.8, 95% CI 1.2-2.7) were associated with cluster. To establish if these differences are due to levels of tuberculosis incidence, more studies are needed, especially in the high/medium burden countries.

More recently, a large population-based molecular study conducted in China was published (C. Yang et al., 2015). The study was conducted from 2009-2012 in five different sites and analyzed data from 2,274 subjects. The tuberculosis clustering proportion was 31% with a range between 21.7% and 36.1% depending on the study site. Interestingly, several potential risk factors were not associated with TB clustering: sex, age, BMI, previous TB history, cavitary presentation, presence of cough, diabetes, diagnosis delay, sputum smear, alcohol use and smoking. There were only two variables that in the multivariable logistic regression analysis showed statistically association with clustering, the drug resistance profile (multi drug resistant strains-MDR- had 86% more chance of being in a cluster as compared to drug susceptible strains) and infection with a Beijing strain, (OR: 1.56, 95% CI 1.23-2.96).

HIV and TB clustering.

A systematic review pooled data from six molecular studies aimed to determine if human immunodeficiency virus (HIV) increases the risk of tuberculosis by increasing the risk of recent infection or by increasing the risk of reactivation (R. Houben et al., 2011). The study included both HIV-infected subjects and non-HIV-infected individuals. Overall, HIV-infected subjects were 25% more likely to being infected with a clustered strain as compared to HIV-negative subjects (OR: 1.25, 95% CI: 1.0-1.5). Perhaps more revealing is that this difference varied according to the age of the subject. In the older group (more than 50 years old), HIV-infected subjects were 2.6 times more likely to belong to a cluster (95% CI 1.4-5.7) whereas in the middle age group (26-50 years old) there were no differences among HIV-infected and HIV-non-infected subjects (OR: 1.0, 95% CI 0.8-1.3). The authors speculate that one potential reason for these findings were that older HIV-infected subjects had been infected by HIV for longer, leading to a higher immunosuppression level and thus increasing the risk of recent infection more than reactivation as compared to younger adults. If true, these results may suggest that health policies may be improved by tuberculosis risk management in HIV-infected individuals (R. Houben et al., 2011).

Methodological considerations regarding molecular studies that aim to monitor transmission.

Several methodological factors can influence the reliability of the results of a molecular study. Among them, we can mention the resolution of the genotyping technique used and the sampling methodology (Mathema et al., 2006; Murray & Alland, 2002).

Bias can occur in molecular epidemiological studies when a low proportion of all tuberculosis cases are sampled (Murray & Alland, 2002). This bias will occur principally in settings where there is a small number of unique cases or when there are a substantial amount of small clusters (Megan Murray, 2002). When this occurs, the total estimate of recent transmission will be underestimated. Moreover, if sampling wat not random, there is risk of selecting more isolates that belong or do not belong to a cluster, leading to either an overestimation of recent transmission or reactivation (Murray & Alland, 2002). To address some of the shortcomings and methodological issues that are particularly relevant for molecular

epidemiological studies, a set of recommendations has been published called “Strengthening the Reporting of Molecular Epidemiology for Infectious Diseases” (STROME-ID) (Field et al., 2014). The aim of STROME-ID is to improve and standardize the reporting of molecular epidemiological studies.

Chapter 3
MATERIALS AND METHODS

In this chapter, the methodology for the specifics aims of this dissertation is presented. Aims 1 and Aims 2 are presented in a joint section as they shared the same study population and settings. Aim 3 is presented in a sole section.

Methodology for Aim 1 and Aim 2

Overarching Aim: To define adequate contact for transmission of *Mycobacterium tuberculosis* in an African urban environment. This overarching aim was addressed through two component aims:

- ✓ Component Aim 1. To define adequate contact for tuberculosis transmission in an African urban environment by examining the interaction within ego-centric networks and develop a score that measures the degree of contact..
- ✓ Component Aim 2. To determine whether the score that was developed covaries with the presence of tuberculosis infection among contacts.

Design Overview

From 2012 to the present, a cross-sectional study in patients with active tuberculosis was conducted in Kampala, Uganda (Community Health and Social Networks of Tuberculosis-COHSONET study-, PI: Dr. Christopher C. Whalen), The proposed study used data collected from the COHSONET study to address the specific aims. As we have described in our literature review, the risk of transmission of tuberculosis depends in several factors, related to the infectious individual, his/her contact and the environment. Trained social scientists and health workers interviewed subjects with tuberculosis disease to capture variables related to the social interaction that might explain what constitutes adequate rate and

variables related to the probability of transmission. A closeness score was built for each of the index case and contact. In Figure 3.1 a schematic representation of the design overview is presented.

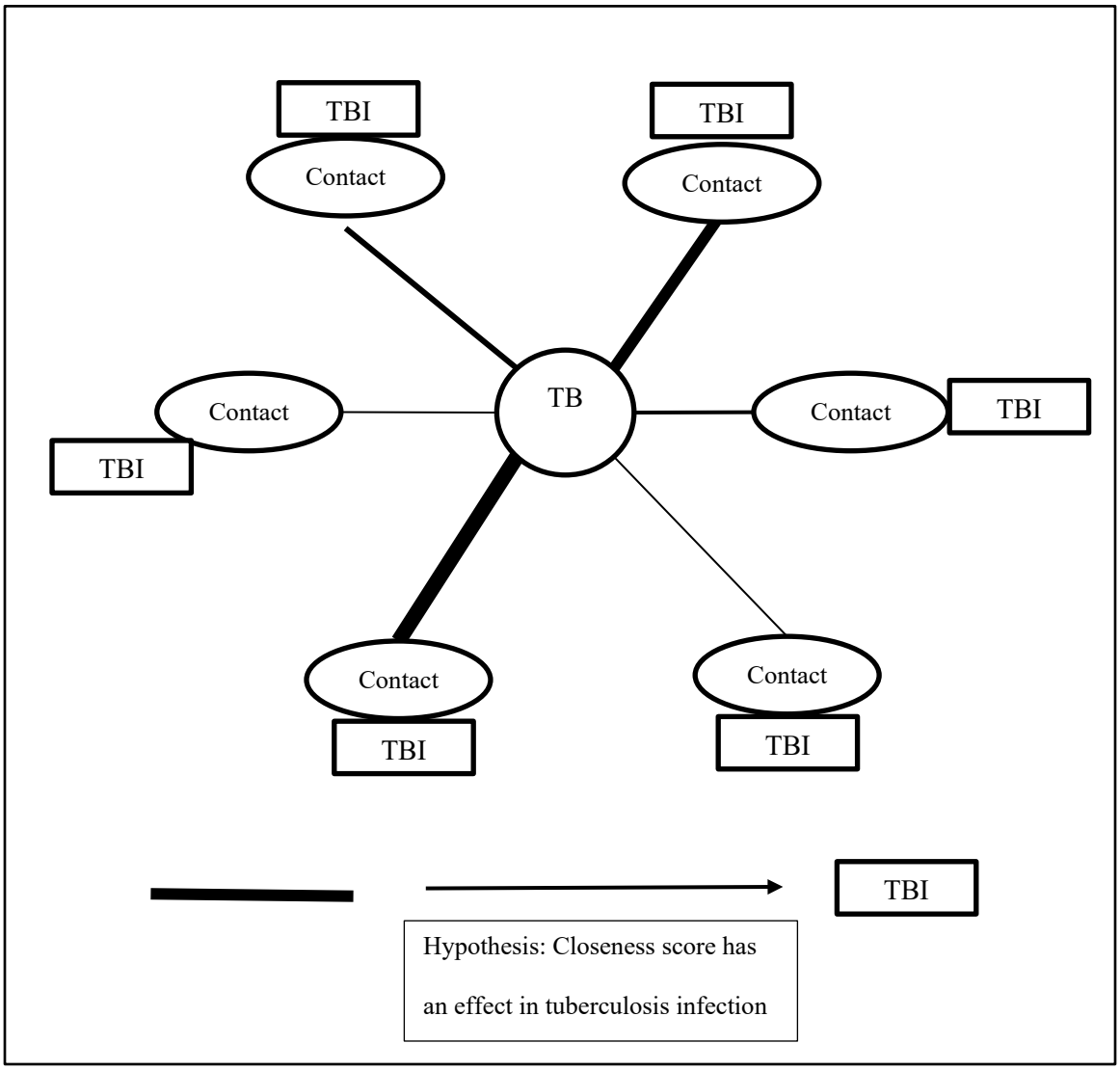


Figure 3.1. Design overview of the closeness score.
The nature of the interaction between a tuberculosis case and his/her contact was examined. A tuberculosis case has different levels of closeness with each of his/her contacts (thickness of the lines exemplifies this phenomenon). A closeness score was be developed for each of these relationships. This closeness score was used as an independent predictor to estimate the probability that the contact has a tuberculosis infection (latent or active disease). As adequate contact is not the only determinant of tuberculosis infection, we adjusted our model to the other explanatory variables.

Study population and subjects

Index cases. The index cases were persons aged 15 years or older, who are residents of Kampala, Uganda; who had signs and symptoms of pulmonary tuberculosis and with at least one positive sputum smear for acid-fast bacilli. All were enrolled in the Community Health and Social Networks of Tuberculosis study (RO1 AI AI09386) from 2012 to the present. Patients were evaluated by the National Tuberculosis Control Program, and after it was confirmed they met inclusion criteria and provided informed consent were recruited.

Contacts. The index case or index control provided a list of their household and community contacts. Members of the field team approached these contacts and if they agreed, they were enrolled in the study.

Study exposure

The study exposure was the "closeness score" between the tuberculosis case and their contacts. As described in the literature review to fully characterize the risk of infection, we need to determine the probability of transmission given contact and also the contact rate. Thus, we captured variables related to social mixing, which provided us with understanding who mixes with who (nature of the relationship, age, sex, etc.) and the frequency and length of the contact. In addition, we captured variables that relates to the concentration of microorganisms (smear status and cavitary presentation of the index case), the ventilation conditions of the setting in which the social interaction occurs and the respiratory rate of the index case (as described in the Wells Riley model). This score was constructed based on the answers and evaluations that the cases and contacts provided to the field staff and these variables were collected in different questionnaires from the COHSONET study.

Study outcome

The study outcome was tuberculosis infection in the contacts. The infection in the contact can be latent or active (disease).

Latent tuberculosis infection was estimated by the use of the tuberculin skin test (TST). In the TST the delayed-type hypersensitivity response is estimated by measuring the diameter of induration (H. Yang, Kruh-Garcia, & Dobos, 2012). This measurement is obtained in a continuous scale, but for convention, the criterion for a positive TST result is 5 mm, 10 mm or 15 mm, depending on the setting and the subject characteristics (Nayak & Acharjya, 2012). For our study, a positive TST result was defined as induration ≥ 10 millimeters as it has shown to be an adequate cut-off in Ugandan setting (Martinez, Sekandi, Castellanos, Zalwango, & Whalen, 2016). Intradermal injection of 5 TU of purified protein derivate was applied in study contacts. After 48-72 hours of the injection, two trained Ugandan technicians using digital calipers, independently measured the induration and recorded as continuous data. The mean measurement was used for the estimation of latent tuberculosis infection (C. C. Whalen, 2014).

Active tuberculosis was defined either as: a) The presence of at least one smear positive for acid fast bacilli, b) Positive culture for *Mycobacterium tuberculosis*, c) A positive molecular result for *Mycobacterium tuberculosis* in a contact with clinical symptoms consistent with tuberculosis disease and d) History of previous tuberculosis disease, informed by the social contact.

Covariates, confounder and effect modifiers

In this analysis, we controlled for all known predictors that might modify the risk of tuberculosis infection, as well for any potential confounder that might mask the true association between closeness and tuberculosis infection. Based on literature review (Ai, Ruan, Liu, & Zhang, 2016; Dheda et al., 2016; Narasimhan et al., 2013) , these can be categorized as related to the tuberculosis case, related to the contact and related to environmental factors (Table 3.1).

Table 3.1 List of potential confounders, covariates and effect modifiers.

Of the tuberculosis case	Of the contact	Environmental factors
Age	Age	Ventilation
Sex	Sex	Crowding
HIV	HIV	
Smear status	Other comorbidities, such as diabetes mellitus	
Social economic status	BCG	
Cavitary presentation	Alcohol intake	
Cough	Smoking	
	Occupational risk	
	BMI	

Statistical Issues

Statistical hypothesis

For specific Aim 2 we present a statistical hypothesis. The null hypothesis is that the closeness score does not have an impact in the proportion of contacts infected with tuberculosis (Equation 3.1). The alternative hypothesis is that the score does have an effect in the level of tuberculosis transmission, which is the proportion of contacts with tuberculosis infection.

H_0= Closeness Score is not an independent predictor for tuberculosis infection

H_A= Closeness Score is an independent predictor for tuberculosis infection

Equation 3.1. Null and alternative hypothesis for Aim 2

Analytic Strategy

This section will present the analytic approach for each of the component aims relating to Aims 1 and Aims 2.

Aim 1. To define adequate contact for tuberculosis transmission in an African urban environment by examining the interaction within ego-centric networks and develop a score that measures the degree of contact.

To address this aim, we conducted a descriptive quantitative analysis of the variables that allowed us to define the relationship between tuberculosis cases and their contacts. An item analysis was conducted in the data set to explore the distribution of the variables. Categorical variables were presented with absolute frequencies and with relative frequencies. Continuous variables were presented with measures of central tendency (median) and with dispersion statistics (interquartile range) (Larson, 2006). Baseline characteristics of tuberculosis cases were estimated as proportions with 95% confidence intervals.

We performed a factor analysis to determine the relationship among variables. The main aim of an exploratory factor analysis is to reduce the number of variables than can explain a response/outcome variable to fewer variables (called factors), as these factors represent a set of underlying constructs present in the data (Härdle & Simar; Joreskog, Olsson, & Wallentin, 2016; Treiblmaier & Filzmoser, 2010). The model can be described as follows (Nardo, Saisana, Saltelli, & Tarantola, 2005) (Equation 3.2).

$x_1 = \alpha_{11}F_1 + \alpha_{12}F_2 + ... + \alpha_{1m}F_m + e_1$

$x_2 = \alpha_{21}F_1 + \alpha_{22}F_2 + ... + \alpha_{2m}F_m + e_2$

...

$x_Q = \alpha_{Q1}F_1 + \alpha_{Q2}F_2 + ... + \alpha_{Qm}F_m + e_Q$

xi= variable with zero mean and unit variance

αi1, αi2, ..., αim = factor loadings related to the variable Xi;

F1, F2,...,Fm = *m* uncorrelated common factors, each with zero mean and unit variance

ei = Q specific factors supposed independently and identically distributed with zero mean

Equation 3.2. Factor Analysis Model.

The rationale behind the use of an exploratory factor analysis in this aim was that several of the variables collected from the tuberculosis cases are highly correlated, for instance, there is a question regarding the level of confidence between a tuberculosis case and their contact, and other question assess if the tuberculosis case has disclosed their tuberculosis status to the contacts. Thus, a variable reduction technique is recommended, and EFA will be a better option than principal component analysis as we believe that the observed variables are a combination of underlying factors and we want to identify them and describe them (Suhr, 2005).

There are four phases in EFA (Barbero Garcia, Vila, & Holgado Tello, 2013) and we followed them as such:

Initial preparation of the data. As first point, a data cleaning of the data was done. Second, a correlation matrix was obtained to confirm there is a high level of correlation among the observed variables. If there was no correlation, it would have been not necessary to conduct an EFA (Barbero Garcia et al., 2013).

Factorability of the original matrix. As we had mixed data (ordinal and binary data) we created a polychoric correlation matrix. EFA models obtained with a polychoric correlation have shown to be more consistent with the measured variables than the Pearson correlation when using ordinal data (Holgado–Tello, Chacón–Moscoso, Barbero–García, & Vila–Abad, 2010). We evaluated the factorability of the items considered to EFA by the Kaiser's Measure of Sampling Adequacy (Kaiser & Rice, 1974). We excluded variables in which this measurement was lower than 0.6 (Taherdoost, Sahibuddin, & Jalaliyoon, 2014).

Factor extraction and Number of factors to retain. We selected the principal axis factoring method to extract our factors. This selection was based on the goal to identify underlying constructs in the data set (Treiblmaier & Filzmoser, 2010). We followed the analysis described by Berghaus and others, in

which three methods were considered: a) keeping factors with eigenvalues[1] greater than 1 (Kaiser Criterion), b) Scree plot- in a graphical representation, retains eigenvalues that appear in the sharp slope of the plot and drop the ones that appear in the leveling area c) variance explained criteria-keep factors that account for 80-90% of the variation.

Factor rotation. As we believed the factors are correlated, an oblique rotation of the factors will be conducted (Gaskin & Happell, 2014). This technique involves the transformation of the factors in order to obtain simpler and more interpretable results (Nardo et al., 2005).

Factor interpretation. After the factors were identified, a name was assigned to each factor, based on their common characteristics and literature review (Barbero Garcia et al., 2013).

Generation of Factor Scores. Factors scores were computed for each individual. To generate the factor scores for an individual, we used the weighted sum scores method, as it allows that the items with the highest loadings[2] to have the highest impact in our factor score (DiStefano, Zhu, & Mindrila, 2009). Factor scores were investigated to check if they met the normality assumption and to determine if they have unimodal distribution (Ameijeiras-Alonso, Crujeiras, & Rodríguez-Casal, 2016). Based on these analyses, the results of the factor scores are presented as median with interquartile ranges.

Association Factor scores with other variables. We wanted to establish the construct validity of the factor scores. To do this, we examined the relationship of these factors with other variables collected from the contacts of the index case who we traced and enrolled in the study. We stratified this group according to type of contact (household and non-household contact) and nature of the relationship between case and contact (spouse, child, sibling, friend, co-worker, other relatives, neighbor, other) The median and interquartile ranges of the identified factor scores were estimated for each stratum. In the process, we examined the variability of the factor scores by gender of contact (men, women), age of contact (0-4

[1] Eigenvalues= variances of the principal components (Nardo et al., 2005).

[2] Loading= 'Correlation between observed variables and factors'(Suhr, 2005)

years, 5-14 years, 15 years and greater), age of index (15-24 years, 25-44 years, 45 years and greater), gender of index case (men and women) and usual place of meeting' (home tuberculosis case, friend's home, relative's home, work place, bar, trading center/shop/kiosk, elsewhere).

To compare the difference in medians among stratified groups we estimated the 95% confidence intervals by bootstrapping, using the package 'boot' for R software (Canty & Ripley, 2017) . We set the number of bootstraps replicates to 10,000 and calculated the intervals with the adjusted bootstrap percentiles (BCa) method. We selected this parameters based on recommendations from Puth and others (Puth, Neuhauser, & Ruxton, 2015).

Sensitivity analysis. We repeated the EFA not in a polychoric correlation matrix but with transformed data. An optimal monotonic transformation of the raw data was performed (PROC PRINQUAL, method=maximum total variance)(SAS Institute Inc, 2017). We selected the 'monotone' transformation, as our variables were ordinal or binary. The transformed variables were used to conduct the factor analysis, following the same procedure previously described.

All analyses were carried out using SAS software v 9.4 (SAS Institute, Cary, NC, US) and R v3.3.1 (R Foundation for Statistical Computing, Vienna, Austria, 2016).

Aim 2. To determine whether the score that was developed covaries with the presence of tuberculosis infection among contacts.

To address this specific aim, we studied the association of the developed scores to the risk of having tuberculosis infection among contacts. An item and exploratory analysis were conducted in the data set to explore the distribution of the exposure variables, covariates and the main outcome (tuberculosis infection). For continuous variables, median values and interquartile ranges were estimated and for categorical variables, proportions with 95% confidence intervals. In addition, visual exploration

was performed- using bar plots, histograms and boxplots- depending of the type of variable. We performed Kendall and polychoric correlation to check correlation among the variables (Knight, 1966; Olsson, 1979). Baseline characteristics of the enrolled contacts were summarized with proportions and measures of central tendency.

We conducted bivariate analysis to explore the relation between each covariate and the exposure variables, and each covariate and the outcome, separately. We initially used Chi-square test (categorical variables) or Wilconxon test (continuous variables) to explore those associations. We also explored the probability of tuberculosis infection against the factor scores, using a loess (locally weighted scatterplot smoothing) model, to obtain a nonparametric smoothed curve (M Friendly, 2015; Michael Friendly & Meyer, 2015).

We estimated the prevalence of contacts with tuberculosis infection with 95% confidence intervals, overall and according to the exposure, confounders and covariates. As the prevalence of tuberculosis infection in this population was over 50%, two regression models were considered as alternatives to logistic regression to obtain a more precise estimate of the association between exposure and the outcome (Coutinho, Scazufca, & Menezes, 2008): a Poisson regression with a robust variance and a log-binomial regression (Wacholder, 1986; Zou, 2004). Unadjusted prevalence ratios were obtained with both methods by exponentiating the coefficients. Similar analyses were done to check the association between potential confounders, covariates with the outcome. As the analyses showed similar results, the stratified and the adjusted prevalence ratios later described were exclusively calculated with the modified Poisson, as it has been shown to be more robust to outliers compared to the log-binomial model (Chen, Shi, Qian, & Azen, 2014).

We calculated the prevalence ratio for the association between increasing scores in the factors and tuberculosis infection, stratified by the other covariates. This stratified analysis informed our regression model, in which we were able to control for multiple covariates.

For model building, a DAG gold-standard change-in-estimate procedure was followed with some modifications (Weng, Hsueh, Messam, & Hertz-Picciotto, 2009). The full model included all potential confounders and the independent predictors that showed to be related to the outcome in our bivariate analyses. For model reduction, the following procedure was as followed. Our crude model solely included the exposure (factors). The final model was created by adding, one at a time, a new variable. The decision to keep a reduced model or to include a new variable was based in the following criteria, in this order: a) Change in the prevalence ratio of the exposure, b) Variable considered an effect modifier based on the exploratory and stratified analysis, c) Quasi-likelihood information criterion (QIC), in which the model with the smallest QIC being preferred (Pan, 2001) and d) having a parsimonious model. After the model with the main effect was constructed, interactions between each of the explanatory variables and the exposure were assessed.

Creation of factor categories and association with prevalence of tuberculosis infection among household and non-household contacts. We used a stratified random sampling to split the data into training data (67%) and test data (33%), stratifying by household versus non-household contact. The partition of the data into 2/3 for training has been shown to be usually robust (Dobbin & Simon, 2011). In the training data, we categorized the factor scores in three categories according to tertiles: Low tertile, Medium tertile and High tertile to explore the association of these scores with the proportion of tuberculosis infection (with 95% confidence intervals) among the overall population, household and non-household contacts. We repeat this analysis in the test data, using the same values obtained in the training data to define the low, medium and high tertiles.

Sensitivity analysis. In a sensitivity analysis, we changed the criteria of TST positivity, considering the HIV status of the contact (≥ 5 millimeters for HIV-infected individuals) and we excluded contacts with history of tuberculosis disease.

All analyses were carried out using SAS software v 9.4 (SAS Institute, Cary, NC, US) and R v3.3.1 (R Foundation for Statistical Computing, Vienna, Austria, 2016).

Ethical Approval

Written informed consent was obtained from all participants prior to study inclusion. Institutional review board clearance was obtained from Ethics Committee at Makerere University School of Public Health and the University of Georgia.

Methodology for Aim 3.

Aim 3: To estimate the level of tuberculosis transmission in Guatemala City, Guatemala by measuring the proportion of clustered tuberculosis cases based on genotypic matching between 2010 and 2014.

Sub-aim 3.1. To identify risk factors associated with these clustered cases in HIV-infected subjects.

Design Overview

Genotypes of *Mycobacterium tuberculosis* (MTB) isolates from HIV-infected and non-HIV infected tuberculosis cases in Guatemala City from 2010-2014 were categorized as clustered or non-clustered depending of their genotype. Clustered strains were considered as evidence of recent transmission. Independent factors associated with having a clustered isolate were investigated in the HIV-infected subjects (Figure 3.2).

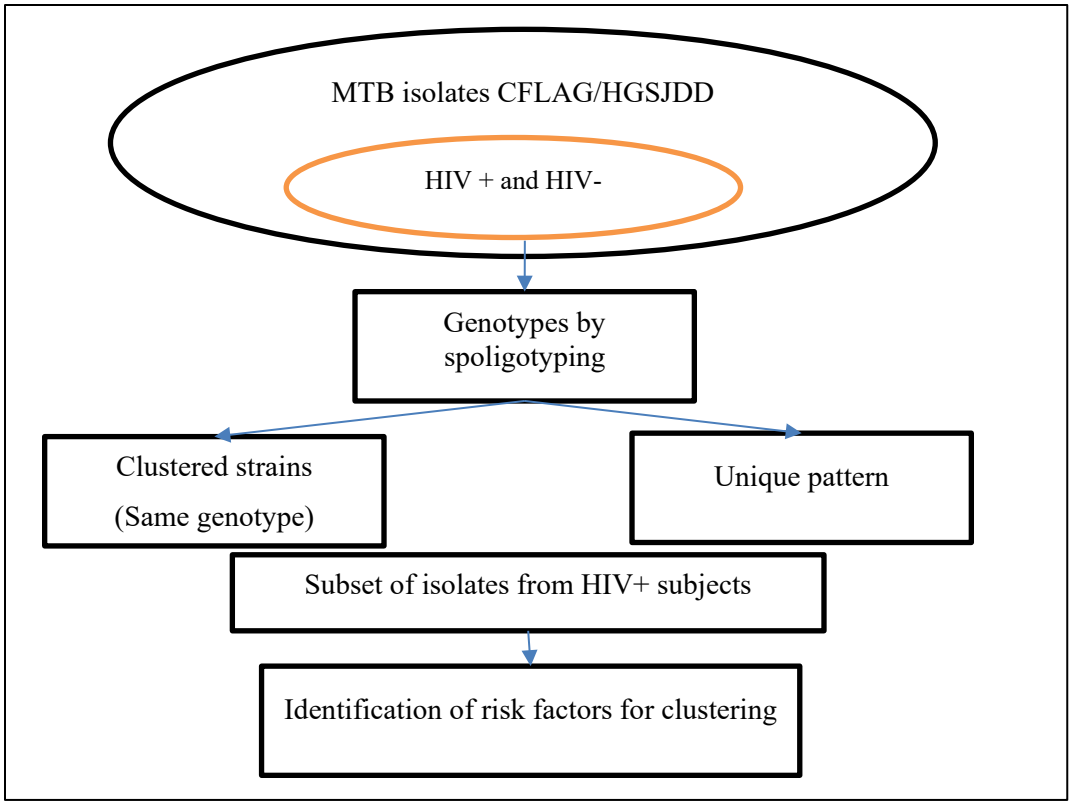


Figure 3.2. Flow diagram of the study

Study population and setting

Study Setting. The study was conducted in Guatemala City, Guatemala at Integral Health Association (ASI). In Guatemala, in 2014, 4,200 new tuberculosis cases were notified for an incidence rate of 25 cases per 100,000/year (World Health Organization., 2016). However, World Health Organization has estimated that the due to the low case detection rate, the incidence could be almost two times higher, with an estimated incidence of 60 cases/100,000/year. Out of these 4,200 individuals with incident TB, 270 were HIV infected subjects.

ASI operates a clinic and a clinical laboratory, which has served over 25 years in Guatemala City, as an institution for the management and treatment of HIV-infected subjects. Currently, there are 2,816 active patients in the clinic. They have been enrolled from 1991 to 2016, with 88% of them being patients from 2006 and later (Figure 3.3).

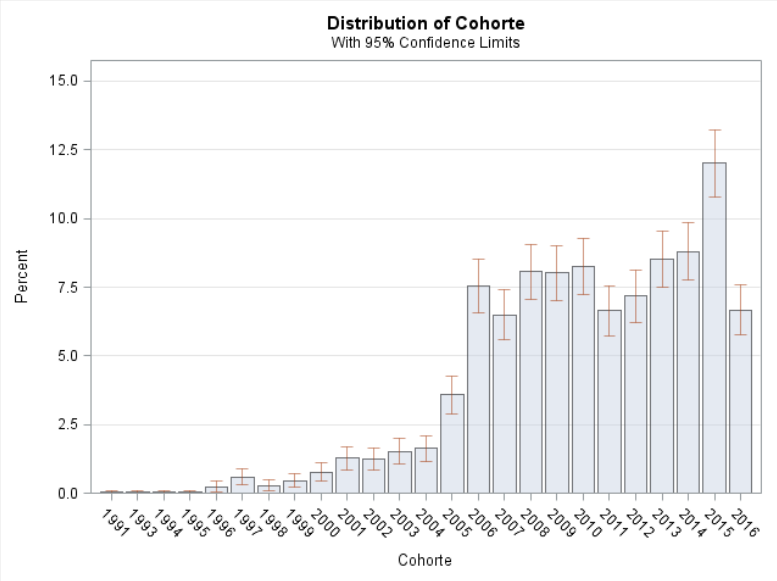


Figure 3.3. Number of patients enrolled in clinic operated by ASI, by year of enrollment.

Most of the 2816 active patients enrolled in the clinic were men (1,762, 63%), followed by women (1027, 34%) and transgender (27, 1%). Male patients tend to be single (59%) and over 25% of them identify as homosexuals. On contrast, over 50% of women are married/live in free union and only 0.2% identify as homosexual. Almost 90% of transgenders are single and homosexual, also they tend to be younger, with 70% of them with an age of 39 years or less. The ethnic group ladino is the most prevalent with over 90% of the patients belonging to it (Data provided by Dr. Blanca Samayoa, Head of Research Unit at the ASI clinic). The current practice at ASI consists in a comprehensive clinical management, laboratory diagnosis and treatment (Dalia Lou, Head of Molecular Laboratory at the ASI clinic, personal communication). The periodicity of the examination and laboratory testing depends on the type of visit that the patient is attending and their individual needs (Table 3.2).

Table 3.2 Work flow for HIV-infected subjects at the ASI clinic.

Type of visit	Clinical management	Laboratory testing	CD4 count	HIV viral load	ART
Baseline	HIV diagnosis/confirmation Weight and height Physical examination	Chemistry blood testing	Yes. If CD4 count <200 cells/mm^3 then other tests are done*.	Yes	Initiation Irrespective of CD4 count**
Year 1	Health check-ups, every 3 months	Just if requested by clinician	Once a year	Twice a year	Yes
Year 2 and more	Health check-ups, every 6 months	Just if requested by clinician	Once a year	Twice a year	Yes

*Lumbar puncture and blood culture

** Initiation of ART irrespective of CD4 count started approximately four years ago. Before, subjects started ART if the CD4 was 200 cells/mm^3.If an opportunistic infection is detected, first this is treated and then ART starts.

The tuberculosis laboratory serves not only HIV population but also receives samples for tuberculosis diagnosis of other centers, in Guatemala City and other regions of the country

Study population and inclusion criteria. DNA isolated from *Mycobacterium tuberculosis* collected from patients who submitted samples for tuberculosis diagnosis to the ASI laboratory and in whom a *Mycobacterium tuberculosis* isolate was detected (2010-2014). MTB isolates should have been confirmed as such by laboratory methods (culture plus specie identification by conventional or molecular methods). Isolates without a clinical record nor a spoligotyping results were excluded. We also excluded isolates with multiple spoligotyping results per isolate but without at least 2 identical results per sample.

Definition of HIV status in the study population. Patients managed and treated in ASI were considered HIV-infected individuals. Patients not managed in ASI but that submitted samples to the ASI laboratory were considered as HIV negative/unknown status.

Genotyping. Isolates have already been analyzed by spoligotyping, as described in detail elsewhere (Kamerbeek et al., 1997). Succinctly, in the direct repeat locus of *Mycobacterium tuberculosis*, direct repeats are interspersed by polymorphic DNA sequences called "spacers". In this technique, both the direct repeats and the spacers are amplified by polymerase chain reaction (PCR). Oligonucleotides that correspond to 43 of these spacers are immobilized into a membrane. The amplified fragments are hybridized to these spacers. The hybridization products are detected by chemiluminescence (Kamerbeek et al., 1997). Depending on the presence and distribution of the spacers, a genotype pattern was obtained for each isolate.

Study exposure

In the case of HIV-infected patients from ASI, selected variables were evaluated as independent factors for the clustering, including demographic, clinical and behavioral variables.

The following demographic variables were analyzed, which were collected at time of enrollment at ASI (baseline), unless specified: Sex, age at time of tuberculosis diagnosis, ethnic group, level of education, place of birth, department of residence, sexual orientation, employment status and civil status.

For the clinical factors the following variables were analyzed: CD4 cell count/mm^3 ± 3 months of tuberculosis diagnosis, history of TB -any episode recorded at least 6 months prior current one, type of sample submitted for tuberculosis diagnosis, smear result, HIV viral copies/mm^3 ± 3 months of tuberculosis diagnosis, HIV clinical stage at enrollment at ASI, HIV clinical stage ± 3 months of tuberculosis diagnosis, discharge motive from ASI during or after tuberculosis diagnosis, year of tuberculosis culture, drug resistance of isolate, multi-drug (MDR) resistance. Multi-drug resistance was defined simultaneous resistance to rifampicin and isoniazid.

For the behavioral factors, the following variables were analyzed, which were collected either before tuberculosis diagnosis or up to 3 months after tuberculosis diagnosis: use of drugs, alcohol consumption, tobacco consumption, had been in a prison.

Study outcome

The study outcome was clustering of a strain. A clustered strain was defined as one that at least share the same genotype with at least one other isolate, regardless of the HIV status of the corresponding patients. Isolates with unique genotype patterns were considered non-clustered. Strains that belong to any cluster were coded as 1 and strains that did not share a genotype pattern with any other isolate were coded 0.

Potential confounders and covariates

In this aim, we identified different risk factors for clustering among HIV-infected subjects. Therefore, most of the variables were considered in the model, as we did not have a single exposure. Nevertheless, we treated age as a potential confounder. Individuals with younger age will always have a higher proportion of recently transmitted cases than those with older age, as reactivation can only occur in individuals that were previously infected (Murray & Alland, 2002).

Statistical issues

Sample size

In population-based molecular epidemiological studies, it is required to collect and genotype as many isolates of *Mycobacterium tuberculosis* as possible, and if possible aiming to coverage of over 60% of the total culture positive cases of the region (Fok et al., 2008). In this aim, all available isolates collected at ASI during the years 2010-2014 were analyzed. We used different scenarios as we lacked information in the sampling rate of the study, so we assumed a sampling rate of 10% or 20% of total active TB cases.

Statistical hypothesis

The null hypothesis was there are no variables than have an effect in the probability of a strain to belong to a cluster (Equation 3.3). The alternative hypothesis was that there were independent factors associated with the clustering of a mycobacterial strain, even after adjustment for other covariates and confounders.

H_0= Demographic, clinical and epidemiological variables among subjects with a clustered strain = Demographic, clinical and epidemiological variables among subjects with a strain of unique pattern ≠

H_A= Demographic, clinical and epidemiological variables among subjects with a clustered strain ≠ Demographic, clinical and epidemiological variables among subjects with a strain of unique pattern

Equation 3.3. Null and alternative hypothesis for Aim 3.

Analytical approach

Aim 3: To characterize the proportion of clustered tuberculosis cases based on genotypic matching in Guatemala City between 2010 and 2014 and to identify risk factors associated with these clustered cases in HIV-infected subjects.

In this Aim, clinical, demographic and behavioral variables were considered the exposure variables and we investigated their effect on the main outcome, presence of a clustered strain. Genotypes of HIV-infected subjects and HIV-non-infected subjects were taken into account to estimate the proportion of clustered strains. Isolates with a unique genotype were considered non-clustered strains and isolates with shared genotype were considered clustered strains.

Descriptive statistics of clustering of mycobacterial isolates. Descriptive statistics were used to present the number and proportion of clustered strains and clusters, distribution of cluster size, mean size of cluster, maximum size of cluster, proportion of cluster with 2 cases, clusters 3-19 cases and clusters ≥20 cases. To be able to compare our findings with the ones previously reported, we included the ones presented and suggested by several reviews, expert comments and recent studies (Fok et al., 2008; Mears, Abubakar, Cohen, McHugh, & Sonnenberg, 2015; Murray & Alland, 2002; C. Yang et al., 2015).

Estimation of the proportion of TB cases due recent transmission. Several methods to estimate the proportion of tuberculosis cases due to recent transmission have been described (Murray & Alland, 2002). We used two: the "n-1" method and a web-based tool based on a regression model.

For the "n-1" method we applied the original formula, developed by Small and others: Recent Transmission Index-$RTI_{n-1} = (n_c - c)/n$ (Small et al., 1994), in which n = total number of cases in the sample, c = is the number of clusters (genotypes represented by at least two cases) and n_c =is the total number of cases in a cluster of two or more . As suggested by Glynn and colleagues (Glynn et al., 2005) because of the length of this study, we re-estimated this proportion using different time windows: 2 years (2010-2011), 3 years (2010-2012) and 4 years (2010-2013).

For the web-based tool, we estimated the recent transmission proportion in the whole time period by using the online tool developed by Kasaie and others (Kasaie et al., 2015). In this technique, several parameters are considered, tuberculosis incidence in the region, sampling coverage, study duration, proportion of clustering and proportion of clusters. These considerations intend to reduce the estimation bias, particularly in areas with a low sampling rate:

i. *Total number of Tuberculosis cases in the sample.*
ii. *Number of clustered cases observed in the sample.* This parameter was estimated by counting the total number of isolates that belong to any cluster.
iii. *Proportion of total active TB cases who have culture and fingerprint data.* This variable is problematic to calculate, because the denominator to estimate this proportion should be the

total number of tuberculosis cases reported in Guatemala City. In the Guatemalan context, as there is limited information regarding this number. Thus, we estimated this proportion using two different scenarios. In the first one, we estimated there were 1,254 tuberculosis cases in Guatemala City during 2010-2014, in the second scenario, it was estimated there were 3,077 tuberculosis cases, by also including neighboring countries as it is likely that many patients have come from these other areas (Data provided electronically by the Health Information Management System of Guatemala). Thus, we estimated this proportion using two different scenarios: 10% or 20% of total active TB cases. *Number of clusters observed in the sample.* This parameter was estimated by counting the total number of clusters that were found in the study sample. One cluster was defined as an identical genotype shared by at least two strains from different patients (Mathema et al., 2006).

iv. *Length of time over which samples were collected (years):* 4 years (2010-2014)

v. *TB Incidence (per 100,000/yrs):* As described earlier, the true incidence of tuberculosis in the country is conflicting. Based on the notification of cases, the incidence rate is 25 cases per 100,000/year (World Health Organization. 2016). However, based on the estimations of the World Health Organization, the incidence rate is 60 cases/100,000/year. The calculation of the recent transmission proportion used both estimates and results were presented accordingly.

Independent factors associated with clustering in the HIV-infected population. Potential factors for clustering were reported for the HIV-population and were compared by Chi-square test (categorical variables), Fisher Test (counts less than 5) or Wilconxon test (continuous variables) in patients with clustered and non-clustered strains. In the categorical variables with more than two classes and when the overall p values were less than 0.20, a Bonferroni correction was conducted: A corrected p value was obtained for the pairwise comparison between a given class and the reference class.

In the variables and classes in which *p* values < 0.20 were obtained, we estimated the crude association between each of these predictors and clustering using regression models. In these regression models, the outcome, or dependent variable, is clustering as defined in section "Study Outcome for Clustering". Since this is a dichotomous outcome variable, a logistic model was initially considered appropriate. However, our sample size was small and there was a very low proportion of one of the events (non-clustered strains). Thus, we estimated the association between exposure and the outcome using the Firth logistic regression method (King & Zeng, 2001). Unadjusted odds ratio and prevalence ratio (with 95% CI) were obtained with this method by exponentiating the regression coefficients. Due to the low proportion of non-clustered strains we did not include conduct multivariate regression models.

All statistical analyses were conducted on SAS software (release 9.4, SAS Institute Inc., Cary, NC, USA).

SENSITIVITY ANALYSIS

We re-analyzed the proportion of clustering and $RTI_{n-1,}$ after exclusion of clusters with a large size, ≥20 isolates. This approach has shown to improve at least partly the specificity of spoligotyping (Scott et al., 2005).

SUPPLEMENTARY ANALYSIS

In the patients that contributed to multiple samples in this study, we described the most relevant clinical, epidemiological and laboratory characteritics.

ETHICAL APPROVAL

Institutional review board clearance was obtained from Zugueme, a Guatemalan independent Ethics Committee, approved by the Ministry of Health of Guatemala and by the University of Georgia.

Chapter 4

DEFINING ADEQUATE CONTACT FOR TRANSMISSION OF MYCOBACTERIUM TUBERCULOSIS IN AN AFRICAN URBAN ENVIRONMENT[3.]

[3] Castellanos ME, Ebell M, Dobbin KK, Quinn F and Whalen CC. To be submitted to BMC Public Health.

Abstract

Background. The risk from exposure to infection increases according to the contact rate, but the definition of adequate contact for transmission is still poorly understood. In this study we aimed to identify factors that can explain the level of contact between tuberculosis cases and their social networks in an African urban environment.

Design/Methods. This was a cross-sectional study conducted in Kampala, Uganda from 2012-2016. We carried out an exploratory factor analysis (EFA) in social network data from tuberculosis cases and their corresponding contacts. We evaluated the factorability of the data to EFA by the Kaiser's Measure of Sampling Adequacy (KMO). The principal axis factoring method and oblique rotation were employed to extract and rotate the factors. We generated factor scores for each interaction case-contact using the weighted sum scores method. The construct validity of the generated factors was evaluated by associating them with other variables related to social mixing.

Results. One hundred and twenty tuberculosis cases provided complete social network information of their interactions with 1,157 social contacts. Thirteen items displayed high intercorrelation (KMO=0.72) and were included for EFA. Two factors were identified, which captured 82% of the variance. The first factor, named 'Setting' involves the type, frequency, duration and ventilation of the usual place of meeting as well the physical proximity among tuberculosis cases and contacts, represented by the sleeping and eating patterns. The second factor, named 'Relationship' was explained by the relationship duration as well as the level of intimacy among cases and contacts, represented by the strength of knowledge of each other, provision of healthcare, and if they were travel buddies. Setting and Relationship scores varied according to the age, gender and nature of the relationship among tuberculosis cases and their contacts.

Conclusions: In this large cross-sectional study from an urban African setting, we identified two factors that can define contact between tuberculosis case and their social networks. These findings also confirm the complexity and heterogeneity of social mixing.

Introduction

Mathematical models of respiratory infectious diseases assume homogeneous mixing in the population, but evidence suggest otherwise (Kong et al., 2016; Mossong et al., 2008; Wallinga et al., 2006). Assuming homogenous mixing among subjects will yield inaccurate estimations of the level of contact and transmission of infectious diseases (Mossong et al., 2008; Wallinga et al., 2006).

The definition of "contact" when studying the spread of respiratory or close-contact transmitted infectious diseases is not standardized but generally, a contact is referred when, at a minimum, a short face-to-face conversation occurs within a short distance and/or physical contact (Dodd et al., 2016; Edmunds et al., 1997; Mossong et al., 2008). Further characterization requires, among others features, knowledge of the order, frequency and duration of the contact (Bansal et al., 2010). Additionally, several factors have been identified to modify the frequency and nature of the contact within an infectious case and their contacts, such as age and gender of the individuals, usual place of interaction and ventilation of the setting (Dodd et al., 2016; Feenstra et al., 2013; Johnstone-Robertson et al., 2011; Melegaro et al., 2011; Mossong et al., 2008).

In order to define adequate contact, all of these factors should be included in studies aiming to understand the dynamics of social mixing among population. These variables tend to be highly correlated, so methods such as principal component analyses and exploratory factor analyses (EFA) have been used to detect the interrelationships among observed variables using data reduction(Pett et al., 2003). EFA is preferred in situations where the aim is also to identify underlying constructs (called 'factors') among variables (Treiblmaier & Filzmoser, 2010).

In the present paper, we carried out an exploratory factor analysis from information collected from a social network survey conducted among tuberculosis cases that assessed their social mixing with contacts within established ego-centric networks, in Rugaba, Uganda. Our main aim was to identify underlying factors that would explain the level of contact among them. As a secondary aim we evaluated

the construct validity of these factors by evaluating their association with other variables related to social mixing.

Study Population and Methods

STUDY POPULATIONS

Index cases. We conducted a cross-sectional study in patients with active tuberculosis in Rugaba, Uganda. The index cases were persons aged 15 years or older, who are residents of Rugaba, Uganda; who had signs and symptoms of pulmonary tuberculosis and that were microbiologically confirmed by a positive sputum smear for acid-fast bacilli, a positive GeneXpert®, a culture or a mixture of these methods. All were enrolled in the Community Health and Social Networks of Tuberculosis study (RO1 AI AI09386), from 2012 to 2016. Patients were evaluated by the National Tuberculosis Control Program, and after it was confirmed they met inclusion criteria and provided informed consent were recruited. Demographic and smear grade status was collected from the tuberculosis cases.

Contacts. The index case provided a list of their household and community contacts- individuals they spent the most time with outside the households. Members of the field team approached these contacts and if they agreed, they were enrolled in the study. In this subset of individuals, we collected demographical variables.

STUDY INSTRUMENT

Information was obtained by trained field visitors using standardized questionnaires that were designed to assess the social networks of index cases. Due to logistic reasons, the index cases restricted this list to the ten closer contacts. The variables related to the social mixing were collected in a social network form (Table S 4.1).

ANALYTICAL STRATEGY

Descriptive Analyses of Index Cases. Baseline characteristics of the tuberculosis cases were summarized with proportions and measures of central tendency.

Factor analysis. We performed an exploratory factor analysis to determine the relationship among variables. The main aim of EFA is to reduce the number of variables than can explain a response/outcome variable to fewer variables (called factors), being these factors a set of underlying constructs present in the data (Härdle & Simar; Joreskog et al., 2016; Treiblmaier & Filzmoser, 2010). The rationale behind the use of an exploratory factor analysis in this aim was that several of the variables collected from the tuberculosis cases are likely to be highly correlated. Thus, a variable reduction technique is necessary, and EFA was a better option than principal component analysis as we believed that the observed variables can be grouped in underlying factors (Suhr, 2005).

Initial preparation of the data and descriptive analyses. First, an item analysis was conducted in the original data set to explore the distribution of each variable. We combined and recoded variables, to obtain a dataset with dichotomous or ordinal values (named 'Recoded dataset'), in which the lowest value represented theoretically less contact and the highest value represented the highest contact (Supplementary Material, Table S 4.2 & Table S 4.3**)**. Items were excluded from EFA **(**Table S 4.3**)** if they were nominal, had limited distribution or did not provide additional information with respect to other items.

We then conducted an item-analysis of the recoded variables that allow us to define the relationship between tuberculosis cases and their contacts. This analysis included the variables considered for EFA and the nominal variables of the data set (independently if they were included or excluded from EFA).

Factorability of the original matrix. As we had mixed data (ordinal and binary data) we created a polychoric correlation matrix. EFA models obtained with a polychoric correlation have shown to be more consistent with the measured variables than the Pearson correlation when using ordinal data (Holgado–Tello et al., 2010). We evaluated the factorability of the items considered to EFA by the Kaiser's Measure of Sampling Adequacy (KMO) (Kaiser & Rice, 1974). We excluded variables in which this measurement was lower than 0.6 (Taherdoost et al., 2014).

Factor extraction and Number of factors to retain. We selected the principal axis factoring method to extract our factors. This selection was based on the goal to identify underlying constructs in the data set (Treiblmaier & Filzmoser, 2010). We followed the analysis described by Berghaus and others, in which three methods were considered: a) keeping factors with eigenvalues[4] greater than 1 (Kaiser Criterion), b) Scree plot- in a graphical representation, retains eigenvalues that appear in the sharp slope of the plot and drop the ones that appear in the leveling area c) variance explained criteria-keep factors that account for 80-90% of the variation. Moreover, only factors with at least three items were retained (Costello & Osborne, 2005).

Factor rotation. As we believed the factors were correlated, an oblique rotation of the factors was conducted (Gaskin & Happell, 2014). This technique involves the transformation of the factors in order to obtain simpler and more interpretable results (Nardo et al., 2005). After the factors were identified, a name was assigned to each factor, based on their common characteristics and literature review (Barbero Garcia et al., 2013).

Generation of Factor Scores. Factors scores were computed for each individual. To generate the factor scores for an individual, we used the weighted sum scores method, as it allows that the items with the highest loadings[5] to have the highest impact in our factor score (DiStefano et al., 2009). Briefly, we multiplied the factor loading for each item to the original score from the Recoded dataset. We then summed together the values obtained for each multiplication for the items grouped in a particular factor to generate the factor score. We excluded items with loadings below 0.30 (Beavers et al., 2013).

Factor scores were investigated to check if they met the normality assumption and to determine if they have unimodal distribution (Ameijeiras-Alonso et al., 2016). Based on these analyses, the results of the factor scores are presented as median with interquartile ranges.

---

[4] Eigenvalues= variances of the principal components (Nardo et al., 2005).

[5] Loading= 'Correlation between observed variables and factors'(Suhr, 2005)

Association Factor scores with other variables. We wanted to establish the construct validity of the factor scores. To do this, we examined the relationship of these factors with other variables collected from the contacts of the index case who we traced and enrolled in the study (See section: 'STUDY POPULATION. Contacts'). We stratified this group according to type of contact (household and non-household contact) and nature of the relationship between case and contact (spouse, child, sibling, friend, co-worker, other relatives, neighbor, other) The median and interquartile ranges of the identified factor scores were estimated for each stratum. In the process, we examined the variability of the factor scores by gender of contact (men, women), age of contact (0-4 years, 5-14 years, 15 years and greater), age of index (15-24 years, 25-44 years, 45 years and greater), gender of index case (men and women) and usual place of meeting' (home tuberculosis case, friend's home, relative's home, work place, bar, trading center/shop/kiosk, elsewhere).

To compare the difference in medians among stratified groups we estimated the 95% confidence intervals by bootstrapping, using the package 'boot' for R software (Canty & Ripley, 2017) . We set the number of bootstraps replicates to 10,000 and calculated the intervals with the adjusted bootstrap percentiles (BCa) method. We selected this parameters based on recommendations from Puth and others (Puth et al., 2015).

Sensitivity analysis. We repeated the EFA not in a polychoric correlation matrix but with transformed data. An optimal monotonic transformation of the raw data was performed (PROC PRINQUAL, method=maximum total variance)(SAS Institute Inc, 2017). We selected the 'monotone' transformation, as our variables were ordinal or binary. The transformed variables were used to conduct the factor analysis, following the same procedure previously described.

ETHICAL APPROVAL

Written informed consent was obtained from all participants prior to study inclusion. Institutional review board clearance was obtained from Ethics Committee at Makerere University School of Public Health and the University of Georgia.

Results

Descriptive characteristics of index cases. We obtained information from 120 index cases with tuberculosis disease regarding their social network. Each case had a median of nine contacts (IQR 8, 12), for 1,179 contacts from the 120 TB cases. For male cases, the median number of household contacts was 4 (IQR 2,6) and non-household contacts was 7 (IQR 4,9). For women, the median number of household contacts was also 4 (IQR 2,6), but they only had a median of 5.5 non-household contacts (IQR 4,8). The majority of index cases were men (83%), young adults between 25-44 years of age (57%), belonged to the Ganda tribe (72%) and had a smear positive **(**Table 4.1).

Descriptive quantitative analysis of the recoded variables. Index cases provided complete social network information for 1,157 out of their 1,179 contacts (98%). The most common type of relationship between contacts and cases were friends (30%), relatives (19%) and siblings (13%) (Table S 4.4). The length of knowing a contact was heterogeneous. Cases have known 25% of the contacts by more than 6 years, whereas 37% of the contacts have been known less than 2 years. Most tuberculosis cases did not adjust the frequency of meetings with their contacts after onset of the cough (89%). In 40% of the cases, the time spent between a case and their contact was between 3.5-28 hours/week. The usual place of meeting was the household of the tuberculosis case (56%), followed by the work place (18%). The ventilation of the usual place of meeting was reported to be full in 48% of the encounters, but it was minimal to poor in a third of them. In 70% of the meetings among tuberculosis cases and contacts, two or more additional people were present (38% 2-4 persons, 22% 5-6 persons, 10% > six persons). Sixty one percent of the TB cases and contacts shared meals at least once a week and 22% of them have slept in the same room or bed. Tuberculosis cases reported to not have shared their tuberculosis diagnosis to more than 55% of their contacts. Six percent of the contacts were reported to have cough.

Factorability of the correlation matrix. We initially considered 15 items from the social network questionnaire for the exploratory factor analysis (Figure S 4.1). There were two variables “Frequency of meeting since onset of cough” and “Number of other people met in addition to contact” with a low

individual KMO (0.36 and 0.12 respectively) and were excluded from the factor analysis (Table 4.2). The 13 items included had an overall KMO 0.72, with individual KMO measurements of >0.60. The visual inspection of the correlation matrix confirmed the exclusion of the aforementioned variables. "Frequency of meeting since onset of cough" did not have a single correlation > 0.15 with any of the other variables and "Number of other people met in addition to contact" only have correlations higher than 0.15 with 3 variables (Figure 4.1, top panel). After excluding these variables, the visual inspection of the polychoric correlation matrix showed an improvement in the degree of correlation among these variables (Figure 4.1,bottom panel). All variables, except "Contact have cough" had correlation values ≥ 0.30 with at least half of the included variables. The variables with the highest number of correlations with the other variables were "Frequency of shared meals since onset cough" and "How well does the case knows contact", each of them correlated with 11/12 variables.

Numbers of factors retained. There were two factors with Eigenvalues higher than one, the first with a value of 5.66 and another with a value of 1.92. These two factors captured 82% of the variance (Table 4.3). Visual inspection of the Scree plot suggested a third factor might be worth to consider (Figure S 4.2) and would have explained an additional 8% of the variance. Thus, promax rotation was conducted choosing first three and then two factors. After rotation of the matrix with three retained factors, six variables were grouped in one Factor and six variables were grouped in a second Factor, with just one single variable (Contact have cough) loaded individually in a third factor (Table 4.4). When two factors were retained, Factor 1 and Factor 2 grouped again the same variables as previously. Variable "Contact have cough" produced low factors loadings in both factors (0.03 and 0.15) implying that this variable might not contribute particularly to any of them.

Based on these results, we decided that our final analysis would retain two factors. Factor 1 grouped variables related to the setting and environment of the contact between the index case and his/her contact, so we named this factor as "Setting factor". The six variables in this category had factor loading of 0.60 or more. Of these six variables, "Nature of ventilation at usual place of meeting", "Frequency of

sleeping in the same room and bed since onset of cough" and "Contact happens indoor or outdoor" had the highest loadings (Table 4.5). Factor 2 grouped variables that corresponded to the intimacy and social relationship of the index case and contact, thus we named it as "Relationship factor", with "Case shared TB diagnosis with contact" and "Care was provided by the contact in the past 3 months" the variables with the highest loadings.

Description of factors scores. We were able to trace and enroll 923 of the 1,157 contacts listed by the index cases. Results from the social network data are indistinct in both groups (Table S 4.4). Factor scores were produced for each of these 923 contacts. In this population Setting and Relationship scores, followed a multimodal distribution (p=0.0 for each factor, unimodality test) (Figure 4.2). Values of setting factor were dispersed between the range of 5.3-18.8, with a median of 10.2 (IQR 6.9, 13.7). In the case of the relationship score, values varied between 4.0 to 14.8, with a median of 7.8 (IQR 6.4, 10.2).

Construct Validity of the Factor Scores. Setting and Relationship scores varied according to the nature of the relationship among a tuberculosis case and their contact (Figure 4.3). Spouses (median: 17.6, IQR 16.3, 18.0) had the highest Setting score, followed by children (median: 14.6, IQR 12.8, 16.3) and siblings (median: 14.3, IQR 12.3, 15.4). In the case of the Relationship factor, spouses (median: 12.8, IQR 10.9-14.1), siblings (median: 11.0, IQR 8.33-12.4) and relatives (median: 9.8, IRQ 7.3, 11.5) had the highest score. For both factors, co-workers, friends, neighbors and other type of contacts had the lowest scores.

Household contacts had a higher setting and relation scores than non-household contacts (Figure 4.4, $p < .0001$). For the setting factor, household contacts had a median value of 14.6 (IQR 12.8-16.2), as opposed to non-household contacts (median 7.4, IQR 6.3, 9.8). The difference in medians in these groups was 7.2 (95% CI 6.8-7.6). The median value of the Relationship factor in household contacts was 9.9 (IQR 7.7, 11.7), higher than non-household contacts (median 7.1, IQR 5.9, 8.6). The difference in medians was 2.9 (95% CI 2.2-3.5).

Stratified analyses of the median values of the factors scores among contacts revealed differences owed to age and gender of the participants (Table 4.6). Contacts of male cases had a lower setting score (median: 9.3, IQR 6.8-12.9) as opposed to contacts of female cases (median: 11.5, IQR 7.7, 15.2). The difference in medians was 2.2 (95% CI 1.3 -3.3). After stratification by gender of contact, the relation male case-male contact had the lowest setting score (median: 9.1, IQR 6.7, 12.4) and the relation female case-male contact showed the highest score (median: 12.0, IQR 8.2, 15.4). The relationship score in contacts of male cases (median: 7.9, IQR 6.3, 10.2) was similar to the one of contacts with female cases (7.7, IQR 6.6, 9.7), with a difference in medians of 0.1 (95% CI -0.2-0.4). These findings were not affected by the gender of the contact (Table 4.6).

Contacts ≤ 4 years old had the highest setting score (median: 13.8, IQR 9.1, 16.2), followed by contacts aged 5-14 years (median: 13.2, IQR 9.9, 15.1) and contacts ≥15 years old (median: 8.8, IQR 6.7, 12.4). The difference in medians among contacts ≤ 4 years of age and contacts ≥ 15 years of age was 4.8 units (95% CI 2.9, 6.2). The combination contacts ≥15 years old and cases aged 24-44 years was the one with the lowest setting score (median: 8.3, IQR 6.6, 12.1). In contrast, the combinations contacts 5-14 years with cases 14-24 years and contacts ≤ 4 years with younger cases (15-24 years and 24-44 years age groups) proved to be the highest (Table 4.6**).**

An inverse association between the age of contact and the relationship score was found. Younger contacts had a median relation score of 6.7 (IQR 5.7, 7.3) as opposed to 7.8 (IQR 7.1, 9.0) among contacts aged 4-14 years and 8.1 (IQR 6.4, 10.7) in contacts aged 15 years or greater. The difference in medians among contacts aged 0-4 years and contacts aged 15 years or greater was 1.5 units (95% CI 1.1, 1.9).

The usual place of meeting among index cases and their household contacts was primarily the household of the tuberculosis case (326/349, 93%). Therefore, we just explored the relation of the setting and relationship scores with the usual place of meeting among non-household contacts and their cases. Among them, meetings in bars resulted in the highest setting score (median: 10.8, IQR 10.5, 11.2),

followed by the home of tuberculosis case (median: 8.2, IQR 7.1, 10.8), work place (7.0, 95% CI 6.2, 8.7) and elsewhere (median: 7.0, IQR 6.1, 8.7). Friend's home (median: 6.4. IQR 5.9, 8.2, relative's home (median: 6.9, IQR 5.8, 8.5) and trading center (median: 6.3, IQR 5.5, 7.2) had the lowest setting scores. On the contrary, the relationship score did not seem to be modified by the usual place of meeting (Figure 4.5).

Sensitivity analysis. The EFA conducted in the transformed data, fourteen items were included as the variable 'Contact have cough' had an individual KMO <0.60 so it was initially excluded. As in our primary analysis, two factors were identified, both with the same variables as previously reported. There were slight differences regarding to the loading values and the relative importance of the variables in each factor (Table S 4.5).

Discussion

In our study, we used exploratory factor analysis to identify two underlying constructs related to the social contact pattern between an infectious case and a contact. The first factor characterized the setting and environment of contact with the index cases and the second described the relationship between the index case and contact. Based on our full analyses, we propose that these two factors can characterize adequate contact among tuberculosis cases and their contacts.

We found that 13 of the 15 variables included in the analyses displayed high intercorrelation. Combined the two factors explained 82% of the variance in the data, with the first factor identified over 60% of the variance. The setting score between a tuberculosis case and a contact is mostly explained by the type, frequency, duration and ventilation of the usual place of meeting as well the physical proximity among cases and contacts, represented by the sleeping and eating patterns. The relationship score is explained by the relationship duration as well as the level of intimacy among cases and contacts, represented by the strength of knowledge of each other, provision of healthcare, and if they are travel buddies. We were also able to establish, in part, the construct validity of these factors. We found that the setting and relationship scores varied according to the nature of the relationship among a tuberculosis case

and their contact and that household contacts had a higher setting and relation scores than non-household contacts, as expected. We observed that family members, especially spouses have the higher values as expected when compared to community members. However, it seems that even among these categories of contacts, there are degrees of proximity that should be considered, suggesting these scores could refine and further characterize the level of contact among contacts of an infectious case.

We found that among non-household contacts, the highest setting score occurred in bars as opposed to other settings. Bars have been implicated in tuberculosis transmission, being suggested that it occurs because close contact happens in a confined space (Classen et al., 1999). Although this association might be confounded by other factors such as smoking, alcohol or drug use, our results seem to confirm there is a high level of contact in this type of location.

Contacts ≤ 4 years old had the highest setting score, particularly with index cases below 45 years. This suggests age assortment of these contacts with their parents or older siblings and will explain the high risk for tuberculosis transmission that these children face in these homes.

Regarding gender, higher setting scores were computed between female cases and their contacts as compared to male cases with their contacts. This will indicate that the main place of interaction of women with their contacts is the household with a high contact rate. Results from a qualitative study conducted in a low-incoming setting showed women tend to stay at home to fulfill their role as caregiver and because of low opportunities for women in the formal job market (Onifade et al., 2010). Moreover, women in Uganda are reported to work 18% more than men in activities at home (Ortiz-Ospina & Tzvetkova, 2017).

The second factor explained an additional 22% of the data set. All the observed variables, except type of transportation, are easily linked to the social and emotional closeness between cases and contacts as well as their length of knowledge. Trust and confidence (measured as sharing tuberculosis diagnosis) were the two variables with the higher loading in the relationship factor. The means of transportation most often used with the contact also was grouped in this factor, but his importance was the lowest, being

the one observed variable not comparable to the others regarding the social closeness. Non-household contacts have low relationship scores whereas household contacts have higher relationship scores. And these scores seem to vary from moderate to high, suggesting different levels of intimacy within the household.

We did not find any gender differences for the relationship factor but there were related to age. In all the age-brackets of the index cases, the higher the age of the contact, the higher the relationship score, confirming the validity of our factor to measure the grade of social closeness.

The relationship factor might modify the adequate contact required for a successful transmission event, so it might be appropriate its study in the dynamics of tuberculosis transmission. Moreover it might be a good indicator of social support, so its measurement could be relevant in an array of disciplines and areas, such as dealing with HIV stigma or mental health interventions(Treiblmaier & Filzmoser, 2010; Tsai et al., 2012). A systematic review of social network analyses in low- and middle-income settings has shown that behavior and health outcomes are associated with the structure and composition of these networks (Perkins, Subramanian, & Christakis, 2015).

EFA has been criticized in the past to produce artificial factors that are not informative (Shapiro, Lasarev, & McCauley, 2002). One of the major strengths of our study is that we minimized this risk by conducting additional analyses that corroborate the robustness of our factors, as it has been recommended (Edefonti et al., 2010). First, our EFA was conducted not only in the polychoric correlation matrix but in the transformed data. Two factors were obtained with identical group of observed variables. Second, the produced scores were consistent with other variables that has been used to describe to certain extent the social network structure (household versus non-household contact, type of the relationship).

There are several limitations of the study. The first one is that because for time and resource constraints we limit the number of reported contacts. However, the number of household contacts per index case that we detected was four, similar to the 3.7 reported among Kampala residents in the Uganda National Household Survey 2016/2017 report (Uganda Bureau of Statistics (UBOS), 2018). A second

limitation would be the risk for recall and response bias as we included self-reported data. Nevertheless, the nature of the questions and the high dispersion and variability of both location and factor scores suggest tuberculosis cases did not felt compel to answer in a particular direction regarding the level of contact with their contacts (no social desirability bias) (Furnham, 1986). Finally, there were some nominal variables that we had to exclude or recoded as binary data from our analyses and which could have been improved our results. However, due to the nature of EFA we just included ordinal and binary data. A principal component method called Factor Analysis of Mixed Data (available under the R package 'FactoMineR'), could be further used to develop dimensions that explore the association of the nominal variables with the other variables (Lê, Josse, & Husson, 2008).

In conclusion, our study identified two factors that can define adequate contact between a tuberculosis case and their contact, explaining 82% of the variance in the observed variables. As a whole, these findings also confirm the complex and heterogeneous social mixing between cases and contacts (Wallinga et al., 2006). In future studies we will evaluate whether these scores might be used to determine the presence of tuberculosis infection among their social networks.

TABLES AND FIGURES

TABLES

Table 4.1. Baseline characteristics of index cases who provided social network form.

Category	No.	(%)
Total number of index cases	**120**	
Male gender	83	69
Age, years, median [IQR[1]]	28 [23-36]	
Age (category)		
15-24	38	32
25-44	68	57
45 or more	14	12
Tribe		
Ganda	86	72
Nyakitara	4	3
Lunyankole	12	10
Lusoga	2	2
Other	14	12
Missing	2	2
Smear grade[2]		
No AFB observed	11	9
10-99 AFB/field	7	6
1-10 AFB/field	17	14
>10 AFB/field	80	67
Missing	5	4

[1] Interquartile range

[2] Number of acid-fast bacilli (AFB)/oil immersion field

Table 4.2 Individual and Overall Kaiser's Measure of Sampling Adequacy.

Initial results with 15 items and final selection with 13 items included in the exploratory factor analysis.

Variable	Kaiser's Measure of Sampling Adequacy 15 items	Kaiser's Measure of Sampling Adequacy 13 items
Overall	**0.55**	**0.72**
Contact happen indoors or outdoors	0.49	0.61
Nature of ventilation at usual place of meeting	0.46	0.63
Case shared TB diagnosis with contact	0.49	0.69
Contact have cough	0.54	0.69
Frequency of shared meals since onset cough	0.55	0.72
Frequency and duration of contact over the past month	0.82	0.73
Care was provided by the contact in the past 3 months	0.53	0.73
Place of usual meeting. Home TB case versus other location.	0.80	0.74
Case trusts contact	0.66	0.75
Length of knowing contact	0.70	0.77
Frequency of sleeping in same room and bed since onset cough	0.54	0.79
How well does the case knows contact	0.54	0.80
Means of transportation used most often with contact. None (walking) versus a type of transportation.	0.79	0.81
Frequency of meeting since onset cough	0.36	NE[1]
Number of other people met in addition to contact	0.12	NE

[1]Not estimated as it was not included in the exploratory factor analysis

Table 4.3. Eigenvalues of the Reduced Correlation Matrix.

Relative proportion of accounted variance and cumulative variance for each factor.

Factor	Eigenvalue	Proportion (%)	Cumulative
1	5.66	61.19	61.19
2	1.92	20.80	82.00
3	0.76	8.23	90.22

Table 4.4. Factor loadings matrix identified by exploratory factor analysis when three factors were retained.

Variable	Factor1	Factor2	Factor3
Nature of ventilation at usual place of meeting	0.64545	-0.07833	0.52392
Frequency of sleeping in same room and bed since onset cough	0.81872	0.06922	-0.00166
Contact happen indoors or outdoors	0.62427	-0.03032	0.54788
Frequency of shared meals since onset cough	0.83986	0.17727	-0.10086
Place of usual meeting. Home TB case versus other location.	0.80195	-0.07569	-0.15572
Frequency and duration of contact over the past month	0.60544	0.11178	0.06195
Case trusts contact	-0.18874	0.93985	0.17896
Case shared TB diagnosis with contact	-0.13698	0.91887	0.20385
Case was provided by the contact in the past 3 months	0.38478	0.68455	-0.15040
Length of knowing contact	0.30005	0.52783	-0.04004
How well does the case knows contact	0.46085	0.50101	-0.04756
Means of transportation used most often with contact. None (walking) versus a type of transportation.	0.11689	0.48138	0.04777
Contact have cough	-0.12275	0.19975	0.38869

Table 4.5. Factor loadings matrix identified by exploratory factor analysis when two factors were retained.

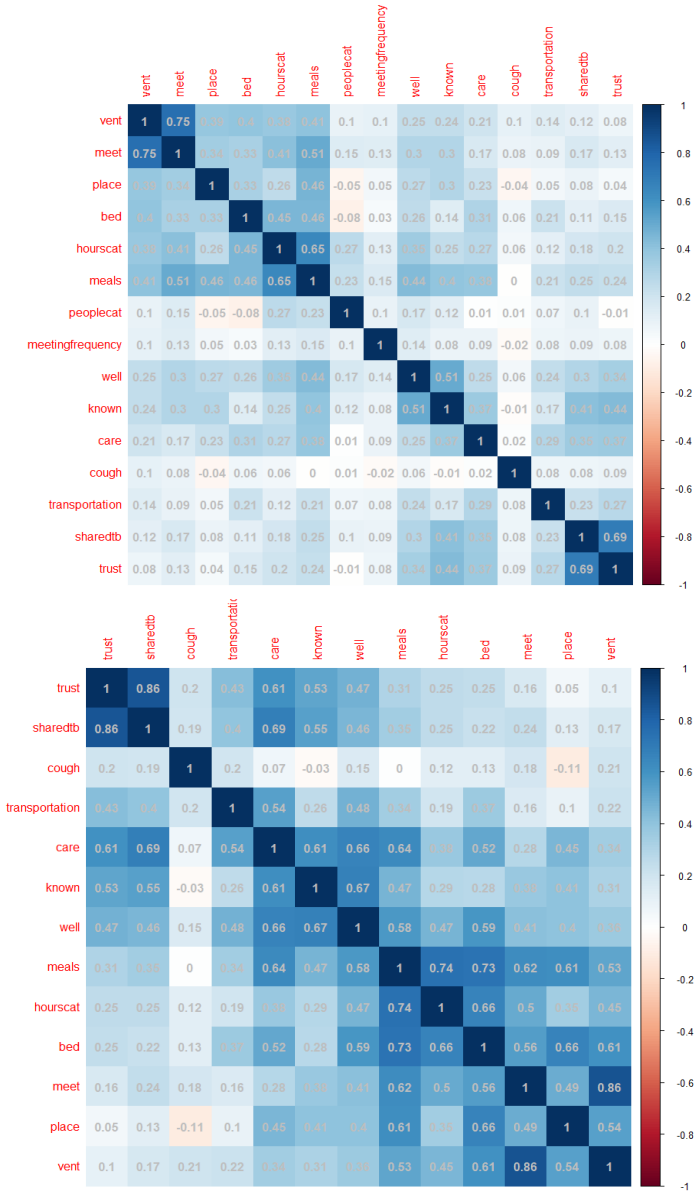
Variable	Factor1 (Setting)	Factor2 (Relationship)
Nature of ventilation at usual place of meeting	0.82523	-0.09963
Frequency of sleeping in same room and bed since onset cough	0.81426	0.07259
Contact happen indoors or outdoors	0.80645	-0.05302
Frequency of shared meals since onset cough	0.76514	0.20167
Place of usual meeting. Home TB case versus Other location.	0.71793	-0.04512
Frequency and duration of contact over the past month	0.62684	0.1081
Case trusts contact	-0.20371	0.94897
Case shared TB diagnosis with contact	-0.13972	0.92397
Care was provided by the contact in the past 3 months	0.2422	0.7234
Length of knowing contact	0.21786	0.56289
How well does the case knows contact	0.38318	0.53055
Means of transportation used most often with contact. None (walking) versus a type of transportation.	0.08628	0.49809
Contact have cough	0.03471	0.15498

Table 4.6. Median Setting and Relationship scores (with interquartile range-IQR) from the sub-set of 923 enrolled contacts with demographic variables collected.

Overall and stratified by gender and age of cases and contacts.

Variable	N (%)	Setting Score Median (IQR)	Relationship Score Median (IQR)
Overall	923 (100)	10.2 (6.9,13.7)	7.8 (6.4,10.2)
Contact with Male Index	649	9.3 (6.8,12.9)	7.9 (6.3,10.2)
Male contact	333	9.1 (6.7,12.4)	7.9 (6.7,9.9)
Female contact	316	10.1 (6.9,12.3)	7.8 (6.0,10.8)
Contact with Female Index	274	11.5 (7.7, 15.1)	7.7 (6.6,9.7)
Male contact	123	12.0 (8.2,15.4)	7.4 (6.4,8.9)
Female contact	151	11.4 (7.2,15.0)	7.8 (6.7,10.4)
Contact with index 15-24 years	288	11.3 (7.3,14.8)	8.0 (6.6,10.6)
00-04 years contact	42	14.2 (9.1,16.2)	6.7 (6.0,7.3)
05-14 years contact	47	14.2 (10.6,16.0)	7.9 (7.7,8.4)
15 years and greater contact	199	10.7 (6.9,13.9)	9.0 (6.5,11.2)
Contact with index 24-44 years	512	9.4 (6.7,13.2)	7.8 (6.3,10.2)
00-04 years contact	44	14.2 (9.4,16.3)	6.8 (5.7,7.3)
05-14 years contact	74	13.0 (9.8,14.9)	7.9 (6.6,10.0)
15 years and greater contact	394	8.3 (6.6,12.1)	8.0 (6.4,10.4)
Contact with index 45 years and greater	123	10.2 (6.9,12.7)	7.7 (6.2,9.7)
00-04 years contact	15	11.0 (6.6,15.4)	6.3 (5.4,6.8)
05-14 years contact	16	12.4(6.9,13.2)	7.7 (7.6,7.9)
15 years and greater contact	92	10.0 (7.0,11.9)	7.7 (6.1,19.4)

FIGURES



Legend	
Vent	Nature of ventilation at usual place of meeting
meet	Contact happen indoors or outdoors
place	Place of usual meeting. Home TB case versus Other location.
bed	Frequency of sleeping in same room and bed since onset cough
hourscat	Frequency and duration of contact over the past month
meals	Frequency of shared meals since onset cough
peoplecat	Number of other people met in addition to contact
Meeting-frequency	Frequency of meeting since onset cough
well	How well does the case knows contact
known	Length of knowing contact
care	Care was provided by the contact in the past 3 months
cough	Contact have cough
transportation	Means of transportation used most often with contact. None (walking) versus a type of transportation.
sharedtb	Case shared TB diagnosis with contact
trust	Case trusts contact

Figure 4.1. Visual representation of polychoric correlation among variables considered for factor analysis.

Top graph: Initial set of variables considered (n=15). Bottom graph: Final set of variables included (n=13). The more intense the blue and pink color the higher the positive (blue) or negative (pink) correlation.

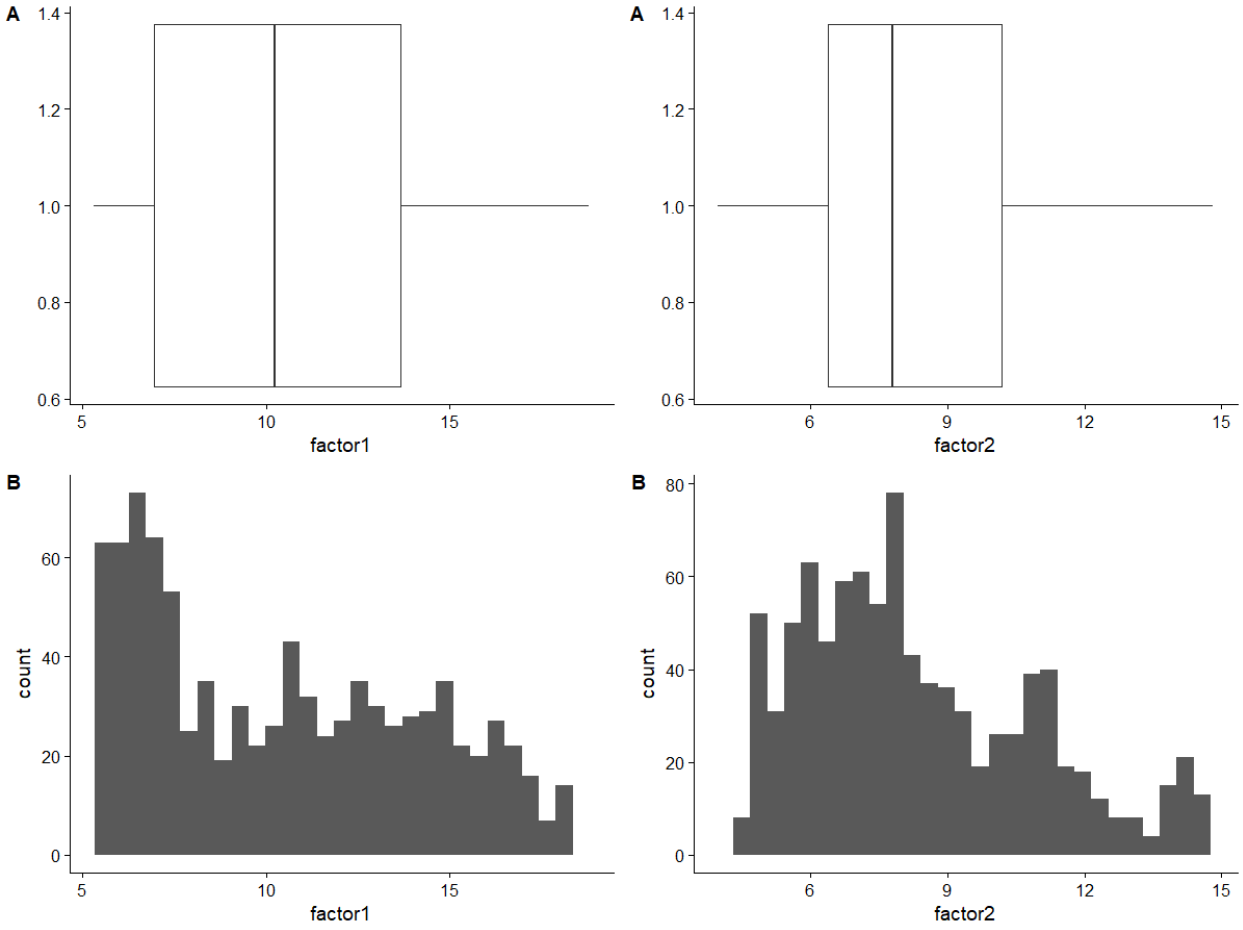


Figure 4.2 Distribution of closeness factors among the study population.

A histogram and a boxplot are shown to study the distribution of the Setting and Relationship factor. Left Panel: Setting Factor. Right Panel: Relationship Factor.

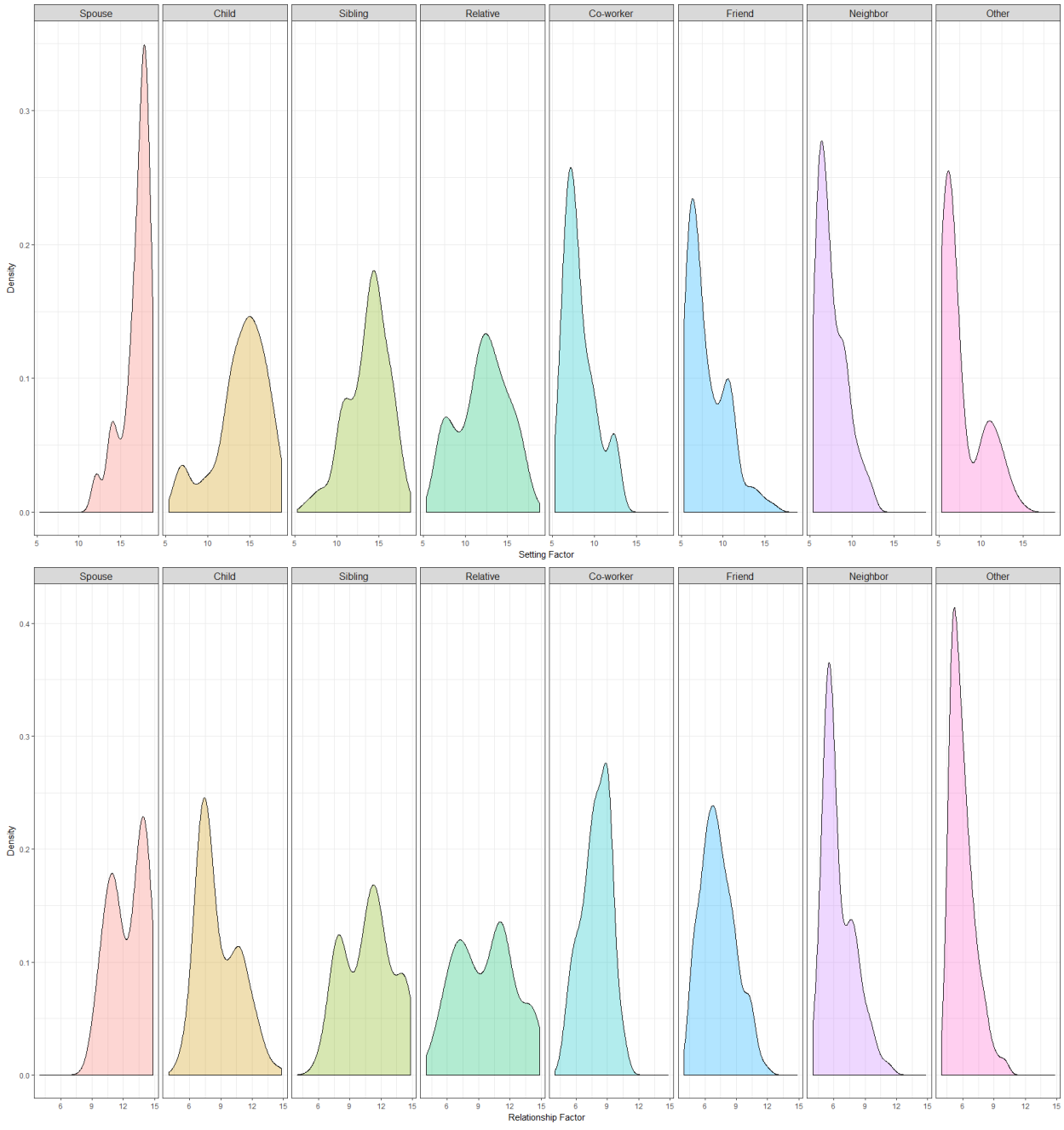


Figure 4.3. Distribution of the Setting and Relationship factors according to the nature of the relationship between a tuberculosis case and their contacts.

A histogram is shown to study the distribution of the Setting and Relationship factor scores, according to the nature of relationship between tuberculosis case and contact. Top Panel: Setting Factor. Bottom Panel: Relationship Factor.

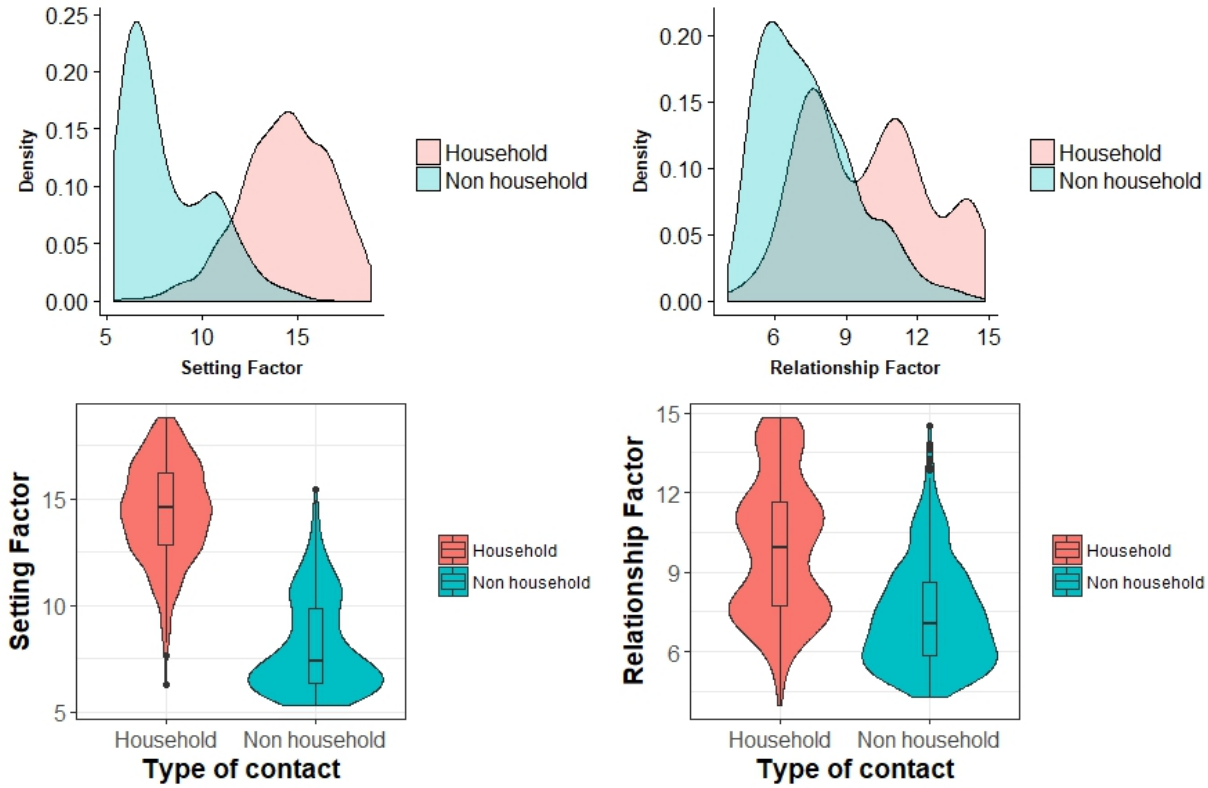


Figure 4.4. Distribution of the Setting and Relationship factors among household and non-household contacts.

A histogram and a violin plot are presented to describe the distribution of the Setting and Relationship Factor Scores among household and non-household contacts of the tuberculosis case. The values on the X-axis of the histogram indicates the score of each factor. The Y-axis indicates the relative frequency (density) at each Factor score. Inside each violin plot a box plot is presented. Left panel: Histogram (top) and violin plot (bottom) for the Setting Factor. Right panel: Histogram (top) and violin plot (bottom) for the Relationship Factor.

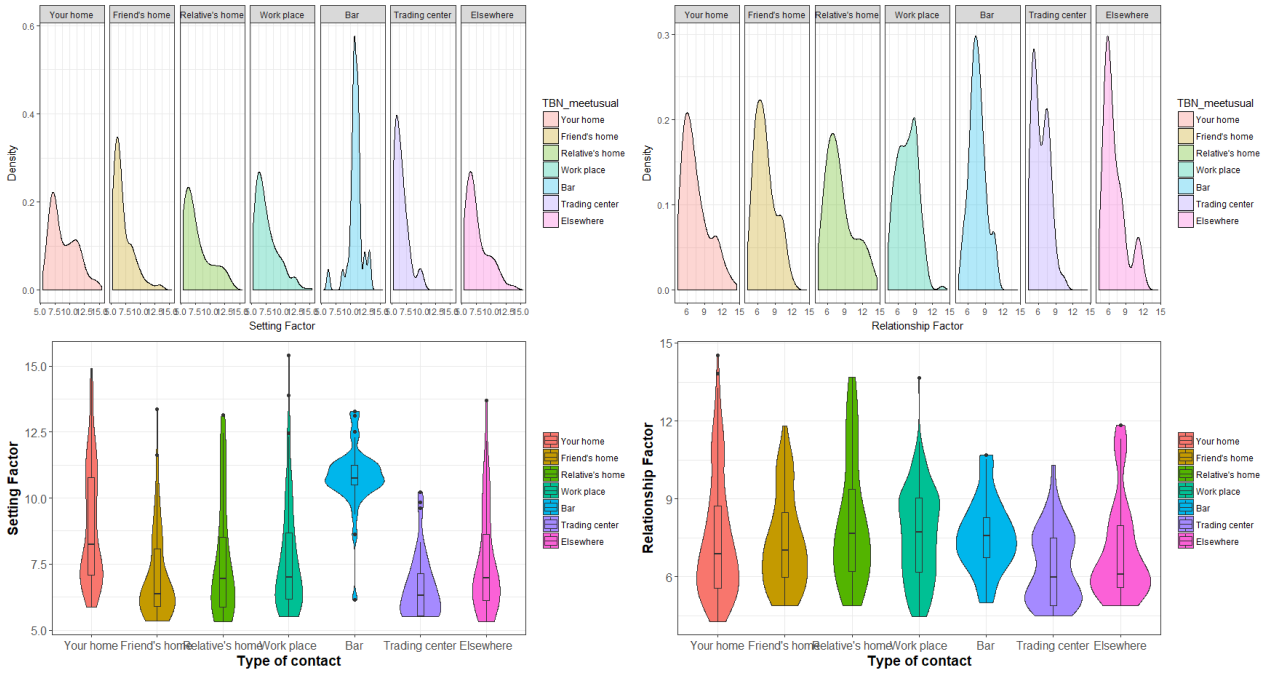


Figure 4.5. Distribution of the Setting and Relationship factors according to the usual place of meeting among non-household contacts.

A histogram and a violin plot are presented to describe the distribution of the Setting and Relationship Factor Scores among non-household contacts of the tuberculosis case. The values on the X-axis of the histogram indicates the score of each factor. The Y-axis indicates the relative frequency (density) at each Factor score. Inside each violin plot a box plot is presented. Left panel: Histogram (top) and violin plot (bottom) for the Setting Factor. Right panel: Histogram (top) and violin plot (bottom) for the Relationship Factor.

SUPPLEMENTARY MATERIAL

Table S 4.1 Social network form conducted among 120 tuberculosis cases.

Question Number	Field Name	Brief Description	Response Option
1	nature	Nature of relationship with contact now	01=Spouse 02=Child 03=Sibling 04=Friend 05=Stranger 06=Co-worker 07=Student colleague 08=Relative 09=Acquaintance 88=Other(specify)
1 Others (specify)	rlnoth	Other specified relationship	Fill in
2	newcont	Is contact new contact?	01=Yes 02=No 77=Don't Remember
3	lgthknow	How long have you known contact	Fill in
3	lgthtme	Unit of duration of knowing contact	01=Days 02=Weeks 03=Months 04=Years
4	lgcont	Duration of having name as contact	Fill in
4	ltrel	Units of duration of having name as contact	01=Days 02=Weeks 03=Months 04=Years
5	bcough	Did you know name before you started to cough	01=Yes 02=No 77=Don't remember
6	rlnchge	Has nature of relationship with contact change since the onset of cough	01=Yes 02=No
7	rln	Nature of relationship before the change	01=Spouse 02=Co-worker 03=Student colleague 04=Relative 05=Friend 06=Acquaintance 88=Other(specify)
7 Others (Specify)	pstrln	Other specified past relationship	Fill in
8	meetfrq	Frequency of contact since the onset of cough?	01=Increased 02=Decreased 03=Remained the same
9	well	How well does the informant know his contact?	01=Very Well 02=Moderately well 03=Somewhat well 04=Not well

Question Number	Field Name	Brief Description	Response Option
			05=Almost do not know him/her
10	discuss	Does case discuss important life issues with contact?	01=Yes 02=No 66=No response
11	confide	Does case confide with control?	01=Yes 02=No 66=No Response
12	sharedtb	Shared TB diagnosis with contact?	01=Yes 02=No
13	frqcont	Frequency of contact over the past 1 month	01=None 02=Less than a day/week 03=1-3 days/week 04=4-6 days/week 05=Daily 77=Don't recall
14	meettime	Duration of usual contact over the last one month	01=Just a short time (Less than or equal to 1hr/day) 02=Part of the day (2-6 hrs./day) 03=Part of the day (7-12 hrs./day) 04=Most of the day(13-18hrs/day) 05=Over 18 hrs/day 77=I don't recall 99=Not Applicable
15a	meals	Shared meals with contact since the onset of cough?	01=Yes 02=No 77=Don't Remember
15b	mealfrq	Frequency of sharing meals with contact	01=< 1 day/week 02=1-3 days/week 03=4-6 days/week 04=Daily 77=Don't recall 99=Not Applicable
16a	sleep	Slept in same room with contact since the onset of cough?	01=Yes 02=No
16b	sleepfrq	Frequency of sleeping in the same room?	01=< 1 day/week 02=1-3 days/week 03=4-6 days/week 04=Daily 77=Don't recall 99=Not Applicable
17a	bed	Slept on the same bed with contact since the onset of cough?	01=Yes 02=No
17b	bedfrq	Frequency of sleeping on same bed	01=< 1 day/week 02=1-3 days/week 03=4-6 days/week 04=Daily 77=Don't recall 99=Not Applicable
18a	care	Provided care by the contact in the past 3 months	01=Yes 02=No
18b	carefrq	Frequency of care provided by contact in the past three months	01=Less than a day/week 02=1-3 days/week 03=4-6 days/week 04=Daily 77=Don't recall 99=Not Applicable

Question Number	Field Name	Brief Description	Response Option
19	meetplc	Do you have a usual meeting place with contact?	01=Yes 02=No
20 Others (specify)	meetusual	Place of usual meeting	01=Your home 02=Friend's home 03=Relative's home 04=Work place 05=School 06=Worship center 07=Club/Association 08=Bar 09=Saloon 10=Gym 11=Trading center/Shop/Kiosk 12=In transit(specify) 88=Elsewhere(specify)
20 Others (specify)	intraspec	Specified usual transit meeting location	Fill in
20	meetoth	Specified other Meeting place with contact	Fill in
21	locdays	Frequency of meeting per week since the onset of work	01=None 02=Less than a day/week 03=1-3 days/week 04=4-6 days/week 05=Daily 77=Don't recall
22	loctime	Usual duration of meeting since the onset of cough	01=Just a short time (Less than or equal to 1hr/day) 02=Part of the day (2-6 hrs./day) 03=Part of the day (7-12 hrs./day) 04=Most of the day(13-18hrs/day) 05=Over 18 hrs/day 77=I don't recall
23 Others (specify)	mostmeet	Most recent meeting location with contact	1=Your home 2=Friend's home 3=Relative's home 4=Work place 5=School 6=Worship center 7=Club/Association 8=Bar 9=Saloon 10=Gym 11=Trading center/Shop/Kiosk 12=In transit (specify) 88=Elsewhere (specify)
23 Others (specify)	transpec	Other specified usual transit of meeting	

Question Number	Field Name	Brief Description	Response Option
23	mostoth	Other location of most recent meeting	Fill in
24	timespent	Time spent in the most recent meeting with contact	01=Just a short time (Less than or equal to 1hr/day) 02=Part of the day (2-6 hrs./day) 03=Part of the day (7-12 hrs./day) 04=Most of the day(13-18hrs/day) 05=Over 18 hrs/day 77=I don't recall
25	meet	Meet contact indoors or outdoors?	01=Mostly Indoors 02=Mostly Outdoors 03=Equally inside and outside
26	vent	Nature of ventilation at usual meeting place	01=POOR: Completely enclosed place (All windows and doors are closed) 02=MINIMAL: Partially enclosed (Some windows and/or doors closed) 03=FAIR: Structure has a roof, enclosed in four walls with large opening typical of a retail shop 04=FULL: Completely outdoor, under a tree, under a roof supported by poles 77=Don't know
27	pple	Meet other people in addition to contact at usual location?	01=Yes 02=No
28	pplenum	Number of other people met in addition to contact	Fill in
29	locoth	Other location of meeting	01=Yes 02=No
30	othloc	Frequency of meeting at the other location	01=None 02=< 1 day/week 03=1-3 days/week 04=4-6 days/week 05=Daily 77=Don't recall
31	trans	Means of transport used together with contact since the onset of cough	01=Motor bike 02=Bodaboda 03=Private vehicle 04=Taxi 05=Lorry 06=Bus 07=Train 08=Plane 09=Boat 10=None 11=Others(specify)

Question Number	Field Name	Brief Description	Response Option
31 Others (specify)	transoth	Other specified means of transportation	Fill in
32	means	Means of transportation used most often with contact	01=Motor bike 02=Bodaboda 03=Private vehicle 04=Taxi 05=Lorry 06=Bus 07=Train 08=Plane 09=Boat 10=None 88=Others(specify)
32 Others (specify)	meansoth	Other specified means of transportation used most often with contact	Fill in
33	cough	Does contact have a cough?	01=Yes 02=No 77=Don't know
34	tb	Does contact have a TB?	01=Yes 02=No 77=Don't know

Source: Codebook from Social network form questionnaire, COHSONET study.

Table S 4.2 List of variables included in the factor Analysis.

Description, variable name and final recoding.

Number	Brief Description	Variable name	Final recoding
1	How long have you known contact? (years)	known	1= Up to 2 years 2= 2-4 years 3= 4-10 years 4= More than 10 years
2	Frequency of contact since the onset of cough?	meetingfrequency	**1**='Decreased' **2**='Same frequency' **3**='Increased';
3	How well does the informant know his contact?	Well	**1**='Not well/almost do not know' **2**='Somewhat well' **3**='Moderately well' **4**='Very well';
4	Does case discuss and confide important life issues with contact?	Trust	**1**='No discuss nor confide' **2**='Discuss but not confide' **3**='Discuss and confide';
5	Shared TB diagnosis with contact?	sharedtb	1=No 2=Yes
6	Frequency and duration of contact over the past 1 month	Hourscat	**1**=' hours <= 3.5/week' **2**='hours 3.5-28 hrs/week' **3**='hours >28-66.5/week' **4**='>66.5 hrs/week/week';
7	Frequency of sharing meals with contact	Meals	**1**='Not shared meals' **2**='Shared meals, less than a day per week' **3**='Shared meals 1-3 days/week' **4**='Shared meals 4-6/week' **5**='Shared meals daily';
8	Frequency of sleeping in the same room and bed?	Bed	**1**='No slept same room nor bed' **2**='Slept same room, but not same bed' **3**='Slept same room and same bed, not daily' **4**='Slept same room and same bed, daily';
9	Frequency of care provided by contact in the past three months	care	**1**='No care by contact' **2**='Provided care, less than a day per week' **3**='Provided care 1-3 days/week' **4**='Provided care 4-6/week' **5**='Provided care daily';
10	Place of usual meeting with contact	Place	1=Not the house of TB case 2=House TB case . =Missing
11	Meet contact indoors or outdoors?	Meet	**1**='Mostly meeting outdoors' **2**='Equally indoors/outdoors' **3**='Mostly meeting indoors';

Number	Brief Description	Variable name	Final recoding
12	Nature of ventilation at usual meeting place	vent	**1**='Full ventilation' **2**='Fair ventilation' **3**='Minimal ventilation' **4**='Poor ventilation';
13	Number of other people met in addition to contact at usual location.	Peoplecat	**1**='< 2 persons/meeting' **2**='2-4 persons/meeting' **3**='5-6 persons/meeting' **4**='>6 persons/meeting';
14	Means of transportation used most often with contact	transportation	**1**='None/walking' **2**='Another type of transportation';
15	Does contact have a cough?	cough	1=No/Don't know 2=Yes

Table S 4.3. List of variables excluded from the Factor Analysis.

Description, variable name, final recoding and Reason to be excluded.

Number	Brief Description	Recoded name	Final recoding	Reason excluded
1	Does contact have a TB?	tb	1=No/Don't know 2=Yes	Low variation (98.05% no, 1.95% yes)
2	Did you know name before you started to cough	bcough	1=No/Don't remember 2=Yes	Low variation (97.20% no, 2.80% yes)
3	Is this contact a new contact?	TBN_newcont		Low variation (98.64% no, 1.36% yes)
4	Has nature of relationship with contact change	rlnchge	01=Yes 02=No	Low variation (99.32% no, 0.17% yes, 0.51% missing)
5	Nature of relationship before the change	rln	01=Spouse 02=Co-worker 03=Student colleague 04=Relative 05=Friend 06=Acquaintance 88=Other(specify)	Low variation (99.83% missing, because answer to previous question). Also, categorical variable.
6	How long have you have this person as contact (years)	contacttime	Continuous variables (in years)	Response equal to answer for 'How long have you known contact (years)?'-in its continuous variable form
7	Nature of relationship with contact now	TBN_nature	01=Spouse 02=Child 03=Sibling 04=Friend 05=Stranger 06=Co-worker 07=Student colleague 08=Relative 09=Acquaintance 10=Neighbor 88=Other(specify)	Categorical variable
8	Most recent meeting location with contact	TBN_mostmeet	**1**="Your home" **2**="Friend's home" **3**="Relative's home" **4**="Work place" **5**="School" **6**="Worship center" **7**="Club/Association/Bar/Saloon/Gym" **8**="Bar" **9**="Saloon" **10**="Gym" **11**="Trading center/Shop/Kiosk" **12**="In transit" **13**="Neigbourhood" **88**="Elsewhere(specify)" **.** ="Missing"	Correlation: 0.91556 with TBN_most usual variable, which was the original question for the recoded "Place" variable.

Number	Brief Description	Recoded name	Final recoding	Reason excluded
9	Time spent in the most recent meeting with contact	TBN_timespent	01=Just a short time (Less than or equal to 1hr/day) 02=Part of the day (2-6 hrs./day) 03=Part of the day (7-12 hrs./day) 04=Most of the day(13-18hrs/day) 05=Over 18 hrs/day . =Missing	Correlation: 0.92574 with Hourscat
10	Other location of meeting	TBN_locoth	01=Yes 02=No	Not informative enough.
11	Frequency of meeting at the other location	TBN_othloc	01=None 02=< 1 day/week 03=1-3 days/week 04=4-6 days/week 05=Daily 77=Don't recall	
12	Means of transport used together with contact since the onset of cough	TBN_Trans	01=Motor bike 02=Bodaboda 03=Private vehicle 04=Taxi 05=Lorry 06=Bus 07=Train 08=Plane 09=Boat 10=None 11=Others(specify)	Not informative enough.
13 and 14*	Frequency and duration of meeting per week since the onset of work	Hoursloccat *This variable represented two questions in the original questionnaire. 1) Frequency and 2) Duration.	**1**=' hours <= 3.5/week' **2**='hours 3.5-28 hrs/week' **3**='hours >28-66.5/week' **4**='>66.5 hrs/week/week';	Correlation: 0.97153 with Hourscat

Table S 4.4 Item analysis questionnaire social network form for the social contacts with complete social network data (n=1,157) and the contacts traced in the study that provided demographic data (n=923).

Included categorical variables and recoded variables considered for the factor analysis.

Variable	1157 contacts with complete social network data		923 traced contacts with other variables collected	
	n	Percent	n	Percent
Nature of relationship with tuberculosis case[1]				
Spouse	34	2.9	25	2.7
Child	140	12.1	122	13.2
Sibling	152	13.1	115	12.5
Friend	343	29.6	270	29.3
Co-workers	86	7.4	67	7.3
Oher relative	216	18.7	173	18.7
Neighbor	100	8.6	78	8.5
Other	86	7.4	73	7.9
Length knowing the contact				
Less than 2 years	429	37.1	351	38.0
2-4 years	206	17.8	163	17.7
5-6 years	236	20.4	197	21.3
More than 6 years	286	24.7	212	23.0
Frequency of meeting since onset cough				
Decreased	63	5.4	44	4.8
Same frequency	1031	89.1	835	90.5
Increased	63	5.4	44	4.8
Hours spent per week with contact				
Less than 3.5 hours/week	335	29.0	265	28.7
Between 3.5-28 hours/week	461	39.8	365	39.5
Between 28-66.5 hours/week	272	23.5	210	22.8
Greater 66.5 hours/week	89	7.7	83	9.0
Location of usual meeting[2] (detailed responses)				
Home of tuberculosis case	645	55.7	514	55.7
Friend's home	66	5.7	54	5.9
Relative's home	73	6.3	55	6.0
Work place	202	17.5	162	17.6
Bar	40	3.5	34	3.7
Trading center/Shop/Kiosk	54	4.7	44	4.8
Elsewhere	77	6.7	60	6.5
Location of usual meeting[2] (binary for EFA)				
Outside home of tuberculosis case	512	44.3	409	44.3
Home of tuberculosis case	645	55.7	514	55.7
Ventilation place of meeting				
Full ventilation	557	48.1	454	49.2
Fair ventilation	219	18.9	169	18.3
Minimal ventilation	183	15.8	144	15.6
Poor ventilation	198	17.1	156	16.9
Indoor or outdoor meeting				
Mostly meeting outdoors	552	47.7	447	48.4

Variable	1157 contacts with complete social network data		923 traced contacts with other variables collected	
	n	Percent	n	Percent
Equally indoors/outdoors	290	25.1	239	25.9
Mostly meeting indoors	315	27.2	237	25.7
Number of other people met in addition to contact				
< 2 persons/meeting	348	30.1	272	29.5
2-4 persons/meeting	435	37.6	342	37.1
5-6 persons/meeting	255	22.0	213	23.1
>6 persons/meeting	119	10.3	96	10.4
Sleeping conditions				
No slept in same room, nor bed	906	78.3	723	78.3
Slept same room, but not same bed	169	14.6	135	14.6
Slept same room and same bed, not daily	21	1.8	16	1.7
Slept same room and same bed, daily	61	5.3	49	5.3
Meals				
Not shared meals	448	38.7	361	39.1
Shared meals, less than a day per week	106	9.2	75	8.1
Shared meals 1-3 days/week	175	15.1	133	14.4
Shared meals 4-6 days/week	66	5.7	58	6.3
Shared meals daily	362	31.3	296	32.1
Case trusts contact				
No discuss nor confide	480	41.5	395	42.8
Discuss but not confide	365	31.5	287	31.1
Discuss and confide	312	27.0	241	26.1
Shared TB diagnosis				
No	643	55.6	521	56.4
Yes	514	44.4	402	43.6
Care by contact				
No care by contact	973	84.1	775	84.0
Care provided, less than a day per week	42	3.6	33	3.6
Provided care 1-3 days/week	55	4.8	38	4.1
Provided care 4-6 days/week	16	1.4	14	1.5
Provided care daily	71	6.1	63	6.8
How well does the case knows contact				
Not well/almost do not know	18	1.6	16	1.7
Somewhat well	159	13.7	129	14.0
Moderately well	269	23.2	214	23.2
Very well	711	61.5	564	61.1
Means of transportation used most often with contact. None (walking) versus a type of transportation.				
None/walking	928	80.2	748	81.0
Another type of transportation	229	19.8	175	19.0
Known if contact has cough				
No	1085	93.8	858	93.0
Yes	72	6.2	65	7.0

[1]The variable ‘‘Nature of the relationship between case and contact’ (spouse, child, sibling, friend, co-worker, relative, neighbor, other) was excluded for EFA but the descriptive analysis is reported.

[2]The categorical variable ‘Location of usual place of meeting’ (Home case, friend’s home, relative’s home, work place, bar, trading center/shop/kiosk, elsewhere) was recoded as a binary variable (Home case, outside home of tuberculosis), and included in the EFA as a dichotomous variable but the descriptive analysis of the original variables is reported here.

Table S 4.5. Sensitivity Analysis: Factor loadings matrix identified by exploratory factor analysis when two factors were retained.

Factor loadings matrix identified by exploratory factor analysis when two factors were retained, using transformed data.

Variable	Factor1 (Setting)	Factor2 (Relationship)
Nature of ventilation at usual place of meeting	0.83323	-0.08066
Frequency of sleeping in same room and bed since onset cough	0.56751	0.09319
Contact happen indoors or outdoors	0.85320	-0.02788
Frequency of shared meals since onset cough	0.74714	0.14831
Place of usual meeting. Home TB case versus Other location.	0.53683	-0.00104
Frequency and duration of contact over the past month	0.59217	0.09052
Case trusts contact	-0.12287	0.87498
Case shared TB diagnosis with contact	-0.08806	0.78001
Case was provided care by the contact in the past 3 months	0.19715	0.48338
Length of knowing contact	0.23812	0.52420
How well does the case knows contact	0.37283	0.37718
Means of transportation used most often with contact. None (walking) versus a type of transportation.	0.09355	0.32840
Frequency since cough has increased/decreased	0.11295	0.11988
Additional people in usual place of meeting	0.22414	0.00264

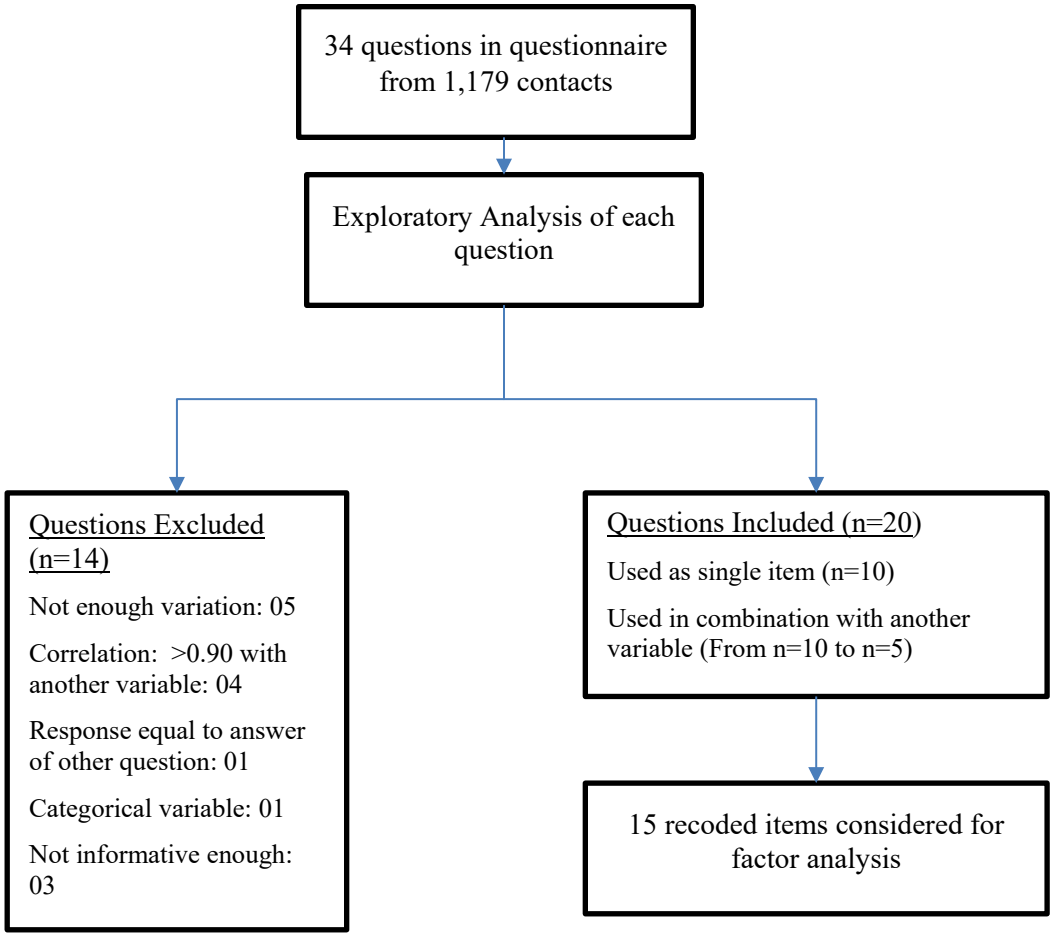


Figure S 4.1. Flow diagram of inclusion criteria for items to be included in the exploratory factor analysis.

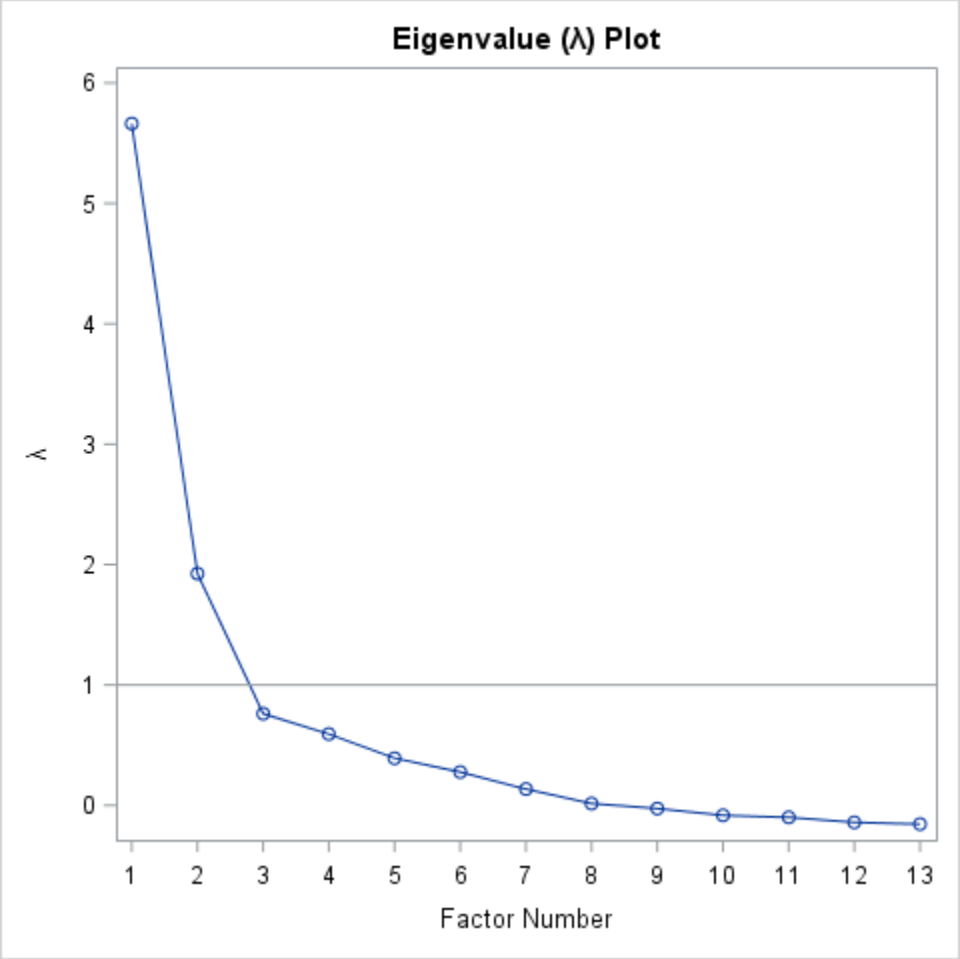


Figure S 4.2. Eigenvalues of thirteen components extracted during factor analysis.
Factors with an eigenvalue ≥ 1 were retained in the model.

Chapter 5

PREVALENCE OF TUBERCULOSIS AMONG CONTACTS OF TUBERCULOSIS: INSIGHTS FROM A CONTACT SCORE. [6]

[6] Castellanos ME, Ebell M, Dobbin KK, Quinn F and Whalen CC. To be submitted to AJE.

Abstract

Background. Exposure to an individual with tuberculosis is necessary for transmission to occur. Previously we developed a score that measures contact between tuberculosis cases and their social networks in an African urban context. This score was built using exploratory factor analysis and identified contact as the conjunction of two domains-Setting and Relationship. Now, our aim is to determine whether this score covaries with the presence of tuberculosis infection among social contacts of tuberculosis cases.

Design/Methods. This was a large cross-sectional study conducted in Kampala, Uganda from 2012-2016. Latent (measured by tuberculin skin test) or active tuberculosis infection was assessed in the social contacts of adult tuberculosis cases. We estimated the prevalence of tuberculosis infection in this population, overall and according to the Setting and Relationship domains, confounders and covariates. We calculated the prevalence ratio for the association between increasing scores in the Setting and Relationship domains and tuberculosis infection, adjusted by other covariates, using a modified Poisson regression model. We further evaluated our scores by exploring the association of these with the proportion of tuberculosis infection among the household and non-household contacts after categorization these scores into three categories according to tertiles.

Results. We enrolled 923 social contacts from 119 tuberculosis cases. The overall prevalence of tuberculosis infection in the social networks was 51% (95% CI 48-54). We found an association of the Setting and Relationship domains with the prevalence of tuberculosis in contacts, with this association being modified by the age of the contact. The effect of the Setting score in the prevalence of tuberculosis was higher among children between 5-14 years (PR=1.24, 95% CI 1.13-1.37) whereas the Relationship score was strongly associated with tuberculosis infection in children of 0-4 years (PR=1.19, 95% CI 1.01-1.41). The prevalence of tuberculosis among non-household contacts of tuberculosis cases was 75% (95% CI 45-100) when part of the high Setting tertile and 68% (95% CI 52-84) when part of the high Relationship tertile.

Conclusion. In this large cross-sectional study from an urban African setting, the Setting and Relationship domains affected the likelihood of infection with *M. tuberculosis* for members of a social network of a tuberculosis case, particularly children contacts.

Introduction

The incidence rate of tuberculosis can be expressed with three factors: contact rate between individuals; transmission probability "*p*", which is the probability that "a contact between an infectious individual and a susceptible host leads to a successful transmission event"; and "P", the probability that an individual contact occurs between a susceptible individual and an infectious individual. Usually, "P" is assumed to be equivalent to the prevalence, i.e. the fraction of infectious individuals in the total population at a given time (Real & Biek, 2007; Thomas & Weber, 2001).

The prevalence of tuberculosis, although challenging, can be measured using several methods, most notably surveillance data from tuberculosis programs or by population-based surveys (Glaziou et al., 2008). Case reports and estimation of secondary attack rates in high endemic areas have shown that the probability of transmission given adequate contact seems to cover a wide spectrum(Kenyon et al., 1996; C. C. Whalen et al., 2011). However, the third component that defines incidence, contact rate, is still poorly understood. Several studies have highlighted the heterogeneity and complexity of the social contact patterns among human populations (Dodd et al., 2016; Mossong et al., 2008; Wallinga et al., 2006). Nevertheless, the quantification of adequate contact between a tuberculosis case and their social network has not been performed in African settings. In a previous work, we developed a contact score that assessed the level of contact between a tuberculosis and their contacts (Chapter 4 of this dissertation).

This score was built using exploratory factor analysis and was based on the information provided by the index case about the nature of their interactions. We did not use any information regarding the tuberculosis status of the contact to developed it. We were able to show that this score has construct validity and that it explains more accurately these relationships that a simple household versus non-household categories. Now, our aim is to determine whether this score covaries with the presence of tuberculosis infection among these contacts.

Study Population and Methods

STUDY POPULATIONS

Index cases. The tuberculosis cases for this Chapter were the same than the ones presented in Chapter 4. Briefly, the index cases were persons aged 15 years or older, who are residents of Kampala, Uganda; who had signs and symptoms of pulmonary tuberculosis and that were microbiologically confirmed by a positive sputum smear. More details of the cases characteristics have been already provided (Chapter 4).

Contacts. The index case or index control provided a list of their contacts as previously described in Chapter 4. These contacts were traced and if enrolled them, demographic and clinical information were collected from them. The sub-set of contacts that were traced and enrolled are the contact population for this Chapter.

EXPOSURE OF INTEREST

The study exposure was a contact score between the tuberculosis case and their contacts. This contact score was composed by two domains. The details of the development of this score are described in Chapter 4 of this dissertation. Briefly, index cases answered questions related to the social mixing between them and each of their social networks. A factor analysis was conducted among these variables, which identified two main domains. One that we called “Setting” as it comprised six variables: a) nature of ventilation at usual place of meeting, b) frequency of sleeping in same room/same bed, c) contact happen indoors or outdoors, d) frequency of shared meals since onset of cough, d) place of usual meeting-home versus other location and e) frequency and duration of contact over the past month. The second domain was called “Relationship”, and six variables were included in it: a) Case trust contact, b) Case shared TB diagnosis with contact, c) Case was provided care by the contact in the past 3 months, d) Length of knowing contact, e) How well does the case knows the contact, f) Means of transportation used most often with contact-none/walking versus a type of transportation. Factor analyses results provided

weights to each of the original responses of the social network data. We used the weighted sum scores method to generate domain scores for each interaction case-contact. This method allows that the items with the highest loadings have the highest impact in our factor score (DiStefano, Zhu, & Mindrila, 2009). We multiplied the factor loading of each item to their original. We excluded items with loadings below 0.30 (Beavers et al., 2013). We summed the weighted answers of the six variables included in each domain. We obtained a Setting and a Relationship scores for each interaction case-contact. We have previously shown these domain scores reliably measure the extent and nature of the contact between an infectious case and susceptible contact. Thus, higher scores correspond with more extensive contact and therefore we hypothesize a higher probability of a transmission event.

OUTCOME OF INTEREST

The study outcome was tuberculosis infection in the contacts. The infection in the contact can be latent or active (disease). Latent tuberculosis infection was estimated using the tuberculin skin test (TST). For our study, a positive TST result was defined as induration ≥ 10 millimeters as it has shown to be an adequate cut-off in the Ugandan setting (Martinez et al., 2016). Intradermal injection of five TU of purified protein derivate was applied in study contacts. After 48-72 hours of the injection, two trained Ugandan technicians using digital calipers, independently measured the induration using digital calipers and recorded as continuous data. The mean of two measurements was used for the estimation of latent tuberculosis infection (C. C. Whalen, 2014).

Active tuberculosis was defined either as: a) The presence of at least one smear positive for acid fast bacilli, b) Positive culture for *Mycobacterium tuberculosis*, c) A positive molecular result for *Mycobacterium tuberculosis* in a contact with clinical symptoms consistent with tuberculosis disease and d) History of previous tuberculosis disease, informed by the social contact.

POTENTIAL CONFONDERS AND COVARIATES

In this analysis, we controlled for known covariates that might modify the risk of tuberculosis infection, as well for any potential confounder that might mask the true association between degree of contact and tuberculosis infection. First, to understand the association between the Setting and Relationship domains and tuberculosis infection the following directed acyclic graph (DAG) was proposed (Figure 5.1). This DAG model hypothesizes that these domains-Setting and Relationship-(exposure) are causally associated with the risk of having a tuberculosis infection (outcome). Based on literature review, the potential confounders of this association are age and sex of contact, age and sex of index case and HIV infection status of the contact (positive or negative) (Dheda et al., 2016; Dodd et al., 2016; Feenstra et al., 2013; Mossong et al., 2008; Narasimhan et al., 2013) (Kizza et al., 2015). A group of potential independent factors were variables related to the social network that were not captured in our factor analysis. These covariates were: TB case knew contact before onset of cough (yes or no), TB case knows contact has tuberculosis infection (yes or no), TB case knows contact ha cough (yes or no), frequency of meeting between contact and case (decreased, increased, stayed the same) and additional people present in usual meeting between contact and TB case. Smear grade of index case (0-1+=low smear grade, 2-3+=high smear grade) is related directly to infectivity of the TB case, so will also be included as a covariate (Narasimhan et al., 2013). Finally, BCG vaccine status in the contact (yes, no, unknown) can induce a false positive tuberculin skin test (Nayak & Acharjya, 2012) so it was be also evaluated.

ANALYTICAL STRATEGY.

An item and exploratory analysis were conducted in the data set to explore the distribution of the exposure variables, covariates and the main outcome (tuberculosis infection). For continuous variables, median values and interquartile ranges were estimated and for categorical variables, proportions with 95% confidence intervals. In addition, visual exploration was performed- using bar plots, histograms and boxplots- depending of the type of variable. We performed Kendall and polychoric correlation to check

correlation among the variables (Knight, 1966; Olsson, 1979). Baseline characteristics of the enrolled contacts were summarized with proportions and measures of central tendency.

We conducted bivariate analysis to explore the relation between each covariate and the exposure variables, and each covariate and the outcome, separately. We initially used Chi-square test (categorical variables) or Wilconxon test (continuous variables) to explore those associations. We also explored the probability of tuberculosis infection against the Setting and Relationship scores, using a loess (locally weighted scatterplot smoothing) model, to obtain a nonparametric smoothed curve (M Friendly, 2015; Michael Friendly & Meyer, 2015).

We estimated the prevalence of contacts with tuberculosis infection with 95% confidence intervals, overall and according to the exposure, confounders and covariates. As the prevalence of tuberculosis infection in this population was over 50%, two regression models were considered as alternatives to logistic regression to obtain a more precise estimate of the association between exposure and the outcome (Coutinho et al., 2008): a Poisson regression with a robust variance and a log-binomial regression (Wacholder, 1986; Zou, 2004). Unadjusted prevalence ratios were obtained with both methods by exponentiating the coefficients. Similar analyses were done to check the association between potential confounders, covariates with the outcome. As the analyses showed similar results, the stratified and the adjusted prevalence ratios later described were exclusively calculated with the modified Poisson, as it has been shown to be more robust to outliers compared to the log-binomial model (Chen et al., 2014).

We calculated the prevalence ratio for the association between increasing scores in the Setting and Relationship domains and tuberculosis infection, stratified by the other covariates. This stratified analysis informed our regression model, in which we were able to control for multiple covariates.

For model building, a DAG gold-standard change-in-estimate procedure was followed with some modifications (Weng et al., 2009). The full model included all potential confounders (described in our DAG), and the independent factors that were shown to be associated with the outcome in our bivariate

analyses. For model reduction, the following procedure was followed. Our crude model solely included the exposure (Setting and Relationship scores). The final model was created by adding, one at a time, a new variable. The decision to keep a reduced model or to include a new variable was based in the following criteria, in this order: a) Change in the prevalence ratio of the exposure, b) Variable considered an effect modifier based on the exploratory and stratified analysis, c) Quasi-likelihood information criterion (QIC), in which the model with the smallest QIC being preferred (Pan, 2001) and d) having a parsimonious model. After the model with the main effect was constructed, interactions between each of the explanatory variables and the exposure were assessed. Age of contact was found to be an interaction term to both Relationship and Setting scores, so interaction terms were added in the final model. In the final model age of contact was categorized in three brackets: 0-4 years, 5-14 years and 15 years and greater to facilitate the interpretation of the interaction terms. We present the results of both the full model and the reduced model.

Creation of domain categories and association with prevalence of tuberculosis infection among household and non-household contacts. We further evaluated our scores by exploring the association of these scores with the proportion of tuberculosis infection among the household and non-household contacts. We used a stratified random sampling to split the data into training data (67%) and test data (33%), using as our strata the household versus non-household contact group. The partition of the data into 2/3 for training has been shown to be usually robust (Dobbin & Simon, 2011). In the training data, we categorized the Setting and Relationship scores in three categories according to tertiles: Low tertile, Medium tertile and High tertile. We calculated the prevalence of tuberculosis infection among contacts (with 95% confidence intervals) in each tertile for the overall, household and non-household population. The change in prevalence of tuberculosis infection by these tertiles was analyzed by the Cochran–Armitage trend test. This analysis was repeated in the test data, using the same values obtained in the training data to define the low, medium and high tertiles.

Sensitivity analysis. In a sensitivity analysis, we changed the criteria of TST positivity, considering the HIV status of the contact (≥ 5 millimeters for HIV-infected individuals) and we excluded contacts with history of tuberculosis disease (Supplementary material, section A).

Association individual variables and prevalence of tuberculosis infection. Finally, we explored the association of the 12 individual variables that comprised the Setting and Relationship factors with the prevalence of tuberculosis infection in the social contacts of tuberculosis cases. The prevalence of tuberculosis infection according to each of the responses is shown with 95% confidence intervals. Results were also estimated by each age category of the contacts. Unadjusted prevalence ratios were obtained by the Poisson regression with a robust variance (Supplementary material, section B).

All analyses were carried out using SAS software v 9.4 (SAS Institute, Cary, NC, US) and R v3.3.1 (R Foundation for Statistical Computing, Vienna, Austria, 2016).

ETHICAL APPROVAL

Written informed consent was obtained from all participants prior to study inclusion. Institutional review board clearance was obtained from Ethics Committee at Makerere University School of Public Health and the University of Georgia. Tuberculin converters were referred for evaluation of tuberculosis by study medical personnel. If tuberculosis was suspected, the participant was referred to the National Tuberculosis Control Program for treatment, otherwise they were offered isoniazid treatment by the study personnel.

Results

Descriptive characteristics of contacts. Contacts of 120 index cases were invited to participate in the study. Both male and female index cases had a median of nine contacts (Range: 4-19 in men and 6-18 in women) for a total of 1,179 contacts. Of these, 962 (82%) were traced and agreed to be enrolled in the study (Figure 5.2). Complete data regarding exposure variables-Setting and Relationship scores- and the

outcome variable-presence of tuberculosis infection-were obtained for 923 of the 962 contacts (96%) out of 119 index cases.

Contacts were similarly distributed in terms of gender (51% women and 49% men). The median age was 23 years (IQR, 13, 32) and 11% of the contacts were children under 5 years of age (Table 5.1). Almost two thirds of the contacts were non-household contacts (62%). Sex assortment differed among sexes. Female contacts had a female index TB case in just 32% of the cases as opposed to male contacts, which had a male index case in 73% of the cases.

Crude and adjusted prevalence ratios of tuberculosis infection among social contacts. The overall prevalence of tuberculosis infection in the social networks was 51% (95% CI 48-54). Data visualization suggested that both Setting, and Relationship scores were positively associated with the presence of tuberculosis infections in the contacts. Contacts with a low Relationship score showed prevalence as low as 25% and individuals with high Relationship score had prevalence as high as 75% (Figure 5.3).

In bivariate analysis, for each unit increase in the Setting and Relationship scores, the prevalence of tuberculosis infection increased by 5% (PR=1.05, 95% CI 1.03-1.07) and 7% (PR=1.07, 95% CI 1.04-1.09) respectively. Other variables positively associated with the risk of tuberculosis infection in the contacts were increasing age, being a household contact and knowledge by the index case that the contact has tuberculosis or cough (Table 5.2). On the contrary, friends, co-workers, neighbors, and other type of distant relationships had lower risks of having tuberculosis infection when compared to the spouse of an index case.

Stratified prevalence of tuberculosis infection among social contacts according to Domain Scores. The association between the Setting and Relationship scores and tuberculosis infection were roughly similar to the crude ones after stratifying by gender of the contact, gender of index, age of index, HIV status of the contact, BCG vaccine (Table 5.3, Figure 5.4 and Figure 5.5). Age of contact was considered an effect modifier as the prevalence ratio varied among different age categories. The effect of the Setting score in the prevalence of tuberculosis was higher among children between 5-14 years (PR=1.26, 95% CI

1.16-1.36) and lower in the older bracket (1.04, 95% CI 1.02-1.06) (Figure 5.4). The Relationship score was strongly associated with tuberculosis infection in children of 0-4 years (PR=1.33, 95% CI 1.19-1.49) and its effect was reduced in individuals older than 15 years (PR=1.05, 95% CI 1.03-1.08) (Figure 5.5).

Adjusted prevalence ratio of tuberculosis infection among social contacts according to Domain Scores. After adjustment by confounders and independent predictors, the Setting score continued to have an impact in the prevalence of tuberculosis infection in contacts. This association was more pronounced in the 5-14 years bracket with a prevalence ratio of 1.24 (95% CI 1.12-1.37) for the reduced regression model and 1.25 (95% CI 1.13-1.39) in the full regression model (Table 5.4 & Figure 5.6). Relationship score was associated with tuberculosis infection exclusively in contacts between 0-4 years (PR=1.22, 95%CI 1.02-1.45 in the reduced model and PR=1.23, 95% CI 1.02-1.49 in the full model).

Construction of tertiles and their relationship with prevalence of tuberculosis infection, stratified by household and non-household contacts. We split the data (n=619) into training and test data (n=304). In the training data, the values of the tertiles for the Setting score were <7.70 for the low tertile, 7.70-12.39 for the medium tertile and 12.43-18.58 for the high tertile. Regarding the Relationship score the values of the tertiles were: <6.92 for the low tertile, 6.92-9.07 for the medium tertile and 9.07-14.80 for the high tertile.

In the training data, the prevalence of tuberculosis infection in individuals in the low tertile for the Setting domain was 46% (95% CI 39-53), roughly similar in the medium tertile (42%, 95% CI 35-49), but lower than the one found in the high tertile (66%, 95% CI 60-73) (Figure 5.7, top panel, left plot & Table 5.5). After stratifying by household and community contacts, this pattern continued. Only one household contact was included in the low Setting tertile, being negative to tuberculosis infection. Household contacts in the medium and high Setting tertile had prevalence of tuberculosis infection of 40% (95% CI 26-54) and 67% (95% CI 60-74) respectively (Figure 5.7, top panel, middle plot). Among non-household contacts, the prevalence of tuberculosis was 46% (95% CI 39-53), 43 (95% CI 35-50) and

58 (95% CI 36-80) for the low, medium and high Setting tertile respectively (Figure 5.7, top panel, right plot).

The results in the test data also showed a higher prevalence of tuberculosis in the high tertile group (Figure 5.7, top panel). In this data, six out of the eight non-household contacts in the high Setting tertile (75%) were infected with tuberculosis (Table 5.5).

In the training data, the prevalence of tuberculosis infection in individuals in the low Relationship tertile was 44% (95% CI 37-50), in the medium tertile was 50% (95% CI 43-57) and in the high tertile was 61%, (95% CI 54-68) (Figure 5.7, bottom panel, left plot). Among household contacts, 27 individuals were in the low tertile and 48% of them were infected (95% CI 29-67%). In the medium tertile group the prevalence was 57% (95% CI 46-69) and in the high tertile group the prevalence was 67% (95% CI 59-75) (Figure 5.7, bottom panel, middle plot). Among non-household contacts, the low tertile Relationship group had a prevalence of tuberculosis infection of 43% (95% CI 36-50) roughly like the one in the medium tertile (45%, 95% CI 37-54) but lower than the one found in the high tertile (51%, 95% CI 40-63) (Figure 5.7, bottom panel, right plot).

Analysis of the test data produced results similarly to the ones of the training data, but in the non-household contacts there was a more pronounced trend of higher the prevalence as higher the tertile (34% low, 45% medium and 68% high, p value for trend=0.0013) (Table 5.5 & Figure 5.7, bottom panel, right plot).

In our sensitivity analysis (Supplementary material, section A)-in which we restricted the analysis to individuals with a numerical results for TST and with a modified definition of TST positivity according to HIV status-, the main results regarding the association of Setting and Relationship factors remained unchanged (Table S 5.1, Table S 5.2 & Table S 5.3). The twelve individual variables that comprised the Setting and Relationship scores showed to different degrees, association with the prevalence of tuberculosis infection among the social contacts of index cases (Supplementary material, section B; Figure S 5.1 & Figure S 5.2).

Discussion

In this large cross-sectional study from a urban African setting, we were able to determine that a score based on two orthogonal domains from a factor analysis was able to assess adequate contact to infectious tuberculosis cases as measured by the probability of tuberculosis infection among their social contacts. Moreover, there was effect modification by age as the Relationship score was only relevant among young children whereas the Setting score associated with infection overall and among contacts younger than 15 years.

Of the two domains that describe the nature of the relationship among a tuberculosis case and his/her contact, the Setting domain proved an important marker for tuberculosis infection, especially among 5-14 years old bracket, in which the prevalence of tuberculosis increased 1.24 times for each increasing unit of Setting score.

Children living with tuberculosis cases are at a higher risk of acquiring tuberculosis as opposed to children living in households without a tuberculosis case (Martinez et al., 2017). Moreover, transmission of tuberculosis in children in Uganda seems to be caused primarily by household transmission (Wampande et al., 2015). Our findings now can further explain the reason that in some households with tuberculosis cases, not all children are infected. There are two domains that affect the likelihood of infection with *Mycobacterium tuberculosis* for members of a social network of an index case. The effect of these two domains varies by age, so they may help explain why some close contacts, especially children, may become infected whereas others do not.

The three variables that provide stronger contribution the Setting domain were the ventilation of the usual place of meeting, the frequency of sleeping with tuberculosis case (room and or/bed) and if the meetings usually happened indoor our outdoors. Previously, these variables have been identified to be associated with risk of tuberculosis infection among contacts of tuberculosis cases (Lienhardt et al., 2003; Nardell, 2016; Rieder, 2001). For instance, it is known that higher ventilation rates decreased quanta concentration, which in turn reduces the number of new incident cases (Beggs, Noakes, Sleigh, Fletcher,

& Siddiqi, 2003). Thus, the inclusion of these variables in our Setting domain supports its strong content validity.

Accordingly, individuals spending significant amounts of time with tuberculosis cases, sharing meals and/or sharing room or bed in indoor-poor ventilated settings were the ones with the higher Setting scores and the ones with the highest prevalence of tuberculosis. By combining these individual variables in one single Setting domain, we were able to capture more precisely the different exposure levels of a contact, furthermore that the typical classification of household and non-household contacts. Our score works both in household and non-household contacts. For instance, in the few community contacts within the high Setting tertile the prevalence of tuberculosis was higher (75%), than among those within the low Setting tertile (41%).

There was an association between the Relationship score and tuberculosis infection in the very young children (0-4 years old) after adjustment by other variables, including the Setting domain. Our findings showed that probably this Relationship domain is reflecting those cases where there is a very intimate relationship (caretaker, mother, older sibling) as opposed to those cases in where there is not such a close relationship (other relative, friend of the family). Contacts who were very well known by the cases had higher prevalence of tuberculosis than contacts that were lesser known. Although BCG vaccine has been shown to be an important confounder to this association, most of our participants were vaccinated so that does not seem to be a concern in our findings.

There were other variables evaluating the nature of the relationship between an index case and a contact, which we asked to the index case that were not included in our Setting and Relationship domains. Only one of them seems to be a relevant independent factor associated with tuberculosis infection among contacts, which is knowledge by the index case of the tuberculosis status of the contact. In our main analysis, this variable was associated with the presence of tuberculosis infection among the contacts. This was not the case for our sensitivity analysis, which only included contacts with latent tuberculosis infection. In other words, index cases are aware of the presence or history of tuberculosis disease among

their contacts. Thus, this simple question to the index case should then be considered highly reliable to find contacts with past and present tuberculosis disease.

The main strength of our study is that we were able to adequately quantify the nature of the relationship between a tuberculosis case and a contact, and then to associate it with the presence of tuberculosis infection in the contacts. Previous studies have used similar methods (Acuna-Villaorduna et al., 2018; Mandalakas et al., 2012). Mandalakas used principal component analysis to measure exposure to tuberculosis and the risk of tuberculosis infection and disease among child household contacts (Mandalakas et al., 2012). Acuna described a modified version of this methodology among adult household contacts in Brazil (Acuna-Villaorduna et al., 2018). Both found an association between the score and infection. Our approach refined and improved upon the approach taken by these studies. The earlier studies were done in household contacts only. We expanded the evaluation to include social network members of an infectious index cases. We used factor analysis in order to identify two domains that may affect infection. With these two domains we were able to partition risk of infection and show how the setting of exposure and the relationship to the index case varied with age. Finally, we used an agnostic approach toward the development of the scores, as we did not try to develop a particular construct, such as proximity for infection, but instead used all variables that describe the nature of contact within the social network.

Our study has some limitations. First, we included contacts with latent tuberculosis, active tuberculosis and history of tuberculosis disease. As this is a cross-sectional study, we cannot know the directionality of the infection in the last two categories. Nevertheless, the sensitivity analysis conducted in the sub-set of contacts with latent tuberculosis infection yielded similar results regarding the effect size of both domains. Second, the variables that originated the Setting and Relationship domains were answered by the tuberculosis case, so recall bias is possible, but it will more likely result in non-differential misclassification (Grimes & Schulz, 2002). Third, it has been seen that social mixing and social behavior is highly context-dependent (Mossong et al., 2008), so our findings might not be applicable in other

settings. Nevertheless, our study contributed to two contexts, the household and non-household contexts of a high-endemic area. We did not have an external dataset, but the split sample approach facilitates the support of the consistency of our scores. Lastly, this score as it is, will likely have limited prognostic value to predict risk for recent tuberculosis transmission. This score was developed to quantitively measure one of the components that drives tuberculosis transmission, adequate contact. However, to estimate accurately the risk for recent tuberculosis transmission we would have to include other variables such as infectivity of the case (i.e. smear grade) or behavioral characteristics of the social contacts that might increase their risk for tuberculosis.

In conclusion, we proved that the Setting and Relationship domains affected the likelihood of infection with *M. tuberculosis* for members of a social network of a tuberculosis case, particularly children contacts. In our work only a small fraction of non-household contacts had high Setting and Relationship scores, signaling a need to further understand the structures and dynamics of the community social networks of tuberculosis cases.

TABLES

Table 5.1. Characteristics of the enrolled contacts of 119 tuberculosis cases from Kampala, Uganda that answered the social network survey.

Characteristic	N	%
Contacts enrolled	962	100
Contacts enrolled and with complete data	923	96
Gender		
Male	456	49
Female	467	51
Type of contact		
Household	349	38
Non-household	574	62
Age, years, median [IQR]	23 [13-32]	
Age (category)		
0-4	101	11
5-14	137	15
15-greater	685	74
Residence		
Lives in Rugaba	889	96
Do not live in Rugaba	33	4
No information available	01	0
HIV Result		
Positive	69	8
Negative	842	91
Refusals/Too young/missing	12	1
BCG vaccine		
Yes (verbal report/immunization card)	791	86
No	79	9
Don't' know	50	5
No information available	3	0
TST Result (TST available)		
TST >=10 mm	403	47
TST < 10 mm	455	53
Tuberculosis infection		
Yes	468	51
TST Result >=10 mm	403	
Previous PPD positive	14	
Previous history of TB	51	
No (TST Result <10 mm)	455	49

Table 5.2. Prevalence and crude prevalence ratio (95% CI) for tuberculosis infection among social contacts of tuberculosis cases by selected potential risk factors.

Category	N	Prevalence tuberculosis infection		Prevalence ratio (Poisson)	Prevalence ratio (log-binomial)
		N	% (95% CI)		
Overall	923	468	51 (48-54)		
Setting score	923			1.05 (1.03-1.07)	1.05 (1.03-1.07)
Relationship score	923			1.07 (1.04-1.09)	1.07 (1.04-1.09)
Gender contact					
Female	467	224	48 (43-52)	1	1
Male	456	244	54 (49-58)	1.08 (0.97-1.21)	1.12 (0.98-1.27)
Gender index case					
Female	274	132	48 (42-54)	1	1
Male	649	336	52 (48-56)	1.07 (0.87-1.32)	1.07 (0.87-1.33)
Sex assortment					
No	438	209	48 (43-59)	1	1
Yes	485	259	53 (49-58)	1.08 (0.97-1.20)	1.12 (0.98-1.28)
Age of contact (continuous)	923			1.01 (1.00-1.01)	1.01 (1.01-1.02)
Age of contact					
0-4	101	42	42 (32-51)	1	1
5-14	137	59	43 (35-51)	1.02 (0.77-1.35)	1.04 (0.75-1.42)
15-greater	685	367	54 (50-57)	1.27 (1.02-1.57)	1.29 (1.02-1.63)
Age of index case (continuous)	923			0.99 (0.98-1.00)	0.99 (0.98-1.00)
Age of index case					
15-24	288	163	57 (51-62)	1	1
25-44	512	244	48 (43-52)	0.85 (0.70-1.03)	0.84 (0.68-1.04)
45 or more	123	61	50 (41-58)	0.85 (0.66-1.09)	0.88 (0.67-1.14)
Type of contact					
Non-household	574	258	45 (41-49)	1	1
Household	349	210	60 (55-65)	1.35 (1.17-1.55)	1.34 (1.15-1.56)
HIV					
No	842	425	50 (47-54)	1	1
Yes	69	37	54 (42-65)	1.01 (0.77-1.33)	1.06 (0.81-1.39)

Category	N	Prevalence tuberculosis infection		Prevalence ratio (Poisson)	Prevalence ratio (log-binomial)
		N	% (95% CI)		
BCG					
No	79	36	46 (35-57)	1	1
Yes	791	404	51(48-55)	1.06 (0.85-1.34)	1.12 (0.88-1.44)
Don't know	50	28	56 (42-70)	1.12 (0.81-1.55)	1.23 (0.87-1.75)
Nature of relationship with tuberculosis case					
Spouse	25	18	72 (54-89)	1	1
Child	122	65	53 (44-61)	0.77 (0.57-1.04)	0.74 (0.54-1.01)
Sibling	115	78	68 (59-76)	0.91 (0.67-1.22)	0.94 (0.70-1.27)
Friend	270	136	50 (44-56)	0.71 (0.54-0.95)	0.70 (0.52-0.93)
Co-workers	67	36	54 (42-66)	0.69 (0.48-0.98)	0.72 (0.53-0.99)
Oher relative	173	90	52 (44-59)	0.77 (0.57-1.04)	0.72 (0.53-0.99)
Neighbor	78	18	23 (14-32)	0.32 (0.20-0.53)	0.32 (0.19-0.53)
Other	27	73	37 (26-48)	0.54 (0.37-0.78)	0.51 (0.35-0.75)
Known contact before cough					
No	25	9	36 (17-55)	1	1
Yes	897	458	51 (48-54)	1.27 (0.83-1.95)	1.42 (0.90-2.25)
Known if contact has cough					
No	858	418	49 (45-52)	1	1
Yes	65	50	77 (67-87)	1.54 (1.28-1.86)	1.58 (1.32-1.89)
Knows if contact has tuberculosis					
No	902	449	50 (46-53)	1	1
Yes	21	19	90 (78-100)	1.60 (1.34-1.90)	1.82 (1.55-2.13)
Frequency of meeting since onset cough					
Decreased	44	18	41 (26-55)	1	1
Same frequency	835	424	51 (47-54)	1.17 (0.77-1.78)	1.24 (0.80-1.91)
Increased	44	26	59 (44-74)	1.39 (0.84-2.29)	1.44 (0.87-2.40)
Number of other people met in addition to contact (continuous)				1.01 (0.98-1.04)	1.01 (0.98-1.04)

Category	N	Prevalence tuberculosis infection		Prevalence ratio (Poisson)	Prevalence ratio (log-binomial)
		N	% (95% CI)		
Number of other people met in addition to contact (categorical)					
<2 persons/meeting	272	137	50 (44-56)	1	1
2-4 persons/meeting	342	169	49 (44-55)	1.01 (0.85-1.21)	0.98 (0.82-1.18)
5-6 persons/meeting	213	105	49 (43-56)	1.05 (0.85-1.30)	0.98 (0.76-1.26)
>6 persons/meeting	96	57	59 (50-69)	1.11 (0.88-1.41)	1.18 (0.91-1.53)
Microscopy Status					
0-1+	122	48	39 (31-48)	1	1
2-3+	772	400	52 (48-55)	1.24 (0.94-1.63)	1.32 (0.98-1.77)

Table 5.3. Prevalence ratio for the association between increasing scores in the Setting and Relationship scores and tuberculosis infection among social contacts of tuberculosis cases.

Overall and stratified analysis by selected key variables.

Variable	N	Prevalence ratio Setting Score	Prevalence ratio Relationship Score
Overall		**1.05 (1.03-1.07)**	**1.07 (1.04-1.09)**
By Gender contact			
Female	467	1.07 (1.04-1.09)	1.07 (1.04-1.11)
Male	456	1.04 (1.02-1.06)	1.06 (1.04-1.09)
Gender index case			
Female	274	1.09 (1.06-1.12)	1.09 (1.05-1.12)
Male	649	1.04 (1.02-1.07)	1.06 (1.03-1.09)
Sex assortment			
No	438	1.06 (1.03-1.09)	1.07 (1.03-1.10)
Yes	485	1.05 (1.02-1.07)	1.07 (1.04-1.10)
Age of contact (continuous)	923	1.06 (1.04-1.08)	1.06 (1.03-1.08)
Age of contact			
0-4	101	1.10 (1.03-1.18)	1.33 (1.19-1.49)
5-14	137	1.26 (1.16-1.36)	1.10 (0.99-1.22)
15-greater	685	1.04 (1.02-1.06)	1.05 (1.03-1.08)
Age of index case (continuous)	923	1.05 (1.03-1.07)	1.07 (1.04-1.09)
Age of index case			
15-24	288	1.06 (1.04-1.09)	1.06 (1.03-1.10)
25-44	512	1.03 (1.00-1.06)	1.06 (1.02-1.10)
45 or more	123	1.11 (1.06-1.16)	1.11 (1.06-1.17)
Type of contact			
Non-household	574	1.05 (1.01-1.09)	1.05 (1.01-1.09)
Household	349	1.08 (1.02-1.15)	1.07 (1.03-1.10)
HIV status contact			
No	842	1.05 (1.03-1.07)	1.06 (1.04-1.09)
Yes	69	1.10 (1.04-1.16)	1.12 (1.05-1.20)
BCG contact			

Variable	N	Prevalence ratio Setting Score	Prevalence ratio Relationship Score
No	79	1.06 (0.99-1.14)	1.06 (0.99-1.14)
Yes	791	1.05 (1.03-1.07)	1.06 (1.03-1.09)
Don't know	50	1.10 (1.05-1.16)	1.14 (1.06-1.22)
Nature of relationship with tuberculosis case			
Spouse	25	1.06 (0.88-1.27)	1.02 (0.88-1.18)
Child	122	1.10 (1.02-1.18)	1.05 (0.98-1.12)
Sibling	115	1.02 (0.95-1.09)	0.96 (0.91-1.02)
Friend	270	1.04 (1.00-1.09)	1.07 (1.00-1.14)
Co-workers	67	0.99 (0.89-1.11)	1.00 (0.83-1.20)
Oher relative	173	1.06 (1.01-1.11)	1.08 (1.03-1.13)
Neighbor	78	0.75 (0.52-1.08)	0.90 (0.73-1.16)
Other	27	1.02 (0.90-1.61)	0.98 (0.74-1.29)
Known contact before cough			
No	25	1.07 (0.91-1.27)	1.18 (0.93.-1.51)
Yes	897	1.05 (1.03-1.07)	1.07 (1.04-1.09)
Known if contact has cough			
No	858	1.05 (1.03-1.07)	1.07 (1.04-1.09)
Yes	65	1.02 (0.99-1.04)	1.04 (0.98-1.10)
Knows if contact has tuberculosis			
No	902	1.05 (1.03-1.07)	1.07 (1.04-1.09)
Yes	21	0.94 (0.88-1.01)	0.96 (0.90-1.03)
Frequency of meeting since onset cough			
Decreased	44	1.01 (0.90-1.12)	1.07 (0.96-1.20)
Same frequency	835	1.05 (1.03-1.07)	1.07 (1.05-1.10)
Increased	44	1.11 (1.01-1.23)	1.04 (0.96-1.12)
Number of other people met in addition to contact	923	1.05 (1.03-1.07)	1.07 (1.04-1.09)

Variable	N	Prevalence ratio Setting Score	Prevalence ratio Relationship Score
Number of other people met in addition to contact (categorical)			
<2 persons/meeting	272	1.03 (1.00-1.05)	1.06 (1.02-1.09)
2-4 persons/meeting	342	1.05 (1.02-1.09)	1.07 (1.02-1.12)
5-6 persons/meeting	213	1.10 (1.05-1.17)	1.06 (1.02-1.11)
>6 persons/meeting	96	1.08 (1.02-1.16)	1.09 (1.00-1.18)
Microscopy Status			
0-1+	122	1.05 (1.00-1.11)	1.08 (1.02-1.15)
2-3+	772	1.05 (1.03-1.07)	1.07 (1.04-1.09)

Table 5.4. Adjusted prevalence ratio for the association between increasing scores in the Setting and Relationship domains and tuberculosis infection among social contacts of tuberculosis cases.

Overall results and stratified by age of contact.

Population, stratified by age of contact (years)	n	Prevalence ratio (95% CI) Reduced model		Prevalence ratio (95% CI) Full model	
		Setting	Relationship	Setting	Relationship
Overall		1.06 (1.03- 1.09)	1.00 (0.96- 1.03)	1.06 (1.04-1.09)	0.99 (0.96-1.02)
0-4	101	1.08 (1.00-1.18)	1.19 (1.01- 1.41)	1.09 (0.99- 1.19)	1.21 (1.01-1.44)
5-14	137	1.24 (1.13-1.37)	1.01 (0.92-1.10)	1.25 (1.13- 1.39)	1.00 (0.92 -1.10)
15-greater	685	1.03 (0.99-1.06)	1.02 (0.98-1.06)	1.03 (1.00- 1.07)	1.02 (0.97- 1.06)

Reduced model: Adjusted by age of contact, knowledge of tuberculosis status of the contact by the index, microscopy status index case and HIV status of contact.

Full model: Adjusted by age of contact, age of index, sex contact, sex index, HIV status of contact, microscopy status index, knowledge of tuberculosis status of the contact by the index case, knowledge of cough status of the contact by the index case, BCG vaccination of the contact

Table 5.5. Prevalence of Tuberculosis infection, according to Setting and Relationship scores categories.

Results are shown for the training (n=619) and test data (n=304), in the overall, household and non-household population.

Category	Score (tertiles)*	Setting score			Relationship score		
		n infected/ n group	Prevalence TB (95% CI)	P value†	n infected/N group	Prevalence TB (95% CI)	P value†
TRAINING DATA							
Overall (N=619)	Low	95/206	46 (39,53)	<.0001	90/206	44 (37,50)	0.0004
	Moderate	87/206	42 (35,49)		103/207	50 (43,57)	
	High	137/207	66 (60,73)		126/206	61 (54,68)	
Household	Low	0/1	0	0.0004	13/27	48 (29,67)	0.0436
(N=234)	Moderate	18/45	40 (26,54)		43/75	57 (46,69)	
	High	126/188	67 (60,74)		88/132	67 (59,75)	
Non-HH‡ (N=385)	Low	95/205	46 (39,53)	0.9246	77/179	43 (36,50)	0.2419
	Moderate	69/161	43 (35,50)		60/132	45 (37,54)	
	High	11/19	58 (36,80)		38/74	51(40,63)	
TEST DATA							
Overall (N=304)	Low	49/121	40 (32,49)	0.0027	39/107	36 (27,46)	0.0005
	Moderate	36/78	46 (35,57)		49/97	50 (40,60)	
	High	64/105	61 (52,70)		61/100	61 (51,71)	
Household	Low	0/1	0	0.1646	8/17	47 (23,71)	0.6299
(N=115)	Moderate	8/17	47 (23,71)		20/32	62 (45,80)	
	High	58/97	60 (50,70)		38/66	58 (46,70)	
Non-HH	Low	49/120	41 (32,50)	0.1124	31/90	34 (24,44)	0.0013
(N=189)	Moderate	28/61	46 (33,58)		29/65	45 (32,57)	
	High	6/8	75 (45,100)		23/34	68 (52,84)	

*****Values of the tertiles for the Setting score: <7.70 for the low tertile, 7.70-12.39 for the medium tertile and 12.43-18.58 for the high tertile. Values of the tertiles for the Relationship score: <6.92 for the low category, 6.92-9.07 for the medium tertile and 9.07-14.80 for the high tertile.

†Cochran-Armitage Trend Test

‡Non-HH=Non-household contacts

FIGURES

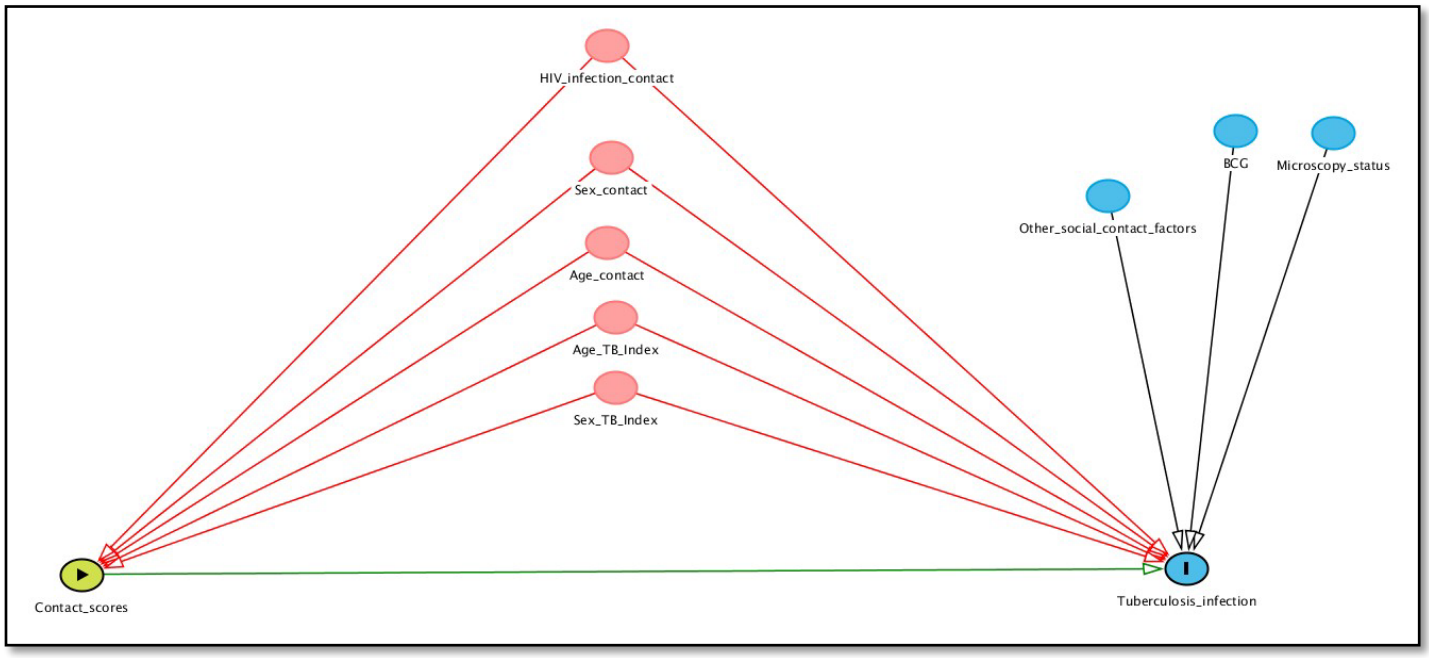


Figure 5.1. Conceptual model of the proposed causal relationship between the Setting and Relationship scores and tuberculosis infection, adjusted by potential confounders and covariates.

To understand the association between the contact domains (Setting and Relationship) and tuberculosis infection (TB) the following directed acyclic graph (DAG) was proposed. This DAG model hypothesizes that the Setting and Relationship scores (exposure) between a tuberculosis case and his/her contact is causally associated with the risk of having a tuberculosis infection (outcome). Potential confounders (pink) and independent factors associated with the outcome (blue) are also included.

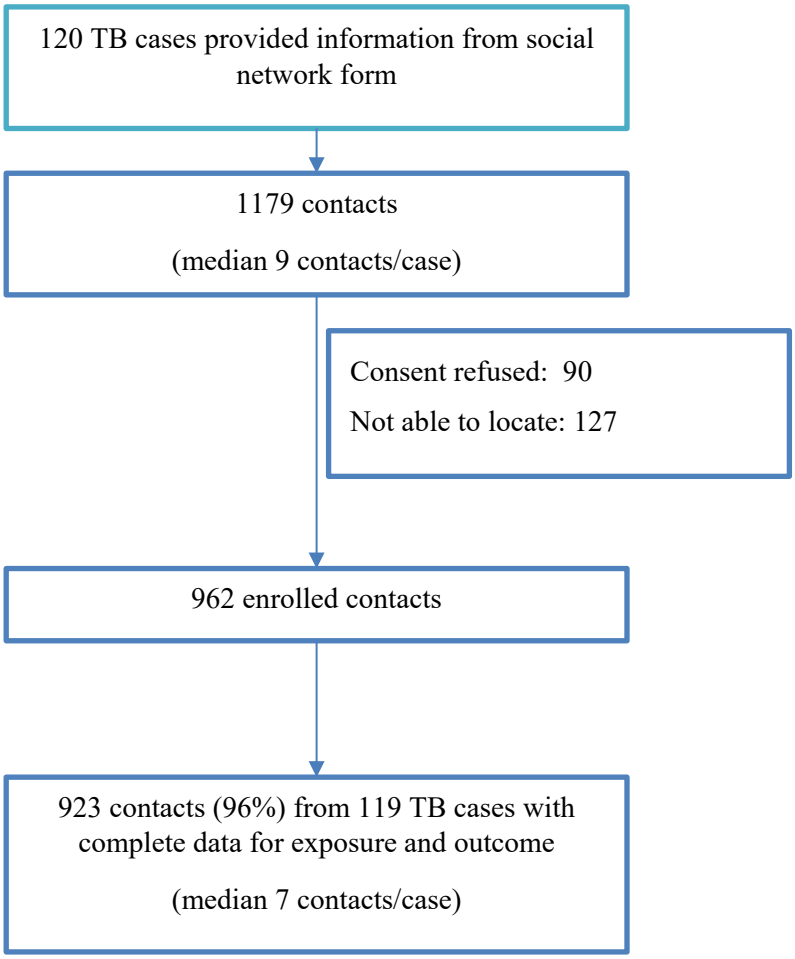


Figure 5.2. Flow diagram of study. 120 tuberculosis cases provided the information to create the Setting and Relationship domain scores for 1179 contacts.

A sub-sample of 923 contacts were evaluated regarding the association of these domains with the presence or absence of tuberculosis infection.

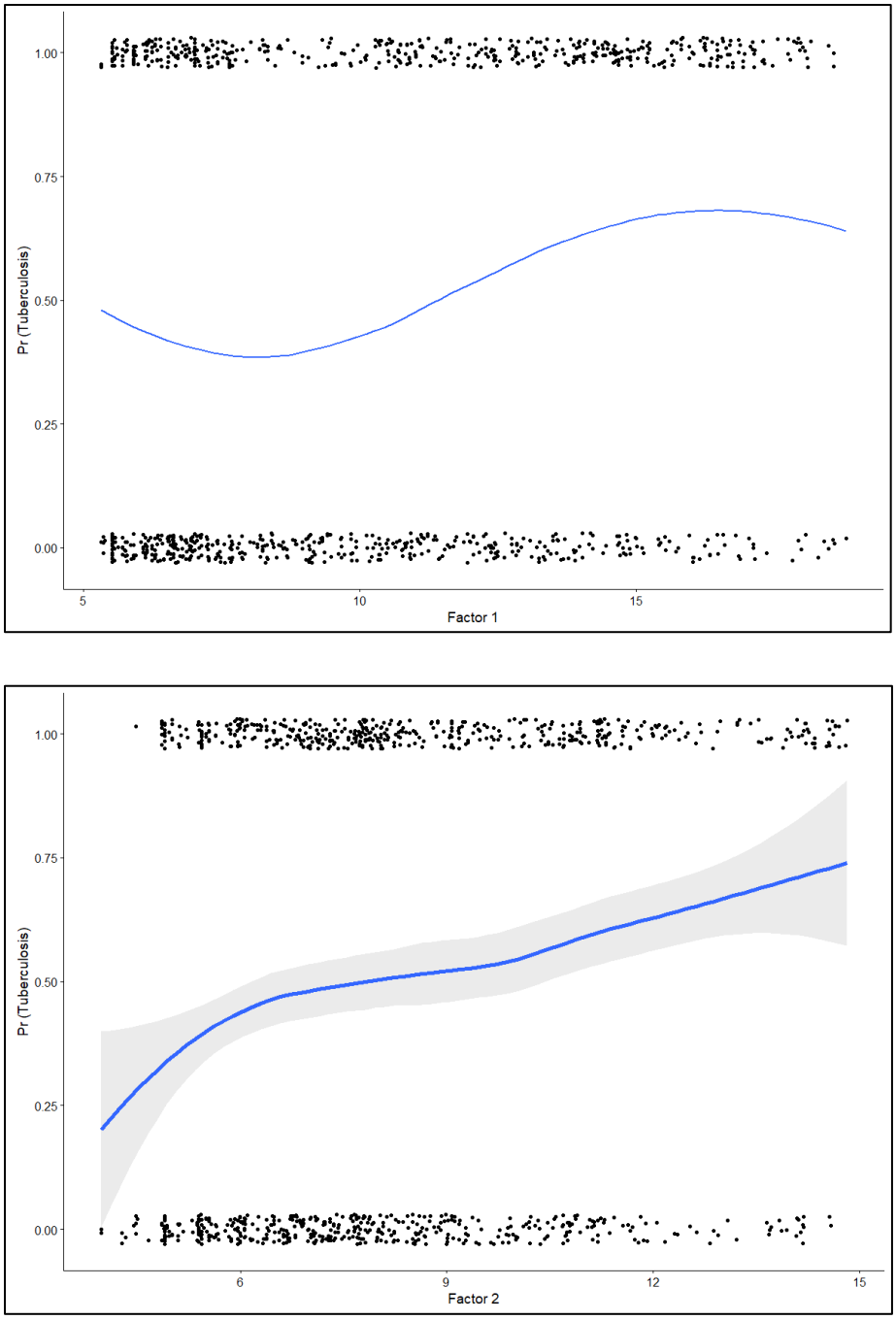


Figure 5.3. Probability of Tuberculosis infection, according to Setting and Relationship Scores.

Nonparametric smoothed curve showing the probability of tuberculosis infection against the Setting and Relationship scores, using a loess (locally weighted scatterplot smoothing) model. Setting Score (Top panel) and Relationship Score (Bottom panel).

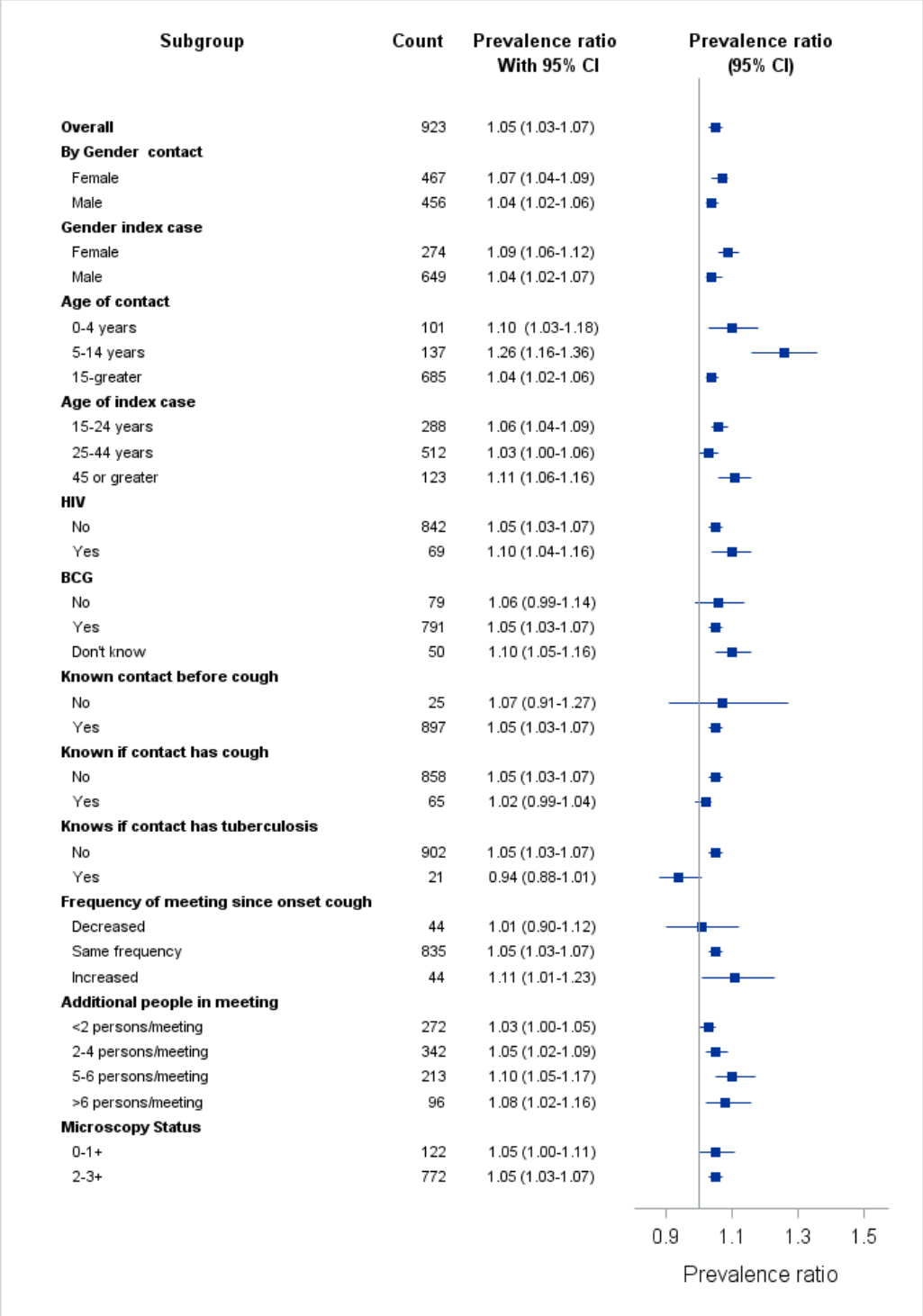


Figure 5.4. Prevalence ratio for the association between increasing scores in the Setting Score and tuberculosis infection.

Overall and stratified analysis by selected key variables. A prevalence ratio > 1 indicates that for each increasing unit of the Setting score, there is a higher prevalence of tuberculosis infection.

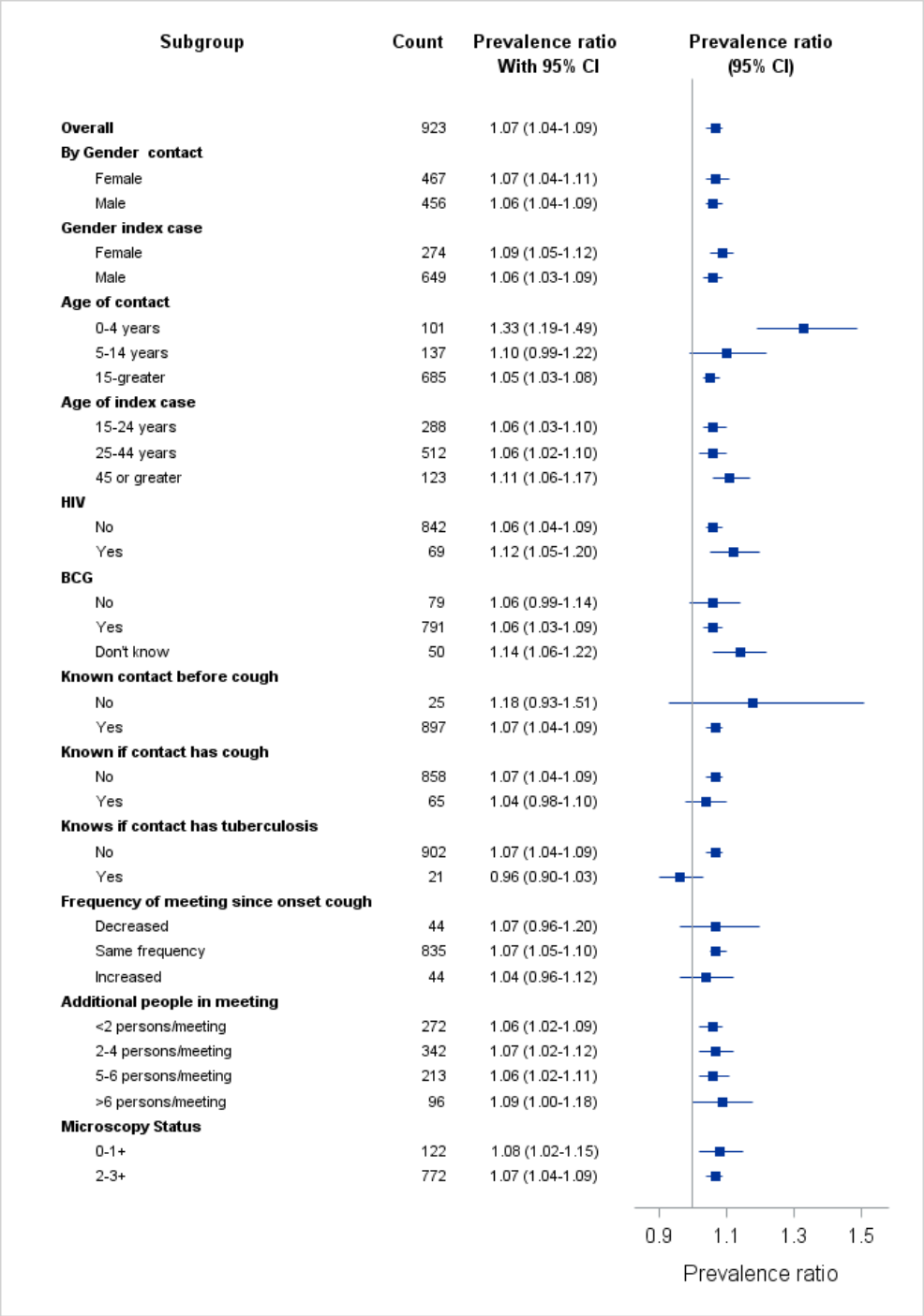


Figure 5.5. Prevalence ratio for the association between increasing scores in the Relationship factor and tuberculosis infection.

Overall and stratified analysis by selected key variables. A prevalence ratio > 1 indicates that for each increasing unit of the Relationship score, there is a higher prevalence of tuberculosis infection.

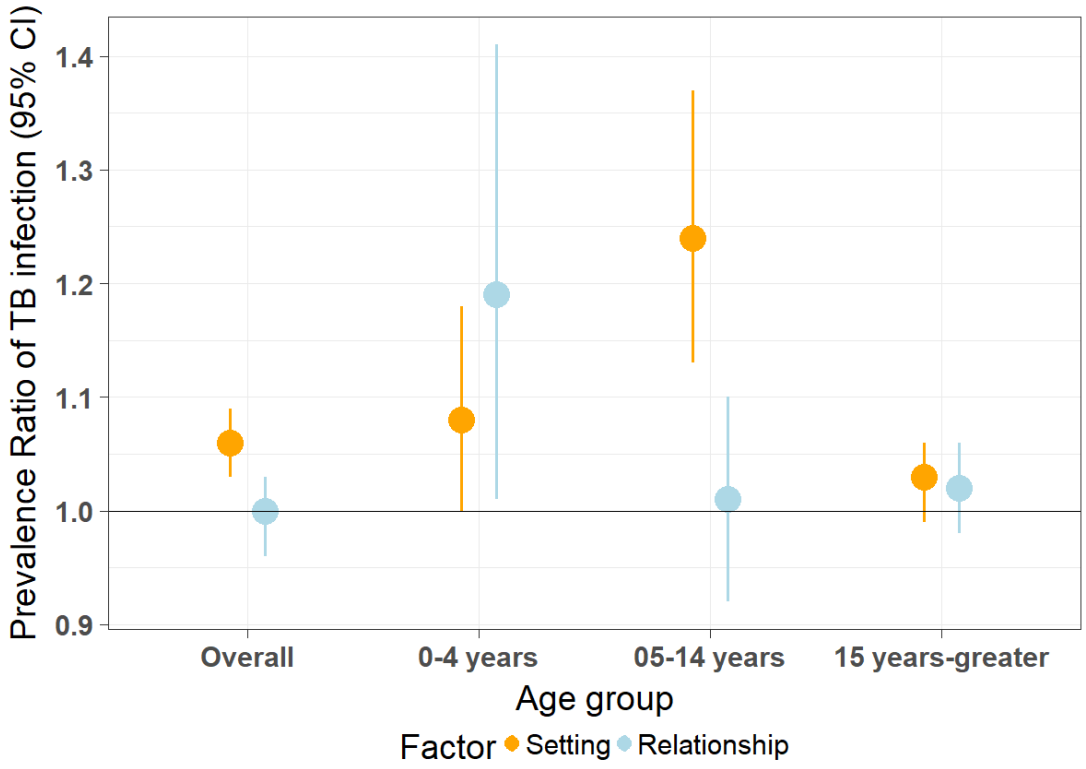
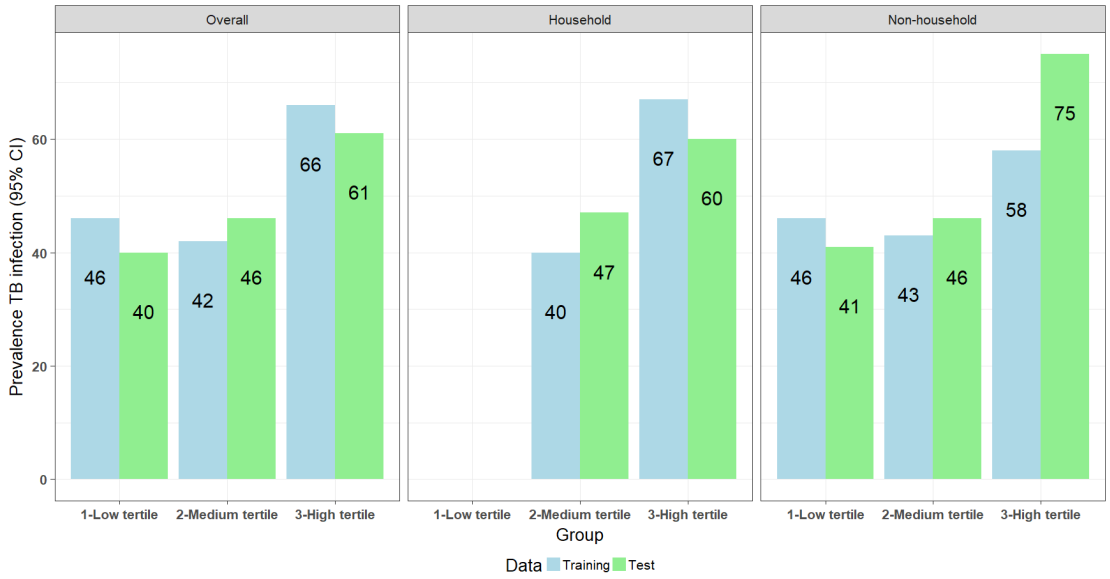


Figure 5.6. Adjusted prevalence ratio for the association between increasing scores in the Setting and Relationship scores and tuberculosis infection.

Overall and stratified by age of contact. An adjusted prevalence ratio > 1 indicates that for each increasing unit of the Setting or Relationship scores, there is a higher prevalence of tuberculosis infection, after adjustment by other covariates. Adjusted by age of contact, knowledge of tuberculosis status of the contact by the index, microscopy status index case and HIV status of contact.

Setting Domain



Relationship Domain

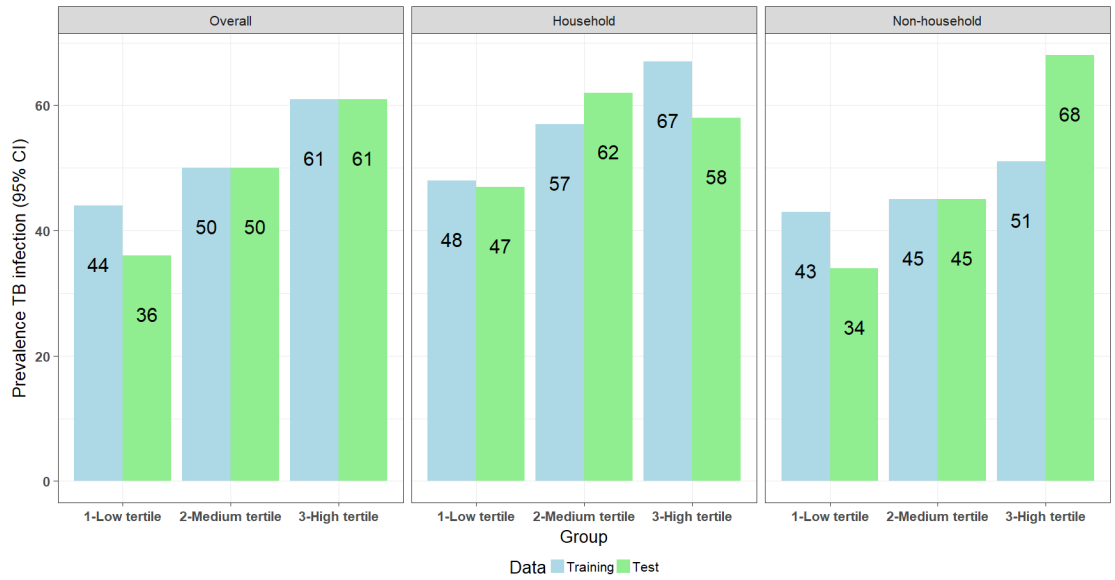


Figure 5.7. Prevalence of Tuberculosis infection, according to Setting and Relationship scores categories.

Prevalence (%) shown in overall population (left panel), household contacts (center panel) and non-household contacts (right panel). Light blue bar represents training data (n=619). Green bar represents test data (n=304). Setting domain: Top three panels. Relationship domain: Bottom three panels.

SUPPLEMENTARY MATERIAL

SECTION A. Sensitivity analysis.

We repeat the main analyses only including 858 subjects with a numerical value for TST. We excluded individuals that were just reported with a "positive TST" and subjects with a previous TB disease. Also, we considered HIV status of our contacts to define TST.

The overall prevalence of latent tuberculosis infection was 47% (95% CI 44-51). The crude prevalence ratio for the Setting score was 1.05 (95% CI 1.03-1.07) and 1.06 (95% CI 1.04-1.09) for the Relationship score (Table S 5.1).

We again observed the effect of age of contact as a modifier of the association between the scores and prevalence of tuberculosis infection (Table S 5.2).

After adjustment, the association of Setting and Relationship scores with presence of tuberculosis infection in the social networks of tuberculosis cases continued to be similar to the one obtained in the main analyses (Table S 5.3). The adjusted effect of the Setting domain in the prevalence of tuberculosis was higher among children between 5-14 years (PR=1.24, 95% CI 1.12-1.37, reduced model) and lower in the older bracket (1.03, 95% CI 1.00-1.07, reduced model). The relationship domain was exclusively associated with tuberculosis infection in children of 0-4 years (PR=1.22, 95% CI 1.02-1.45, reduced model).

SECTION B. Individual variables that comprise the Setting and Relationship domains and their association with tuberculosis infection among contacts of tuberculosis cases.

Except for transportation, all the variables that comprised the domain scores shown association with the prevalence of tuberculosis infection in the contacts of tuberculosis cases (Table S 5.4). In particular, the prevalence of tuberculosis was higher among contacts who spent more than 66.5 hours/week with the case (75%), contacts who provided care to the case daily (75%) and individuals that slept in the same room and/or the same bed that the cases (range: 64-75% depending of frequency). The

social networks with the lowest prevalence were the ones that the tuberculosis cases reported to not known well (12%) and contacts that spent less than 3.5 hours/week with the case (38%).

We observed that in the 05-14-year-aged contacts, variables that suggested low Setting exposure to the tuberculosis cases resulted in very low prevalence of tuberculosis infection and vice versa, explaining the strong association of the Setting Score with tuberculosis infection in this age category (Figure S 5.1).

The six variables that conformed the Relationship domain were informative to the presence of tuberculosis infection among contacts, but it seems that there were not differences in prevalence among answers showing low exposure as compared to answers with moderate exposure (Figure S 5.2).

Table S 5.1. Prevalence and crude prevalence ratio (95% CI) for tuberculosis infection by selected potential risk factors. Sensitivity analysis.

Category	N	Prevalence tuberculosis infection		Prevalence ratio (Poisson)
		N	% (95% CI)	% (95% CI)
Overall	858	405	47 (44-51)	
Setting score				1.05 (1.03-1.07)
Relationship score				1.06 (1.04-1.09)
Gender contact				
Female	440	198	45 (40-50)	1
Male	418	207	50 (45-54)	1.07 (0.94-1.21)
Gender index case				
Female	256	114	45 (38-51)	1
Male	602	291	48 (44-52)	1.07 (0.86-1.32)
Sex assortment				
No	415	187	45 (40-50)	1
Yes	443	218	49 (45-54)	1.06 (0.93-1.20)
Age of contact (continuous)				1.01 (1.00-1.01)
Age of contact				
0-4	98	39	40 (30-50)	1
5-14	132	54	41 (32-49)	1.04 (0.77-1.41)
15-greater	628	312	50 (46-54)	1.25 (0.98-1.58)
Age of index case (continuous)				0.99 (0.98-1.00)
Age of index case				
15-24	262	137	52 (46-58)	1
25-44	484	217	45 (40-49)	0.86 (0.70-1.06)
45 or more	112	51	46 (36-55)	0.82 (0.61-1.11)
Type of contact				
Non-household	325	187	58 (52-63)	1
Household	533	218	41 (37-45)	1.41 (1.20-1.65)
HIV				
No	793	376	47 (44-51)	1
Yes	54	24	44 (31-58)	0.88 (0.64-1.21)
BCG				

Category	N	Prevalence tuberculosis infection		Prevalence ratio (Poisson)
		N	% (95% CI)	% (95% CI)
No	73	30	41 (30-52)	1
Yes	738	352	48 (44-51)	1.08 (0.83-1.39)
Don't know	44	23	52 (37-67)	1.18 (0.82-1.68)
Nature of relationship with tuberculosis case				
Spouse	23	16	70 (51-88)	1
Child	120	63	52 (44-61)	0.77 (0.56-1.07)
Sibling	96	59	61 (52-71)	0.88 (0.63-1.22)
Friend	244	110	45 (39-51)	0.67 (0.49-0.92)
Co-workers	66	35	53 (41-65)	0.71 (0.48-1.03)
Oher relative	160	78	49 (41-57)	0.74 (0.52-1.06)
Neighbor	78	18	23 (14-32)	0.32 (0.19-0.54)
Other	71	26	37 (25-48)	0.54 (0.37-0.81)
Known contact before cough				
No	24	8	33 (14-52)	1
Yes	833	396	48 (44-51)	1.30 (0.82-2.05)
Known if contact has cough				
No	815	377	46 (43-50)	1
Yes	43	28	65 (51-79)	1.36 (1.01-1.84)
Knows if contact has tuberculosis				
No	853	402	47 (44-51)	1
Yes	5	3	60 (17-100)	1.14 (0.53-2.44)
Frequency of meeting since onset cough				
Decreased	42	16	38 (23-53)	1
Same frequency	775	366	47 (44-51)	1.21 (0.74-1.96)
Increased	41	23	56 (41-71)	1.46 (0.82-2.58)
Number of other people met in addition to contact (continuous)				1.01 (0.97-1.04)

Category	N	Prevalence tuberculosis infection		Prevalence ratio (Poisson)
		N	% (95% CI)	% (95% CI)
Number of other people met in addition to contact (categorical)				
<2 persons/meeting	256	122	48 (42-54)	1
2-4 persons/meeting	322	150	46 (41-52)	1.02 (0.85-1.22)
5-6 persons/meeting	200	92	46 (39-53)	1.04 (0.83-1.31)
>6 persons/meeting	80	41	51 (40-62)	1.06 (0.81-1.37)
Microscopy Status				
0-1+	109	35	32 (23-41)	1
2-3+	727	357	49 (45-53)	1.46 (1.04-2.06)

Table S 5.2. Prevalence ratio for the association between increasing scores in the Setting and Relationship scores and tuberculosis infection among social contacts of tuberculosis cases. Sensitivity analysis.

Overall and stratified analysis by selected key variables.

Variable	N	Prevalence ratio Factor Setting	Prevalence ratio Factor relationship
Overall		1.05 (1.03-1.07)	1.07 (1.04-1.09)
By Gender contact			
Female	440	1.07 (1.04-1.09)	1.07 (1.03-1.10)
Male	418	1.04 (1.01-1.06)	1.06 (1.03-1.10)
Gender index case			
Female	256	1.09 (1.06-1.12)	1.09 (1.05-1.13)
Male	602	1.04 (1.01-1.07)	1.06 (1.02-1.09)
Sex assortment			
No	415	1.06 (1.03-1.09)	1.06 (1.02-1.10)
Yes	443	1.04 (1.02-1.07)	1.07 (1.03-1.10)
Age of contact (continuous)		1.06 (1.04-1.08)	1.06 (1.03-1.08)
Age of contact			
0-4	98	1.10 (1.03-1.17)	1.34 (1.18-1.51)
5-14	132	1.26 (1.16-1.37)	1.11 (1.00-1.23)
15-greater	628	1.04 (1.02-1.06)	1.04 (1.02-1.07)
Age of index case (continuous)	858	1.05 (1.03-1.07)	1.06 (1.04-1.09)
Age of index case			
15-24	262	1.06 (1.03-1.09)	1.06 (1.02-1.11)
25-44	484	1.03 (0.99-1.06)	1.06 (1.01-1.10)
45 or more	112	1.12 (1.06-1.18)	1.11 (1.05-1.18)
Type of contact			
Non-household	325	1.02 (0.98-1.07)	1.05 (1.01-1.09)
Household	533	1.08 (1.02-1.15)	1.04 (1.00-1.09)
HIV status contact			
No	793	1.11 (1.03-1.20)	1.06 (1.03-1.09)

Variable	N	Prevalence ratio Factor Setting	Prevalence ratio Factor relationship
Yes	54	1.05 (1.03-1.07)	1.16 (1.05-1.27)
BCG contact			
No	73	1.06 (0.99-1.15)	1.06 (0.98-1.15)
Yes	738	1.05 (1.02-1.07)	1.06 (1.03-1.09)
Don't know	44	1.10 (1.03-1.17)	1.13 (1.05-1.22)
Nature of relationship with tuberculosis case			
Spouse	23	1.14 (0.87-1.49)	1.01 (0.86-1.19)
Child	120	1.12 (1.04-1.20)	1.05 (0.98-1.12)
Sibling	96	1.02 (0.93-1.13)	0.96 (0.89-1.03)
Friend	244	1.02 (0.96-1.09)	1.04 (0.95-1.14)
Co-workers	66	0.98 (0.87-1.01)	0.97 (0.80-1.17)
Oher relative	160	1.06 (1.01-1.18)	1.08 (1.02-1.14)
Neighbor	78	0.75 (0.52-1.08)	0.90 (0.72-1.12)
Other	71	0.99(0.86-1.14)	0.94 (0.71-1.25)
Known contact before cough			
No	24	1.09 (0.91-1.29)	1.22 (0.93-1.62)
Yes	833	1.05 (1.03-1.07)	1.06 (1.03-1.09)
Known if contact has cough			
No	815	1.05 (1.03-1.07)	1.07 (1.04-1.10)
Yes	43	1.02 (0.97-1.06)	1.04 (0.94-1.16)
Knows if contact has tuberculosis			
No	853	1.05 (1.03-1.07)	1.07 (1.04-1.10)
Yes	5	0.73 (0.53-1.02)	0.91 (0.65-1.27)
Frequency of meeting since onset cough			
Decreased	42	1.00 (0.88-1.13)	1.05 (0.92-1.21)
Same frequency	775	1.05 (1.03-1.07)	1.07 (1.04-1.10)
Increased	41	1.14 (1.03-1.26)	1.04 (0.96-1.13)

Variable	N	Prevalence ratio Factor Setting	Prevalence ratio Factor relationship
Number of other people met in addition to contact	856	1.05 (1.03-1.07)	1.06 (1.04-1.09)
Number of other people met in addition to contact (categorical)			
<2 persons/meeting	256	1.02 (0.99-1.06)	1.05 (1.01-1.09)
2-4 persons/meeting	322	1.05 (1.01-1.09)	1.07 (1.01-1.12)
5-6 persons/meeting	200	1.11 (1.05-1.18)	1.07 (1.02-1.13)
>6 persons/meeting	80	1.09 (1.01-1.16)	1.10 (0.99-1.21)
Microscopy Status			
0-1+	109	1.05 (0.97-1.13)	1.06 (0.99-1.12)
2-3+	727	1.05 (1.02-1.07)	1.06 (1.03-1.10)

Table S 5.3. Adjusted prevalence ratio for the association between increasing scores in the Setting and Relationship domains and tuberculosis infection among social contacts of tuberculosis cases. Sensitivity analysis.

Overall results and stratified by age of contact.

Population, stratified by age of contact (years)	n	Prevalence ratio (95% CI) Reduced model		Prevalence ratio (95% CI) DAG model	
		Setting	Relationship	Setting	Relationship
Overall		1.07 (1.03-1.10)	0.99 (0.95-1.03)	1.07 (1.04-1.10)	0.98 (0.94-1.02)
0-4	98	1.08 (0.99-1.18)	1.22 (1.02-1.45)	1.08 (0.99-1.18)	1.23 (1.02-1.49)
5-14	132	1.24 (1.12-1.37)	1.01 (0.92-1.10)	1.25 (1.13-1.39)	1.00 (0.91-1.10)
15-greater	628	1.03 (1.00-1.07)	1.01 (0.97-1.06)	1.04 (0.99-1.18)	1.00 (0.96-1.05)

Overall Reduced model: Adjusted by age of contact, microscopy status index case and HIV status of contact.

Overall DAG mode: Adjusted by age of contact, age of index, sex contact, sex index, HIV status of contact, microscopy status index, knowledge of cough status of the contact by the index case, BCG vaccination of the contact

Table S 5.4. Univariate analysis. Association between individual variables that comprised the Setting and Relationship domains and Prevalence of tuberculosis infection among contacts of tuberculosis cases.

Variable		Prevalence TB		Prevalence ratio
	N in group	n infected	% (95% CI)	95% CI
Length knowing the contact				
Less than 2 years	351	158	45 (40-50)	1
2-4 years	163	76	47 (39-54)	1.0 (0.9-1.3)
5-6 years	197	107	54 (47-61)	1.2 (1.0-1.5)
More than 6 years	212	127	60 (53-66)	1.3 (1.1-1.6)
Hours spent per week with contact				
Less than 3.5 hours/week	265	102	38 (33-44)	1
Between 3.5-28 hours/week	365	176	48 (43-53)	1.3 (1.0-1.6)
Between 28-66.5 hours/week	210	128	61 (54-68)	1.5 (1.2-1.8)
Greater 66.5 hours/week	83	62	75 (65-84)	2.0 (1.6-2.5)
Location of usual meeting				
Outside home of tuberculosis case	409	192	47 (42-52)	1
Home of tuberculosis case	514	276	54 (49-58)	1.2 (1.0-1.4)
Ventilation place of meeting				
Full ventilation	454	201	44 (40-49)	1
Fair ventilation	169	90	53 (46-61)	1.1 (0.9-1.4)
Minimal ventilation	144	93	65 (57-72)	1.4 (1.2-1.7)
Poor ventilation	156	84	54 (46-62)	1.2 (1.0-1.5)
Indoor or outdoor meeting				
Mostly meeting outdoors	447	193	43 (39-48)	1
Equally indoors/outdoors	239	125	52 (46-59)	1.2 (1.0-1.5)
Mostly meeting indoors	237	150	63 (57-69)	1.4 (1.2-1.6)
Sleeping conditions				
No slept in same room, nor bed	723	338	47 (43-50)	1
Slept same room, but not same bed	135	86	64 (56-72)	1.3 (1.1-1.5)
Slept same room and same bed, not daily	16	12	75 (54-96)	1.4 (0.9-2.2)
Slept same room and same bed, daily	49	32	65 (52-79)	1.5 (1.2-1.8)

Variable	N in group	Prevalence TB n infected	Prevalence TB % (95% CI)	Prevalence ratio 95% CI
Meals				
Not shared meals	361	160	44 (39-49)	1
Shared meals, less than a day per week	75	29	39 (28-50)	1.0 (0.8-1.4)
Shared meals 1-3 days/week	133	63	47 (39-56)	1.2 (1.0-1.5)
Shared meals 4-6 days/week	58	31	53 (40-66)	1.4 (1.0-1.8)
Shared meals daily	296	185	62 (57-68)	1.5 (1.2-1.7)
Case trusts contact				
No discuss nor confide	395	179	45 (40-50)	1
Discuss but not confide	287	142	49 (44-55)	1.0 (0.9-1.2)
Discuss and confide	241	147	61 (55-67)	1.3 (1.1-1.5)
Shared TB diagnosis				
No	521	233	45 (40-49)	1
Yes	402	235	58 (54-63)	1.3 (1.2-1.5)
Care by contact				
No care by contact	775	379	49 (45-52)	1
Care provided, less than a day per week	33	16	48 (31-66)	1.1 (0.8-1.6)
Provided care 1-3 days/week	38	20	53 (37-69)	1.2 (0.9-1.6)
Provided care 4-6 days/week	14	6	43 (17-69)	0.9 (0.6-1.5)
Provided care daily	63	47	75 (64-85)	1.5 (1.3-1.8)
How well does the case knows contact				
Not well/almost do not know	16	2	12 (0-28)	1
Somewhat well	129	54	42 (33-50)	2.4 (1.0-6.0)
Moderately well	214	89	42 (35-48)	2.2 (0.9-5.4)
Very well	564	323	57 (53-61)	3.1 (1.2-7.5)
Means of transportation used most often with contact. None (walking) versus a type of transportation.				
None/walking	748	380	51 (47-54)	1
Another type of transportation	175	88	50 (43-58)	1.1 (0.9-1.3)

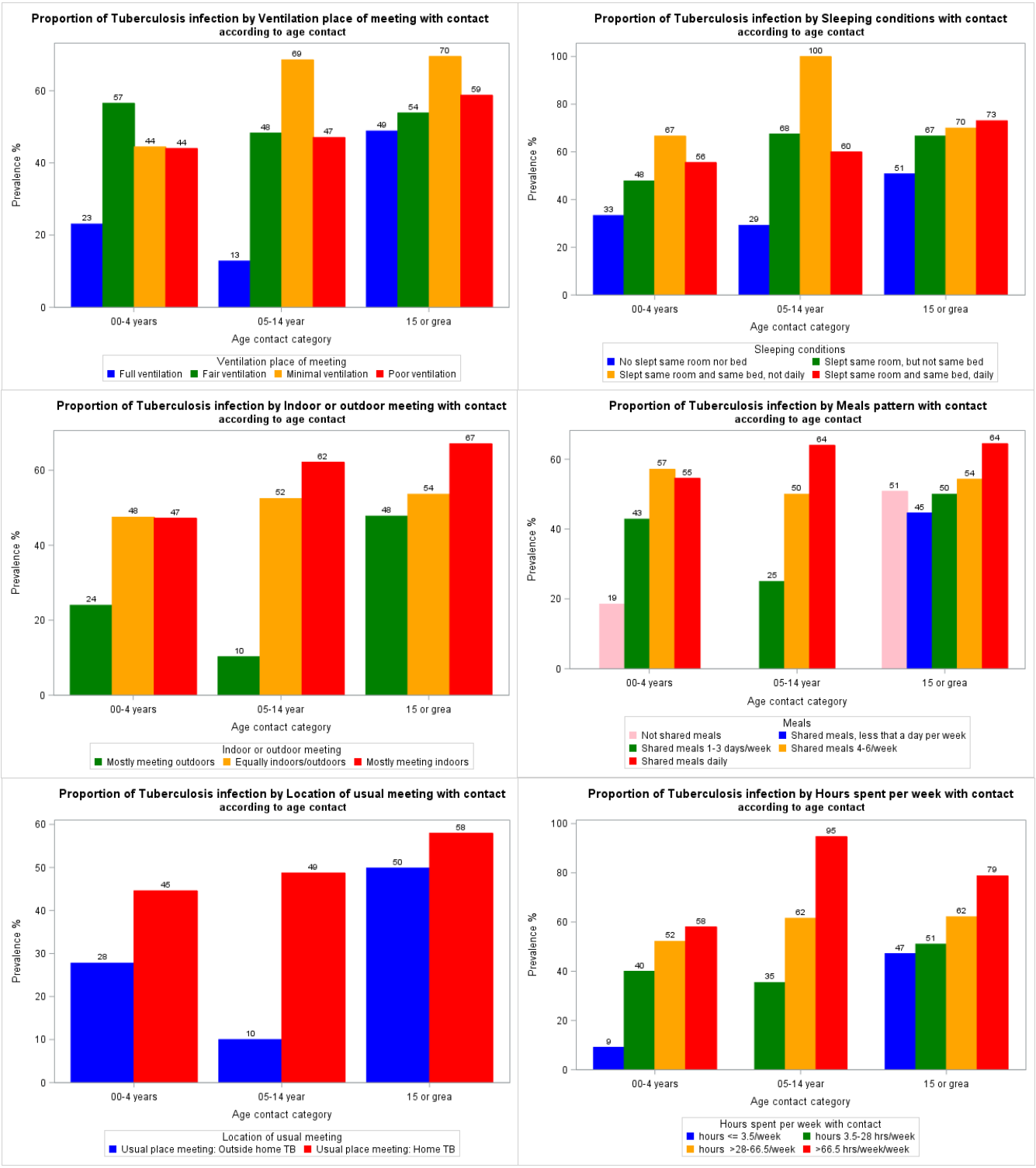


Figure S 5.1. Prevalence of Tuberculosis infection in Contacts, according to the six individual variables that comprise the Setting domain.

Results shown stratified by age of contact.

`

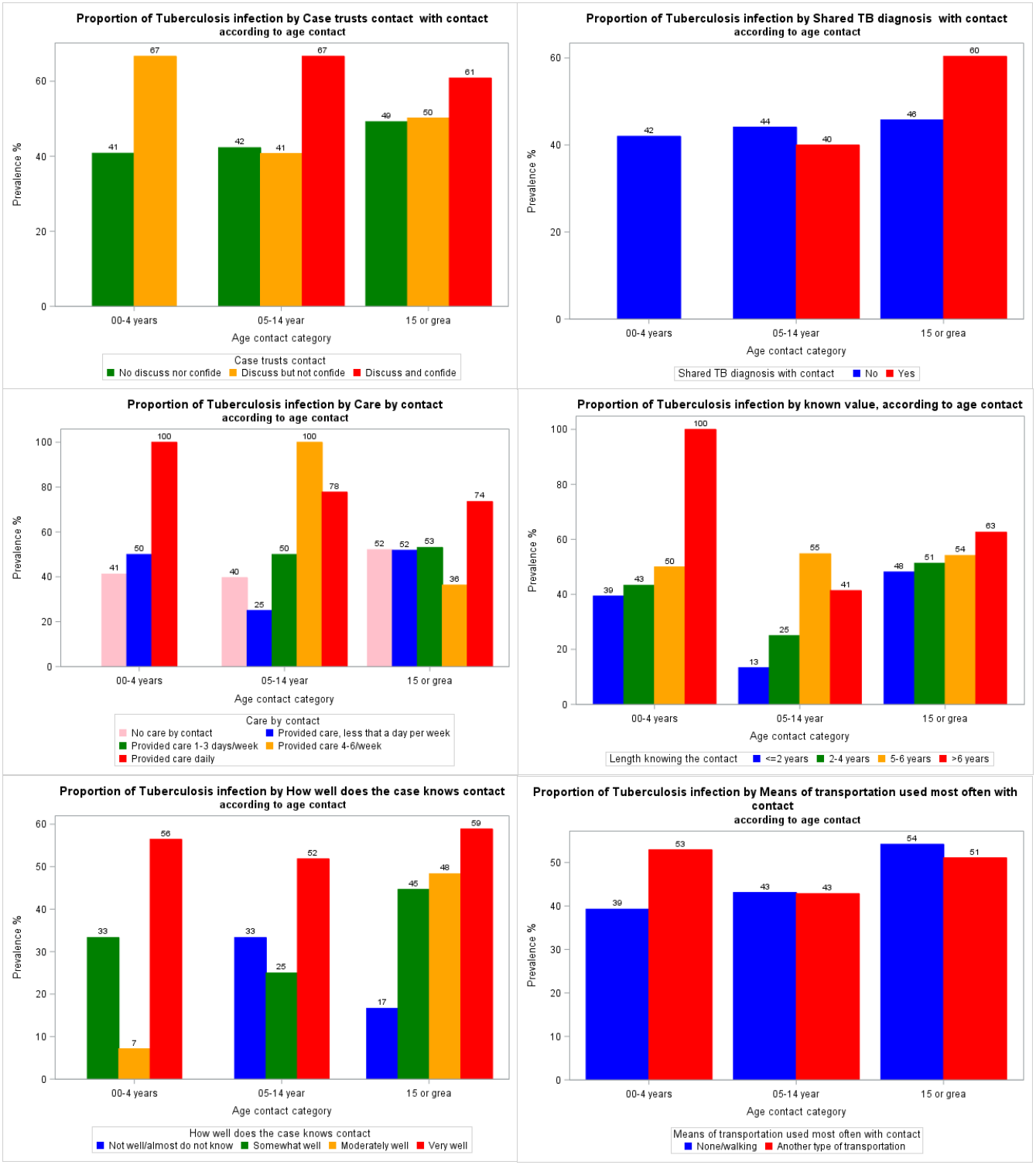


Figure S 5.2. Prevalence of Tuberculosis infection in Contacts, according to six individual variables that comprise Relationship domain.

Results shown stratified by age of contact.

Chapter 6

CHARACTERIZATION OF THE PROPORTION OF CLUSTERED TUBERCULOSIS CASES: INSIGHTS FROM A MOLECULAR EPIDEMIOLOGY STUDIES IN GUATEMALA BETWEEN 2010 AND 2014. [7]

[7] Castellanos ME, Lau D, Ebell M, Dobbin KK, Quinn F, Samayoa B and Whalen CC. To be submitted to AJE.

Abstract

Background. There is little information about the proportion of clustering of tuberculosis cases from low-income settings, which can represent ongoing transmission events. We investigated for the first time the proportion of clustered tuberculosis cases based on genotypic matching in Guatemala City, Guatemala between 2010 and 2014 and potential risk factors associated with these clustered cases in HIV-infected subjects. Moreover, the genetic diversity of *M. tuberculosis* isolates in this country is presented.

Design and methods. This study was a retrospective observational study conducted in *Mycobacterium tuberculosis* isolates from HIV-infected and non-HIV infected tuberculosis cases that submitted samples to a referral tuberculosis laboratory in Guatemala City, Guatemala from 2010-2014. Genotyping results were compared with the international spoligotyping database, SITVIT2 and classified accordingly. We generated a spoligoforest using the SpolTool program. We categorized spoligotype patterns as clustered or non-clustered depending of their genotype and estimated the proportion of clustering and the recent transmission index (RTI_{n-1}). We analyzed the crude association between demographic, clinical and behavioral variables and clustering in the HIV-population.

Results. From 2010 to 2014, a total of 479 patients were confirmed as tuberculosis cases by culture at the study site. Spoligotype patterns were available from 413 patients (86%), ten of them with two isolates included in the study. Overall, the most frequent spoligotyping families were LAM (39%), followed by T (23%), Haarlem (14%), X (13%), Beijing (3%), East African-Indian-EAI (3%) and Unknown (3%) representing 98% of the isolates. We detected 91 spoligotype patterns and 23 of them had not been previously reported in the international spoligotyping database. Out of the 423 isolates, 371 strains (88%) were grouped in 39 clusters (range: 2-92). The recent transmission index (RTI_{n-1}) was estimated to be in the range of 78% to 90%. Two factors were associated with clustering in the 120 HIV-infected group with available data: education (OR: 4.0, 95% CI 1.2-13.8) and pulmonary tuberculosis (OR=5.4, 95% CI 1.6-17.9).

Conclusion. There might be high levels of ongoing transmission of *M. tuberculosis* in Guatemala City, Guatemala as indicated by clustering in a convenience sample. Among HIV-infected patients, clustering was more likely in pulmonary disease and when individuals have some level of education as opposed to no education. Moreover, we detected previously unreported strains of *M. tuberculosis* that contribute to tuberculosis morbidity in the country.

Introduction

Mycobacterium tuberculosis strains are considered clustered when the genotypes between two or more organisms are the same. Clustered strains represent a chain of transmission, and may represent ongoing or recent transmission if the strains are sampled during a short period of time, such as one to two years. Thus, when the sample collection interval is narrow enough, and in a well-defined geographical area (J. Glynn et al., 1999), we can infer that clustered cases represent recent transmission; the assumption being that clusters are "epidemiologically linked chains of recently transmitted disease" (Murray & Nardell, 2002). Strains with unique genotypes are thought to represent reactivation of an old tuberculosis infection and are considered non-clustered.

The distribution of clustered strains is affected by several host and population-level characteristics (Fok et al., 2008; M. Murray, 2002). Individual host characteristics that influence levels of tuberculosis clustering include place of birth, pulmonary tuberculosis disease (rather than extrapulmonary), and alcohol abuse (Fok et al., 2008). Population-level characteristics that affect clustering are age structure of the population, prevalence of latent tuberculosis infection and HIV prevalence, to name a few (M. Murray, 2002). Several studies have been performed evaluating levels of tuberculosis clustering and potential risk factors for increased cluster risk. Most of these studies, however, were conducted in low-incidence tuberculosis settings and little is known about relevant risk factors for recent transmission in HIV-infected patients in settings with a medium tuberculosis burden, such as Guatemala.

The main aim of this study was to characterize the proportion of clustered tuberculosis cases based on genotypic matching in Guatemala City, Guatemala between 2010 and 2014 and to identify risk factors associated with these clustered cases in HIV-infected subjects. As a secondary aim, we present for the first time the genetic diversity of *M. tuberculosis* isolates from patients in Guatemala. The international genotyping database SITVIT2, which contains data from 111,635 isolates from 169 countries, was employed to classify and relate the Guatemalan strains to the global framework of

tuberculosis (Couvin, David, Zozio, & Rastogi, 2018). Knowledge of existing circulating genotypes and the identification of risk factors associated with recent transmission will allow an evidenced-based approach for health policymakers to direct and concentrate targeted tuberculosis control measures to high-risk populations

Study population and methods

DESIGN OVERVIEW

This is a retrospective observational study. Genotypes of *Mycobacterium tuberculosis* (MTB) isolates from HIV-infected and non-HIV infected tuberculosis cases in Guatemala City from 2010-2014 were categorized as clustered or non-clustered depending of their genotype. Clustered strains were considered as evidence of recent transmission. Independent potential factors associated with having a clustered isolate were investigated in the HIV-infected subjects.

STUDY POPULATION, SETTING AND DATA SOURCES

Study Setting. The study was conducted in Guatemala City, Guatemala at Integral Health Association (ASI). In Guatemala, in 2014, 4,200 new tuberculosis cases were notified for an incidence rate of 25 cases per 100,000/year (World Health Organization., 2016). However, the World Health Organization has estimated that due to the low case detection rate, the incidence could be almost two times higher, with an estimated incidence of 60 cases/100,000/year. Out of these 4,200 individuals with incident TB, 270 were HIV infected subjects.

ASI operates a clinic and a clinical laboratory, which has served over 25 years in Guatemala City, as an institution for the management and treatment of HIV-infected subjects. Currently, there are 2,816 active patients in the clinic. The tuberculosis laboratory at ASI serves as a referral center, receiving samples not only from the HIV population treated by ASI staff but also from other centers, in Guatemala City and other regions of the country.

Study Population: Patients with tuberculosis who submitted samples for the diagnosis of tuberculosis at the laboratory of ASI in Guatemala City during 2010-2014 were included in this study.

Inclusion and Exclusion Criteria. We only included individuals in whom a *Mycobacterium tuberculosis* isolate was detected during 2010-2014. *Mycobacterium tuberculosis* isolates should have been confirmed as such by laboratory methods (culture plus specie identification by conventional or molecular methods). Isolates without a clinical record nor a spoligotyping results were excluded. We also excluded isolates with multiple spoligotyping results per isolate but without at least two identical results per sample.

Definition of HIV status in the study population. Patients managed and treated in ASI were considered HIV-infected individuals. Patients not managed in ASI but that submitted samples to the ASI laboratory were considered as HIV negative/unknown status.

Genotyping technique. Isolates were analyzed by spoligotyping, as described in detail elsewhere (Kamerbeek et al., 1997). Succinctly, in the direct repeat locus of *Mycobacterium tuberculosis*, direct repeats are interspersed by polymorphic DNA sequences called "spacers". In this technique, both the direct repeats and the spacers are amplified by polymerase chain reaction (PCR). Oligonucleotides that correspond to 43 of these spacers are immobilized into a membrane. The amplified fragments are hybridized to these spacers. The hybridization products are detected by chemiluminescence (Kamerbeek et al., 1997). Depending on the presence and distribution of the spacers, a genotype pattern is obtained for each isolate.

Data sources. Data regarding time of tuberculosis diagnosis, type of sample, spoligotyping results and drug susceptibility of the isolates were extracted from the laboratory records of the ASI Tuberculosis Laboratory for all study patients. For the HIV-infected individuals, demographic, clinical and behavioral data was extracted by ASI staff from the electronic database "MANGUA" (Garcia et al., 2015).

STUDY OUTCOME FOR CLUSTERING

The study outcome was the clustering of a strain. A clustered strain was defined as one that at least share the same genotype with another isolate, regardless of the HIV status of the corresponding patients. Isolates with unique genotype patterns were considered non-clustered/unique. Strains that belong to any cluster were coded as one (1) and strains that do not share a genotype pattern with any other isolate were coded as zero (0).

STUDY EXPOSURES FOR CLUSTERING

In the case of HIV-infected patients from ASI, an extensive set of demographics, clinical and behavioral variables were evaluated as independent potential factors for clustering. The complete list and categorization of these variables are presented in Table S 6.1

ANALYSTICAL STRATEGY

Data preparation and cleaning. From the main analysis, we excluded all isolates without spoligotyping results (Figure 6.1). We deleted exact duplicates. When an isolate had multiple spoligotyping results, we kept the spoligotype that had at least two identical results. Several patients presented with multiple samples. For each patient, the first available sample was selected. We excluded patients with different spoligotypes if the strains were isolated in a time period of less than six months and the samples came from the same source (pulmonary, extra-pulmonary). If the difference was more than six months or if the samples came from different source (pulmonary vs extra-pulmonary), we considered them as distinct patients regardless of the spoligotyping result. After the cleaning, each patient had a unique spoligotype, and the source of the sample(s) was categorized as pulmonary or extra-pulmonary. Patients with pulmonary and extra-pulmonary samples collected at the same time, were considered as extra-pulmonary cases.

We transformed the octal format of the spoligotypes to binary format. It should contain 43 digits, representing the presence (1) or absence (0) of 43 spacers, but for some laboratory issues at the study site only 42 digits were captured. We assumed that all samples have presence of the last spacer, and we added that digit, as this as recommended to us by Dr. Leen Rigouts from the Institute of Tropical Medicine, Antwerp, Belgium, an expert in molecular epidemiology of tuberculosis (personal communication, October 2018). Her advice was based on the fact that overwhelmingly majority of spoligotype patterns have the last spacer present. We confirmed these results checking the frequency of the pattern with or without the 43th spacer in SpolSimilaritySearch, an online tool with a collection of over 100,000 isolates from 169 countries of origin of the tuberculosis patients (http://www.pasteur-guadeloupe.fr:8081/SpolSimilaritySearch/index.jsp). We searched the five most common genotypes of our study and we found that the spoligotype patterns with the presence of the last spacer represented ≥99.5% of the isolates collected (Table S 6.2).

Study population. We estimated the proportion of tuberculosis cases with spologytping results out of the total number of patients with a positive *M. tuberculosis culture*. For all patients, we estimated the type of sample source (pulmonary and extrapulmonary) and the drug susceptibility patterns of their isolates (resistant, non resistant) using proportions. We stratified these results by HIV status of the patients.

Assignment of shared international type and spoligotyping families. We assigned each spoligotype a shared international type (SIT) and family using the international database SITVIT2, and update of SITVITWEB (http://www.pasteur-guadeloupe.fr:8081/SITVIT2/) (Couvin et al., 2018). If there was no SIT listed, but the pattern was reported in the database as an orphan isolate, we named the pattern as 'Pseudo-XXU/C' whereby the XX represents a two-digit number that we assigned. The U represents a unique spoligotype in our dataset and C represents a pattern found in cluster in our dataset. If the pattern had never been previously reported in SITVIT2 we named the pattern as 'GUA-XXU/C, with the same nomenclature previously described. Spoligotypes without an assigned family were analyzed with the

SPOTCLUST online tool (http://tbinsight.cs.rpi.edu/run_spotclust.html) to assign the most probably family. We estimated the frequency of these SITs and families using proportions, in the overall population and per year of isolation.

Characteristics of the patients with two samples included in the study. We described the major characteristics of the patients that provided two samples for the study: HIV status, sex, age at time of tuberculosis diagnosis, type of sample, time between tuberculosis diagnosis, acid-fast bacilli result, drug susceptibility patterns, SIT, spoligotyping family, CD4 count, previous history of tuberculosis, behavioral risk factors for clustering and record of death.

Visualization of relationship among spoligotypes. We constructed a spoligoforest, a visualization of relationships among spoligotypes (Reyes, Francis, & Tanaka, 2008), based on mutation events, using the SpolTool online programs available at http://spoltools.emi.unsw.edu.au/. (Tang, Reyes, Luciani, Francis, & Tanaka, 2008). In this type of graph, each node represents a spoligotype. The size of each node represents the number of strains that belongs to a given spoligotype. Each node is labelled with the SIT/study name. The number in brackets in the node reflects the cluster size. Directed edges indicate single-event deletion that relates to two clusters, the arrowheads pointing to the descendants. In this model, multiple inbound edges are reduced to one inbound edge, the one with the maximum weight as described previously (Reyes et al., 2008). Edges with weights less than 0.5 are presented as dotted lines. Edges with weights ≥ 0.5 and <1 are shown as dashed lines. Edges with weights of 1 are represented as solid lines. We selected the layout based on a Fruchterman–Reingold algorithm. The spoligoforest was colored by spoligotyping family using the open source graph visualization software Graphviz (https://www.graphviz.org/. We used the same colors that were used from the authors of the SITVIT2 in the most recent report (Couvin et al., 2018).

Descriptive statistics of clustering of mycobacterial isolates. Descriptive statistics were used to present the number and proportion of clustered strains and clusters, distribution of cluster size, mean size of cluster, maximum size of cluster, proportion of cluster with 2 cases, clusters 3-19 cases and clusters

≥20 cases. We selected these statistics based on previous literature (Fok et al., 2008; Mears et al., 2015; Murray & Alland, 2002; Smith, Maguire, Anderson, Macdonald, & Hayward, 2017; C. Yang et al., 2015)

Estimation of the proportion of TB cases due recent transmission. Several methods to estimate the proportion of tuberculosis cases due to recent transmission have been described (Murray & Alland, 2002). We used two methods to estimate the proportion of tuberculosis cases due to recent transmission: the traditional "n-1" method and a web-based tool based on a regression model.

For the "n-1" method we applied the original formula, developed by Small and others: Recent Transmission Index, $RTI_{n-1} = (n_c - c)/n$ (Small et al., 1994), in which n = total number of cases in the sample, c = is the number of clusters (genotypes represented by at least two cases) and n_c =is the total number of cases in a cluster of two or more . As suggested by Glynn and colleagues (Glynn et al., 2005) because of the length of this study, we re-estimated the proportion of tuberculosis cases due to recent transmission using different time windows: 2 years (2010-2011), 3 years (2010-2012) and 4 years (2010-2013).

We also estimated the recent transmission proportion in the whole time period by using a web-based method developed by Kasaie and others (Kasaie et al., 2015). In this technique, several parameters are considered, tuberculosis incidence in the region, sampling coverage, study duration, proportion of clustering and proportion of clusters. These considerations intend to reduce the estimation bias, particularly in areas with a low sampling rate. We used different scenarios as we lacked information in the real incidence of tuberculosis cases in the country (25 and 60 cases/1000,000 pop/yr.) and the sampling rate of the study (10% or 20% of total active TB cases who had spoligotyping data). The details of this estimation are shown in Appendix 1.

Independent factors associated with clustering in the HIV-infected population. Potential risk factors for clustering (previously described in "Study exposures for clustering" section) were reported for the HIV-population and were compared by Chi-square test (categorical variables), Fisher Test (counts less than 5) or Wilconxon test (continuous variables) in patients with clustered and non-clustered strains. In

the categorical variables with more than two classes and when the overall *p* values were less than 0.20, a Bonferroni correction was conducted: A corrected *p* value was obtained for the pairwise comparison between a given class and the reference class.

In the variables and classes in which *p* values < 0.20 were obtained, we estimated the crude association between each of these predictors and clustering using regression models. In these regression models, the outcome, or dependent variable, is clustering as defined in section "Study Outcome for Clustering". Since this is a dichotomous outcome variable, at logistic model was initially considered appropriate. However, our sample size was small and there was a very low proportion of one of the events (non-clustered strains). Thus, we estimated the association between exposure and the outcome using the Firth logistic regression method (King & Zeng, 2001). Unadjusted odds ratio and prevalence ratio (with 95% CI) were obtained with this method by exponentiating the regression coefficients. Due to the low proportion of non-clustered strains we did not conduct multivariate regression models.

All statistical analyses were conducted on SAS software (release 9.4, SAS Institute Inc., Cary, NC, USA).

SENSITIVITY ANALYSIS

We re-analyzed the proportion of clustering and $RTI_{n-1,}$ after exclusion of clusters with a large size, ≥20 isolates. This approach has shown to improve at least partly the specificity of spoligotyping (Scott et al., 2005).

ETHICAL APPROVAL

Institutional review board clearance was obtained from Zugueme, a Guatemalan independent Ethics Committee, approved by the Ministry of Health of Guatemala and by the University of Georgia.

Results

Study population. From 2010 to 2014, a total of 479 patients were confirmed as tuberculosis cases by culture at the study site. Four hundred thirteen patients (86%) had isolates with spoligotyping results. Of

these, there were 8 patients who provided two samples collected at different time points (> 6 months). In addition, there were 2 patients who presented mixed infections, with samples collected concurrently from different sites (pulmonary and extra-pulmonary) and different spoligotype pattern (later described in detail).

Based on our methodology, we considered these 10 cases as 20 patients, thus the final sample size of the study was 423 patients (Figure 6.1). Of these patients, 140 (33%) were confirmed as HIV-infected (33%). Most patients provided pulmonary samples (67%). Overall, 30% of the patients had strains resistant to at least one of the anti-tuberculosis drugs and 2% of them were MDR-tuberculosis cases (Table 6.1).

Spoligotype patterns and families. We detected 91 spoligotype patterns, with 16 genotypes being present in ≥ 5 samples and comprising 74% of the sample (Table 6.2). The five most frequent genotypes were SIT 33 (22%) from the Latin American-Mediterranean (LAM) family, SIT 53 (13%) from the ill-defined family T, SIT 42 (6%) from the LAM family, SIT 50 (6%) from the Haarlem family and SIT 119 (5%) from the X family. The complete list of the 91 spoligotypes identified among the 423 cases is presented in Table S 6.3. Overall, the most frequent spoligotyping families were LAM (n=163, 39%), followed by T (n=96, 23%), Haarlem (n=59, 14%), X (n=56, 13%), Beijing (n=13, 3%), East African-Indian-EAI (n=13, 3%) and Unknown (n=13, 3%) representing 98% of the isolates. The Ural, Family 33, MANU and S families were uncommon (5 or less isolates). The frequency of the most common spoligotyping families did not vary substantially by year of isolation (Figure 6.2).

There were 23 genotypes not previously reported in the international spoligotyping database belonging to 35 strains in the study (Table 6.3). Among these, the most frequent spoligotyping families were EIA (n=12, 34%), followed by LAM (n=7, 20%) and T family (n=6, 17%). Five of these genotypes were in clusters, four with a cluster size of 2 and one with a cluster size of 9 (named “GUA01C”) from the EIA family.

Characteristics of the ten patients with two samples whom were considered as distinct 20 patients. In the case of the two patients with mixed infections, their two samples presented different spoligotyping families. For study ID 97, the sputum sample had a spoligotype belonging to the Haarlem family, whereas

the bone marrow sample had a spoligotype from the X family. For study ID 203, the sputum sample had a LAM family, whereas the stool sample had a spoligotype from the T family (Table 6.4). The latter patient had a co-infection with HIV, with a CD4 count of 28 cells/mm^3. In the eight patients with two tuberculosis episodes with a difference of > 6 months, seven of them presented the same spoligotype in the different time points, whereas only in one case (study ID 334) the samples had different spoligotype patterns (SIT 53, T family and SIT 42, LAM family). She was a woman that had been in prison. Also, it is worth to note that nine of these ten patients were HIV-infected individuals. A comprehensive description of the major characteristics of these patients and theirs samples in presented (Table 6.4).

Visualization of relationship among spoligotypes. Based on the results of the spoligoforest (Figure 6.3), SIT 53 was the root node and the most likely oldest spoligotype. SIT 42 was a descendant of SIT 33 and was the precursor of the biggest cluster in the study (SIT 33). These three SITs, 53, 33 and 42 had many descendants, suggesting these genotypes had been circulating for enough time to generate mutations. Only 18 genotypes out of the 91 were orphan nodes. The 'GUA01C' genotype seems to be a descendant of a X family member, SIT 137. Isolates from the Beijing family seemed to have evolved independently for the largest component in the spoligoforest. The major findings of the spoligoforest were similar when stratifying by HIV status (Figure 6.4).

Descriptive statistics of clustering of mycobacterial isolates. Out of the 423 isolates, 371 strains (88%) were grouped in 39 clusters (range: 2-92). There were 15 clusters (38%) with size 2, 19 clusters with size 3-19 (49%) and 5 clusters with size ≥20 (13%) (Figure 6.5). The proportion of clustering was similar in the HIV-population (89%) as compared with the patients with a HIV negative/unknown status (87%).

Estimation of the proportion of TB cases due recent transmission. The proportion of tuberculosis cases due to recent transmission (RTI_{n-1}) by the traditional 'n-1' method was 78% in the five years of the study. This proportion minimally decreased when using different time windows, 75% for the 2-year window, 77% for the 3-year window and 78% for the 4-year window (Table 6.5). When using the online

tool developed by Kasaie and others (Kasaie et al., 2015), the RTI_{n-1} was around 90%, this proportion being similar across different scenarios of local tuberculosis incidence and sampling rates (Table S 6.4).

In the sensitivity analysis conducted after exclusion of clusters with ≥ 20 isolates, the proportion of clustering was 75% (152/204). The RTI_{n-1} was 58% by the traditional 'n-1' method and around 73% (72-74% depending scenario) by the Kasaie method.

Independent factors associated with clustering in the HIV-infected individual. Out of the 140 patients with HIV confirmation, we were able to collect the majority of demographic, behavioral and clinical variables for 120 patients. These cases were primarily men (77%), with a median age of 35.5 years (IQR 28-44). At the time of the tuberculosis diagnosis, 65% where at HIV clinical stage 3, and 63% had a CD4 count of less than 200 cells/mm^3. The proportion of clustering in the isolates of this sub-group was 89% (107/120) similar to the proportion in the whole HIV-population.

In bivariate analysis, of all the demographic variables investigated (Table 6.6), only education level seemed to be of relevance. Patients with some education had a proportion of clustering of 92%, higher than the one detected in patients with some level of education (74%, OR=4.0, 95% CI 1.2-13.8).

Patients that submitted pulmonary samples had clustered isolates in 95% of the cases, whereas patients who submitted extra-pulmonary samples had a lower proportion (77%, OR=5.4, 95% CI 1.6-17.9). Similarly, patients with a positive smear result had a higher proportion of clustering (98%) than patients with a negative smear result (85%, OR=5.2, 95% CI 0.9-30.5). We also noted that patients with a HIV viral load of < 4.0 $_{log10}$ copies/mm^3 had clustered isolates in 100% of the cases, whereas patients with higher viral loads (≥ 4.0 $_{log10}$ copies/mm^3) had lower proportion (81%, OR=7.5, 95% CI 0.4-148.4); but these findings did not reach statistical significance. Previous history of tuberculosis, prison, drug use, alcohol and tobacco consumption were not associated with the proportion of clustering (Table 6.7).

Discussion

We uncovered a high level of recent transmission among tuberculosis cases from Guatemala, Central America between 2010-2014. The proportion of clustered cases detected in this study, 88% was comparable to other spoligotyping-based studies reported among tuberculosis cases in the South of Mexico (78%) and in Honduras (84%), both neighboring areas of Guatemala (Nava-Aguilera et al., 2011; Rosales, Pineda-Garcia, Ghebremichael, Rastogi, & Hoffner, 2010).

A systematic review analyzing data from 36 studies conducted in 17 countries estimated the median tuberculosis clustering proportion at 39%, but a wide range was detected (7-72%) (Fok et al., 2008). In those studies, IS6110-RFLP was the main method for genotyping, in contrast of our use of spoligotyping. It has been shown that spoligotyping, owed to its low-resolution power, can overestimate the number of clustered isolates by up to 50% (Gori et al., 2005). If that was the case, the proportion of clustering in this study might be at least of 44%, still suggesting a high rate of transmission.

We did not find differences in the level of recent transmission using different time windows. Moreover, the proportion of clustered isolates did not vary according to HIV status. HIV status has been reported to be of importance exclusively in patients ≥45 years of age (Glynn et al., 2005). We did not observe this association in our results, probably owed to the high levels of recent transmission in the whole population.

We found that risk factors for clustering in the HIV population were pulmonary tuberculosis and having smear-positive tuberculosis. Our findings resemble the ones obtained in other molecular studies (Gonzalez et al., 2003; Hamblion et al., 2016; Ong et al., 2004). In San Francisco, US cervical lymphatic tuberculosis and non-respiratory tuberculosis were associated with low clustering proportion (adjusted odds ratio 0.55, 95% CI 0.31–0.96, and 0.55, 95% CI 0.37–0.83 respectively) (Ong et al., 2004). Similar results were reported in Houston, US, where 65% of patients with pulmonary tuberculosis had clustered strains in contrast with 58% of patients with extra-pulmonary tuberculosis (Gonzalez et al., 2003). Moreover, in a systematic review of risk factors associated with recent transmission of tuberculosis,

sputum smear positivity was identified as one of them (pooled odds ratio 1.4, 95% CI 1.2-1.6) (Nava-Aguilera et al., 2009). These results suggest that HIV-infected cases with a pulmonary tuberculosis and/or smear-positive tuberculosis are part of recent transmission chains, confirming previous findings showing that HIV-infected tuberculosis cases are equal infectious than HIV-seronegative cases when they are smear-positive (Martinez et al., 2016).

Patients with no education had a lower proportion of clustering. This association could be modified or associated by other factors. For instance, living in urban areas as opposed to living in rural areas has been shown to be a risk factor for recent transmission (pooled OR 1.5, 95% CI 1.3-1.7) (Nava-Aguilera et al., 2009). Study patients with no education might be living in rural or more isolated areas, thus decreasing the chance of having clustered strains.

We found evidence of mixed infection in two patients who were infected with different strains in their concurrent pulmonary and extra-pulmonary samples. The proportion of mixed infections in tuberculosis is still unclear, the limited evidence shows proportions from 8% to 57% in high burden areas (Tarashi, Fateh, Mirsaeidi, Siadat, & Vaziri, 2017). A previous report from Uganda shown that 26/51 (51%) of HIV-infected individuals with MTB strains in both sputum and blood had discordant genotypes (Ssengooba et al., 2015). Low CD4 T-cell count has been associated with the presence of mixed tuberculosis infections (Shin et al., 2015), a feature present in the HIV-infected individual with a mixed infection from our study. These mixed infections might be the result of a single transmission event, multiple transmission events or if a second infection causes the relapse of the first infection (McIvor, Koornhof, & Kana, 2017). There is an urgent need to further our research in this areas as mixed infections can impact strategies aiming to control and treat tuberculosis (T. Cohen et al., 2012)

As a second aim we described for the first time, the molecular epidemiology of tuberculosis in Guatemala. The most common spoligotypes families found in this study, LAM, T and Haarlem are similar to the ones obtained in South America, although their proportion varied according to the country as recently reviewed (Balcells et al., 2015; Woodman, Haeusler, & Grandjean, 2019). Nevertheless, the

X family accounted for less than 7% in the majority of these studies and we found a prevalence of 13%. In Honduras, the only Central American country with genotyping studies, LAM family accounted for 55% of the isolates, followed by Haarlem (16%), T (16%) and X (6%) (Rosales et al., 2010).

The most prevalent genotypes identified in this study belong to the Euro-American phylogenetic lineage 4, according to a genotype classification recently published (Wiens et al., 2018). Lineage 4 is one of the seven lineages in which are categorized *M. tuberculosis* complex strains and is the one with the largest distribution in the world (Banuls, Sanou, Anh, & Godreuil, 2015; Stucki et al., 2016). Its introduction in the Americas is considered to be owed to the European colonization of the continent after the 1500s (Woodman et al., 2019).

Before this work, only 21 isolates from Guatemala have been documented in the international spoligotype SITVIT2 database, which currently comprises 103, 856 strains with 9, 658 patterns (http://www.pasteur-guadeloupe.fr:8081/SITVIT2/). The addition of 23 spoligotype-patterns not previously described will increase our knowledge of the circulating genotypes in Central America. Moreover, the detection of a cluster of nine isolates from a never reported genotype from the East African-India family requires further research to understand its transmission dynamics. This cluster was found in the tip of the spoligoforest, which might indicate a strain that it is transmitting faster as compared to others (Reyes et al., 2008). It might be a recent introduction in the country, or a local genotype recently emerged and still limited to this geographical region.

This study has several important limitations. We were not able to use an additional second genotyping to better discriminate our spoligotypes, which more likely resulted in an overestimation of the level of transmission. To confirm our results a second genotyping method such as mycobacterial interspersed repetitive-unit–variable-number tandem-repeat (MIRU-VNTR) typing or whole genome sequencing (WGS) is warranted (Jagielski et al., 2016). Second, our study population was restricted to one health center so ascertainment bias could have occurred (Delgado-Rodriguez & Llorca, 2004), limiting the generalizability of our results. Third, we utilized a convenience sampling with a low sampling rate, which

understates the level of transmission (Glynn, Vynnycky, & Fine, 1999; Murray & Alland, 2002; Van Soolingen, 2001). Fourth, information bias is possible owed to the quality of our data (Delgado-Rodriguez & Llorca, 2004). Several of the potential risk factors for clustering were collected months of years before the tuberculosis diagnosis or they had incomplete information on the level of exposure. Nevertheless, if this is bias is present is more likely a non-differential misclassification bias. Based on these limitations, we believe our results are exploratory in nature but still suggest high levels of transmission in Guatemala and might explain the lack of reduction of tuberculosis in the last ten years in the country A recent molecular study, using whole genome sequencing identified an outbreak of the Beijing family in a poor area in a neighborhood in Guatemala City close to a prison center (Saelens et al., 2015) for at least 2 years, showcasing the risk of outbreaks that might last a long time and the ongoing transmission in the community.

In conclusion, there might be high levels of ongoing transmission of *M. tuberculosis* in Guatemala as indicated by clustering in a convenience sample. Among HIV-infected patients, clustering was more likely in pulmonary disease and when individuals have some level of education as opposed to no education. Moreover, we detected previously unreported strains of *M. tuberculosis* that contribute to tuberculosis morbidity in the country. Further prospective studies in Guatemala and neighboring countries with novel genotyping techniques and larger sampling fractions are urgently needed to further characterize the molecular diversity and transmission dynamics of tuberculosis in the Central American region.

Table 6.1 Type of sample and drug resistance pattern of MTB isolates from patients that attended ASI, Guatemala City, Guatemala from 2010-2014.

Results are shown for the overall population and stratified by HIV status.

Characteristic	All N=423		HIV infected n=140		HIV status negative/unknown n=283	
	N	%	N	%	N	%
Type of sample						
Pulmonary	285	67	94	67	191	67
Extrapulmonary	138	33	46	33	92	33
Isoniazid (n=320)						
Sensitive	278	87	119	94	159	82
Resistant	42	13	8	6	34	18
Rifampicin (n=321)						
Sensitive	308	96	121	95	187	97
Resistant	13	4	7	5	6	3
Pyrazinamide (n=108)						
Sensitive	102	94	85	94	17	94
Resistant	6	6	5	6	1	6
Streptomycin (n=321)						
Sensitive	253	79	106	83	147	76
Resistant	68	21	22	17	46	24
Ethambutol (n=321)						
Sensitive	308	96	122	95	186	96
Resistant	13	4	6	5	7	4
Any drug resistance (n=321)						
No	224	70	92	72	132	68
Yes	97	30	36	28	61	32
MDR† resistance (n=321)						
No	315	98	125	98	190	98
Yes	05	2	2	2	3	2

†MDR=Multidrug-resistant, defined as resistant to at least isoniazid and rifampicin

Table 6.2. Most frequent spoligotypes of M. tuberculosis strains from tuberculosis cases in Guatemala City, Guatemala.

Distribution of the 16 most frequent spoligotypes identified in the study. For each spoligotype it is presented the shared international type (SIT), spoligotype family, binary format, octal code and frequency. Results are shown for the overall population and stratified by HIV status. These 16 genotypes comprised 74% of the total sample.

SIT†	Family *	Binary format	Octal code	Overall n=423		HIV negative /unknown n=283		HIV positive n=140	
				n	Percent	n	Percent	n	Percent
33	LAM	■■■■■■■■□□□■■■■■■■■■■□□□□■■■■■■■■■□□□□■■■■■■■■	776177607760771	92	22	60	21.2	32	22.9
53	T	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■	777777777760771	56	13	36	12.7	20	14.3
42	LAM	■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■■	777777607760771	26	6	18	6.4	8	5.7
50	Haarlem	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■■■■■■■■	777777777720771	25	6	15	5.3	10	7.1
119	X	■■■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■■■■□□□□■■■■■■■■	777776777760771	20	5	13	4.6	7	5.0
1	Beijing	□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■■■	000000000003771	13	3	8	2.8	5	3.6
137	X	■■■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■■■■□□□□■■□□□□■	777776777760601	12	3	6	2.1	6	4.3
2	Haarlem	□□□□□□□□□□□□□□□□□□□□□□□□□□■□□□□□□■□□□□■■■■■■■	000000004020771	12	3	10	3.5	2	1.4
92	X	■■■□□□□□□□□■■■■■□■■■■■■■■■■■■■■■□□□□■■■■■■■■	700076777760771	11	3	10	3.5	1	0.7
GUA01C	EAI	■■■■■■■■■■■■■■■■■□■■■■■■■□□□□□□□□□□□□□□□□□□■	777776770000001	9	2	8	2.8	1	0.7
Psd01C	Haarlem	□□□□□□□□□□□□□□□□□□□□□□□□□□□■□□■□□■□□□□■■■□■■■	000000004620731	9	2	7	2.5	2	1.4
39	T	■■■■■■■■■■■■■■■■■■■□■■■□□■■■■■■■■□□□□■□□■■■■■	777777347760471	7	2	5	1.8	2	1.4
3101	LAM	■■■■■□■■■■□□□■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■■	756177607760771	6	1	1	0.4	5	3.6
425	X	■■■□□□□□□□□□■■■■■□■■■■■■■■□□□□■■■□□□□■■■■■■■■	700076774160771	5	1	4	1.4	1	0.7
47	Haarlem	■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□■□□□□■■■■■■■■	777777774020771	5	1	3	1.1	2	1.4
Psd02C	Unknown	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□□□□□□□■	777777777700001	5	1	4	1.4	1	0.7

*SIT=Shared international type *EAI=East African-Indian, LAM=Latin-America and Mediterranean

Table 6.3. Genotypes not previously reported in the international database.

There were 23 genotypes not previously reported in the spoligotyping database belonging to 35 strains in the study. Of these 35 strains, 24 belonged to patients with negative/unknown status and 11 to HIV-infected patients.

Study name	Family†	Octal code	Overall		HIV negative /unknown		HIV positive	
			n	Percent	n	Percent	n	Percent
GUA01C	EAI	777776770000001	9	26	8	33	1	9
GUA02C	X	700075774160771	2	6	2	8	0	0
GUA03C	LAM	776177607760721	2	6	0	0	2	18
GUA04C	LAM	777347607760331	2	6	2	8	0	0
GUA05C	EAI	777700777413701	2	6	1	4	1	9
GUA06U	LAM	372177607760771	1	3	1	4	0	0
GUA07U	EAI	376177600000001	1	3	1	4	0	0
GUA08U	Family33	676501777177601	1	3	1	4	0	0
GUA09U	T	700000037760771	1	3	0	0	1	9
GUA10U	X	700175777760771	1	3	1	4	0	0
GUA11U	LAM	710777607760771	1	3	1	4	0	0
GUA12U	T	740000037760771	1	3	1	4	0	0
GUA13U	Haarlem	757677777720771	1	3	1	4	0	0
GUA14U	LAM	774167607760631	1	3	0	0	1	9
GUA15U	T	776000000060771	1	3	0	0	1	9
GUA16U	Haarlem	777017777720771	1	3	1	4	0	0
GUA17U	T	777377777700001	1	3	0	0	1	9
GUA18U	T	777417777760701	1	3	1	4	0	0
GUA19U	X	777600377360771	1	3	0	0	1	9
GUA20U	X	777774077560001	1	3	1	4	0	0
GUA21U	T	777775777760671	1	3	1	0	0	0
GUA22U	Family33	777777407756771	1	3	0	0	1	9
GUA23U	Family33	777777760776071	1	3	0	0	1	9

†EAI=East African-Indian, LAM=Latin-America and Mediterranean

Table 6.4. Characteristics of the ten patients with two samples whom were considered as distinct 20 patients.

Patients in bold font had isolates with different genotypes. Patients in plain font had the same genotype with samples submitted in different time periods (> 6 months).

Study ID	HIV[γ]	Sex	Age (yrs)	Type of sample	AFB†	Drug resistance‡	Months between samples	SIT* and family	CD4 count/mm^3	Risk factors	Previous history TB	Death record
97	**Neg/**	**NA**	**NA**	**Bone marrow**	**1+**	**ISNr**	**1**	**70, X**	**NA**	**NA**	**NA**	**NA**
	unk			**Sputum**	**3+**	**NA**		**2, Haarlem**	**NA**			
144	Pos	Man	26	Sputum	3+	Sen	15	91, X	NA		Yes, pulmonary	MTB death
			27	Sputum	No AF	RMPr			23			
148	Pos	Man	39	Sputum	NA	Sen	12	39, T	10	Drugs, alcohol, tobacco		
			40	Node biopsy	3+	Sen			14			
170	Pos	Man	32	Sputum	1+	ISNr, RMPr	29	3101, LAM	NA			
			34	Sputum	2+	ISNr, RMPr, STMr			352			
203	**Pos**	**Man**	**36**	**Stool**	**No AF**	**Sen**	**2**	**Psd02C, Unknown**	**28**		**Yes, pulmonary**	**MTB death**
			36	**Sputum**	**No AF**	**Sen**		**42, LAM**	**28**			
241	Pos	Woman	30	Sputum	No AF	Sen	20	33, LAM	224			
			32	Sputum	No AF	Sen			174			
328	Pos	Man	27	CSF	NA	Sen	7	GUA03C, LAM	NA	Drugs, alcohol, tobacco	Yes, extrapulmonary or disseminated	
			28	CSF	NA	Sen			NA			
334	**Pos**	**Woman**	**24**	**Sputum**	**No AF**	**ISNr**	**22**	**53, T**	**378**	**Prison, alcohol, tobacco, UDVP**		
			26	**Sputum**	**No AF**	**Sen**		**42, LAM**	**NA**			
381	Pos	Man	27	Sputum	No AF	Sen	13	33, LAM	148	Alcohol, tobacco		Suspicion MTB Death
			28	Sputum	No AF	Sen			53			
436	Pos	Man	43	Sputum	1+	Sen	10	50, Haarlem	154	NA		
			44	Sputum	2++	PZAr			NA			

[γ]=Neg/unk=Negative/unknown status, Pos=Positive. †AFB=Acid-fast bacilli result *SIT=Shared international type

‡ Sen=Sensitive all drugs tested. ISN=Isoniazid, RMP=rifampicin, PZA=pyrazinamide, STM=Streptomycin, r=resistant.

NA=Data not available.

Table 6.5. Descriptive statistics comparing clustering of mycobacterial isolates depending years of isolation.

Period years	Length (years)	Total isolates analyzed	Number of clusters	Number of clustered strains	Mean size of clusters (±SD)	Range of cluster	Maximum cluster size	Clusters of size 2 (%)	Proportion of clustering (%)	RTI† "n-1"
2010-2011	2	205	25	179	7.2 (±10.2)	2-49	49	9/25 (36)	87.3	75.1%
2010-2012	3	274	28	238	8.5 (±13.4)	2-65	65	11/28 (39)	86.9	76.6%
2010-2013	4	363	37	319	8.6 (±15.0)	2-81	81	15/37 (40)	87.9	77.7%
2010-2014	5	423	39	371	9.5 (±16.9)	2-92	92	15/39 (38)	87.7	78.5%

† **RTI "n-1"** =recent transmission index. In this table this measurement was done using the traditional 'n-1' method.

Table 6.6. Demographic characteristics of patients with non-clustered and clustered isolates, determined by spoligotyping.

Characteristics	Overall N=120 (%) †	Clustered N=107 (%) *	Non-clustered N=13 (%) *	Overall *p* value¶	Corrected *p* value‡
Gender				0.30	
Male	92 (77)	80 (87)	12 (13)		
Female	28 (23)	27 (96)	1 (4)		
Age, years, continuous					
Age, years, median [IQR]	35.5 [28-44]	35 [28-44]	37 [28-42]	0.75	
Age, years, category				1.0	
<45 years	96 (80)	85 (88)	11 (11)		
≥ 45 years	24 (20)	22 (92)	2 (8)		
Country of birth				1.0	
Guatemala	111 (93)	99 (89)	12 (11)		
Other Central American countries	8 (7)	7 (88)	1 (12)		
No data	1 (1)	1 (100)	0		
Department of residency				0.32	
Guatemala	89 (74)	81 (91)	8 (9)		
Other	31 (26)	26 (84)	5 (16)		
Civil status				0.63	
Single	60 (50)	52 (87)	8 (13)		
Married	14 (12)	14 (100)	0		
Free union	33 (28)	29 (88)	4 (12)		
Widowed	6 (5)	5 (83)	1 (17)		
Separated	6 (5)	6 (100)	0		
No data	1 (1)	1 (100)	0		
Ethnic group				0.34	

Characteristics	Overall N=120 (%) †	Clustered N=107 (%) *	Non-clustered N=13 (%) *	Overall *p* value¶	Corrected *p* value‡
Ladino	61 (51)	54 (88)	7 (11)		
Mayan	14 (12)	11 (79)	3 (21)		
Other	1 (1)	1 (100)	0		
Missing	44 (37)	41 (93)	3 (7)		
Sexual orientation				0.77	
Heterosexual	93 (78)	82 (88)	11 (12)		
Homosexual	14 (12)	12 (86)	2 (14)		
Bisexual	9 (8)	9 (100)	0		
No data	4 (3)	4(100)	0		
Education level				0.08	
None	19 (16)	14 (74)	5 (26)		Ref
Some education	98 (82)	90 (92)	8 (8)		0.07
No data	3 (3)	3 (100)	0		1.00
Employment status				0.71	
Informal sector	54 (45)	46 (85)	8 (15)		
Housewife	20 (17)	19 (95)	1 (5)		
Unemployed	21 (18)	20 (95)	1 (5)		
Permanent employment	11 (9)	9 (82)	2 (18)		
Casual employment	9 (8)	8 (89)	1 (11)		
Student	1 (1)	1 (100)	0		
No data	4 (3)	4 (100)	0		

†Column percentage. * Row percentage per each category.

¶ Chi square (categorical variables), Fisher Exact text (categorical variables, expected counts less than 5) or Wilcoxon test (continuous variables)

‡ *p* value with the Bonferroni correction for the pairwise comparison between given category and the reference (ref) group

Table 6.7. Behavioral and clinical characteristics of patients with non-clustered and clustered isolates, determined by spoligotyping.

Characteristics	Overall N=120 (%) †	Clustered N=107 (%) *	Non-clustered N=13 (%) *	Overall *p* value¶	Corrected *p* value‡‡
Drugs				0.55	
No	93 (78)	84 (90)	9 (10)		
Yes	23 (19)	19 (83)	4 (17)		
Don't' answer/No data	4 (3)	4 (100)	0		
Alcohol consumption				0.22	
No	37 (31)	35 (95)	2 (5)		
Yes	64 (53)	57 (89)	7 (11)		
Don't' answer/No data	19 (16)	15 (79)	4 (21)		
Tobacco consumption				0.14	
No	54 (45)	51 (94)	3 (6)		Ref
Yes	46 (38)	40 (87)	6 (13)		0.60
Don't' answer/No data	20 (17)	16 (80)	4 (20)		0.16
Had been in prison				0.31	
No	67 (56)	62 (93)	5 (7)		
Yes	12 (10)	11 (92)	1 (8)		
Don't' answer/No data	41 (34)	34 (83)	7 (17)		
Smear result				0.04	
Negative	67 (56)	57 (85)	10 (15)		Ref
Positive	43 (36)	42 (98)	1 (2)		0.10
No data	10 (8)	8 (80)	2 (20)		1.0

Characteristics	Overall N=120 (%) †	Clustered N=107 (%) *	Non-clustered N=13 (%) *	Overall *p* value¶	Corrected *p* value‡‡
Type of sample submitted to TB diagnosis				0.005	
Pulmonary	81 (68)	77 (95)	4 (5)		
Extra-pulmonary	39 (33)	30 (77)	9 (23)		
Previous history of TB				0.61	
No records	109 (91)	96 (88)	13 (12)		
Yes	11 (9)	11 (100)	0 (0)		
CD4 count (±6 months TB diagnosis)				0.21	
Median count [IQR], cells/mm3	62 [28-148]	73 [27-170]	33 [28-66]		
CD4 count (±6 months TB diagnosis), category				0.21	
CD4 <200 cells, mm3	76 (63)	65 (86)	11 (14)		
CD4 ≥200 cells, mm3	17 (14)	17 (100)	0		
No data	27 (23)	25 (93)	2 (7)		
Viral load (±6 months TB diagnosis), continuous				0.12	
Median [IQR], $\log_{10}$ viral copies/mm^3	5.2 [4.5-5.7]	5.1 [3.2-5.7]	5.4 [5.2-5.5]		
Viral load (±6 months TB diagnosis), category				0.03	
Low viral load (<4.0 $\log^{10}$/mm^3)	15 (13)	15 (100)	0		Ref
High viral load (≥4.0 $\log^{10}$/mm^3)	53 (44)	43 (81)	10 (19)		0.20

Characteristics	Overall N=120 (%) †	Clustered N=107 (%) *	Non-clustered N=13 (%) *	Overall *p* value¶	Corrected *p* value‡‡
No data	52 (43)	49 (94)	3 (6)		1.0
HIV clinical stage (Baseline)				0.30	
Stage 1	21 (18)	21 (100)	0		
Stage 2	35 (29)	31 (89)	4 (11)		
Stage 3	63 (53)	54 (86)	9 (14)		
No data	1 (1)	1 (1)	0		
HIV clinical stage (Time visit)				0.30	
Stage 1	7 (6)	7 (100)	0		
Stage 2	24 (20)	23 (96)	1 (4)		
Stage 3	78 (65)	66 (85)	12 (15)		
No data	11 (9)	11 (100)	0		
Discharge motive				0.80	
Not recorded	47 (39)	43 (91)	4 (9)		
Death	42 (35)	37 (88)	5 (12)		
Loss of contact	23 (19)	20 (87)	3 (13)		
Transfer	6 (5)	5 (83)	1 (17)		
Patient decision	2 (2)	2 (100)	0		
Isolation MTB culture				0.20	
2010	29 (24)	23 (79)	6 (21)		Ref
2011	24 (20)	23 (96)	1 (4)		0.45
2012	25 (21)	24 (96)	1 (4)		1.00
2013	24 (20)	20 (83)	4 (17)		1.00

Characteristics	Overall N=120 (%) †	Clustered N=107 (%) *	Non-clustered N=13 (%) *	Overall *p* value¶	Corrected *p* value‡
2014	18 (15)	17 (94)	1 (6)		0.90
Drug resistance				0.89	
Any resistance	28 (23)	26 (93)	2 (7)		
No resistance	82 (68)	72 (88)	10 (12)		
Missing	10 (8)	9 (90)	1 (10)		
MDRγ resistance				1.0	
Yes	2 (2)	2 (100)	0		
No	107 (89)	95 (89)	12 (11)		
Missing	11 (9)	10 (91)	1 (9)		

†Column percentage. * Row percentage per each category.

¶ Chi square (categorical variables), Fisher Exact text (categorical variables, expected counts less than 5) or Wilcoxon test (continuous variables)

‡ *p* value with the Bonferroni correction for the pairwise comparison between given category and the reference (ref) group.

γ MDR=Multidrug-resistant, defined as resistant to at least isoniazid and rifampicin

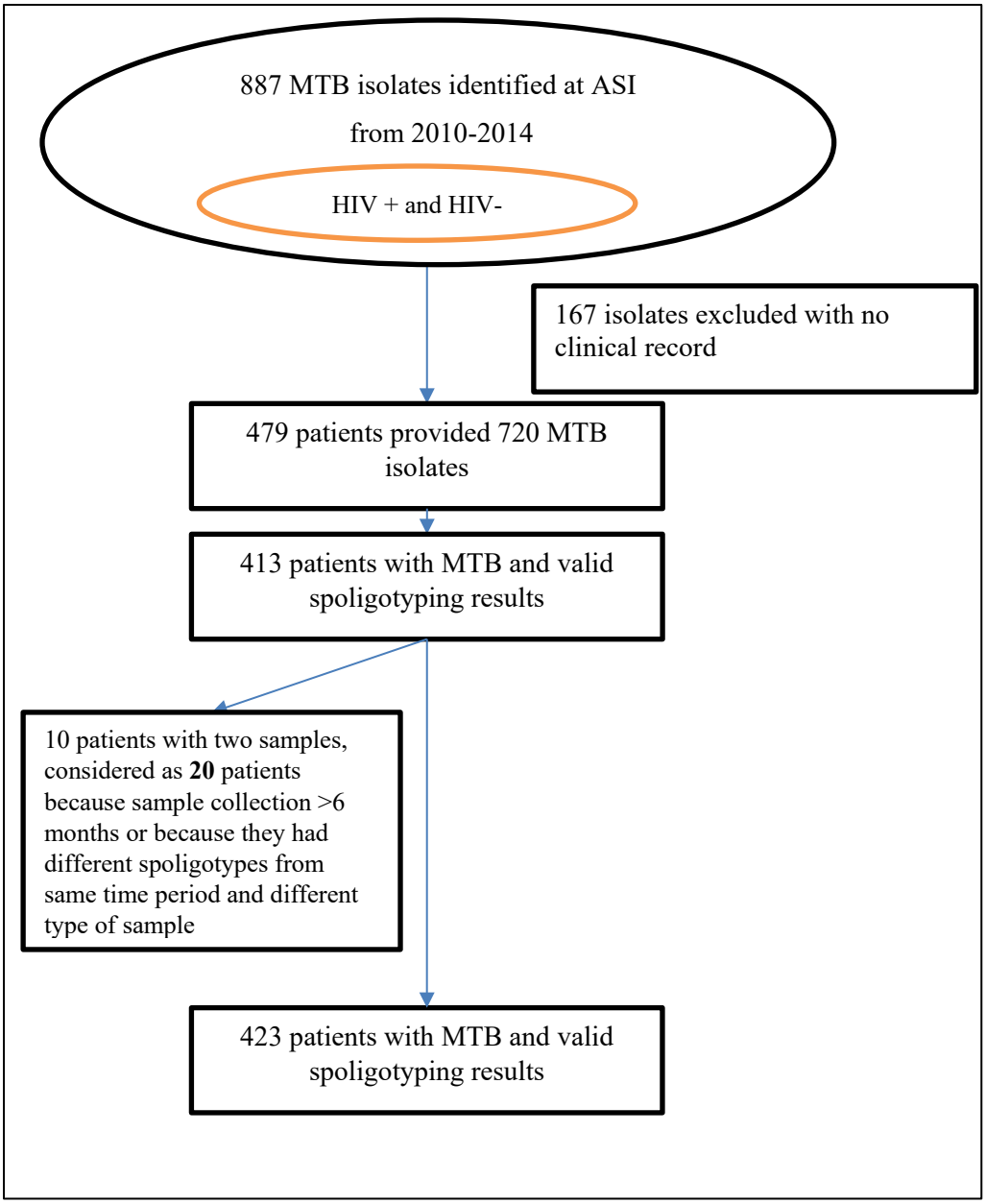


Figure 6.1. Flow diagram of the study. From 2010-2014, 887 M. tuberculosis isolates were identified at the tuberculosis laboratory at ASI, Guatemala City, Guatemala.

Out of these, 720 MTB isolates belonged to 479 tuberculosis cases. After data cleaning, we included in this study 423 tuberculosis cases with a valid spoligotyping result and estimated the proportion of clustering cases.

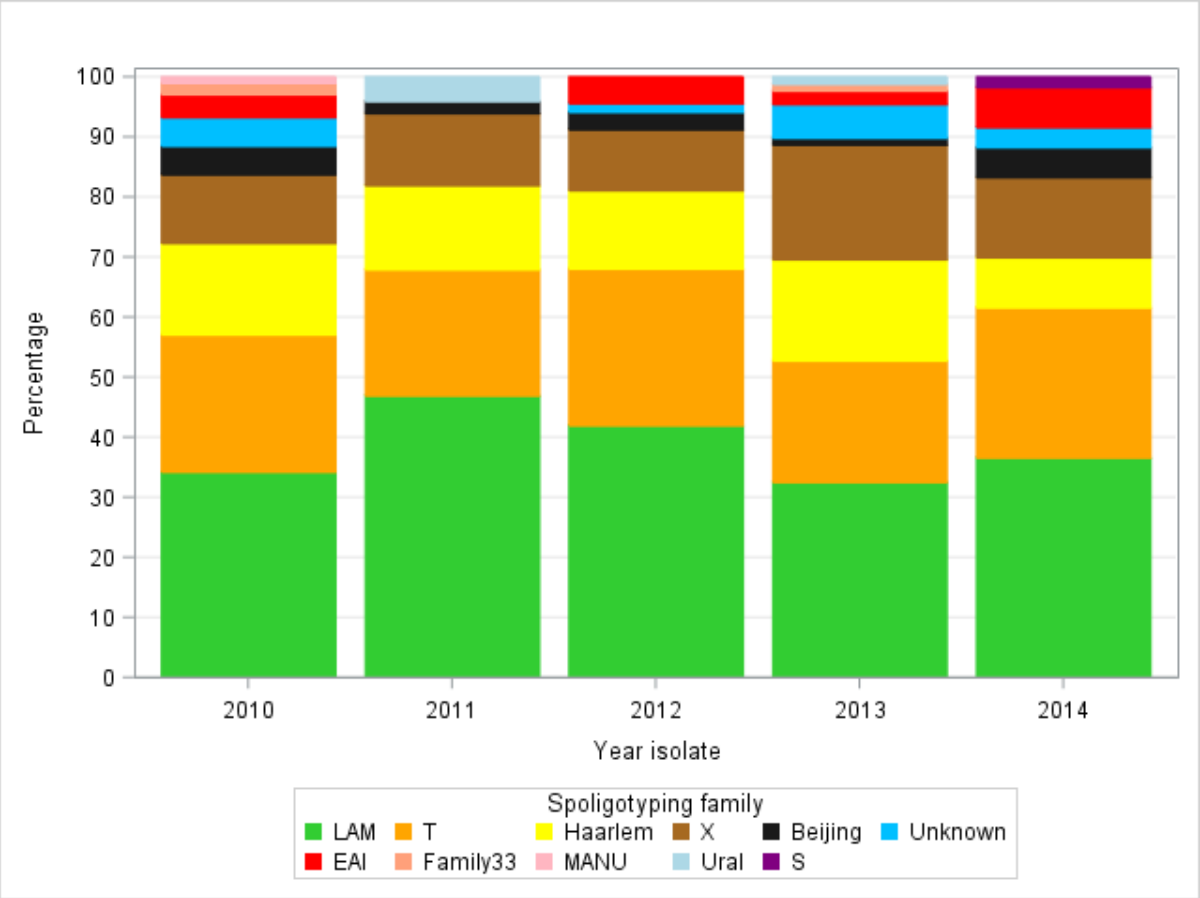
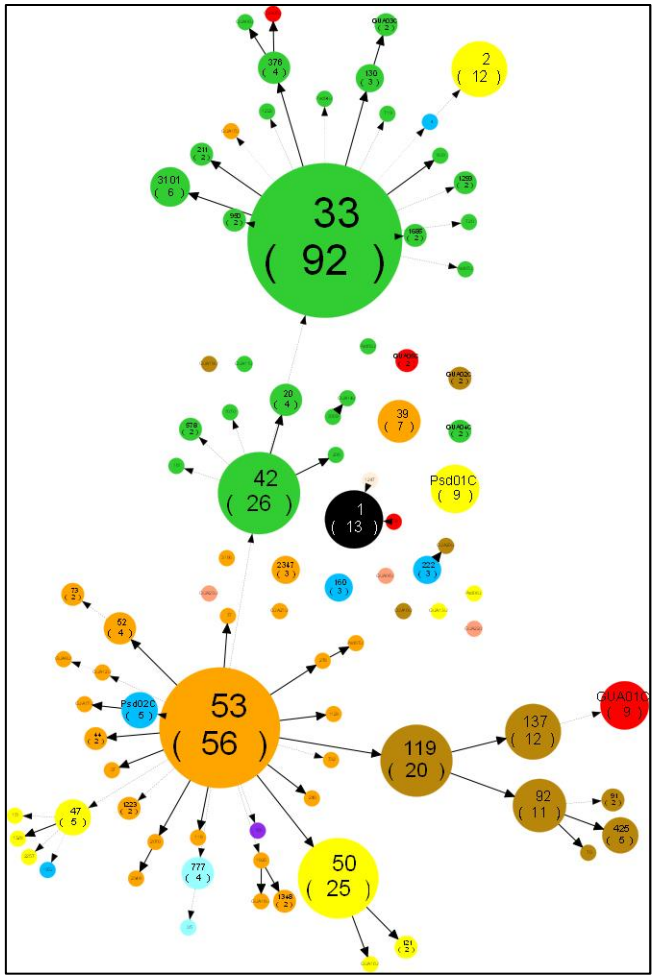


Figure 6.2. Spoligotypes lineages from MTB isolates of patients from the TB laboratory at ASI, Guatemala City, Guatemala, 2010-2014.

Each bar is divided into colored segments indicating the relative percentage of each spoligotyping family detected each year.



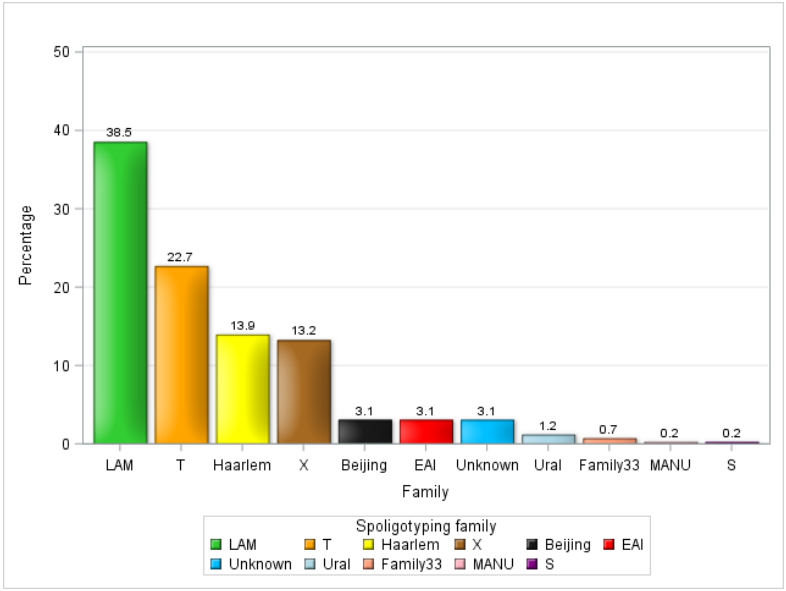


Figure 6.3. Spoligoforest tree based on spoligotypes collected from 2010-2014 (n = 423 isolates).

Each node represents a spoligotype. The size of each node represents the number of strains that belongs to a given spoligotype. Each node is labelled with the SIT/study name. The number in brackets reflects the cluster size. Directed edges indicate single-event deletion that relates to two clusters, the arrowheads pointing to the descendants. Edges with weights less than 0.5 are presented as dotted lines. Edges with weights ≥0.5 and <1 are shown as dashed lines. Edges with weights of 1 are represented as solid lines. The color of the nodes indicates the family of the cluster, as indicated in the pie chart. The percentages in the bar chart indicate the proportion of the given spoligotyping family in the study sample.

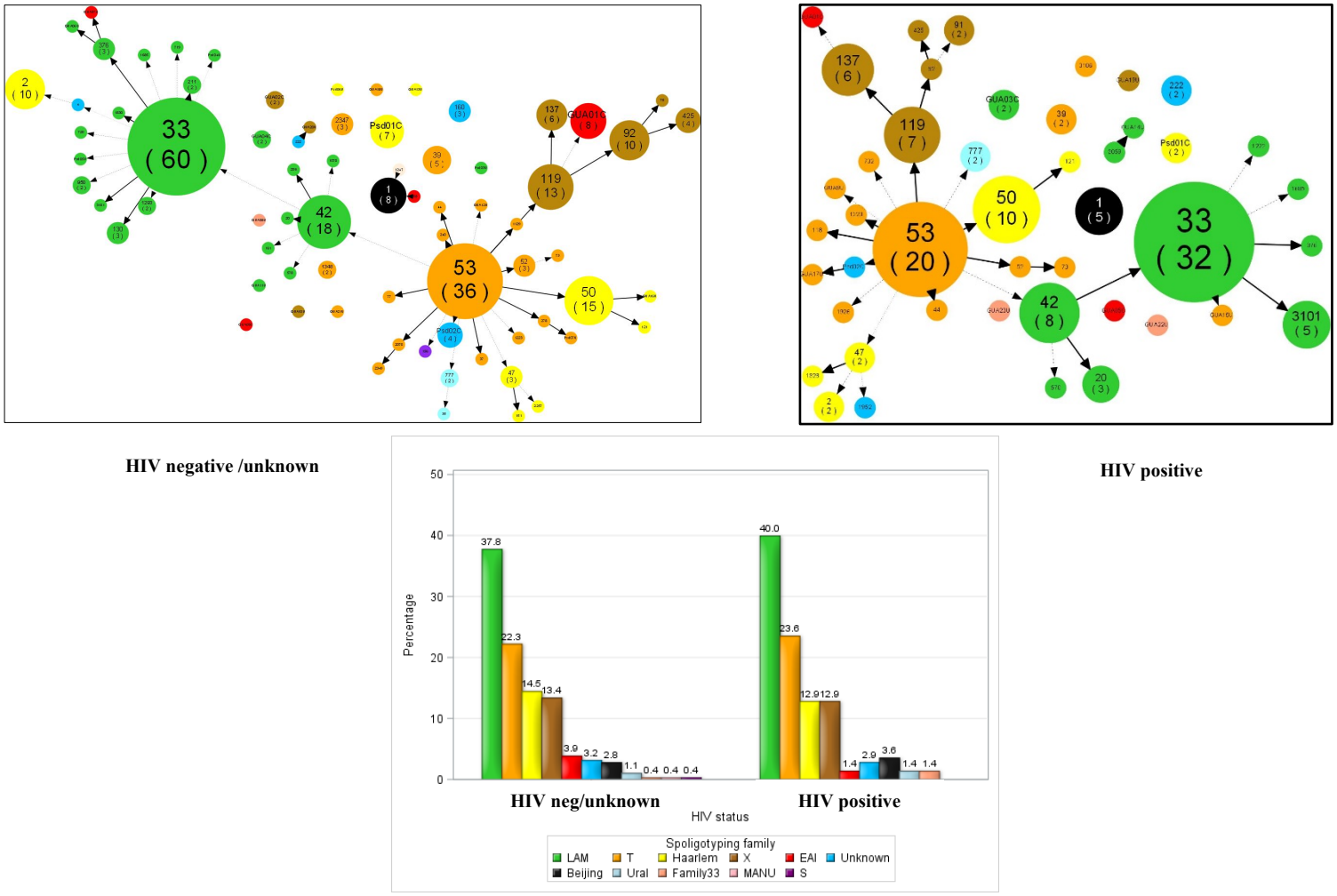


Figure 6.4. Spoligoforest trees based on spoligotypes collected from 2010-2014, stratified by HIV status.

Each node represents a spoligotype. The size of each node represents the number of strains that belongs to a given spoligotype. Each node is labelled with the SIT/study name. The number in brackets reflects the cluster size. Directed edges indicate single-event deletion that relates to two clusters, the arrowheads pointing to the descendants. Edges with weights less than 0.5 are presented as dotted lines. Edges with weights ≥0.5 and <1 are shown as dashed lines. Edges with weights of 1 are represented as solid lines. The color of the nodes indicates the family of the cluster, as indicated in the pie chart. **Left panel:** HIV status negative or unknown. **Right panel:** HIV status positive. **Bottom panel:** Proportion of the given spoligotyping families stratified by HIV status.

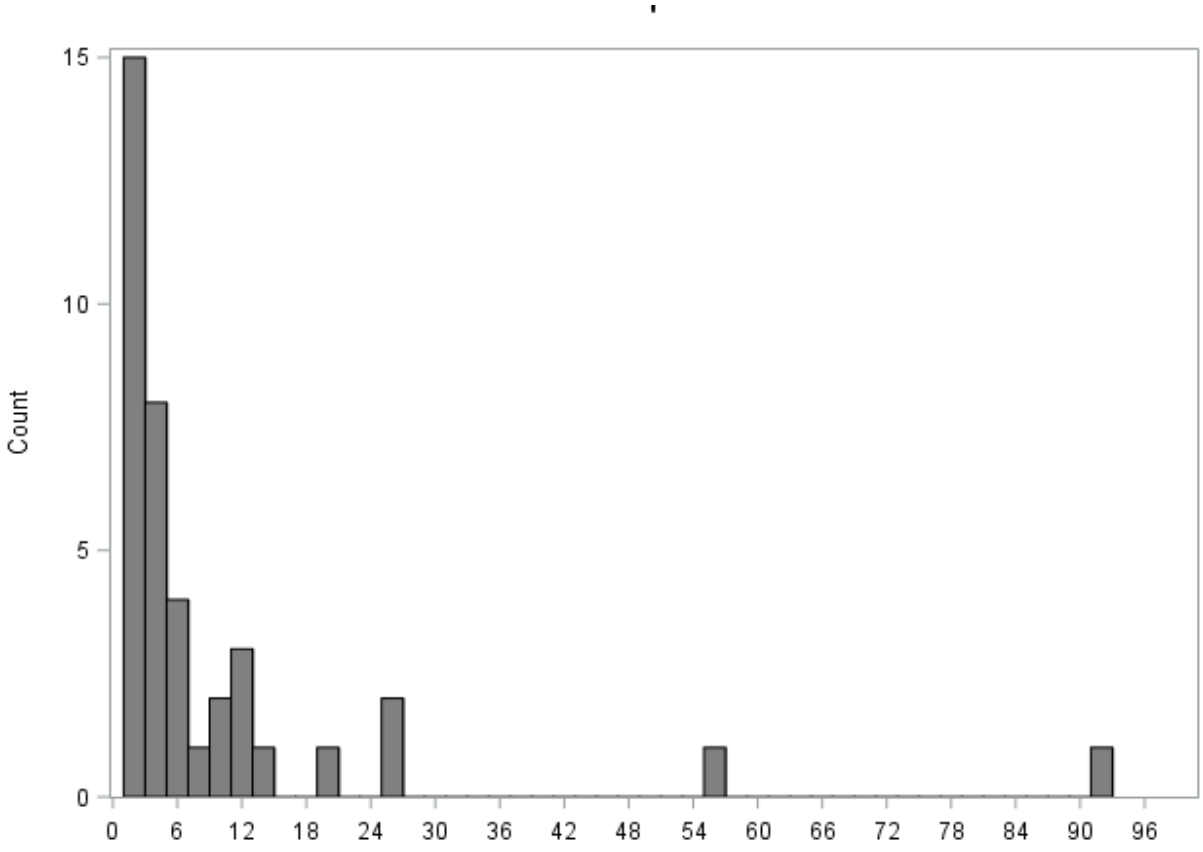


Figure 6.5. Cluster size distribution among the patients with clustered mycobacterial strains isolated from 2010-2014.

In the X axis, the cluster size is given. The Y axis represent the frequency of the given cluster size. The range of the cluster size was 2-92 isolates.

SUPPLEMENTARY MATERIAL

Table S 6.1List of demographics, clinical and behavioral variables collected for the HIV-infected individuals.

Variables	Categories	Time collection	Comment
DEMOGRAPHIC			
Sex	Male, female	Baseline	
Age at time of tuberculosis diagnosis	Continuous and categorized as < 45 yrs old and ≥ 45 yrs old	Time of tuberculosis diagnosis	It has been found that persons ≥ 45 yrs old had higher proportion of clustering (Glynn et al., 2005).
Ethnic group	Ladino, Mayan, other	Baseline	
Level of education	None, some level	Baseline	
Place of birth	Guatemala, other Central American countries	Baseline	
Department of residence	Guatemala, other	Baseline	
Sexual orientation	Heterosexual, homosexual, bisexual	Baseline	
Employment status	Informal sector, housewife, unemployed, permanent employment, causal employment, student	Baseline	
Civil status	Single, married, free union, widowed, separated	Baseline	
CLINICAL			
CD4 cell count/mm^3	Continuous and categorized as CD4 <200 cells, CD4 ≥200 cells	± 3 months of tuberculosis diagnosis	
History of Tuberculosis	Yes, no record	any episode recorded at least 6 months prior current one	
Type of sample submitted for tuberculosis diagnosis	Pulmonary or extra pulmonary	Time of tuberculosis diagnosis	Patients with both pulmonary and extra pulmonary samples were included in the extra pulmonary group.
Smear result	Negative, positive		
HIV viral log_{10} copies/mm^3	Continuous and categorized as <4.0 log_{10} copies, ≥ 4.0 log_{10} copies	± 3 months of tuberculosis diagnosis	
HIV clinical stage	Stage 1, stage 2, stage 3	Baseline	

Variables	Categories	Time collection	Comment
HIV clinical stage	Stage 1, stage 2, stage 3	± 3 months of tuberculosis diagnosis	
Discharge motive from ASI	Death, loss of contact, transfer, patient decision	During or after tuberculosis diagnosis	
Year of tuberculosis culture	2010, 2011, 2012, 2013, 2014	Time of tuberculosis diagnosis	
Drug resistance of isolate	Yes to at least one the following drugs if tested-isoniazid, rifampicin, streptomycin, pyrazinamide, ethambutol-, no resistance	Time of tuberculosis diagnosis	
Multi-drug (MDR)	Yes, no	Time of tuberculosis diagnosis	Multi-drug resistance was defined simultaneous resistance to rifampicin and isoniazid.
BEHAVIORAL			
Use of drugs	Yes, no	Collected either before tuberculosis diagnosis or up to 3 months after tuberculosis diagnosis	
Alcohol consumption	Yes, no		
Tobacco consumption	Yes, no		
Had been in a prison	Yes, no		

Table S 6.2. Frequency of the five most common spoligotypes present in the study in the SpolSimilaritySearch† tool, according to the presence or absence of the last spacer.

Binary pattern in the study, with the last spacer represented as X	Frequency spoligotypes in SpolSimilaritySearch tool		
	Last spacer present	Last spacer missing	Total isolates
11111111000111111111000011111111000011111X	1,113 (99.8)	2 (0.2)	1,115 (100)
11111111111111111111111111111111000011111X	6,152 (99.8)	15 (0.2)	6,167 (100)
11111111111111111111000011111111000011111X	3,275 (99.9)	4 (0.1)	3,279 (100)
11111111111111111111111111111010000111111X	3,309 (99.8)	7 (0.2)	3,316 (100)
11111111111111111011111111111111000011111X	1,077 (99.6)	4 (0.4)	1,081 (100)

†Available at: http://www.pasteur-guadeloupe.fr:8081/SpolSimilaritySearch/index.jsp

Table S 6.3. Complete list of 91 spoligotypes identified among 423 tuberculosis cases from Guatemala City, Guatemala (2010-2014).

The shared international type (SIT), binary format of the spoligotype pattern, frequency and relative percentage in this study are presented.

SIT†	Binary pattern	Frequency	Percent
33	1111111100011111111100001111111100001111111	92	21.7
53	1111111111111111111111111111111100001111111	56	13.2
42	1111111111111111111000011111111000011111111	26	6.1
50	1111111111111111111111111111101000011111111	25	5.9
119	1111111111111111011111111111111000011111111	20	4.7
1	0000000000000000000000000000000000111111111	13	3.1
137	1111111111111111011111111111111000011000001	12	2.8
2	0000000000000000000000010000001000011111111	12	2.8
92	1110000000001111101111111111111000011111111	11	2.6
GUA01C	1111111111111111011111100000000000000000001	9	2.1
Psd01C	0000000000000000000000010011001000011101111	9	2.1
39	1111111111111111101110011111111000010011111	7	1.7
3101	1111011100011111111100001111111100001111111	6	1.4
425	1110000000001111101111110000111000011111111	5	1.2
47	1111111111111111111111110000001000011111111	5	1.2
Psd02C	1111111111111111111111111111100000000000001	5	1.2
20	1101111111111111111100001111111100001111111	4	0.9
376	0111111100011111111100001111111100001111111	4	0.9
52	1111111111111111111111111111111100001110111	4	0.9
777	1111111111111111111111111111000100001111111	4	0.9
130	1111111100011111111100001111111100001110111	3	0.7
160	1111111111111111110011111110000000000000001	3	0.7
222	1111111111111111000001111111011100001111111	3	0.7
2347	1111111100000000000000000111101100001111111	3	0.7
121	1111111111111111111111110111101000011111111	2	0.5
1223	1111111100001111111111111111111111000011111111	2	0.5
1293	1101111100011111111100001111111100001111111	2	0.5
1348	1111111000000111111111111111111100001110001	2	0.5

SIT†	Binary pattern	Frequency	Percent
1685	11111111000011111111000011111111000001111111	2	0.5
211	11111111000101111111000011111111000001111111	2	0.5
44	11111111111111111111101111111111000001111111	2	0.5
578	11001111111111111111000011111111000001111111	2	0.5
73	11111111111101111111111111111111100001110111	2	0.5
91	11100000000001111011111111111111000001111111	2	0.5
950	11100000000001111111000011111111000001111111	2	0.5
GUA02C	11100000000011110111111100001110000111111	2	0.5
GUA03C	11111111000111111111000011111111000001110101	2	0.5
GUA04C	11111111101110011111000011111111000000110111	2	0.5
GUA05C	11111111111100000011111111110000101111110001	2	0.5
1070	11111111111111111111000011111111000000111111	1	0.2
1129	11111111011111111111111111111111000001111111	1	0.2
118	11111111111111011111111111111111000001111111	1	0.2
1222	11111100000111111111000011111111000001111111	1	0.2
1247	11111111111111111111000011111111001111111111	1	0.2
1328	11111111111111111111000011100000010000111111	1	0.2
151	11111111111111011111111111000000100001111111	1	0.2
156	11111111000111111111111111111111000001111111	1	0.2
161	11111111111111111111000011111111000001111111	1	0.2
1830	11111111000111111101000011111111000001111111	1	0.2
1926	11111111111111111111111111111111100001110001	1	0.2
1952	11111111111111111111111100000000000001111111	1	0.2
206	11110000011111111111000011111111000001111111	1	0.2
2070	11111111111111111111001111111111000001111111	1	0.2
2257	11111111111100000000000010000001000001111111	1	0.2
2349	11001111111111111111001111111111000001111111	1	0.2
240	11111111111111111111111111111111000001111111	1	0.2
278	11111111111111111111111111111111100001111101	1	0.2
3059	11111110000111011111000011111111000001111111	1	0.2
3186	11111111111111111111111010011111000001111111	1	0.2
35	11111111111101111111111111111000100001111111	1	0.2
37	11111111111101111111111111111111000001111111	1	0.2

SIT†	Binary pattern	Frequency	Percent
4	00000000000000000000000011111111000011111111	1	0.2
458	11111111111111111111111111110000001111111111	1	0.2
70	11100000000011111011111111111111000011011111	1	0.2
719	11111111000111111111000001111111100001111111	1	0.2
720	11111111000111111111000011111111000001111111	1	0.2
732	11111111111111001111111111111111000011111111	1	0.2
77	11110001111111111111111111111111000011111111	1	0.2
GUA06U	01111101000111111111000011111111000011111111	1	0.2
GUA07U	01111111000111111111000000000000000000000001	1	0.2
GUA08U	11011111010100000111111111100111111111100001	1	0.2
GUA10U	11100000000111110111111111111111000011111111	1	0.2
GUA11U	11100100011111111111000011111111000011111111	1	0.2
GUA12U	11110000000000000000011111111111000011111111	1	0.2
GUA13U	11110111111011111111111111111101000011111111	1	0.2
GUA14U	11111110000111011111000011111111000011001111	1	0.2
GUA15U	11111111000000000000000000000011000011111111	1	0.2
GUA16U	11111111100000111111111111111101000011111111	1	0.2
GUA17U	11111111101111111111111111111000000000000001	1	0.2
GUA18U	11111111110000111111111111111111000011100001	1	0.2
GUA19U	11111111111000000001111111101111000011111111	1	0.2
GUA20U	11111111111111110000011111110111000000000001	1	0.2
GUA21U	11111111111111110111111111111111000011011111	1	0.2
GUA22U	11111111111111111100000111111101110111111111	1	0.2
GUA23U	11111111111111111111110000111111111000011111	1	0.2
GUA9U	11100000000000000000011111111111000011111111	1	0.2
Psd03U	11010111111111111111000011100111000011111111	1	0.2
Psd04U	11100000000000000000000011111111000011111111	1	0.2
Psd05U	11111100000001111111000011111111000011111111	1	0.2
Psd06U	11111111111111111111111110011001000011111111	1	0.2
Psd07U	11111111111111111111111111111111000011011101	1	0.2

†SIT=Shared international type

Table S 6.4. Proportion of tuberculosis cases due recent transmission, using online tool developed by Kasaie and others.

Different scenarios according to estimated incidence in the region and sampling rate. In the Appendix 1 the calculations for these findings are shown.

Estimated incidence 100,000/yrs	Estimated sampling rate	RTI "n-1†
25	10% cases	92%
25	20% cases	90%
60	10% cases	92%
60	20% cases	90%

† **RTI "n-1"** =recent transmission index. In this table this measurement was done using the online tool developed by Kasaie and others.

Appendix 1. Proportion of tuberculosis cases due recent transmission, using online tool developed by Kasaie and others.

- ✓ The equation is presented and an example of the calculation step by step is shown for the scenario of 25 cases/100,000/year tuberculosis incidence with a sampling rate of 20% of the cases.
- ✓ The notation and parameters need to solve the equation are also presented.
- ✓ The results for the four scenarios are also show.

▪ Equation: Y = P (-0.208) + D (-0.00759) + c (0.879) + I (-0.0000266) + n (-0.794) + 0.281												
▪ Y= 0.2 (-0.208) + 5(-0.00759) + 0.877 (0.879) + 25(-0.0000266) + 0.0922 (-0.794) + 0.281												
▪ Y= -0.0416 -0.03795 + 0.771 -6.66X 10-4 -0.07045 + 0.281										c	0.877069	C/SS
Y = 0.892 (89%)										n	0.092199	N/SS
Total number of TB cases in the sample:				SS		423						
Number of clustered cases observed in the sample:				C		371						
Number of clusters observed in the sample:				N		39						
Proportion of total active TB cases who have culture and fingerprint data				P		0.1	0.2					
Length of time over which samples were collected (years):				D		5						
TB Incidence (per 100,000/yrs):				I		25	60					
	Equation	n-1	Scenario									
	Y	0.919323	10% cases, 25/100,000/yrs incidence									
	Y	0.898523	20% cases, 25/100,000/yrs incidence									
	Y	0.918392	10% cases, 60/100,000/yrs incidence									
	Y	0.897592	20% cases, 60/100,000/yrs incidence									
						▪ c=C/SS =371/423 = 0.877						
						▪ n=N/SS = 39/423 =0.0922						

Chapter 7
SYNTHESIS OF RESULTS AND MAJOR CONCLUSIONS

This dissertation was designed to further our understanding of what factors contribute to tuberculosis incidence in low-income settings. We founded our work in the model of the cycle of tuberculosis transmission, studying factors that affect both individual and population levels (Figure 7.1).

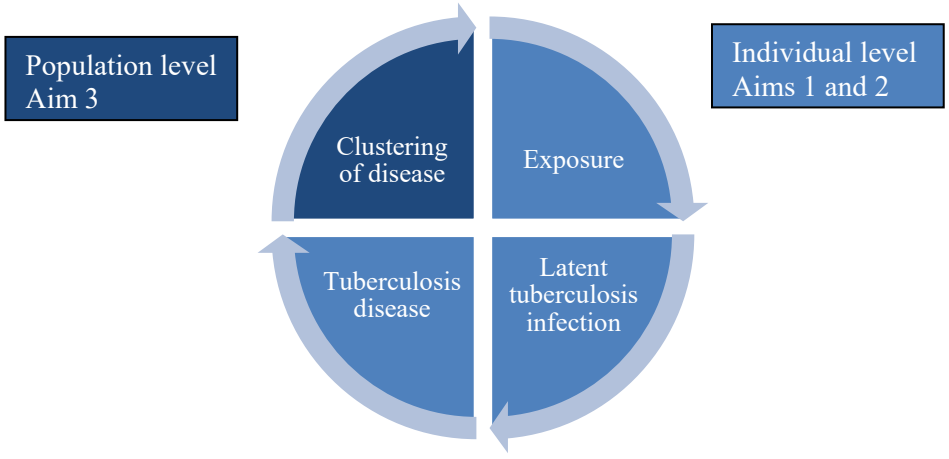


Figure 7.1. Cycle of tuberculosis transmission and aims of this study

At the individual level, exposure leads to infection and then disease in a sub-set of the infected individuals (Mathema et al., 2017). Incident tuberculosis cases represents new cases of tuberculosis disease that are either reactivation events of past infections or recent transmission events which progressed in primary disease in individuals who were not able to contain recent infection (Christopher C Whalen, 2016). The transmission of tuberculosis is complex as well as its monitoring, but we know that the risk from exposure to tuberculosis infection increases according to the level of contact between an infective individual and a susceptible host (Mathema et al., 2017; Thomas & Weber, 2001; Turner et al.,

2017). However, the definition and measurement of adequate contact for transmission of tuberculosis is still poorly understood. In the first aim of the study we defined adequate contact for tuberculosis transmission in Rubaga Division, Uganda, an African urban environment, by examining the interaction within ego-centric networks and develop a score that measures the degree of contact. For our second aim we determined whether the score correlates with the prevalence of tuberculosis infection in the social network of the tuberculosis cases. At the population level, clustering of cases might indicate a recent transmission chain that feeds again into new infections. In the third aim of this research we estimated the level of tuberculosis transmission by measuring the proportion of clustered tuberculosis cases based on genotypic matching in Guatemala City, Guatemala between 2010 and 2014 and identified risk factors associated with recent transmission in HIV-infected individuals, providing for the first time an insight in the molecular epidemiology of tuberculosis in this middle-burden country.

MAJOR FINDINGS AND THEIR SIGNIFICANCE

In our first aim we described the social interaction of 120 tuberculosis cases and their social networks, identifying twelve variables highly inter-correlated. Subsequently, we used exploratory factor analysis to identify domains of these variables. We identified and named two domains, the setting and relationship domains. The setting domain involves the type, frequency, duration and ventilation of the usual place of meeting as well the physical proximity among tuberculosis cases and contacts, represented by the sleeping and eating patterns. The relationship domain was explained by the relationship duration as well as the level of intimacy among cases and contacts, represented by the strength of knowledge of each other, provision of healthcare, and if they were travel buddies. Based on the information provided by the tuberculosis cases about the nature of their interactions we developed domain scores. We were able to show that this score has construct validity and that confirms the complex and heterogeneous social mixing between tuberculosis cases and their social network.

For our second aim, we proved that these two domains affected the likelihood of infection with *M. tuberculosis* for members of a social network of a tuberculosis case, particularly children contacts.

Children aged 5-14 years with high scores in the setting domain and children aged 0-4 years with high scores in the relationship domain had higher probability of tuberculosis infection. Our score worked in both the household and non-household contexts. Non-household contacts with high scores in the setting and relationship domains were found to have higher prevalence of tuberculosis than in the general population.

The setting and relationship domains that we identified might be considered component causes from a Causal Pie, the model proposed to understand the multi-factorial nature of many diseases (Dicker, Coronado, Koo, & Parrish, 2006; Rothman, 1976). In this model, each of the individual factors that increase the risk for having a disease are called 'component causes' and are represented as a slide of the pie. A sufficient cause is the set of component causes that will complete a causal pie, leading to an outcome (Wensink, Westendorp, & Baudisch, 2014). A necessary cause would be a component cause that it is always required for the outcome to occur (Parascandola & Weed, 2001).

In the case of tuberculosis infection, the necessary cause would be the presence of the infectious agent, *M. tuberculosis*, but there are many host and environmental factors that might be considered component causes (Narasimhan et al., 2013; Nelson & Williams, 2006). High scores in the setting or relationship domain correspond with more extensive contact between tuberculosis case and a susceptible host. As the effect of these two factors varied by age, we suggest the setting domain represents a component cause for children aged 5-14 years and the relationship domain represents a component cause for the infant population.

The major findings of our third aim show a high level of ongoing transmission of *M. tuberculosis* as indicated by clustering in a convenience sample of tuberculosis patients. Among HIV-infected patients, clustering was more likely in pulmonary disease and when individuals have some level of education as opposed to no education. Moreover, we also described for the first time, the molecular epidemiology of tuberculosis in Guatemala, a middle-burden country. The most prevalent genotypes identified in this study belong to the Euro-American phylogenetic lineage 4, like the ones found in other

Latin American countries. We described previously unreported strains of *M. tuberculosis* that contribute to tuberculosis morbidity in Guatemala. Finally, our study shown the presence of mixed infections in patients with both pulmonary and extra-pulmonary tuberculosis.

This dissertation has provided a deeper insight into components of the tuberculosis transmission cycle. First, to our knowledge is the first time that a method is developed to quantitatively measure adequate contact between infectious cases and susceptible hosts. In the past, it has been assumed that social contacts of tuberculosis cases, particularly household contacts have similar intensities of exposure, leading to homogenous risk of infection. We have shown there is heterogeneity in the contact patterns among social contacts of tuberculosis cases. A strength of our work is that we used an agnostic approach toward the development of the domain scores, although validation in other populations and areas is still needed.

Second, and despite its exploratory nature owed to the low discrimination of the genotyping technique utilized, the molecular epidemiology study offers some insight into the reason for the stagnant incidence rates in Guatemala. This country like other Latin American countries, has focused primarily its activities for tuberculosis control in the detection and treatment of the cases (PAHO., 2016). However, the median delay in diagnosis of pulmonary tuberculosis in low and middle-income settings has been estimated to be 67-70 days (Getnet, Demissie, Assefa, Mengistie, & Worku, 2017; Sekandi et al., 2015). The implication of this is crucial, the more time an infectious tuberculosis patient goes undetected, the higher the transmission probability of infection. In developing countries, the risk to progress from tuberculosis infection to tuberculosis disease is increased by social and economic factors such as HIV and malnutrition (Lonnroth & Raviglione, 2008). The Replacement Principle of tuberculosis prescribes that 'Only by preventing new cases of disease will be able to reduce and ultimately extinguish an epidemic" (Christopher C Whalen, 2016). The findings of our molecular study appear to support this argument and call for a change in the current local policies, incorporating new measures that either interrupt transmission or focus in the detection and treatment of latent tuberculosis infection.

FUTURE DIRECTIONS

A natural progression of this work is to validate our method to measure adequate contact in settings outside of Africa. Social mixing might differ according to cultures (Auld, Shah, Cohen, Martinson, & Gandhi, 2018), so further research is welcomed to understand its performance in other contexts and other populations from middle- and high-burden countries. We would propose a similar approach like ours, a social network data collection from tuberculosis cases and exploratory factor analysis to first identify the domains more relevant for adequate contact in those areas. As a second step, the agreement of these domains with tuberculosis infections in the social contacts should be performed.

Likewise, tuberculosis seems to be driven by the interactions of tuberculosis cases with their social contacts in the community, not in the household (Martinez et al., 2017). Nevertheless, in our work only a small fraction of non-household contacts had high setting and relationship scores. This indicates there is still a need to further understand the structures and dynamics of the non-household social networks of tuberculosis cases. The Mobility and Tuberculosis Transmission Study-MATTRS study (Whalen CC, http://grantome.com/grant/NIH/R01-AI093856-06A1), which be launched this year, will trace the movement of pulmonary cases in the community before their diagnosis, using cellular telephone metadata and whole genome sequencing. The potential identification of 'hot-spots' for tuberculosis transmission in the community is warranted to develop health policies that more effectively control the infection. Other areas that need further research includes the evaluation of the role and proportion that causal contacts and super-spreaders have in regard to tuberculosis transmission in the community (McCreesh & White, 2018; Wang et al., 2014)

To confirm the levels of transmission of tuberculosis in Guatemala future research should utilize more discriminatory techniques such as whole genome sequence (WGS) or mycobacterial interspersed repetitive-unit–variable-number tandem-repeat (MIRU-VNTR) typing. In addition, our data did not allow us to identify risk factors in the HIV-seronegative population, who are most of the tuberculosis cases in the country. In order to halt the transmission events that lead to new cases we need to identify risk factors

that can be modified or lead to concrete actions by the public health authorities Finally, there is an urgent need to further our research in the area of mixed infections and its role in the spread and dissemination of the disease in both the HIV-seropositive and HIV-seronegative populations.

CONCLUSION

With the slogan ‘United to end tuberculosis: an urgent global response to a global epidemic’, the United Nations (UN) General Assembly adopted a political declaration of the UN high-level meeting on the fight against tuberculosis on October 2018. The Heads of State and Government confirmed its commitment to end tuberculosis globally by 2030. To achieve this goal, it is required that by 2020 the tuberculosis incidence rate be declining at 4-5% per year. Currently the tuberculosis incidence rate is decreasing at around 2% per year (World Health Organization, 2018). Clearly, the current tuberculosis control measurements which focus exclusively in the passive detection and treatment of active cases are having limited efficacy. There is an urgent need to shift our attention to the prevention of new cases. Along those lines, our work has provided an innovative approach to estimate contact for tuberculosis transmission, especially in the pediatric population. This can lead to a better identification of high-risk populations that might be benefited by preventive therapy. Finally, our molecular findings confirm previous work regarding the primary role of recent infection in the burden of tuberculosis transmission in low-income settings. Although there is still more research necessary to completely understand the transmission dynamics of tuberculosis, there are already measures that can be undertaken. Active case-findings has recently been shown to be feasible and affordable in limited-resources countries (Karki, Kittel, Bolokon Jr, & Duke, 2017).

To eliminate tuberculosis in eleven years, the world needs the commitment of governments, international agencies, civil sector, researchers and funding agencies to devote their resources in evidence-based strategies, of which the reduction of tuberculosis transmission should be in the frontline.

REFERENCES

Acuna-Villaorduna, C., Jones-Lopez, E. C., Fregona, G., Marques-Rodrigues, P., Gaeddert, M., Geadas, C., . . . Dietze, R. (2018). Intensity of exposure to pulmonary tuberculosis determines risk of tuberculosis infection and disease. *Eur Respir J, 51*(1). doi:10.1183/13993003.01578-2017

Ai, J. W., Ruan, Q. L., Liu, Q. H., & Zhang, W. H. (2016). Updates on the risk factors for latent tuberculosis reactivation and their managements. *Emerg Microbes Infect, 5*, e10. doi:10.1038/emi.2016.10

Aït-Khaled, N., Alarcón, E., Armengol, R., Bissell, K., Boillot, F., & Cameniro, J. (2010). Management of tuberculosis: a guide to the essentials of good practice. *Paris: International Union Against Tuberculosis and Lung Disease*.

Ameijeiras-Alonso, J., Crujeiras, R. M., & Rodríguez-Casal, A. (2016). Mode testing, critical bandwidth and excess mass. *arXiv preprint arXiv:1609.05188*.

Andrews, J. R., Morrow, C., Walensky, R. P., & Wood, R. (2014). Integrating social contact and environmental data in evaluating tuberculosis transmission in a South African township. *Journal of Infectious Diseases, 210*(4), 597-603.

Auld, S. C., Shah, N. S., Cohen, T., Martinson, N. A., & Gandhi, N. R. (2018). Where is tuberculosis transmission happening? Insights from the literature, new tools to study transmission and implications for the elimination of tuberculosis. *Respirology*. doi:10.1111/resp.13333

Balcells, M. E., Garcia, P., Meza, P., Pena, C., Cifuentes, M., Couvin, D., & Rastogi, N. (2015). A first insight on the population structure of Mycobacterium tuberculosis complex as studied by spoligotyping and MIRU-VNTRs in Santiago, Chile. *PLoS One, 10*(2), e0118007. doi:10.1371/journal.pone.0118007

Bansal, S., Read, J., Pourbohloul, B., & Meyers, L. A. (2010). The dynamic nature of contact networks in infectious disease epidemiology. *J Biol Dyn, 4*(5), 478-489. doi:10.1080/17513758.2010.503376

Banuls, A. L., Sanou, A., Anh, N. T., & Godreuil, S. (2015). Mycobacterium tuberculosis: ecology and evolution of a human bacterium. *J Med Microbiol, 64*(11), 1261-1269. doi:10.1099/jmm.0.000171

Barbero Garcia, M. I., Vila, E., & Holgado Tello, F. P. (2013). *Introducción básica al análisis factorial*: Editorial UNED.

Barnes, P. F., & Cave, M. D. (2003). Molecular epidemiology of tuberculosis. *New England Journal of Medicine, 349*(12), 1149-1156.

Beavers, A. S., Lounsbury, J. W., Richards, J. K., Huck, S. W., Skolits, G. J., & Esquivel, S. L. (2013). Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research & Evaluation, 18*.

Beggs, C. B., Noakes, C. J., Sleigh, P. A., Fletcher, L. A., & Siddiqi, K. (2003). The transmission of tuberculosis in confined spaces: an analytical review of alternative epidemiological models. *Int J Tuberc Lung Dis, 7*(11), 1015-1026.

Borgdorff, M. W., & van Soolingen, D. (2013). The re-emergence of tuberculosis: what have we learnt from molecular epidemiology? *Clin Microbiol Infect, 19*(10), 889-901. doi:10.1111/1469-0691.12253

Canty, A., & Ripley, B. (2017). Package 'boot'.

Chen, W., Shi, J., Qian, L., & Azen, S. P. (2014). Comparison of robustness to outliers between robust poisson models and log-binomial models when estimating relative risks for common binary outcomes: a simulation study. *BMC Med Res Methodol, 14*(1), 82.

Classen, C. N., Warren, R., Richardson, M., Hauman, J. H., Gie, R. P., Ellis, J. H., . . . Beyers, N. (1999). Impact of social interactions in the community on the transmission of tuberculosis in a high incidence area. *Thorax, 54*(2), 136-140.

Cohen, J. e., Powderly, W. G. e., & Opal, S. M. e. (2017). *Infectious diseases* (Fourth edition. ed.).

Cohen, T., van Helden, P. D., Wilson, D., Colijn, C., McLaughlin, M. M., Abubakar, I., & Warren, R. M. (2012). Mixed-strain mycobacterium tuberculosis infections and the implications for tuberculosis treatment and control. *Clin Microbiol Rev, 25*(4), 708-719. doi:10.1128/CMR.00021-12

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis&58; four recommendations for getting the most from your analysis. *Practical Assessment, 10*(7), 1-9.

Coutinho, L., Scazufca, M., & Menezes, P. R. (2008). Methods for estimating prevalence ratios in cross-sectional studies. *Revista de saude publica, 42*(6), 992-998.

Couvin, D., David, A., Zozio, T., & Rastogi, N. (2018). Macro-geographical specificities of the prevailing tuberculosis epidemic as seen through SITVIT2, an updated version of the Mycobacterium tuberculosis genotyping database. *Infect Genet Evol*. doi:10.1016/j.meegid.2018.12.030

Davies, P. D. O., Gordon, S. B., & Davies, G. (2014). *Clinical tuberculosis* (Fifth edition. ed.). Boca Raton: CRC Press, Taylor & Francis Group.

Delgado-Rodriguez, M., & Llorca, J. (2004). Bias. *Journal of epidemiology and community health, 58*(8), 635-641.

Dheda, K., Barry, C. E., 3rd, & Maartens, G. (2016). Tuberculosis. *Lancet, 387*(10024), 1211-1226. doi:10.1016/S0140-6736(15)00151-8

Dicker, R. C., Coronado, F. t., Koo, D., & Parrish, R. G. (2006). Principles of epidemiology in public health practice; an introduction to applied epidemiology and biostatistics.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation, 14*(20), 2.

Dobbin, K. K., & Simon, R. M. (2011). Optimally splitting cases for training and testing high dimensional classifiers. *BMC Med Genomics, 4*, 31. doi:10.1186/1755-8794-4-31

Dodd, P. J., Looker, C., Plumb, I. D., Bond, V., Schaap, A., Shanaube, K., . . . White, R. G. (2016). Age- and Sex-Specific Social Contact Patterns and Incidence of Mycobacterium tuberculosis Infection. *Am J Epidemiol, 183*(2), 156-166. doi:10.1093/aje/kwv160

Edefonti, V., Bravi, F., Garavello, W., La Vecchia, C., Parpinel, M., Franceschi, S., . . . Decarli, A. (2010). Nutrient-based dietary patterns and laryngeal cancer: evidence from an exploratory factor analysis. *Cancer Epidemiol Biomarkers Prev, 19*(1), 18-27. doi:10.1158/1055-9965.EPI-09-0900

Edmunds, W. J., O'callaghan, C., & Nokes, D. (1997). Who mixes with whom? A method to determine the contact patterns of adults that may lead to the spread of airborne infections. *Proceedings of the Royal Society of London B: Biological Sciences, 264*(1384), 949-957.

Esmail, H., Barry, C., Young, D., & Wilkinson, R. (2014). The ongoing challenge of latent tuberculosis. *Phil. Trans. R. Soc. B, 369*(1645), 20130437.

Feenstra, S. G., Nahar, Q., Pahan, D., Oskam, L., & Richardus, J. H. (2013). A qualitative exploration of social contact patterns relevant to airborne infectious diseases in northwest Bangladesh. *Journal of health, population, and nutrition, 31*(4), 424.

Field, N., Cohen, T., Struelens, M. J., Palm, D., Cookson, B., Glynn, J. R., . . . MacCannell, D. (2014). Strengthening the Reporting of Molecular Epidemiology for Infectious Diseases (STROME-ID): an extension of the STROBE statement. *The Lancet Infectious Diseases, 14*(4), 341-352.

Fok, A., Numata, Y., Schulzer, M., & FitzGerald, M. J. (2008). Risk factors for clustering of tuberculosis cases: a systematic review of population-based molecular epidemiology studies. *Int J Tuberc Lung Dis, 12*(5), 480-492.

Friendly, M. (2015). Visualizing GLMs for binary outcomes. Retrieved from http://ddar.datavis.ca/pages/extra/titanic-glm-ex.pdf

Friendly, M., & Meyer, D. (2015). *Discrete data analysis with R: visualization and modeling techniques for categorical and count data* (Vol. 120): CRC Press.

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and individual differences, 7*(3), 385-400.

Garcia, J. I., Samayoa, B., Sabido, M., Prieto, L. A., Nikiforov, M., Pinzon, R., . . . Study Group, T. M. (2015). The MANGUA Project: A Population-Based HIV Cohort in Guatemala. *AIDS Res Treat, 2015*, 372816. doi:10.1155/2015/372816

Gaskin, C. J., & Happell, B. (2014). On exploratory factor analysis: A review of recent evidence, an assessment of current practice, and recommendations for future use. *International journal of nursing studies, 51*(3), 511-521.

Getnet, F., Demissie, M., Assefa, N., Mengistie, B., & Worku, A. (2017). Delay in diagnosis of pulmonary tuberculosis in low-and middle-income settings: systematic review and meta-analysis. *BMC Pulm Med, 17*(1), 202. doi:10.1186/s12890-017-0551-y

Glaziou, P., Van der Werf, M., Onozaki, I., & Dye, C. (2008). Tuberculosis prevalence surveys: rationale and cost [Educational series: prevalence surveys. Serialised guidelines. Assessing tuberculosis prevalence through population-based surveys. Number 1 in the series]. *The international journal of tuberculosis and lung disease, 12*(9), 1003-1008.

Glynn, J., Bauer, J., De Boer, A., Borgdorff, M., Fine, P., Godfrey-Faussett, P., & Vynnycky, E. (1999). Interpreting DNA fingerprint clusters of Mycobacterium tuberculosis [Position Paper]. *The international journal of tuberculosis and lung disease, 3*(12), 1055-1060.

Glynn, J. R., Crampin, A. C., Yates, M. D., Traore, H., Mwaungulu, F. D., Ngwira, B. M., . . . Fine, P. E. (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. *J Infect Dis, 192*(3), 480-487. doi:10.1086/431517

Glynn, J. R., Vynnycky, E., & Fine, P. E. (1999). Influence of sampling on estimates of clustering and recent transmission of Mycobacterium tuberculosis derived from DNA fingerprinting techniques. *Am J Epidemiol, 149*(4), 366-371.

Gonzalez, O. Y., Adams, G., Teeter, L. D., Bui, T. T., Musser, J. M., & Graviss, E. A. (2003). Extra-pulmonary manifestations in a large metropolitan area with a low incidence of tuberculosis. *Int J Tuberc Lung Dis, 7*(12), 1178-1185.

Gori, A., Bandera, A., Marchetti, G., Degli Esposti, A., Catozzi, L., Nardi, G. P., . . . Franzetti, F. (2005). Spoligotyping and Mycobacterium tuberculosis. *Emerg Infect Dis, 11*(8), 1242-1248. doi:10.3201/eid1108.040982

Grimes, D. A., & Schulz, K. F. (2002). Bias and causal associations in observational research. *Lancet, 359*(9302), 248-252. doi:10.1016/S0140-6736(02)07451-2

Hamblion, E. L., Le Menach, A., Anderson, L. F., Lalor, M. K., Brown, T., Abubakar, I., . . . Public Health England Strain Typing Project, B. (2016). Recent TB transmission, clustering and predictors of large clusters in London, 2010-2012: results from first 3 years of universal MIRU-VNTR strain typing. *Thorax, 71*(8), 749-756. doi:10.1136/thoraxjnl-2014-206608

Härdle, W., & Simar, L. *Applied multivariate statistical analysis* (Vol. 22007): Springer.

Heemskerk, D., Caws, M., Marais, B., & Farrar, J. (2015). *Tuberculosis in Adults and Children*: Springer Science+ Business Media.

Holgado–Tello, F. P., Chacón–Moscoso, S., Barbero–García, I., & Vila–Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity, 44*(1), 153.

Houben, R., Crampin, A., Ndhlovu, R., Sonnenberg, P., Godfrey-Faussett, P., Haas, W., . . . Bruchfeld, J. (2011). Human immunodeficiency virus associated tuberculosis more often due to recent infection than reactivation of latent infection [Review article]. *The international journal of tuberculosis and lung disease, 15*(1), 24-31.

Houben, R. M., & Dodd, P. J. (2016). The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling. *PLoS Med, 13*(10), e1002152. doi:10.1371/journal.pmed.1002152

Jagielski, T., Minias, A., van Ingen, J., Rastogi, N., Brzostek, A., Zaczek, A., & Dziadek, J. (2016). Methodological and Clinical Aspects of the Molecular Epidemiology of Mycobacterium tuberculosis and Other Mycobacteria. *Clin Microbiol Rev, 29*(2), 239-290. doi:10.1128/cmr.00055-15

Johnstone-Robertson, S. P., Mark, D., Morrow, C., Middelkoop, K., Chiswell, M., Aquino, L. D., . . . Wood, R. (2011). Social mixing patterns within a South African township community: implications for respiratory disease transmission and control. *Am J Epidemiol*, kwr251.

Joreskog, K. G., Olsson, U. H., & Wallentin, F. Y. (2016). *Multivariate analysis with LISREL*: Springer.

Kaiser, H. F., & Rice, J. (1974). Little jiffy, mark IV. *Educational and psychological measurement, 34*(1), 111-117.

Kamerbeek, J., Schouls, L., Kolk, A., Van Agterveld, M., Van Soolingen, D., Kuijper, S., . . . Goyal, M. (1997). Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. *J Clin Microbiol, 35*(4), 907-914.

Karki, B., Kittel, G., Bolokon Jr, I., & Duke, T. (2017). Active Community-Based Case Finding for Tuberculosis With Limited Resources: Estimating Prevalence in a Remote Area of Papua New Guinea. *Asia Pacific Journal of Public Health, 29*(1), 17-27.

Kasaie, P., Mathema, B., Kelton, W. D., Azman, A. S., Pennington, J., & Dowdy, D. W. (2015). A Novel Tool Improves Existing Estimates of Recent Tuberculosis Transmission in Settings of Sparse Data Collection. *PLoS One, 10*(12), e0144137. doi:10.1371/journal.pone.0144137

Kenyon, T. A., Valway, S. E., Ihle, W. W., Onorato, I. M., & Castro, K. G. (1996). Transmission of multidrug-resistant Mycobacterium tuberculosis during a long airplane flight. *New England Journal of Medicine, 334*(15), 933-938.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political analysis, 9*(2), 137-163.

Kizza, F. N., List, J., Nkwata, A. K., Okwera, A., Ezeamama, A. E., Whalen, C. C., & Sekandi, J. N. (2015). Prevalence of latent tuberculosis infection and associated risk factors in an urban African setting. *BMC infectious diseases, 15*(1), 165.

Knight, W. R. (1966). A computer method for calculating Kendall's tau with ungrouped data. *Journal of the American Statistical Association, 61*(314), 436-439.

Kong, L., Wang, J., Han, W., & Cao, Z. (2016). Modeling Heterogeneity in Direct Infectious Disease Transmission in a Compartmental Model. *International journal of environmental research and public health, 13*(3), 253.

Larson, M. G. (2006). Descriptive statistics and graphical displays. *Circulation, 114*(1), 76-81. doi:10.1161/CIRCULATIONAHA.105.584474

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: an R package for multivariate analysis. *Journal of statistical software, 25*(1), 1-18.

Levy, S. (2012). The Evolution of Tuberculosis: Genetic analysis offers new insight on the spread of an ancient disease. *BioScience, 62*(7), 625-629. doi:10.1525/bio.2012.62.7.3

Lienhardt, C., Fielding, K., Sillah, J., Tunkara, A., Donkor, S., Manneh, K., . . . Bennett, S. (2003). Risk factors for tuberculosis infection in sub-Saharan Africa: a contact study in The Gambia. *Am J Respir Crit Care Med, 168*(4), 448-455. doi:10.1164/rccm.200212-1483OC

Lin, P. L., & Flynn, J. L. (2010). Understanding latent tuberculosis: a moving target. *The Journal of Immunology, 185*(1), 15-22.

Lonnroth, K., & Raviglione, M. (2008). Global epidemiology of tuberculosis: prospects for control. *Semin Respir Crit Care Med, 29*(5), 481-491. doi:10.1055/s-0028-1085700

Mandalakas, A. M., Kirchner, H. L., Lombard, C., Walzl, G., Grewal, H. M., Gie, R. P., & Hesseling, A. C. (2012). Well-quantified tuberculosis exposure is a reliable surrogate measure of tuberculosis infection. *Int J Tuberc Lung Dis, 16*(8), 1033-1039. doi:10.5588/ijtld.12.0027

Martinez, L., Sekandi, J. N., Castellanos, M. E., Zalwango, S., & Whalen, C. C. (2016). Infectiousness of HIV-Seropositive Patients with Tuberculosis in a High-Burden African Setting. *Am J Respir Crit Care Med, 194*(9), 1152-1163. doi:10.1164/rccm.201511-2146OC

Martinez, L., Shen, Y., Mupere, E., Kizza, A., Hill, P. C., & Whalen, C. C. (2017). Transmission of Mycobacterium Tuberculosis in Households and the Community: A Systematic Review and Meta-Analysis. *Am J Epidemiol, 185*(12), 1327-1339. doi:10.1093/aje/kwx025

Mathema, B., Andrews, J. R., Cohen, T., Borgdorff, M. W., Behr, M., Glynn, J. R., . . . Wood, R. (2017). Drivers of Tuberculosis Transmission. *J Infect Dis, 216*(suppl_6), S644-S653. doi:10.1093/infdis/jix354

Mathema, B., Kurepina, N. E., Bifani, P. J., & Kreiswirth, B. N. (2006). Molecular epidemiology of tuberculosis: current insights. *Clin Microbiol Rev, 19*(4), 658-685. doi:10.1128/CMR.00061-05

McCreesh, N., & White, R. G. (2018). An explanation for the low proportion of tuberculosis that results from transmission between household and known social contacts. *Sci Rep, 8*(1), 5382. doi:10.1038/s41598-018-23797-2

McIvor, A., Koornhof, H., & Kana, B. D. (2017). Relapse, re-infection and mixed infections in tuberculosis disease. *Pathog Dis, 75*(3). doi:10.1093/femspd/ftx020

Mears, J., Abubakar, I., Cohen, T., McHugh, T. D., & Sonnenberg, P. (2015). Effect of study design and setting on tuberculosis clustering estimates using Mycobacterial Interspersed Repetitive Units-Variable Number Tandem Repeats (MIRU-VNTR): a systematic review. *BMJ Open, 5*(1). doi:10.1136/bmjopen-2014-005636

Melegaro, A., Jit, M., Gay, N., Zagheni, E., & Edmunds, W. J. (2011). What types of contacts are important for the spread of infections? Using contact survey data to explore European mixing patterns. *Epidemics, 3*(3), 143-151. doi:https://doi.org/10.1016/j.epidem.2011.04.001

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., . . . Edmunds, W. J. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med, 5*(3), e74. doi:10.1371/journal.pmed.0050074

Moström, P., Gordon, M., Sola, C., Ridell, M., & Rastogi, N. (2002). Methods used in the molecular epidemiology of tuberculosis. *Clinical microbiology and infection, 8*(11), 694-704.

Murray, M. (2002). Determinants of cluster distribution in the molecular epidemiology of tuberculosis. *Proc Natl Acad Sci U S A, 99*(3), 1538-1543. doi:10.1073/pnas.022618299

Murray, M. (2002). Sampling bias in the molecular epidemiology of tuberculosis. *Emerg Infect Dis, 8*(4), 363-369.

Murray, M., & Alland, D. (2002). Methodological problems in the molecular epidemiology of tuberculosis. *Am J Epidemiol, 155*(6), 565-571.

Murray, M., & Nardell, E. (2002). Molecular epidemiology of tuberculosis: achievements and challenges to current knowledge. *Bull World Health Organ, 80*(6), 477-482.

Narasimhan, P., Wood, J., Macintyre, C. R., & Mathai, D. (2013). Risk factors for tuberculosis. *Pulm Med, 2013*, 828939. doi:10.1155/2013/828939

Nardell, E. A. (2016). Indoor environmental control of tuberculosis and other airborne infections. *Indoor Air, 26*(1), 79-87. doi:10.1111/ina.12232

Nardo, M., Saisana, M., Saltelli, A., & Tarantola, S. (2005). Tools for Composite Indicators Building-EUR 21682 EN. *Institute for the Protection and Security of the Citizen, Ispra*.

Nava-Aguilera, E., Andersson, N., Harris, E., Mitchell, S., Hamel, C., Shea, B., . . . Morales-Perez, A. (2009). Risk factors associated with recent transmission of tuberculosis: systematic review and meta-analysis. *Int J Tuberc Lung Dis, 13*(1), 17-26.

Nava-Aguilera, E., Lopez-Vidal, Y., Harris, E., Morales-Perez, A., Mitchell, S., Flores-Moreno, M., . . . Andersson, N. (2011). Clustering of Mycobacterium tuberculosis cases in Acapulco: Spoligotyping and risk factors. *Clin Dev Immunol, 2011*, 408375. doi:10.1155/2011/408375

Nayak, S., & Acharjya, B. (2012). Mantoux test and its interpretation. *Indian Dermatol Online J, 3*(1), 2-6. doi:10.4103/2229-5178.93479

Nelson, K. E., & Williams, C. M. (2006). *Infectious disease epidemiology : theory and practice* (2nd ed. ed.). Sudbury, Mass. ; London: Jones and Bartlett Publishers.

Noakes, C. J., & Sleigh, P. A. (2008). *Applying the Wells-Riley equation to the risk of airborne infection in hospital environments: The importance of stochastic and proximity effects.* Paper presented at the Indoor Air 2008: The 11th International Conference on Indoor Air Quality and Cl.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika, 44*(4), 443-460.

Ong, A., Creasman, J., Hopewell, P. C., Gonzalez, L. C., Wong, M., Jasmer, R. M., & Daley, C. L. (2004). A molecular epidemiological assessment of extrapulmonary tuberculosis in San Francisco. *Clin Infect Dis, 38*(1), 25-31. doi:10.1086/380448

Onifade, D. A., Bayer, A. M., Montoya, R., Haro, M., Alva, J., Franco, J., . . . Evans, C. A. (2010). Gender-related factors influencing tuberculosis control in shantytowns: a qualitative study. *BMC Public Health, 10*(1), 381. doi:10.1186/1471-2458-10-381

Ortiz-Ospina, E., & Tzvetkova, S. (2017). Working women: Key facts and trends in female labor force participation.

PAHO. (2018). Tuberculosis in the Americas, 2018.

PAHO. (2016). Framework for Tuberculosis Control in Large Cities in Latin America and the Caribbean.

Pan American Health Organization. (2013). Tuberculosis in the Americas: regional report 2014. Epidemiology, control and financing: PAHO Washington DC.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics, 57*(1), 120-125.

Parascandola, M., & Weed, D. L. (2001). Causation in epidemiology. *J Epidemiol Community Health, 55*(12), 905-912.

Perkins, J. M., Subramanian, S. V., & Christakis, N. A. (2015). Social networks and health: a systematic review of sociocentric network studies in low- and middle-income countries. *Soc Sci Med, 125*, 60-78. doi:10.1016/j.socscimed.2014.08.019

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis : the use of factor analysis for instrument development in health care research*. Thousand Oaks, Calif. ; London: SAGE.

Potter, G. E., Handcock, M. S., Longini Jr, I. M., & Halloran, M. E. (2011). Estimating within-household contact networks from egocentric data. *The annals of applied statistics, 5*(3), 1816.

Potter, G. E., Handcock, M. S., Longini Jr, I. M., & Halloran, M. E. (2012). Estimating within-school contact networks to understand influenza transmission. *The annals of applied statistics, 6*(1), 1.

Puth, M. T., Neuhauser, M., & Ruxton, G. D. (2015). On the variety of methods for calculating confidence intervals by bootstrapping. *J Anim Ecol, 84*(4), 892-897. doi:10.1111/1365-2656.12382

Real, L., & Biek, R. (2007). Infectious disease modeling and the dynamics of transmission *Wildlife and emerging zoonotic diseases: the biology, circumstances and consequences of cross-species transmission* (pp. 33-49): Springer.

Reyes, J. F., Francis, A. R., & Tanaka, M. M. (2008). Models of deletion for visualizing bacterial variation: an application to tuberculosis spoligotypes. *BMC Bioinformatics, 9*, 496. doi:10.1186/1471-2105-9-496

Rieder, H. L. (2001). Risk of travel-associated tuberculosis. *Clin Infect Dis, 33*(8), 1393-1396. doi:10.1086/323127

Rosales, S., Pineda-Garcia, L., Ghebremichael, S., Rastogi, N., & Hoffner, S. E. (2010). Molecular diversity of Mycobacterium tuberculosis isolates from patients with tuberculosis in Honduras. *BMC Microbiol, 10*, 208. doi:10.1186/1471-2180-10-208

Rothman, K. J. (1976). Causes. *Am J Epidemiol, 104*(6), 587-592.

Rudnick, S., & Milton, D. (2003). Risk of indoor airborne infection transmission estimated from carbon dioxide concentration. *Indoor Air, 13*(3), 237-245.

Saelens, J. W., Lau-Bonilla, D., Moller, A., Medina, N., Guzman, B., Calderon, M., . . . Tobin, D. M. (2015). Whole genome sequencing identifies circulating Beijing-lineage Mycobacterium tuberculosis strains in Guatemala and an associated urban outbreak. *Tuberculosis (Edinb), 95*(6), 810-816. doi:10.1016/j.tube.2015.09.001

Samayoa-Peláez, M., Ayala, N., Yadon, Z. E., & Heldal, E. (2016). Implementation of the national tuberculosis guidelines on culture and drug sensitivity testing in Guatemala, 2013. *Revista Panamericana de Salud Pública, 39*(1), 44-50.

SAS Institute Inc. (2017). The PRINQUAL Procedure *SAS/STAT® 14.3 User's Guide*.

Scott, A. N., Menzies, D., Tannenbaum, T. N., Thibert, L., Kozak, R., Joseph, L., . . . Behr, M. A. (2005). Sensitivities and specificities of spoligotyping and mycobacterial interspersed repetitive unit-variable-number tandem repeat typing methods for studying molecular epidemiology of tuberculosis. *J Clin Microbiol, 43*(1), 89-94. doi:10.1128/JCM.43.1.89-94.2005

Sekandi, J. N., Zalwango, S., Martinez, L., Handel, A., Kakaire, R., Nkwata, A. K., . . . Whalen, C. C. (2015). Four degrees of separation: social contacts and health providers influence the steps to final diagnosis of active tuberculosis patients in Urban Uganda. *BMC infectious diseases, 15*(1), 361.

Shapiro, S. E., Lasarev, M. R., & McCauley, L. (2002). Factor analysis of Gulf War illness: what does it add to our understanding of possible health effects of deployment? *Am J Epidemiol, 156*(6), 578-585.

Shin, S. S., Modongo, C., Ncube, R., Sepako, E., Klausner, J. D., & Zetola, N. M. (2015). Advanced immune suppression is associated with increased prevalence of mixed-strain Mycobacterium tuberculosis infections among persons at high risk for drug-resistant tuberculosis in Botswana. *J Infect Dis, 211*(3), 347-351. doi:10.1093/infdis/jiu421

Small, P. M., Hopewell, P. C., Singh, S. P., Paz, A., Parsonnet, J., Ruston, D. C., . . . Schoolnik, G. K. (1994). The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. *N Engl J Med, 330*(24), 1703-1709. doi:10.1056/NEJM199406163302402

Smith, C. M., Maguire, H., Anderson, C., Macdonald, N., & Hayward, A. C. (2017). Multiple large clusters of tuberculosis in London: a cross-sectional analysis of molecular and spatial data. *ERJ Open Res, 3*(1). doi:10.1183/23120541.00098-2016

Ssengooba, W., Cobelens, F. G., Nakiyingi, L., Mboowa, G., Armstrong, D. T., Manabe, Y. C., . . . de Jong, B. C. (2015). High Genotypic Discordance of Concurrent Mycobacterium tuberculosis Isolates from Sputum and Blood of HIV-Infected Individuals. *PLoS One, 10*(7), e0132581. doi:10.1371/journal.pone.0132581

Stein, M. L., Van Steenbergen, J. E., Buskens, V., Van Der Heijden, P. G., Chanyasanha, C., Tipayamongkholgul, M., . . . Kretzschmar, M. E. (2014). Comparison of contact patterns relevant for transmission of respiratory pathogens in Thailand and the Netherlands using respondent-driven sampling. *PLoS One, 9*(11), e113711.

Sterling, T. R., Pope, D. S., Bishai, W. R., Harrington, S., Gershon, R. R., & Chaisson, R. E. (2000). Transmission of Mycobacterium tuberculosis from a cadaver to an embalmer. *New England Journal of Medicine, 342*(4), 246-248.

Stucki, D., Brites, D., Jeljeli, L., Coscolla, M., Liu, Q., Trauner, A., . . . Gagneux, S. (2016). Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. *Nat Genet, 48*(12), 1535-1543. doi:10.1038/ng.3704

Suhr, D. D. (2005). Principal component analysis vs. exploratory factor analysis. *SUGI 30 proceedings, 203*, 230.

Sulis, G., Roggi, A., Matteelli, A., & Raviglione, M. C. (2014). Tuberculosis: epidemiology and control. *Mediterranean journal of hematology and infectious diseases, 6*(1), 2014070.

Supply, P. (2005). Multilocus variable number tandem repeat genotyping of Mycobacterium tuberculosis. *Technical Guide*.

Supply, P., Allix, C., Lesjean, S., Cardoso-Oelemann, M., Rusch-Gerdes, S., Willery, E., . . . van Soolingen, D. (2006). Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of Mycobacterium tuberculosis. *J Clin Microbiol, 44*(12), 4498-4510. doi:10.1128/JCM.01392-06

Sze To, G., & Chao, C. (2010). Review and comparison between the Wells–Riley and dose-response approaches to risk assessment of infectious respiratory diseases. *Indoor Air, 20*(1), 2-16.

Taherdoost, H., Sahibuddin, S., & Jalaliyoon, N. (2014). Exploratory factor analysis: Concepts and theory. *Advances in Pure and Applied Mathematics*.

Tang, C., Reyes, J. F., Luciani, F., Francis, A. R., & Tanaka, M. M. (2008). spolTools: online utilities for analyzing spoligotypes of the Mycobacterium tuberculosis complex. *Bioinformatics, 24*(20), 2414-2415. doi:10.1093/bioinformatics/btn434

Tarashi, S., Fateh, A., Mirsaeidi, M., Siadat, S. D., & Vaziri, F. (2017). Mixed infections in tuberculosis: The missing part in a puzzle. *Tuberculosis (Edinb), 107*, 168-174. doi:10.1016/j.tube.2017.09.004

Thomas, J. C., & Weber, D. J. (2001). *Epidemiologic methods for the study of infectious diseases*: Oxford University Press.

Treiblmaier, H., & Filzmoser, P. (2010). Exploratory factor analysis revisited: How robust methods support the detection of hidden multivariate data structures in IS research. *Information & management, 47*(4), 197-207.

Tsai, A. C., Bangsberg, D. R., Frongillo, E. A., Hunt, P. W., Muzoora, C., Martin, J. N., & Weiser, S. D. (2012). Food insecurity, depression and the modifying role of social support among people living with HIV/AIDS in rural Uganda. *Soc Sci Med, 74*(12).

Turner, R. D., Chiu, C., Churchyard, G. J., Esmail, H., Lewinsohn, D. M., Gandhi, N. R., & Fennelly, K. P. (2017). Tuberculosis Infectiousness and Host Susceptibility. *J Infect Dis, 216*(suppl_6), S636-S643. doi:10.1093/infdis/jix361

Uganda Bureau of Statistics (UBOS). (2018). *Uganda National Household Survey 2016/2017*

Van Soolingen, D. (2001). Molecular epidemiology of tuberculosis and other mycobacterial infections: main methodologies and achievements. *J Intern Med, 249*(1), 1-26.

Wacholder, S. (1986). Binomial regression in GLIM: estimating risk ratios and risk differences. *Am J Epidemiol, 123*(1), 174-184.

Wallinga, J., Teunis, P., & Kretzschmar, M. (2006). Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *Am J Epidemiol, 164*(10), 936-944.

Wampande, E. M., Mupere, E., Jaganath, D., Nsereko, M., Mayanja, H. K., Eisenach, K., . . . Tuberculosis Research, U. (2015). Distribution and transmission of Mycobacterium tuberculosis complex lineages among children in peri-urban Kampala, Uganda. *BMC Pediatr, 15*, 140. doi:10.1186/s12887-015-0455-z

Wang, W., Mathema, B., Hu, Y., Zhao, Q., Jiang, W., & Xu, B. (2014). Role of casual contacts in the recent transmission of tuberculosis in settings with high disease burden. *Clin Microbiol Infect, 20*(11), 1140-1145. doi:10.1111/1469-0691.12726

Weng, H. Y., Hsueh, Y. H., Messam, L. L., & Hertz-Picciotto, I. (2009). Methods of covariate selection: directed acyclic graphs and the change-in-estimate procedure. *Am J Epidemiol, 169*(10), 1182-1190. doi:10.1093/aje/kwp035

Wensink, M., Westendorp, R. G., & Baudisch, A. (2014). The causal pie model: an epidemiological method applied to evolutionary biology and ecology. *Ecol Evol, 4*(10), 1924-1930. doi:10.1002/ece3.1074

Whalen, C. C. (2014). *PROTOCOL. COMMUNITY HEALTH AND SOCIAL NETWORKS OF TUBERCULOSIS*. University of Georgia.

Whalen, C. C. (2016). The Replacement Principle of Tuberculosis. Why Prevention Matters: Am Thoracic Soc.

Whalen, C. C., Zalwango, S., Chiunda, A., Malone, L., Eisenach, K., Joloba, M., . . . Mugerwa, R. (2011). Secondary attack rate of tuberculosis in urban households in Kampala, Uganda. *PLoS One, 6*(2), e16137. doi:10.1371/journal.pone.0016137

Wiens, K. E., Woyczynski, L. P., Ledesma, J. R., Ross, J. M., Zenteno-Cuevas, R., Goodridge, A., . . . Hay, S. I. (2018). Global variation in bacterial strains that cause tuberculosis disease: a systematic review and meta-analysis. *BMC Med, 16*(1), 196. doi:10.1186/s12916-018-1180-x

Woodman, M., Haeusler, I. L., & Grandjean, L. (2019). Tuberculosis Genetic Epidemiology: A Latin American Perspective. *Genes (Basel), 10*(1). doi:10.3390/genes10010053

World Health Organization. (2018). Global Tuberculosis Report 2018. Geneva: WHO; 2018.

World Health Organization. (2016). *Global tuberculosis report 2016*. Geneva: World Health Organization.

Yang, C., Shen, X., Peng, Y., Lan, R., Zhao, Y., Long, B., . . . Qiao, K. (2015). Transmission of Mycobacterium tuberculosis in China: A population-based molecular epidemiology study. *Clinical Infectious Diseases*, civ255.

Yang, H., Kruh-Garcia, N. A., & Dobos, K. M. (2012). Purified protein derivatives of tuberculin--past, present, and future. *FEMS Immunol Med Microbiol, 66*(3), 273-280. doi:10.1111/j.1574-695X.2012.01002.x

Yates, T. A., Khan, P. Y., Knight, G. M., Taylor, J. G., McHugh, T. D., Lipman, M., . . . Abubakar, I. (2016). The transmission of Mycobacterium tuberculosis in high burden settings. *Lancet Infect Dis, 16*(2), 227-238. doi:10.1016/S1473-3099(15)00499-5

Zou, G. (2004). A modified poisson regression approach to prospective studies with binary data. *Am J Epidemiol, 159*(7), 702-706.