Distribution Agreement

Signature:

______________________________ ____________________
Julia Interrante Date

The Role of Exogenous Reinfection in Patients with Recurrent TB Disease

in the United States

By

Julia Interrante

Master of Public Health

Department of Epidemiology

Neel Gandhi, MD

Committee Chair

Maryam Haddad

Committee Member

The Role of Exogenous Reinfection in Patients with Recurrent TB Disease in the United States

By

Julia Interrante

B.A., University of Virginia, 2009

Thesis Committee Chair: Neel Gandhi, MD

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Public Health

in Epidemiology

2014

Abstract

The Role of Exogenous Reinfection in Patients with Recurrent TB Disease
in the United States

By Julia Interrante

Purpose. Traditionally, recurrent tuberculosis (TB) has been assumed to result from endogenous reactivation. Genotyping now allows us to determine how much of recurrence is actually due to exogenous reinfection. To determine the extent of and to better understand factors leading to reinfection rather than reactivation, we analyzed patients in the United States with two episodes of TB disease during 1993 to 2011.

Methods. The study population was drawn from all TB cases in the 50 states, Puerto Rico, and the District of Columbia, as reported of June 25, 2012. We identified recurrent cases by matching on date of birth, sex, race, country of origin, state, and year of first episode. Genotyping was used to distinguish between reinfection and reactivation. Selection required time from treatment completion in first episode be ≥12 months before the start of second episode. To statistically evaluate the effects of predictors on reinfection, a logistic regression model was fit.

Results. Among patients with recurrent TB who completed treatment during their first episode, 136 patients were identified, involving 116 reactivations and 20 reinfections. Reinfection occurred in 15% of the population with recurrent TB. Three factors were statistically significant for reinfection after adjustment, including being black or Hispanic (odds ratio (OR) 4.4, 95% confidence interval (CI) 1.1-17.2), living ≤12 years in the United States (OR 3.5, 95% CI 1.0-11.9), and having received treatment exclusively by directly observed therapy (DOT) during first episode (OR 4.8, 95% CI 1.2-19.8).

Conclusions. In persons who experience two episodes of TB, genotyping evidence suggests that the majority of second episodes are reactivation of the first episode. However, minorities, those more recently immigrated to the United States, and those with more rigorous treatment regimens during first episodes have a greater risk for exogenous reinfection. This suggests that these populations are being successfully treated for TB, but other risk factors for recent transmission increase their risk of reinfection. Public health interventions should continue to focus on these populations and their areas of residence, work, and recreation for evidence of recent TB transmission to prevent further and future spread of TB.

The Role of Exogenous Reinfection in Patients with Recurrent TB Disease in the United States

By

Julia Interrante

B.A., University of Virginia, 2009

Thesis Committee Chair: Neel Gandhi, MD

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Public Health

in Epidemiology

2014

ACKNOWLEDGEMENTS

I thank the Division of TB Elimination at the Centers for Disease Control and Prevention for providing the dataset used for this analysis. Specifically, I would like to thank Lindsay Kim, M.D., M.P.H. for her work in validating the matching algorathim used in this study, Steve Kammerer, M.B.A. for retrieving and matching the genotyping data used in this analysis, and Juliana Grant, M.D., M.P.H. for introducing me to and teaching me about tuberculosis and genotyping processes. Most importantly, I would like to thank Dr. Neel Gandhi, M.D. of Emory University and Maryam B. Haddad, M.S.N., M.P.H. of the Division of TB Elimination at CDC for their time, guidance, and mentorship.

TABLE OF CONTENTS

Chapter I: Literature Review 1

Section 1: Overall Epidemiology of Tuberculosis 1

1. What is Tuberculosis 1
2. Signs and Symptoms 1
3. Transmission 2
4. TB Infection and Disease 2
5. Detection of TB 3
6. Treatment 5
7. Global TB Prevalence 6

Section 2: Tuberculosis in the United States 7

1. Incidence and Decline 7
2. High-Risk Populations 8
3. CDC and TB Reporting 10

Section 3: Second Episodes of Tuberculosis Disease 11

1. What is Recurrence? 11
2. Risk Factors for Recurrence 13
3. High-Incidence Characteristics 15
4. Low-Incidence Characteristics 16
5. Immigration 17

Section 4: Genotyping 18

1. Definition and Uses 18

2. Typing Methods 19

3. Strain Discrimination 21

4. National Tuberculosis Genotyping and Surveillance Network 22

Section 5: Immunology 23

1. Why reinfection is Possible 23

2. Protective and Partial Immunity 24

3. The Effects of First Infection Strain Type on Second Infection Strain Type 26

4. Strain Virulence and Prevalence 26

5. Mixed Infections 28

Section 6: Treatment Implications 29

1. Development of Drug Resistance 29

2. Treatment Choices 29

3. Vaccine Development 31

Chapter II: Manuscript 33

A. Title, Authors, Abstract 33

B. Introduction 34

C. Methods 36

Data Source and Study Population 36

Genotyping 37

Analysis 38

Ethical Approval 40

D. Results..40

Participants..40

Reinfection versus Reactivation 41

Multivariable Analysis..42

Unrestricted Recurrent Population Comparison 42

E. Discussion..43

Limitations..46

Conclusions..47

References..48

Tables..57

Table 1a. Demographic characteristics of patients with two episodes of TB, among those completing treatment during first episode, in the United States, 1993–2011..57

Table 1b. Clinical and treatment characteristics of patients with two episodes of TB, among those completing treatment during first episode, in the United States, 1993–2011 59

Table 2a. Demographic characteristics of patients with two episodes of tuberculosis in the United States, 1993–2011..61

Table 2b. Clinical and treatment characteristics of patients with two episodes of tuberculosis in the United States, 1993–2011 63

Table 3. Predictors of tuberculosis recurrence among 136 patients completing treatment during the first episode. Bivariate and multivariate analysis. United States, 1993–2011 65

Figures..66

Figure 1. Description of the study cohort 66

Figure 2. Distribution of years living in the United States at second episode among patients with recurrent TB completing treatment in first episode, 1993–2011..67

Chapter III: Public Health Implications and Future Directions 68

Appendices...70

Appendix A: IRB Approval 70

Appendix B: Individual Variable Tables 71

Appendix C: Models 91

Appendix D: SAS code 100

Matching Algorithm...100

Data Cleaning and New Variable Creation..103

Data Analysis 118

CHAPTER I: LITERATURE REVIEW

Section 1: Overall Epidemiology of Tuberculosis

1. What is Tuberculosis

A chronic bacterial infection, tuberculosis (TB) is a major cause of morbidity and mortality globally (1, 2). TB is caused by the bacillus *Mycobacterium tuberculosis* and is characterized by the formation of small nodules in the lungs or other locations throughout the body (2-4). The *M. tuberculosis* complex is a slow-growing acid-fast bacilli that contains over 4,000 genes (1, 3). In humans, TB disease begins by exposure to other infectious human sources. The exposed human must first develop infection, and then in some cases, infection progress to active disease (1).

2. Signs and Symptoms

Tuberculosis disease can occur in the lungs (pulmonary), outside of the lungs (extrapulmonary), or systemically, infecting multiple organs throughout the body (miliary) (5). Pulmonary TB often presents through cough, fatigue, fever, night sweats, chills, weight loss, chest pain, and in some advanced cases, through pleuritic pain and hemoptysis (1, 5). These symptoms can vary from mild to severe (5). Presentation of a cough is the most common symptom and is accompanied with or without bloody sputum (1, 5). Like pulmonary TB, extrapulmonary symptoms include fatigue and night sweats, but instead of the characteristic cough produced by pulmonary disease, extrapulmonary symptoms are often associated with the infected organs (5). Signs of TB can be identified through abnormal chest radiographs and other abnormal chest imaging as well as through clinical evidence of disease occurrence (2). If chest radiographs reveal cavitations produced by TB in the upper segments of the lung lobes, the disease is considered cavitary TB (1).

3. Transmission

Tuberculosis is most commonly and most efficiently transmitted through airborne transmission during exposure to droplet nuclei expelled by a diseased individual (1). Bacilli are discharged through the acts of coughing, sneezing, talking, or singing, increasing opportunities for the bacteria to become aerosolized (1). These sputum droplets then evaporate in the air leaving droplet nuclei. *M. tuberculosis* enters the body through droplet nuclei when they are inhaled by susceptible persons, initiating new TB infections (1, 6). TB is transmitted by people who have pulmonary or high respiratory tract TB (1). The risk of infection increases with the length of exposure, proximity to source, amount of ventilation, and the contagiousness of the source (1, 6).

4. TB Infection and Disease

The clinical manifestation of tuberculosis is complex. Infection begins in the latent form, referred to as latent tuberculosis, or LTBI. People latently infected are asymptomatic; therefore, most are unaware that they are infected and identification of new cases both by health care workers and the infected individuals is difficult (1, 5). Infection can produce extremely small lesions in the lungs that quickly heal and leave little evidence of infection (1). Infection can progress to active disease owing to an assortment of factors (5). LTBI can remain dormant for years before active disease develops, if disease develops at all (1, 3, 7). Identifying and treating LTBI before the occurrence of active disease can decrease patients' overall risk of progressing to disease (1).

The time from infection to progression to disease can vary greatly, and predicting progression can be complex (8, 9). Faster progression is usually age-dependent and more common in immunocompromised individuals (e.g., HIV-infected persons), infants, people who are underweight or undernourished, and people with diabetes or certain types of cancers (1, 10). Less than 10% of people with LTBI develop active disease and, of those, half progress to disease within the first two years after infection (1, 8). However, this risk remains elevated throughout the first five years after exposure (8). This decline of risk over time is seen in the United States, where the risk for developing active TB disease is 1% within the first year, compared to 0.07% at 8 to 10 years later (8).

In the disease stage, referred to as active tuberculosis, symptoms are a result of the host's response to the replication of tubercle bacilli within the body (3). Signs and symptoms of active disease are often nonspecific and can overlap with other pulmonary and systemic diseases (3, 5). In pulmonary TB, the most common site of active TB, tubercle bacilli affect the lungs and respiratory tract. Pulmonary TB presents in 80% of infected individuals (2, 3, 5). However, this proportion can differ based on other health-related factors. For example, among people with an immunodeficiency and in children, extrapulmonary infections are more common (5).

5. Detection of TB

Detecting latent TB infections and active TB disease through awareness campaigns, testing, and treatment are crucial in the drive for global TB elimination (4). In industrialized countries, these efforts are especially important in specific sub-populations including foreign-born persons, racial and ethnic minorities, and disproportionately affected groups (4). There are many methods of testing for both latent and active tuberculosis, including tuberculin skin tests (TST) and interferon-gamma release assays (IGRA) (1, 4). TST is the most common form of testing. TST is a delayed-type hypersensitivity test that looks for a cellular immune response to an intracutaneous injection of purified protein derivative (PPD). PPD consists of a solution prepared from cultures of tubercle (3).

Both TST and IGRA can detect infection within two to six weeks after exposure (1). However, such tests are not always sensitive enough to identify all positive cases. Up to 25% of people with active disease do not have reactions to PPD tuberculin, resulting in false-negative TST (1). This means that a negative TST cannot rule out infection or disease and, therefore, signs and symptoms consistent with TB must be taken into account when identifying cases (1).

Diagnosis of active TB disease is based on epidemiologic evaluations of risks, clinical symptoms and findings, and laboratory tests (3). These tests include TST, IGRA, chest radiographs, microscopic examinations, and sputum or culture of biopsy specimen (3). Chest radiography is used as a follow-up test for patients with positive TST results but is also used in highly suspected cases or in cases that might illicit false-negative TST results. Abnormal chest radiography can signify pulmonary tuberculosis, but normal chest radiography does not rule out active disease. For example, patients coinfected with HIV often have normal chest radiographs (11).

Cases of active TB are confirmed through the isolation of the *M. tuberculosis* complex on culture (1). Taken from sputum or biopsy tissues, culture is a slow process that can sometimes lead to misdiagnosis. Levy et al. found cultures to be 81.5% sensitive, which is greater sensitivity than sputum smears, but low sensitivity means that some cases will remain unidentified (12). Acid-fast bacilli (AFB) sputum smear tests can confirm tuberculosis diagnosis, but there are also problems with this test. If the number of tubercle bacilli present in the sputum sample are low, as is often the case in noncavitary disease, the test can provide low sensitivity. Levy et al. found sputum smear sensitivity to be as low as 53.1% (12). Highly infectious cases of TB are often sputum-smear positive for AFB, and 65% of these cases result in death if left untreated (1). More recent studies conducted in the United States and Canada have found that even sputum-smear negative but culture-positive cases can be highly infectious as well (1).

6. Treatment

Antimicrobial chemotherapy is the treatment of active TB disease through a combination of antituberculosis drugs (1). This treatment can help to eliminate contagiousness within two to four weeks after the start of therapy (1). Treatment is available for both active and latent tuberculosis, but because active growing and latent bacilli work differently, treatment is complex. Drugs that are capable of killing actively replicating bacilli are often unable to kill bacilli in the resting phase. Treatment continues for a minimum of six months to allow latent organisms to be exposed and killed by the drugs during metabolic activity. Unfortunately, this extended treatment period provides the opportunity for the organism to develop drug resistance. The use of multiple drugs during treatment can reduce this risk. Common drugs for treatment include isoniazid, rifampicin, pyrazinamide, and ethambutol (3). In the United States, TB programs focus on treating patients with LTBI to prevent the development of active disease, a method that is often too expensive for implementation in developing countries (13).

7. Global TB Prevalence

TB infections and disease are found in virtually every country (1). Today, one-third of the world's population is latently infected with TB, and approximately 8.6 million new cases of active TB occurred in 2012 (14). TB is the eighth most prominent cause of death in the world, causing 1.3 million deaths annually (1, 14). The epidemiology of TB in developed countries, which are low-incidence areas, varies drastically in comparison to TB in developing countries, where TB disease is often endemic and 80% of cases occur (1, 15, 16). In 2012, Southeast Asia had the highest number of cases, at 2.3 million, while Africa had the highest per capita rate of TB, and 255 per 100,000 (14). Many developing countries also face pandemic levels of human immunodeficiency virus (HIV), increasing the risk and severity of TB disease. In sub-Saharan Africa, where the majority of the world's HIV cases are located, more than 30% of patients with TB are co-infected with HIV (15). Drug resistance has become more common among TB patients, especially in developing countries where TB incidence is high. Multi-drug resistant TB, or MDR TB, involves resistance to at a minimum isoniazid and rifampicin. MDR is common among HIV-positive cases and is associated with high infectiousness and mortality (1). In 2008, the World Health Organization reported that 4.8% of TB cases are MDR TB (1). More recently, extensively drug-resistant TB, or XDR TB, has been identified due to poorly managed TB therapy. XDR involves MDR with additional resistance to fluoroquinolone drugs and any of amikacin, kanamycin, or capreomycin (1).

Section 2: Tuberculosis in the United States

1. Incidence and Decline

Globally, TB prevalence is the lowest in the Americas and in Europe (17). While prevalence is lower in the United States than in developing countries, certain populations remain at high risk for infection, disease, and transmission (1). In the 1980s, funding for and action by TB control programs globally, including in the United States, began to dwindle. Control efforts deteriorated because it was believed that TB was no longer a major threat (1). This reduction in TB control programs coincided with the emergence of HIV, increased immigration from high-incidence to low-incidence regions, and poorer socio-economic conditions among the most impoverished, and as a result, TB reemerged as a major public health concern (1, 3, 4). In an attempt to curb this trend, efforts to fight TB in the United States have double since 1993 through a formative influx in government commitment and investment in TB control and treatment programs (3, 4, 15). Since this time there has again been steady decline in TB rates in the United States (4, 15).

In 2012, 9,951 new cases of tuberculosis were reported in the United States, a 6.1% decrease since 2011 and the first time the United States reached a prevalence below 10,000 during the 21st century (4). This rate of decline is slower among certain populations, especially among immigrants. TB rates from 2011 to 2012 were higher in 17 states but lower in 33 states (4). Four states in the United States report half of the total TB cases in the US in 2012: California, Texas, New York, and Florida; accounting for 4,967 cases (4).

2. High-Risk Populations

Tuberculosis is a social disease. It occurs disproportionately among the disadvantaged: the homeless, the malnourished, and the overcrowded. Risk factors for TB in low-incidence areas like the United States include age, poverty, being a minority, and living in urban areas (3, 4, 8). Locations where TB continues to spread in the United States include nursing homes, homeless shelters, hospitals, schools, and prisons (1). The populations where new cases are found are often among immigrants and foreign-born persons, racial and ethnic minorities, HIV-positive persons, people spending time in correctional facilities, substance abusers, and the homeless (4). Asians, blacks, and Hispanics have rates between 6 and 25 times as high as whites in the United States (4). Among people born in the United States, the greatest disparities in TB rates are found between whites and blacks. Blacks are 5.8 times more likely to be infected with TB than are whites (4).

In an outbreak in South Carolina in 2007, 25% of the cases were found to have been incarcerated and substance abuse was found as one of the two primary factors contributing to the cluster's growth (18). In 2012, 4.2% of TB cases were in a correctional facility at diagnosis, and from 1993 to 2010, 24% of all national US-born primary cases had a history of substance abuse involving excessive alcohol use, injection drug use, and/or non-injection drug use (4, 19). This proportion increases to 31% among US-born recurrent cases (20). In 2012, 12.1% of TB cases had a self-reported history of excessive alcohol abuse (4). Among 473 patients in Los Angeles in the early 1990s, 25.6% were found to be HIV-positive and 62% were involved in TB clusters (21).

Among persons with known HIV status in 2010 (known in 80% of reported cases), 7.7% of TB cases were also HIV-positive (4).

As of 2012, 5.6% of persons with TB in the United States have been homeless in the past year (4). Homeless populations in the United States face an assortment of health problems. Homeless persons often suffer from serious health problems including mental health, substance abuse, injuries, assault, skin problems, and more. Chronic conditions are poorly controlled in homeless populations, and infectious diseases occur at higher rates than in the general population (including TB, HIV, Hepatitis, and sexually transmitted infections), and they have higher mortality rates (22). Diseases in the homeless also pose a threat to the general population. Shelters can provide environments favorable to disease transmission because of crowding, which drives people into more frequent contact (22). Large clusters of TB cases have been linked to homeless shelters and soup kitchens (23). Homelessness has also been linked to poor treatment adherence, which can lead to TB recurrence (20).

The persistence of TB cases in the United States is largely concentrated among foreign-born individuals (4, 15, 24). While TB rates in foreign-born persons has also declined since the 1990s, foreign-born persons have become an increasing proportion of TB cases in the United States, representing 63% of all cases (4). In 2012, the rate of TB among this group was 11.5 times that of US-born persons (4). The majority of foreign-born cases originate from a select few countries immigrating into the United States; 21% from Mexico, 12% from the Philippines, 9% from India, 7% from Vietnam, and 6% from China (4).

Drug resistance is not as common in the United States as it is in developing countries, but some MDR and XDR cases have been reported. In the United States, drug-susceptibility tests for at least isoniazid and rifampin were conducted for approximately 97% of TB cases in 2011, and 1.6% of those cases were found to be MDR (4). This was a increase from the 1.3% of cases found to be MDR in 2010 (4). Of those 127 cases with MDR in 2011, 86% of those were foreign-born persons (4). Only one case of XDR was reported in the United States in 2012 (4).

3. CDC and TB Reporting

The TB control strategy in the United States consists of case detection through focused testing in high-risk populations, treating latent infections to prevent the development of active disease, case reporting and case management, surveillance, evaluation, and outbreak control (25). This strategy of targeting testing focuses in high-risk and congregate settings were TB transmission is more likely, such as jails, homeless shelters, and long-term care and health care facilities (25).

Standardized national reporting to the Centers for Disease Control and Prevention (CDC) from state and local health departments in all 50 states and the District of Columbia began in 1953 (4). All cases that meet the CDC and the Council of State and Territorial Epidemiologists' case definition for a verified case of TB should be reported to the CDC through the submission of a Report of a Verified Case of Tuberculosis (4). The case definition for TB has changed several times over the century (3). Currently, cases are considered reportable if they meet either the clinical case definition or if they are laboratory confirmed (2). The clinical criteria for reporting a verified case of TB in the United States requires a positive TST or IGRA, signs and symptoms consistent with

TB, a complete diagnostic evaluation for TB, and being treated with at least two antituberculosis chemotherapy drugs (2). Cases must meet all four of the criteria listed above to be considered a verified cases of TB (2). For laboratory confirmation, cases only need to meet at least one of the three requirement (2). These requirements involve a positive culture of *M. tuberculosis* from any clinical specimen, representation of *M. tuberculosis* during nucleic acid amplification, or a positive smear for acid-fast bacilli (2, 3).

In the instance of TB recurrence, there are further considerations that must be met in order for states to count patients as a verified new case of TB. According to current official CDC practice, cases should not be counted and reported as a new verified case of TB if the new case arises from the same patient within 12 months of their first case or since treatment completion (2). Cases should only be reported if more than 12 months pass after the patient completed therapy or if more than 12 months pass after the patient was lost to supervision during treatment (2, 3).

Section 3: Second Episodes of Tuberculosis Disease

1. What is Recurrence?

The definition of recurrent TB disease has not been well established and is not consistent among current literature. Most studies have varying criteria for defining those who fit within their own definition of recurrence, while others chose to use different terms to describe secondary episodes of TB disease. Some researchers argue that for a truly secondary episode of TB disease to occur, the first episode should conclude with treatment completion. However, even within this line of thought, there is no consensus about what is entailed in treatment completion. These requirements range from simply affirming treatment success (26, 27), completing the full treatment course subscribed (28-31), or having accomplished curative therapy (9, 32-35) to either having a positive culture after bacteriologically confirmed cure or correctly/successfully completing treatment (36-38). Other researchers are more specific in their requirements including both culture conversion and treatment completion (39, 40) and, additionally, clinical recovery after the first episode (24, 41, 42). Further still, some researcher provide a more inclusive view of second episodes, only requiring either a positive culture (9), based on the time period between episodes (20, 43, 44), or more widely, including any second episode, even if the first episode resulted in treatment failure (45-47).

Among researchers with more inclusive views of secondary episodes of TB disease, other terms have been used to describe recurrence. Jasmer et al. define recrudescent TB disease as a positive culture found four months after treatment for the first episode began and before treatment completion (39). Middelkoop et al. define retreatment TB as an episode occurring in patients who had previous TB disease treatment, including those who completed treatment or who interrupted or failed treatment (37). Verver et al. distinguish between explained recurrent TB based on treatment failure and unexplained recurrent TB based on treatment completion (47). The required length of time from the end of the first episode to the initiation of a new episode to be considered a recurrent TB case varies in different studies, ranging from three to 12 months (20, 24, 26, 29, 31, 33, 37, 43). In accordance with CDC reporting in the United States, the time from the end of the first episode before the subsequent episode occurs must be more than 12 months to be counted as a second case of verified TB disease (48).

Recurrent TB cases are generally understood as either exogenous or external reinfection or endogenous, meaning internal, reactivation. Commonly, endogenous reactivation reflects previous episodes that were incompletely cured and thus return – usually represented by the same genotype in both episodes, and exogenous reinfection reflects previous episodes that were cured but a new infection occurs – usually represented by different genotypes in each episode. Again, these definitions have not been consistent among studies distinguishing second episodes of TB disease. The definition of reinfection has remained fairly stable, but some authors use the terms recurrence and relapse interchangeably (33, 38, 43) while others use relapse and reactivation interchangeably (9, 24, 26, 28, 36, 39, 45).

2. Risk Factors for Recurrence

Recurrence rates are not only an expression of treatment efficacy and the effectiveness of a country's TB control program, but also represent the strength of an individual's immune system (49). Significant risk factors for recurrence include substance abuse, birth origin and immigration status, and clinical symptoms: pulmonary disease that is sputum-smear positive and have a cavitary site of infection (18, 26, 29-31, 37, 41, 49, 50). Other studies also found that age 25-44 years (30), prior treatment regimen with isoniazid and rifampin only (30, 31), gender (37), and chronic lung disease (50) were risk factors for recurrent TB. Recurrence is more likely with drug-susceptible TB in the initial episode than it is with MDR or XDR TB (29). Recurrence was also found to lead to poorer outcomes than single TB episodes, including a higher risk for the development of multi-drug resistant TB in the second episode as well as higher mortality rates (26, 30, 50). Other risk factors identified for recurrent disease include HIV positivity and drug abuse (24, 29, 43). HIV positivity is a very prominent feature of recurrent cases, and HIV positive patients have twice the risk of developing a recurrent episode (24, 29).

Endogenous reactivation is often associated with inadequate treatment in the first TB episode (29, 46, 51). Risk factors associated with reactivation include cavitary disease and HIV (24, 26). Reactivation was also associated with MDR in the second episode (29). Exogenous reinfection cases are often younger, more alcoholic, and more often female than reactivation cases (43). Risk of reinfection was higher among HIV positive patients (26, 29-31, 41, 42, 46, 47, 49, 50, 52). The time from treatment completion until the second episode is a factor that varies between reactivation and reinfection, where risk of reinfection persists over a longer period of time than reactivation. In one study, time from treatment completion to reactivation averaged 1.8 years (29). Bang et al. found that reactivation rates increased up to four years after treatment completion and then decreased to a low level after 4 years (26). While the risk for reactivation decreases over time, the risk for reinfection increases over time (26). Although proportions of reinfection remain lower than reactivation, they persist at steady rates over time while reactivation decreases (26). Dobler et al. found that time to reinfection after treatment was finished ranged from eight months to 57 month, averaging 1.4 years (29). Bang et al. found that treatment completion was less of a factor in reinfection than it was in reactivation and rates continue to increase after four years (26).

Tuberculosis treatment failure, as defined by the World Health Organization, occurs when a TB patient undergoing antituberculosis therapy continues to have a positive smear and/or culture or reverts back to a positive result after five months from treatment initiation or who converts to a positive result for the first time after two months post treatment initiation (53). Treatment failure was often believed to result from poor treatment adherence, incorrect treatment regimens, poor drug quality, malabsorption by the patient, or from ineffective TB control programs. The use of genotyping technologies in TB studies has shown that this is not always the case. Some instances of assumed treatment failure are actually attributable to unidentified reinfection (9).

3. High-Incidence Characteristics

TB recurrence in areas with a high incidence of TB present differently than in low-incidence areas. This difference also exists among endogenous reactivation and exogenous reinfection. High-incidence areas, including many countries in Africa and Asia, have TB recurrence rates ranging from 0.4 to 18% (29, 47). Verver et al. found that 14% of recurrent cases in high-incidence areas emerged after successful treatment, while 28% of recurrences presented after treatment failure, not representing a significant difference between rates of recurrence and treatment success or failure in the initial episode (47).

Exogenous reinfection is believed to be the more common type of recurrence in high-incidence TB regions of the world. Risk of reinfection was found more often in high-incidence settings where intensive TB transmission is occurring (26, 29-31, 41, 42, 46, 47, 49, 50, 52). People in high-incidence areas that had previous TB episodes were found to be at greater risk for developing active disease when reinfected (47). In high-incidence areas, reinfection rates were found to be seven times that of general TB incidence rates and four times that of age adjusted rates for initial TB infection, at 2.2 per 100 person-years (47). Prevalence of reinfection in high-incidence regions averaged at

77% (29, 47). Data available for identifying risk factors for reinfection in high-incidence areas were limited. However, factors that were found not to be significantly associated with reinfection in these areas included age, sex, HIV, immigration, multidrug resistance, smear positivity, and rate of infection (24, 47).

4. Low-Incidence Characteristics

In areas with a low incidence of TB, such as the United States, Canada, and much of Western Europe, the prevalence of recurrence was found to range from 1% - 7% (29). In the United States specifically, recurrence rates were observed between 4% - 6% in 2010 (20). In low-incidence areas, directly observed therapy (DOT) was found to result in lower rates of recurrence, suggesting that successful treatment is more effective in preventing recurrent episodes in low-incidence areas (49).

Reactivation is believed to be the more prevalent type of recurrent TB in low-incidence areas, ranging from 73-92% of recurrent cases (24, 26, 29, 30). While reinfection is much less common in low-incidence areas, it still contributes to TB recurrence and should be addressed in TB control strategies (29). Prevalence of reinfection in low-incidence areas ranged from 1.9 to 33% overall, and 26-27% of recurrent cases (26, 29, 43, 47). In the United States in 2010, reinfection was found in 8.3% of recurrent cases (20). In low-incidence areas, reinfection has been focused among select populations such as substance abusers, homeless shelter residents, and advanced HIV infected patients (24, 47). The fact that reinfection is more common in high-incidence areas than in low-incidence areas supports the claim that greater prevalence of *M. tuberculosis* represents greater risk for reinfection (24).

5. Immigration

Immigrants moving from high-incidence countries to low-incidence countries are major contributors to both new TB incidence as well as rates of recurrence (24, 29, 54). Codecasa et al. found that 3% of immigrants had previous cases of TB (54). More than 80% of TB cases in low-incidence Australia arose among immigrants (29). Kim et al. found that in the United States, 53.4% of recurrent cases were US-born while 46.3% were foreign-born (20). Undocumented immigration, a major point of concern in the United States, can lead to delayed diagnosis and treatment, increasing the risk of TB transmission (54).

Risk factors for recurrent TB among immigrants vary based on the length of time since immigration. Studies have found that risks are greatest in the time immediately following immigration due to factors such as stress and malnutrition (49). Then, new TB incidence decreases sharply over the first two years after arrival (49). After the initial two year decline, these elevated risks persist for at least a decade after arrival (49). The time from immigration until the second episode is a factor that varies between reactivation and reinfection. Reactivation for immigrants is more drawn out, continuing beyond two years following immigration (49).

Recurrence rates among immigrants occur at similar rates to those of initial infections in their home countries (29). In the Netherlands, immigrants make up 85% of TB cases and have similar rates to overall incidence rates for immigrants, suggesting that the majority of TB among immigrants occurs because of recurrence (29). Continuing risks for recurrence among immigrants include continued transmission from fellow immigrants and continued transit between new and home countries (49). Other reasons immigrants remain at higher risk for reinfection while in their new countries is due to various factors that have been linked to reinfection in low-incidence areas, including socioeconomics, overcrowding (represented by higher *M. tuberculosis* strain circulation in close proximity), and poor hygienic conditions (24, 49).

Section 4: Genotyping

1. Definition and Uses

Molecular genetic typing (genotyping) is used by laboratories to analyze the genetic material of *Mycobacterium tuberculosis* (7, 55). Sections of the genetic content of *M. tuberculosis* form patterns that vary in differing strains of *M. tuberculosis* (55). This variation helps to distinguish *M. tuberculosis* strains between as well as within individuals. Genotyping of *M. tuberculosis* has multiple applications, including assisting in the identification of TB outbreaks, providing evidence for recent transmission, testing for drug susceptibility, identifying laboratory cross-contamination, evaluating second episodes of disease, and guiding treatment regimens (45, 55). Along with epidemiologic links, genotyping can be used to detect the presence of outbreaks because patients who present with the same strain of *M. tuberculosis* may be related (55). In the 1990s, genotyping was used to estimate that 20% to 50% of TB cases were due to recent transmission in the urban areas of Los Angeles and San Francisco (21, 23). Genotyping is also used to both support and refute the presence of cross-contamination, lending to the discovery that about three percent of isolates genotyped were false-positives, supporting the development of better laboratory contamination controls (29, 45).

Genotyping has arguably become one of the most important factors in the evaluation of second episodes of tuberculosis. In recurrent TB disease, genotyping can help researchers better understand the role of endogenous reactivation and exogenous reinfection, can help guide treatment courses, and can help identify treatment failure (7, 45). Previously, the majority of TB cases were believed to arise as a result of endogenous reactivation of past infections. However, genotyping advancements have challenged this belief (45). Recent enhancements in genotyping techniques and technologies have made distinguishing reinfection from reactivation fast and relatively easy (29, 55). Discriminating between reactivation and reinfection in recurrent TB cases is important because endogenous reactivation can be an indication of treatment failure, requiring alterations in treatment plans, while exogenous reinfection often represents recent transmission and strain circulation (29, 45).

2. Typing Methods

M. tuberculosis is collected through clinical specimen samples originating from sources such as sputum, bronchial washes, urine, blood, and tissues from suspected TB patients (55). Isolates are *M. tuberculosis* that grows in culture, which are then sent to laboratories for genotyping (55). DNA is extracted from cultured isolates and sections of the *M. tuberculosis* genome are analyzed. Currently, there are three major methods used in the identification of *M. tuberculosis* strains: restriction-fragment-length polymorphism (RLFP), spacer oligonucleotide (spoligotyping), and mycobacterial interspersed repeat units – variable-number tandem repeat (MIRU-VNTR); the combination of the three providing the most accurate level of distinction (29).

RFLP is the standard approach to genotyping and was the only method available for much of the 1990s (29, 45). RFLP analyzes the distribution of a section of the *M. tuberculosis* genome at insertion sequence IS*6110* of the strain (45, 55). RFLP typing requires a large amount of DNA, meaning that the specimen must be cultured for a long period of time (29, 45). DNA must then be purified, cut into fragments, probed for IS*6110*, and then captured on film (55). RFLP patterns containing seven or more bands are considered sufficient for accurate discrimination between strains (45, 55, 56). If fewer than seven bands are present, additional genotyping techniques can be used in tandem with RFLP for analysis (45, 55, 56). These genotype results are available in datasets shared between laboratories and used during outbreak investigations (45, 57).

Spoligotyping is the use of the patterns of spacers in a *M. tuberculosis* strain and can be used for genotyping (58, 59). This typing method looks for the presence or absence of spacer sequences, which differs between strains, in a direct-repeat region of the genome (45, 55). While spoligotyping alone has less discriminatory power than RFLP, it takes less time to develop results because it requires smaller amounts of DNA, can be used on clinical samples, and can be expressed digitally (29, 45, 59). Spoligotyping is often used in combination with one of the other typing methods for further discrimination between TB strains.

Both spoligotyping and MIRU-VNTR are polymerase-chain reaction (PCR) based methods, meaning that these typing methods can be used on stored DNA samples and produce digital results (29, 55). Like spoligotyping, MIRU-VNTR does not require DNA purification, is simpler than RFLP, and can simultaneously analyze large numbers of strains (45). With a total of 41 loci reported, *M. tuberculosis* genomes contain many

MIRU that vary in length and sequence (45, 55). These differences are used in MIRU-VNTR for the discrimination of strains (55). Most laboratories use 12-loci MIRU in combination with spoligotyping to report strain differentiation and are used in the identification of recurrent disease (29, 45, 55). The addition of larger numbers of loci in MIRU-VNTR analysis helps to increase its discriminatory power over RFLP techniques and is useful in identifying clusters of cases faster than RFLP, helping to thwart potential outbreaks (45, 52).

3. Strain Discrimination

In exogenous reinfection, *M. tuberculosis* strains are considered different if they are clearly distinct by any of the common genotyping methods, usually greater than a few bands for RFLP typing or multiple loci differences in MIRU-VNTR typing (29, 47). While many previous studies examine risk factors for recurrent TB, relatively few studies have genotyping data available for a large portion of their recurrent cases, preventing the studies from distinguishing between reactivation and reinfection (29, 37). In the studies that do include genotyping information on recurrent TB cases, the limited number of TB cases with genotype data available was cited as a limitation by many authors, making it difficult to generalize findings or perform further data analysis (29, 30, 47). Distinction of recurrent TB cases, whether reactivation or reinfection, is important because of its potential treatment and policy implications as well as its recognition in developing more comprehensive and effective TB control programs. For example, because of a greater understanding of reinfection, patients are no longer routinely started on second-line therapy at their second episode (29).

4. National Tuberculosis Genotyping and Surveillance Network

In 1993 the CDC funded six regional laboratories through the National TB Surveillance System (NTSS) with the purpose of establishing a national genotyping database involving epidemiologic, geographic, and clinical data (7, 60). This network was expanded in 1996 to include sentential surveillance and continued to expand into the current network comprised of the CDC, seven labs, and seven regional surveillance sites, known as the National Tuberculosis Genotyping and Surveillance Network (NTGSN), with the goal of using DNA fingerprinting to guide TB control program activities (7). Contracting with laboratories in California and Michigan, the National Tuberculosis Genotyping Service (NTGS) began in January 2004, with the goal of increasing rapid genotyping of isolates for availability within 10 days using the PCR-based methods of MIRU-VNTR and spoligotyping (52, 56, 60).

Originally, NTGS laboratories utilized 12-locus MIRU-VNTR, along with spoligotype, assigned as a PCRType ("PCR" plus a sequentially assigned set of five numbers) to all isolates genotyped in the United States (52, 56, 60). In addition to being assigned a PCRType, NTGS laboratories assign state-specific designations to each combination genotype result (55). When further discrimination is needed for distinguishing because of closely related strains, NTGS laboratories use the addition of RFLP methods for added analysis (56). In April 2009, NTGS laboratories expanded from 12 to 24 loci to increase discriminatory capacity, assigning each 24-locus MIRU-VNTR and spoligotype combination a unique GENType ("GEN" plus a sequentially assigned set of five numbers) (52). This increased discriminatory power helps to better identify transmission patterns and better differentiate between endogenous reactivation and exogenous reinfection (52).

Many isolates collected before April 2009 have been reanalyzed by the NTGS laboratories to include the expanded 24-locus MIRU and accompanying GENTypes, meaning that isolates can have the same PCRType but different GENTypes (52). When NTGS was initiated in 2004, only about 50% of culture-positive cases were genotyped (55, 60). This number increased to 86% in 2007 and to 88% nationally in 2010 (55, 60). Since 2004, over 70,000 isolates have been genotyped through the genotyping network (60). From 2008 to 2010, 23,108 TB cases had reached to goal of having at least one genotyped isolate per case (60).

Section 5: Immunology

1. Why Reinfection is Possible

Multiple theories exist about why it is possible for reinfection to exist. Reinfection of *M. tuberculosis* in individuals is possible because initial infections do not provide a fully protective immunity against subsequent infections. In this theory of the existence of reinfection, recurrence occurs because TB transmissibility is high enough to promote infections and recurrence is sustained (51) Other theories on why reinfection is attainable involve ideas that certain individuals have a higher predisposition for TB infection and disease. In this theory it is argued that having had a previous episode of TB increases the individual's vulnerability to other strains of *M. tuberculosis* (27). Reinfection can occur on a continuum around the initial episode (9). On this continuum, the reinfection can occur before the first episode of disease develops and persist in the latent form, undetectable during the active disease stage of the other strain. Alternatively, the reinfection can occur during the treatment of the first episode of disease and delay presentation. The reinfection can also occur after the first episode of disease has concluded with curative treatment (9).

Before genotyping techniques were developed and applied to TB cases, it was believed that recurrent disease only developed through reactivation of the initial episode, but studies like Wolleswinkel and van den Bosch, who in 1992 examined declining TB rates in the Netherlands, changed this understanding (43). Wolleswinkel and van den Bosch argued that if people were only becoming infected through reactivation, TB rates would remain the same among older age group while there would be a decline in the overall population. Instead they found that TB was declining in the general population as well as in older populations, showing evidence that some disease must be resulting from reinfection (43).

2. Protective and Partial Immunity

Before the development of genetic typing, reinfection was believed to be a rare event (24, 35). It was generally assumed that protective immunity was gained through initial infections, making reinfection impossible or extremely rare (24). These beliefs lead to poor outcomes with regard to treatment recommendations (45). Without knowledge of reinfection, individuals concluding treatment in first episodes were assumed to have gained resistance to reinfection through protective immunity against subsequent infections, and were not instructed to avoid future exposure or be reevaluated for further exposure (9, 45, 57).

Arguments for protective immunity acquired by the initial episode are challenged by evidence of simultaneous strain infections as well as the high prevalence of reinfection found in many populations (9, 35, 49, 61). Protective immunity is further complicated in HIV-infected patients, making it more difficult to control disease (34, 57). Failure of protective immunity helps to explain the somewhat ineffectual results of bacilli Calmette-Guérin (BCG), providing evidence that a TB vaccine must go beyond evoking immune system response to natural infection in order to be effective (45, 51).

In contrast to protective immunity, partial immunity means that individuals are protected against reinfection while they are infected with their initial episode, but regain all or some of their prior susceptibility after curative treatment is completed (51). Some researchers have argued in favor of the existence of select partial immunity in TB infections (51). Sutherland described partial immunity in the 1970s from a study in which he argued that 63% of Dutch men and 81% of Dutch women developed protective immunity against active disease in a second infection because of distant past infections (62). Other studies supporting partial immunity suggest that reinfection rates in high-incidence areas prove that initial episodes only provide partial immunity (49). Partial immunity also helps to explain why BCG has variable outcomes (51).

Other studies argue against any kind of partially protective immunity due to previous episodes (43). They argue that if previous infections provided partial immunity then reinfection rates would be lower than new incidence rates, which was found not to be the case in a study conducted in South Africa in which reinfection rates were higher than incidence rates (47). This study is important because of its implications for TB control activities and vaccine development (27). Furthermore, other studies suggest that previous episodes actually increase immune susceptibility to reinfection, depending on the individual since infection affects people differently (36). Gomes argues that even if previous infections infer some amount of partial immunity, it would be extremely difficult to measure the amount of protection provided because individuals with reinfection are naturally overrepresented in studies because they are already at higher risk for their initial infection (36).

3. The Effects of First Infection Strain Type on Second Infection Strain Type

Few studies evaluate whether the type of strain from an initial infection affects the type of strain in a second episode in recurrent TB cases because of the complexity of the evaluation and because of the lack of tools available for analysis. In high-incidence areas, people successfully treated for TB are at greater risk for developing active disease due to reinfection, but it is unclear if this is due to the high risk of reinfection in high-incidence areas or because there exists a subgroup of people who are intrinsically more vulnerable to TB disease, potentially due to environmental or genetic factors (47, 63). The problem with elucidating this uncertainty is the lack of molecular tests available that can differentiate between old and new infections (47). While not much may be understood at this time about one infection's impact on another, some evidence shows that factors influencing individuals' genes are associated with susceptibility to certain TB strains or families of TB strains (63).

4. Strain Virulence and Prevalence

Before applications of genotyping, the common belief was that *M. tuberculosis* strains were equally virulent (21, 23, 45, 64). However, genotyping analyses have proven that a small number of strains cause the majority of TB cases, indicating that some strains are more effective at causing transmission or progression from infection to disease than others (21, 23, 45, 64). Genotyping has also provided evidence that certain strains interact differently with their hosts with varying potential for transmission potentials (23). In a 2008 study conducted in South Africa, one spoligotype was responsible for the majority of multidrug-resistant cases, while another was responsible for most of the extensively drug-resistant cases (50). Increased virulence of a strain might also coincide with an enhanced ability to develop within hosts as a recurrent infection or disease.

Great variability is found in the number and type of genotypes throughout epidemiologically unrelated cases (45). However, the distinct interaction of strains with their host means that some strains, while more virulent, are also more prevalent (45). Beijing strain, a member of the East Asian Lineage group, are often prevalent both in the United States and globally (45). Theories as to why Beijing strains are so prevalent include the hypothesis that this occurs because Beijing strains were initially introduced in multiple locations before other strains were present in those areas, allowing Beijing greater time for transmission, and the hypothesis that its prevalence is because of biologically enhanced transmission potential (45). Some believe Beijing strains might be more easily aerosolized, enhancing their abilities to establish infection and progress to TB disease (45).

In Australia, Beijing strain was related to greater drug resistance and one case of simultaneous infection with two different Beijing strains (29, 65, 66). In a 2012 Morbidity and Mortality Weekly Report in the United States, the most common genotype was seen in 4% of all genotyped cases, found in 43 of the 51 NTGSN reporting areas (60). In the United States, PCR00002, of the East-Asian Beijing family, is the most nationally and globally prevalent, presenting in cases in every state in the United States (18). A South Carolina cluster found that this strain was involved with recent transmission due to recurrent disease (18).

5. Mixed Infections

When multiple strain infections develop, the ability to differentiate between treatment failure and reinfection is further complicated. In this instance, a strain of TB that was dormant during the initial treatment stage could become active during treatment and could result in the misclassification of the cause of the recurrence, often attributed to treatment failure if genotyping is not conducted (39). Sources of multiple strain infections can arise during treatment within hospital facilities when patients come into contact with other patients infected with different virulent strains and become cross-infected during the course of their treatment (53).

The prevalence of multiple TB strain infections is not well established. These types of infections contribute to the limitations in distinguishing reactivation from reinfection (9). Risks for mixed infections include HIV and time in corrections facilities, exposing people to multiple sources of infection (67). There is a theoretical possibility that some people with reinfection were actually infected with multiple strains at the time of the first episode, but only one strain presented at the time of the first disease and the second episode is really a reactivation of the second initial infection presenting itself at a later time (29). There is also a theoretical possibility that reinfection with the same strain is underestimating the prevalence rates for reinfection. Situations such as these mean that an underestimation of reinfection rates might exist. Warren et al. argue that this underestimation would probably not be large in areas with high strain diversity, increasing the chance that new infections are in fact new strains (47). However, this might not be the case in low-incidence areas where the opposite could be true and the underestimation of reinfection due to second episodes presenting with the same strain is much higher than expected.

Section 6: Treatment Implications

1. Development of Drug Resistance

The emergence of drug-resistance due directly to treatment failure has not been adequately proven, and the development of drug-resistance is sometimes confused with reinfection if genotyping techniques are not applied (8, 53). In a study conducted by Andrews et al., all of the patients in the study with MDR or XDR acquired the resistant strains through reinfection rather than through developed resistance due to treatment failure or inadequate treatment compliance (50). In another study by Kruuner et al., the researchers came to the same conclusion during extended follow-up periods. In the study, six patients continued to test positive for drug-susceptible strains during 110 months of follow up. The five patients in the study who developed MDR TB exhibited a different strain than their initial infection at the same point in time as their development of drug-resistance (53).

2. Treatment Choices

The outcome of a case of active TB disease is dependent on the effectiveness and the administration of their antituberculosis therapy (41). Focusing solely on preventing the development of drug-resistance during treatment ignores the problem presented by the possibility of reinfection and has the potential to lead to harmful outcomes for the patient (50). In Krunner et al.'s study, all five of the patients who developed reinfections with MDR TB strains were initially considered by their physicians to have development drug resistance due to "unsupervised drug administration, poor patient compliance with therapy, and errors in medical prescriptions of drug regiments" (53). The use of genotyping along with a better understanding of reinfection in these cases could have prevented delayed treatment for the new drug-resistant strains or even prevented the development of secondary episodes of active disease.

Recurrence due to reactivation and reinfection are often clinically indistinguishable. This is problematic because extending follow-up time after treatment completion to look for evidence of recurrence could affect the rates of recurrence (9). For example, recognizing that second episodes might be due to a new strain of *M. tuberculosis* would mean that some patients could still be treated with isoniazid even if it had been administered during their first episode (45). This could also mean that treatment regimens for reinfection could better reflect current drug-resistant strains in circulation in a person whose first episode would require a different treatment regimen than one based on past drug-resistance patterns at the time of their first episode (39, 45).

Recommendations for the treatment of recurrent cases include testing specimens from patients at multiple points throughout the treatment process, including pre- and post-treatment periods, to look for different strains of infection (53). If risk factors for reactivation and reinfection are better understood, patients could benefit by earlier identification and treatment of recurrence and, ideally, prevention of the recurrence (41). Retreatment of TB is expensive; therefore, if new strains are identified in second episodes and considered before treatment regimens are selected, second episode strains could be treated with the same first-line drugs as were used during first episode treatment, saving money when therapy for drug-resistance is not necessary (41). Better understanding of recurrence could lead to the prevention or decreased duration and severity of second episodes through intensifying drug regimens in initial treatment plans, extending treatment periods, or following up with secondary drugs after treatment completion (41). Clinicians should be aware of recurrence and appropriate evaluations should be conducted, including genotyping tests to distinguish between treatment failure, reactivation, and reinfection (39).

3. Vaccine Development

Vaccine development for tuberculosis has been difficult and relatively ineffective. Studies on the efficacy of Bacille Calmette-Guérin (BCG) show estimates between 0% and 80% (68). The failure of BCG to fully protect against TB infection and disease is partially explained by the failure of the human body to develop natural protection to reinfection through immunity after a primary episode of TB (9, 43, 45). In order for a TB vaccine to be effective, it must produce a greater than natural immunity (9, 45, 68). Susceptibility to and predictors of reinfection must be better understood to develop a vaccine that is fully protective against TB (68).

Areas of TB control that should be further studied to prevent future infections include better estimations of exogenous reinfection, better tests that can further distinguish between strains during the infection stage, and further studies into the treatment and prevention of recurrent cases (9, 39). Greater knowledge of the risks and contribution of reinfection and reactivation should be conducted and used in the development of new TB control strategies (9).

Currently there are no molecular tests available that can distinguish between primary infection strains and secondary reinfection strains of TB (9). This is problematic because it leads to a lack of knowledge as to the cause of first episodes of active disease, creating the potential for complicating treatment courses and for the potential selection of an ineffective therapy regimen. This is also problematic because it makes it impossible to predict the development of a secondary episode that stems from a previous reinfection, making the source of the new case difficult to identify. Studies into the optimal duration of treatment for reinfection to prevent subsequent episodes can also help to reduce the burden of recurrence in the United States (39).

CHAPTER II: MANUSCRIPT

The Role of Exogenous Reinfection in Patients with Recurrent TB Disease in the United States

Authors: Julia Interrante, Dr. Neel Gandhi, Maryam Haddad

ABSTRACT

Purpose. Traditionally, recurrent tuberculosis (TB) has been assumed to result from endogenous reactivation. Genotyping now allows us to determine how much of recurrence is actually due to exogenous reinfection. To determine the extent of and to better understand factors leading to reinfection rather than reactivation, we analyzed patients in the United States with two episodes of TB disease during 1993 to 2011.

Methods. The study population was drawn from all TB cases in the 50 states, Puerto Rico, and the District of Columbia, as reported of June 25, 2012. We identified recurrent cases by matching on date of birth, sex, race, country of origin, state, and year of first episode. Genotyping was used to distinguish between reinfection and reactivation. Selection required time from treatment completion in first episode be ≥12 months before the start of second episode. To statistically evaluate the effects of predictors on reinfection, a logistic regression model was fit.

Results. Among patients with recurrent TB who completed treatment during their first episode, 136 patients were identified, involving 116 reactivations and 20 reinfections.

Reinfection occurred in 15% of the population with recurrent TB. Three factors were statistically significant for reinfection after adjustment, including being black or Hispanic (odds ratio (OR) 4.4, 95% confidence interval (CI) 1.1-17.2), living ≤12 years in the United States (OR 3.5, 95% CI 1.0-11.9), and having received treatment exclusively by directly observed therapy (DOT) during first episode (OR 4.8, 95% CI 1.2-19.8).

Conclusions. In persons who experience two episodes of TB, genotyping evidence suggests that the majority of second episodes are reactivation of the first episode. However, minorities, those more recently immigrated to the United States, and those with more rigorous treatment regimens during first episodes have a greater risk for exogenous reinfection. This suggests that these populations are being successfully treated for TB, but other risk factors for recent transmission increase their risk of reinfection. Public health interventions should continue to focus on these populations and their areas of residence, work, and recreation for evidence of recent TB transmission to prevent further and future spread of TB.

INTRODUCTION

In 2012, approximately 8.6 million new cases of tuberculosis (TB) developed globally (14). TB is the eighth most prominent cause of death in the world, causing 1.3 million deaths annually (1, 14). While TB incidence and prevalence is lower in the United States than in developing countries, with 9,951 new cases reported in 2012, certain populations remain at high risk for infection, disease, and transmission, including those impoverished, minorities, immigrants, HIV-positive persons, inmates, substance abusers, and the homeless (1, 48). Certain states have higher background incidence of TB than others, and four states reported half of the total TB cases in the United States in 2012 (48).

Recurrent TB is a second episode of TB disease occurring in patients with a previous episode of disease who completed, interrupted, or stopped treatment during their first episode. In the United States, a case of TB is considered recurrent if more than 12 months elapsed before the second episode is recognized. TB recurrence in countries with a high incidence of TB presents differently than in low-incidence countries. In high-incidence countries, prevalence of recurrence was found between 18% and 32% of TB cases (34, 46, 47). In countries with a low incidence of TB, such as the United States, Canada, and Europe, the prevalence of recurrence was found ranged 1%–7% (29).

The etiology of recurrent TB cases is generally conceptualized as either endogenous, the consequence of internal reactivation of previous disease, or exogenous, representing external reinfection or new disease. Genotyping has served an important role in evaluating the etiology of second episodes of TB. Rather than making the assumption that TB recurrence reflects incomplete cure of the first TB episode, clinicians and public health personnel can now ascertain whether recurrent cases are due to endogenous reactivation or exogenous reinfection, by comparing genotypes from the first episode of disease with those from the second episode (29, 45).

Endogenous reactivation is typically represented by the same genotype in both episodes, whereas different genotypes between episodes are more likely to represent exogenous reinfection. Exogenous reinfection is believed to be more common, 36%–77% of recurrent TB cases, in high-incidence TB settings worldwide, where intensive TB transmission is occurring (26, 34, 41, 42, 46, 47, 49). While reinfection is thought to be less common in low-incidence countries, 8%–27% of recurrent TB cases, it also should be addressed in TB control strategies (26, 29, 30, 43, 47). In low-incidence areas, reinfection is thought to be focused among select populations including substance abusers, homeless shelter residents, and HIV-infected patients (24, 47).

In the past, it was typically assumed that most recurrent TB was due to endogenous reactivation. However, with genotyping now available, it is possible to determine the proportion of recurrence that is actually due to exogenous reinfection. While previous studies have examined risk factors for recurrent TB, relatively few studies have had any genotyping available or had genotyping available for both episodes, preventing them from distinguishing between reactivation and reinfection and impeding the generalization of findings and additional analyses (29, 30, 37, 47). To determine the extent of reinfection and to better understand the factors that lead to exogenous reinfection compared to endogenous reactivation, we analyzed patients in the United States with two episodes of TB during the period 1993 to 2011. Genotyping was used to distinguish between reinfection and reactivation.

METHODS

Data Source and Study Population

This is a retrospective, observational study utilizing data from the United States National Tuberculosis Surveillance System to access the role of exogenous reinfection in recurrent TB cases in the United States. Standardized national reporting of TB cases in the United States began in 1953 (4). All cases that meet the CDC and Council of State and Territorial Epidemiologists case definition are reported to the National Tuberculosis Surveillance System through a Report of a Verified Case of Tuberculosis (4). This study population was drawn from all TB cases in the 50 states, Puerto Rico, and the District of Columbia during 1993 to 2011, as reported of June 25, 2012.

This study used a matching algorithm to identify recurrent cases. Recurrent cases were identified by comparing date of birth, sex, race, and country of origin, as well as state and year of the first episode among TB cases in the National Tuberculosis Surveillance System. This matching algorithm has been validated among recurrent cases occurring from 1993 to 2006 and was found to have a positive predictive value of 97.8% (30).

Recurrent cases were eligible for the study if genotyping data were available for both first and second TB episodes (Figure 1). The time between treatment completion or last follow-up from the first episode and the start of the second episode was required to be more than 12 months. Additionally, cases were required to have completed therapy in the first episode treatment course to be eligible for the principal analysis. A supplemental analysis was also performed to examine the impact of including those who did not complete therapy during their first TB episode.

Genotyping

In 2004, the National Tuberculosis Genotyping Service was created with the intention of offering rapid polymerase-chain reaction (PCR)-based genotyping of isolates from every culture-positive TB case in the United States (52, 56, 60). Initially, 12-locus mycobacterial interspersed repeat units–variable number tandem repeat (MIRU-VNTR), along with spacer oligonucleotide (spoligotyping), was performed for each isolate (52, 56, 60). In 2009, MIRU-VNTR analysis expanded from 12 to 24 loci, and still used in conjunction with spoligotyping, to increase the discriminatory power for identifying transmission patterns and differentiating between endogenous reactivation and exogenous reinfection (52). Prior to the National Tuberculosis Genotyping Service, patient isolates were genotyped based on IS*6110*-based restriction fragment-length polymorphism (RFLP) analysis patterns. The genotyping coverage of culture-positive cases has increased from 51% in 2004 to 93.5% in 2012 (55, 60, 69).

For this study, three methods were used to compare genotypes between TB episodes. For older cases in which PCR-based methods were not available for typing in both episodes, RFLP analysis patterns were used to determine the type of recurrence, comparing the number and length of specific DNA fragments. Otherwise, cases had to match by spoligotype and whatever number of MIRU-VNTR loci were examined (i.e., 12 or 24) for both episodes. One-band differences in RFLP and one-locus MIRU-VNTR differences were accepted as matching results.

In this study, we defined endogenous reactivations as cases in which both episodes of TB were represented by the same genotype. Exogenous reinfections were defined as cases in which the two episodes of TB were represented by different genotypes.

Analysis

The outcome of interest was exogenous reinfection. The primary predictors of interest were age at second episode, sex, race, state, having been born in Mexico, years living in the United States at time of second episode, HIV co-infection, having received treatment exclusively by directly observed therapy (DOT) during first episode, and treatment duration at first episode.

Other variables examined as predictors included country of origin; history of homelessness, corrections, or substance abuse; disease site, cavitary disease, or smear positivity at first episode; treatment success at first episode (i.e., treatment completion and sputum conversion); TB strain lineage at second episode; and time between episodes. Time between episodes was calculated from the time patients stopped therapy at first episode to the time patients began therapy at second episode. If treatment was not initiated or dates were missing, the date of testing for drug susceptibility was used as a proxy.

All data were analyzed using SAS 9.3 (Cary, NC). Descriptive statistics and chi-square tests of association were calculated for all predictors to examine bivariate levels of association with reinfection. To statistically evaluate the effects of the predictors on reinfection, a logistic regression model was fit. All factors were considered potential predictors since there was no main exposure of interest; therefore, no variables were considered confounders in this analysis. Predictors with P <0.1 were screened into the initial model. Independence was assumed based on study protocol.

In order to increase model specificity (i.e., to include only those persons considered “cured” and thus less expected to have recurrent TB), the final model used for analysis included only those who completed treatment during their first TB episode. Interactions between all predictors were assessed and were dropped from the model using backward P-value based elimination. Individual variables were dropped using change-in-estimate elimination. The final model contained sex, race, living in State A, years living in the United States, HIV co-infection, and treatment type and duration. Unadjusted and adjusted odds ratios (OR), along with 95% confidence intervals (CI), were computed. Statistical significance for all calculations was assessed at $P < 0.05$.

Further analysis was conducted on all patients with two episodes of TB in order to examine the differences of the distribution and effects of predictors on reinfection without the restriction of treatment completion in the first episode. In this analysis, all outcomes of first episodes, except death, were accepted, including treatment completion, loss to follow-up, refusal, or unknown.

Ethical Approval

Data collected by the National Tuberculosis Surveillance System and the National Tuberculosis Genotyping Service are part of national routine TB surveillance. Therefore, a determination of "no Institutional Review Board review required" was given by the Emory IRB for this retrospective analysis of existing data. However, approval was received through the Analytic Steering Committee for the Division of Tuberculosis Elimination at the Centers for Disease Control and Prevention (CDC) for access to and analysis of the de-identified dataset.

RESULTS

Participants

Out of the 3,039 potentially recurrent cases identified, 136 patients with two episodes of TB who completed treatment in the first episode were identified for inclusion in the study (Figure 1). The median age at second episode was 50 years (IQR 37-60), 35 (26%) were female, 38 (28%) were black, non-Hispanic, and 33 (24%) were Hispanic

(Table 1a). Recurrent cases occurred in 36 states in the United States and Puerto Rico, and 59 (43%) of the study cases occurred in four states with high background TB incidence. Among those who were foreign-born (n=58, 43%), Mexico was the most common country of origin (n=19, 33%), and the median number of years living in the United States at second episode was 12 years (IQR 7-20).

The most common social risk factor was substance abuse (n=57, 42%), and 18 patients were HIV-infected at the time of the second episode (13%). During first episode, 88 (65%) received treatment exclusively by DOT, whereas 34 (25%) received a combination of DOT and self-administered treatment, and 13 (10%) patients received self-administered treatment alone (Table 1b). The median months of treatment were 7 (IQR 6-10). The majority of recurrent cases occurred between 1 and 2 years after the first episode (43%), and the most prevalent strain of TB during second episodes was EuroAmerican (63%).

Reinfection versus Reactivation

Of the 136 patients who completed treatment, 116 (85%) patients were considered reactivation, and 20 (15%) were considered reinfection based on comparison of genptyping patterns from both episodes. Reactivation and reinfection patients varied on a few distinct factors. Reinfection patients were a median of 7 years younger at second episode, were more often Black or Hispanic, and more often lived in State A (Table 1a). Among foreign-born patients, reinfection was more common among those of Mexican-origin (58% versus 26%) and reinfection patients had lived in the United States a median of 5 fewer years at second episode (Figure 2). Reinfection patients were more often HIV-infected (20% versus 12%), were more likely to have been treated exclusively with DOT

(85% versus 61%) during the first episode, and received a median of two more months of treatment (Table 1b). Reinfection and reactivation patients had similar proportions of pulmonary, cavitary, and smear-positive disease. The median number of months between episodes was equivalent at 26 months and the majorities of TB strains were EuroAmerican at the second episodes.

Multivariable Analysis

Among patients with two episodes of TB who completed treatment, four factors were individually significant for reinfection on bivariate analysis (Table 3). Being black or Hispanic (OR 4.4, 95% CI 1.4-14.1), living in State A at diagnosis (OR 3.6, 95% CI 1.3-10.5), being of Mexican birth (OR 4.7, 95% CI 1.6-14.0), and having lived in the United States for a shorter period of time (OR 3.7, 95% CI 1.4-10.1) greatly increased the odds that a patient developed reinfection rather than reactivation. After adjusting for demographics, clinical features, and treatment factors, three factors remained statistically significant for reinfection over reactivation. Being black or Hispanic (OR 4.4, 95% CI 1.1-17.2), living ≤12 years in the United States (OR 3.5, 95% CI 1.0-11.9), and receiving treatment by DOT exclusively during the first episode (OR 4.8, 95% CI 1.2-19.8) greatly increased the odds of developing reinfection.

Unrestricted Recurrent Population Comparison

When we did not require treatment completion during the first TB episode, a total of 188 patients with two episodes of TB were identified, involving 164 (87%) reactivation patients and 24 (13%) reinfection patients (Figure 1). Thus, reinfection occurred in 13% of the full population with two episodes of TB (i.e., 24 of 188), as compared with 15% among those who had completed treatment in the first episode (i.e., 20 of 136) and 8% in those who had not completed treatment in the first episode (i.e., 4 of 52).

Distributions of most predictors between reactivation and reinfection from the full population experiencing two episodes were similar to those when treatment completion was required. However, slight differences in distribution are of note. In the restricted population, race/ethnicity (*P*=0.11) and years living in the United States at second episode (*P*=0.08) were no longer individually significant for reinfection (Tables 2a and 2b). Of patients in the full population, 17 (52%) who were Hispanic and 9 (41%) who were foreign-born and had lived in the United States for ≥12 years did not completed therapy in the first episode. Other factors also changed between full and restricted populations but did not affect individual significance with reinfection, including 17 (44%) who were homeless, 13 (76%) in corrections, and 11 (38%) who were HIV-infected that were dropped in the final model because they did not complete treatment during the first episode.

DISCUSSION

In this study, we aimed to ascertain how potentially predictive factors of exogenous reinfection of TB disease varied from those of endogenous reactivation among patients with two episodes of TB in the United States during 1993 and 2011. We used genotyping to discriminate between the two types of recurrence; endogenous reactivation and exogenous reinfection. Reinfection was found in 15% of patients who completed treatment during their first episode. After adjusting for covariates, being black or Hispanic, receiving treatment exclusively by DOT during the first episode, and living in the United States less than 12 years were significantly associated with reinfection rather that reactivation among those who completed treatment. Living in State A, being of Mexican birth, and receiving ≥9 months of treatment were also associated with reinfection, but did not remain statistically significant on multivariable analysis.

Previous studies have been conducted examining the prevalence and risk factors for recurrent TB disease. While many of these studies examined risk factors for recurrence, relatively few studies had genotyping data available, preventing the studies from distinguishing between reactivation and reinfection. Among studies with genotyping, the median sample size was 32 (range 8–184) patients with a median of 5 reinfection patients (range 1–33); therefore, many of these studies did not have the statistical power to run comparative tests between reinfection and reactivation (24, 26, 28-30, 33, 39, 42, 43). Additionally, the definition of recurrent TB disease has not been well established and is not consistent among current literature. This unsettled definition results in great variation in the prevalence of both recurrence and reinfection. In studies that included all second episodes of TB, the prevalence of reinfection ranged 16%–26% of recurrent cases (26, 43), while those only involving patients who completed treatment prior to the second episode found reinfection prevalence ranged 8%–27% of recurrent cases (24, 28-30, 33, 39, 42).

This study assessed the prevalence of reinfection among both populations; patients with two episodes of TB who completed treatment in the first episode as well as all patients with two episodes of TB, regardless of treatment outcome in the first episode. We found that among all patients with two episodes, reinfection presented in 13% of recurrent cases, and among only those patients completing therapy, reinfection presented in 15% of recurrent cases.

The strongest association with TB reinfection was receiving treatment exclusively by DOT during the first episode. First episode treatment duration was also longer for this group. One potential interpretation for this finding is that patients who self-administered treatment and were treated for a shorter length of time might have been less likely to achieve cure of the first episode, and thus, more likely to experience a second episode characterized by reactivation. Likewise, the fact that reinfection patients were more often treated exclusively by DOT and for longer periods of time might signify that these patients received complete and successful treatment and were cured of their first episode. Hence, their reinfection was more likely an indication of the patient's association with people and places where TB transmission was still occurring rather than that of unsuccessful treatment.

While race, living in State A, and being of Mexican birth were individually significantly associated with reinfection, only being black or Hispanic remained significant in the multivariable model. These three features might be closely related and, therefore, measure similar factors; likely being the reason why only one remained significant. Other studies found immigration status, a factor also related to the above features, significantly associated with reinfection (24).

Reinfection, as a marker of recent transmission, is plausibly more likely in areas where many strains of TB are circulating (i.e., in areas of high background incidence of TB). In this study, exogenous reinfection was associated with having lived a shorter amount of time in the United States. Social factor similarities among more recent immigrants (e.g., residing, working, or socializing in settings with other recently arriving immigrants) could increase their exposure to new TB strains, increasing their risk of a novel TB infection despite prior successfully treated TB disease. In TB epidemiology in the United States, recent immigration is typically perceived as a risk factor for recent transmission in one's country of origin (i.e., for imported infection); thus, this finding was somewhat surprising yet reassuring in that it suggests TB diagnosed and treated soon after immigration was likely adequately treated (70).

Limitations

The matching process used for this study employed the algorithm described and developed by Kim et al., which had 97.8% positive predictive value for correctly identifying recurrent cases (30). A limitation of this matching process is the negative predictive value of the algorithm; recurrent cases were unable to be included in this study if both episodes occurred in different states or if the previous episode occurred overseas. This exclusion was necessary because we do not have a systematic source of information about patients' TB history before they arrive in the United States, and, to protect patient identities, the algorithm validation process required patients to be in the same state for diagnosis of both episodes. However, it is unlikely that enough people with recurrent TB moved states to largely affect the results of the study.

The variables used to determine the length of time patients had been in the United States at the time of diagnosis are somewhat unreliable. Immigrants might be hesitant to accurately report their length of time living in the United States for fear of deportation or other unknown reasons. To address this issue, the reported dates and time periods were examined in both episodes to most accurately determine the correct measurements for each patient.

Lastly, HIV was the most commonly related factor with reinfection in previous studies (9, 33, 34, 39, 41, 42). However, this study did not find HIV to be significantly associated with reinfection. This was likely due to small numbers; nonetheless, not all states record and report certain variables in the Report of a Verified Case of Tuberculosis in the same way. Some states, notably California, only record positive HIV results, not reporting when HIV status is negative, whereas other states typically report both positive and negative results. There were also a small number of patients who refused or were not offered HIV tests. In order to address this issue, HIV was dichotomized as positive versus other.

Conclusions

While genotyping evidence suggests that the majority of persons who experience two episodes of TB undergo reactivation of the first episode, minorities and those more recently immigrated to the United States have a greater risk for exogenous reinfection. This suggests that these populations are being successfully treated for TB, yet other factors are leading to their exposure to further transmission and increasing their risk of reinfection. Public health interventions should continue to focus on these populations and their areas of residence, work, and recreation for evidence of TB transmission, to prevent further and future spread of TB. Future studies should examine these exposures and risk factors that are common among those who are successfully treated yet later become reinfected with TB.

REFERENCES

1. Heymann DL, ed. Control of Communicable Diseases Manual. 19th ed. Washington, DC: American Public Health Association; 2008.
2. Tuberculosis(TB)(Mycobacterium tuberculosis): 2009 Case Definition. In. Atlanta, GA: National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention (U.S.), Division of Tuberculosis Elimination; 2009.
3. Nelson KE W, CM, eds. Infectious disease epidemiology: theory and practice. 2nd ed. ed. Sudbury, Mass.: Jones and Bartlett Publishers; 2007.
4. Trends in tuberculosis--United States, 2012. MMWR Morb Mortal Wkly Rep 2013;62(11):201-5.
5. Pitchenik AE, Fertel D, Bloch AB. Mycobacterial disease: epidemiology, diagnosis, treatment, and prevention. Clin Chest Med 1988;9(3):425-41.
6. Riley RL. Aerial dissemination of pulmonary tuberculosis. Am Rev Tuberc 1957;76(6):931-41.
7. Castro KG, Jaffe HW. Rationale and methods for the National Tuberculosis Genotyping and Surveillance Network. Emerg Infect Dis 2002;8(11):1188-91.
8. Comstock GW. Frost revisited: the modern epidemiology of tuberculosis. Am J Epidemiol 1975;101(5):363-82.
9. Chiang CY, Riley LW. Exogenous reinfection in tuberculosis. Lancet Infect Dis 2005;5(10):629-36.
10. Vynnycky E, Fine PE. The natural history of tuberculosis: the implications of age-dependent risks of disease and the role of reinfection. Epidemiol Infect 1997;119(2):183-201.

11. Churchyard GJ, Fielding KL, Lewis JJ, Chihota VN, Hanifa Y, Grant AD. Symptom and chest radiographic screening for infectious tuberculosis prior to starting isoniazid preventive therapy: yield and proportion missed at screening. AIDS 2010;24 Suppl 5:S19-27.

12. Levy H, Feldman C, Sacho H, van der Meulen H, Kallenbach J, Koornhof H. A reevaluation of sputum microscopy and culture in the diagnosis of pulmonary tuberculosis. Chest 1989;95(6):1193-7.

13. Targeted tuberculin testing and treatment of latent tuberculosis infection. This official statement of the American Thoracic Society was adopted by the ATS Board of Directors, July 1999. This is a Joint Statement of the American Thoracic Society (ATS) and the Centers for Disease Control and Prevention (CDC). This statement was endorsed by the Council of the Infectious Diseases Society of America. (IDSA), September 1999, and the sections of this statement. Am J Respir Crit Care Med 2000;161(4 Pt 2):S221-47.

14. WHO G. Global Tuberculosis Control, WHO Report 2013. Geneva, Switzerland: World Health Organization; 2013.

15. Jassal MS, Bishai WR. Epidemiology and challenges to the elimination of global tuberculosis. Clin Infect Dis 2010;50 Suppl 3:S156-64.

16. WHO. Tuberculosis: Fact Sheet. In. Geniva: World Health Organization; 2013.

17. WHO G. Global Tuberculosis Control, WHO Report 2005. Geneva, Switzerland: World Health Organization; 2005.

18. Buff AM, Moonan PK, Desai MA, McKenna TL, Harris DA, Rogers BJ, et al. South Carolina tuberculosis genotype cluster investigation: a tale of substance abuse and recurrent disease. Int J Tuberc Lung Dis 2010;14(10):1347-9.

19. Oeltmann JE, Kammerer JS, Pevzner ES, Moonan PK. Tuberculosis and substance abuse in the United States, 1997-2006. Arch Intern Med 2009;169(2):189-97.

20. Kim L, Moonan PK, Yelk Woodruff RS, Kammerer JS, Haddad MB. Epidemiology of recurrent tuberculosis in the United States, 1993-2010 [Short communication]. Int J Tuberc Lung Dis 2013;17(3):357-60.

21. Small PM, Hopewell PC, Singh SP, Paz A, Parsonnet J, Ruston DC, et al. The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. N Engl J Med 1994;330(24):1703-9.

22. Hwang SW, Tolomiczenko G, Kouyoumdjian FG, Garner RE. Interventions to improve the health of the homeless: a systematic review. Am J Prev Med 2005;29(4):311-9.

23. Barnes PF, Yang Z, Preston-Martin S, Pogoda JM, Jones BE, Otaya M, et al. Patterns of tuberculosis transmission in Central Los Angeles. JAMA 1997;278(14):1159-63.

24. Bandera A, Gori A, Catozzi L, Degli Esposti A, Marchetti G, Molteni C, et al. Molecular epidemiology study of exogenous reinfection in an area with a low incidence of tuberculosis. J Clin Microbiol 2001;39(6):2213-8.

25. Chorba T. Organization of TB Programs in the United States. In. Emory University, Rollins School of Public Health: Epidemiology 542; 2013.

26. Bang D, Andersen A, Thomsen V, Lillebaek T. Recurrent tuberculosis in Denmark: relapse vs. re-infection The International Journal of Tuberculosis and Lung Disease 2010 14(4):447-453.

27. Yew WW, Leung CC. Are some people not safer after successful treatment of tuberculosis? Am J Respir Crit Care Med 2005;171(12):1324-5.

28. Cacho J, Perez Meixeira A, Cano I, Soria T, Ramos Martos A, Sanchez Concheiro M, et al. Recurrent tuberculosis from 1992 to 2004 in a metropolitan area. Eur Respir J 2007;30(2):333-7.

29. Dobler CC, Crawford AB, Jelfs PJ, Gilbert GL, Marks GB. Recurrence of tuberculosis in a low-incidence setting. Eur Respir J 2009;33(1):160-7.

30. Kim L, Moonan PK, Yelk Woodruff RS, Kammerer JS, Haddad MB. Factors associated with recurrent tuberculosis among persons who completed treatment in the United States. In. Unpublished Manuscript.

31. Pascopella L, Deriemer K, Watt JP, Flood JM. When tuberculosis comes back: who develops recurrent tuberculosis in california? PLoS One 2011;6(11):e26541.

32. Fine PE, Small PM. Exogenous reinfection in tuberculosis. N Engl J Med 1999;341(16):1226-7.

33. Lambert ML, Hasker E, Van Deun A, Roberfroid D, Boelaert M, Van der Stuyft P. Recurrence in tuberculosis: relapse or reinfection? Lancet Infect Dis 2003;3(5):282-7.

34. Sonnenberg P, Murray J, Glynn JR, Shearer S, Kambashi B, Godfrey-Faussett P. HIV-1 and recurrence, relapse, and reinfection of tuberculosis after cure: a cohort study in South African mineworkers. Lancet 2001;358(9294):1687-93.

35. van Rie A, Warren R, Richardson M, Victor TC, Gie RP, Enarson DA, et al. Exogenous reinfection as a cause of recurrent tuberculosis after curative treatment. N Engl J Med 1999;341(16):1174-9.

36. Gomes MG, Aguas R, Lopes JS, Nunes MC, Rebelo C, Rodrigues P, et al. How host heterogeneity governs tuberculosis reinfection? Proc Biol Sci 2012;279(1737):2473-8.

37. Millet JP, Orcau A, de Olalla PG, Casals M, Rius C, Cayla JA. Tuberculosis recurrence and its associated risk factors among successfully treated patients. J Epidemiol Community Health 2009;63(10):799-804.

38. Sahadevan R, Narayanan S, Paramasivan CN, Prabhakar R, Narayanan PR. Restriction fragment length polymorphism typing of clinical isolates of Mycobacterium tuberculosis from patients with pulmonary tuberculosis in Madras, India, by use of direct-repeat probe. J Clin Microbiol 1995;33(11):3037-9.

39. Jasmer RM, Bozeman L, Schwartzman K, Cave MD, Saukkonen JJ, Metchock B, et al. Recurrent tuberculosis in the United States and Canada: relapse or reinfection? Am J Respir Crit Care Med 2004;170(12):1360-6.

40. Shen G, Xue Z, Shen X, Sun B, Gui X, Shen M, et al. The study recurrent tuberculosis and exogenous reinfection, Shanghai, China. Emerg Infect Dis 2006;12(11):1776-8.

41. Panjabi R, Comstock GW, Golub JE. Recurrent tuberculosis and its risk factors: adequately treated patients are still at high risk. Int J Tuberc Lung Dis 2007;11(8):828-37.

42. Pettit AC, Kaltenbach LA, Maruri F, Cummins J, Smith TR, Warkentin JV, et al. Chronic lung disease and HIV infection are risk factors for recurrent tuberculosis in a low-incidence setting. Int J Tuberc Lung Dis 2011;15(7):906-11.

43. de Boer AS, Borgdorff MW, Vynnycky E, Sebek MM, van Soolingen D. Exogenous re-infection as a cause of recurrent tuberculosis in a low-incidence area. Int J Tuberc Lung Dis 2003;7(2):145-52.

44. Garcia de Viedma D, Marin M, Hernangomez S, Diaz M, Ruiz Serrano MJ, Alcala L, et al. Tuberculosis recurrences: reinfection plays a role in a population whose clinical/epidemiological characteristics do not favor reinfection. Arch Intern Med 2002;162(16):1873-9.

45. Barnes PF, Cave MD. Molecular epidemiology of tuberculosis. N Engl J Med 2003;349(12):1149-56.

46. Middelkoop K, Bekker LG, Shashkina E, Kreiswirth B, Wood R. Retreatment tuberculosis in a South African community: the role of re-infection, HIV and antiretroviral treatment. Int J Tuberc Lung Dis 2012;16(11):1510-6.

47. Verver S, Warren RM, Beyers N, Richardson M, van der Spuy GD, Borgdorff MW, et al. Rate of reinfection tuberculosis after successful treatment is higher than rate of new tuberculosis. Am J Respir Crit Care Med 2005;171(12):1430-5.

48. Reported Tuberculosis in the United States, 2011. Atlanta, GA: CDC; October 2012.

49. Cohen T, Murray M. Incident tuberculosis among recent US immigrants and exogenous reinfection. Emerg Infect Dis 2005;11(5):725-8.

50. Andrews JR, Gandhi NR, Moodley P, Shah NS, Bohlken L, Moll AP, et al. Exogenous reinfection as a cause of multidrug-resistant and extensively drug-resistant tuberculosis in rural South Africa. J Infect Dis 2008;198(11):1582-9.

51. Gomes MG, White LJ, Medley GF. Infection, reinfection, and vaccination under suboptimal immune protection: epidemiological perspectives. J Theor Biol 2004;228(4):539-49.

52. GENType: New Genotyping Terminology to Integrate 24-locus MIRU-VNTR. In: National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention (U.S.), Division of Tuberculosis Elimination; 2012.

53. Kruuner A, Pehme L, Ghebremichael S, Koivula T, Hoffner SE, Mikelsaar M. Use of molecular techniques to distinguish between treatment failure and exogenous reinfection with Mycobacterium tuberculosis. Clin Infect Dis 2002;35(2):146-55.

54. Codecasa LR, Porretta AD, Gori A, Franzetti F, Degli Esposti A, Lizioli A, et al. Tuberculosis among immigrants from developing countries in the province of Milan, 1993-1996. Int J Tuberc Lung Dis 1999;3(7):589-95.

55. Tuberculosis Genotyping. In: National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention (U.S.), Division of Tuberculosis Elimination; 2008.

56. New CDC Program for Rapid Genotyping of Mycobacterium tuberculosis Isolates. MMWR Morb Mortal Wkly Rep 2005;54(2):47.

57. Small PM, Shafer RW, Hopewell PC, Singh SP, Murphy MJ, Desmond E, et al. Exogenous reinfection with multidrug-resistant Mycobacterium tuberculosis in patients with advanced HIV infection. N Engl J Med 1993;328(16):1137-44.

58. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, et al. Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. J Clin Microbiol 1997;35(4):907-14.

59. Dale JW, Brittain D, Cataldi AA, Cousins D, Crawford JT, Driscoll J, et al. Spacer oligonucleotide typing of bacteria of the Mycobacterium tuberculosis complex: recommendations for standardised nomenclature. Int J Tuberc Lung Dis 2001;5(3):216-9.

60. Tuberculosis genotyping--United States, 2004-2010. MMWR Morb Mortal Wkly Rep 2012;61(36):723-5.

61. Warren RM, Victor TC, Streicher EM, Richardson M, Beyers N, Gey van Pittius NC, et al. Patients with active tuberculosis often have different strains in the same sputum specimen. Am J Respir Crit Care Med 2004;169(5):610-4.

62. Canetti G, Sutherland I, Svandova E. Endogenous reactivation and exogenous reinfection: their relative importance with regard to the development of non-primary tuberculosis. Bull Int Union Tuberc 1972;47:116-34.

63. Moller M, Hoal EG. Current findings, challenges and novel approaches in human genetic susceptibility to tuberculosis. Tuberculosis (Edinb) 2010;90(2):71-83.

64. van Soolingen D, Borgdorff MW, de Haas PE, Sebek MM, Veen J, Dessens M, et al. Molecular epidemiology of tuberculosis in the Netherlands: a nationwide study from 1993 through 1997. J Infect Dis 1999;180(3):726-36.

65. van Soolingen D, Hermans PW, de Haas PE, Soll DR, van Embden JD. Occurrence and stability of insertion sequences in Mycobacterium tuberculosis complex strains: evaluation of an insertion sequence-dependent DNA

polymorphism as a tool in the epidemiology of tuberculosis. J Clin Microbiol 1991;29(11):2578-86.

66. Yang Z, Barnes PF, Chaves F, Eisenach KD, Weis SE, Bates JH, et al. Diversity of DNA fingerprints of Mycobacterium tuberculosis isolates in the United States. J Clin Microbiol 1998;36(4):1003-7.

67. Chaves F, Dronda F, Alonso-Sanz M, Noriega AR. Evidence of exogenous reinfection and mixed infection with more than one strain of Mycobacterium tuberculosis among Spanish HIV-infected inmates. AIDS 1999;13(5):615-20.

68. Gomes MG, Franco AO, Gomes MC, Medley GF. The reinfection threshold promotes variability in tuberculosis epidemiology and vaccine efficacy. Proc Biol Sci 2004;271(1539):617-23.

69. Ghosh S, Moonan PK, Cowan L, Grant J, Kammerer S, Navin TR. Tuberculosis genotyping information management system: enhancing tuberculosis surveillance in the United States. Infect Genet Evol 2012;12(4):782-8.

70. Cain KP, Benoit SR, Winston CA, Mac Kenzie WR. Tuberculosis among foreign-born persons in the United States. JAMA 2008;300(4):405-12.

TABLES

Table 1a. Demographic characteristics of patients with two episodes of TB, among those completing treatment during first episode, in the United States 1993–2011

		Recurrent TB		
	Total Recurrent (n=136)	Reactivation (n=116)	Reinfection (n=20)	
Characteristics	n (%)	n (%)	n (%)	*P* value
Demographics				
Age (years) at second episode				0.61
< 19	3 (2.2)	3 (2.6)	0 (0)	
19-29	15 (11)	13 (11)	2 (10)	
30-39	22 (16)	18 (16)	4 (20)	
40-49	27 (20)	20 (17)	7 (35)	
50-59	33 (24)	30 (26)	3 (15)	
60-69	13 (9.6)	11 (9.5)	2 (10)	
70 +	23 (17)	21 (18)	2 (10)	
median (IQR)	50 (37-60)	51 (37-61)	44 (34-52)	0.17
Female	35 (26)	32 (28)	3 (15)	0.23
Race/ethnicity				0.02 *
Black, non-Hispanic	38 (28)	30 (26)	8 (40)	
Hispanic	33 (24)	25 (22)	8 (40)	
White, non-Hispanic and Other[1]	65 (48)	61 (53)	4 (20)	
Rate of background TB incidence				
High-incidence states	59 (43)	48 (41)	11 (55)	0.26
State A	22 (37)	15 (31)	7 (64)	0.02 *
State B	25 (43)	21 (44)	4 (36)	0.76
State C	7 (12)	7 (15)	0 (0)	0.59
State D	5 (8.5)	5 (10)	0 (0)	1.00
Low-incidence states[2]	77 (57)	68 (59)	9 (45)	
Foreign born	58 (43)	46 (40)	12 (60)	0.09
Foreign born region				<0.01 *
Mexican	19 (33)	12 (26)	7 (58)	
African[3]	5 (8.6)	2 (4.4)	3 (25)	
Other[4]	34 (59)	32 (70)	2 (17)	
Years living in U.S. at second episode				0.11
< 2	1 (1.7)	1 (2.2)	0 (0)	
2-5	10 (17)	6 (13)	4 (33)	
6-10	16 (28)	13 (28)	3 (35)	
11-20	18 (31)	13 (28)	5 (42)	
21 +	13 (22)	13 (28)	0 (0)	
median (IQR)	12 (7-20)	14 (8-26)	9 (4-13)	0.04 *
Risk Factors[5]				
Homeless	22 (16)	17 (15)	5 (25)	0.32
Corrections	4 (2.9)	3 (2.6)	1 (5.0)	0.47
Substance Abuse[6]	57 (42)	49 (42)	8 (40)	0.83
Alcohol[7]	48 (35)	41 (35)	7 (35)	0.91
IDU[6]	7 (5.2)	5 (4.3)	2 (10)	0.28
Non-IDU[8]	31 (23)	29 (25)	2 (10)	0.16

Notes for Table 1a.

*Statistically significant at $p < 0.05$.

[1] Other includes Asian, American Indian/Alaska Native, Hawaiian/Pacific Islander, Multiracial non-Hispanic.

[2] Low-incidence background TB includes 32 states and Puerto Rico.

[3] African includes Ethiopia, Sudan, Somalia, Zimbabwe.

[4] Other includes Bosnia-Herzegovina, Poland, Russia, Ukraine, China, Hong Kong, Republic of Korea, Democratic Peoples Republic of Korea, Taiwan, Ecuador, Guatemala, Honduras, Haiti, El Salvador, Trinidad and Tobago, Indonesia, Cambodia, Laos, Philippines, Thailand, Vietnam.

[5] Yes if ever involved in the risk behavior as recorded at time of either episode.

[6] Missing in one reactivation patients.

[7] Missing in three reactivation patients.

[8] Missing in two reactivation patients.

TB = tuberculosis; IDU = intravenous drug use.

Table 1b. Clinical and treatment characteristics of patients with two episodes of TB, among those completing treatment during first episode, in the United States 1993–2011

		Recurrent TB		
	Total Recurrent (n=136)	Reactivation (n=116)	Reinfection (n=20)	
Characteristics	n (%)	n (%)	n (%)	*P* value
Clinical Features				
HIV positive by second episode [1]	18 (13)	14 (12)	4 (20)	0.30
Disease site at first episode				0.32
Both	10 (7.4)	7 (6.0)	3 (15)	
Pulmonary Only	117 (86)	101 (87)	16 (80)	
Extrapulmonary Only	9 (6.6)	8 (6.9)	1 (5.0)	
Cavitary disease at first episode	47 (35)	39 (34)	8 (40)	0.58
Smear positive at first episode	92 (68)	77 (66)	15 (75)	0.45
Treatment Factors				
Treatment type at first episode [2]				0.10
Both	34 (25)	31 (27)	3 (15)	
DOT only	88 (65)	71 (61)	17 (85)	
SAT only	13 (9.6)	13 (11)	0 (0)	
Change in drug resistance between episodes	17 (13)	13 (11)	4 (20)	0.28
Treatment duration (months) at first episode				0.57
Never treated	0 (0)	0 (0)	0 (0)	
< 2	0 (0)	0 (0)	0 (0)	
2-6	51 (38)	46 (40)	5 (25)	
7-9	48 (35)	39 (34)	9 (45)	
10-12	28 (21)	23 (20)	5 (25)	
13 +	9 (6.6)	8 (6.9)	1 (5.0)	
median (IQR)	7 (6-10)	7 (6-10)	9 (7-10)	0.16
Treatment success at first episode [3]	99 (73)	83 (72)	16 (80)	0.69
Completed therapy	136 (100)	116 (100)	20 (100)	n/a
Sputum conversion[4]	99 (73)	83 (72)	16 (80)	0.69
Other Factors				
Years between episodes				0.19
1-2	59 (43)	50 (43)	9 (45)	
>2 - 3	35 (26)	32 (28)	3 (15)	
>3 - 4	13 (9.6)	12 (10)	1 (5.0)	
>4 - 5	19 (14)	16 (14)	3 (15)	
> 5	10 (7.4)	6 (5.2)	4 (20)	
median months (IQR)	26 (19-44)	26 (19-43)	26 (17-60)	0.82
Lineage at second episode [5]				0.08
EuroAmerican	86 (63)	71 (61)	15 (75)	
East Asian	31 (23)	30 (26)	1 (5.0)	
IndoOceanic	3 (2.2)	3 (2.6)	0 (0)	
East African Indian	4 (2.9)	2 (1.7)	2 (10)	
Bovis	4 (2.9)	3 (2.6)	1 (5.0)	

Notes for Table 1b.

[1] Records positivity versus other (missing, unknown, not offered, or refused in 26 reactivation and four reinfection patients).

[2] Unknown or missing in one reactivation patient.

[3] Recorded both treatment completion and sputum conversion; missing one or both qualifying records in 21 reactivation and three reinfection patients.

[4] Missing in 21 reactivation and three reinfection patients.

[5] Unknown in seven reactivation and one reinfection patients.

TB = tuberculosis; HIV = human immunodeficiency virus; DOT = directly observed therapy.

Table 2a. Demographic characteristics of patients with two episodes of tuberculosis in the United States 1993–2011

		Recurrent TB		
	Total Recurrent	Reactivation	Reinfection	
	(n=188)	(n=164)	(n=24)	
Characteristics	n (%)	n (%)	n (%)	*P* value
Demographics				
Age (years) at second episode				0.47
< 19	3 (1.6)	3 (1.8)	0 (0)	
19-29	23 (12)	20 (12)	3 (13)	
30-39	31 (16)	26 (16)	5 (21)	
40-49	41 (22)	32 (20)	9 (38)	
50-59	41 (22)	38 (23)	3 (13)	
60-69	18 (9.6)	16 (9.8)	2 (8.3)	
70 +	31 (16)	29 (18)	2 (8.3)	
median (IQR)	49 (36-60)	51 (37-61)	43 (33-51)	0.06
Female	47 (25)	43 (26)	4 (17)	0.31
Race/ethnicity				0.11
Black, non-Hispanic	54 (29)	45 (27)	9 (38)	
Hispanic	50 (27)	41 (25)	9 (38)	
White, non-Hispanic and Other[1]	84 (45)	78 (48)	6 (25)	
Rate of background TB incidence				
High-incidence states	83 (44)	68 (41)	15 (63)	0.05
State A	32 (39)	24 (35)	8 (53)	0.04 *
State B	32 (39)	26 (38)	6 (40)	0.26
State C	12 (14)	12 (18)	0 (0)	0.37
State D	7 (8.4)	6 (8.8)	1 (6.7)	1.00
Low-incidence states[2]	105 (56)	96 (59)	9 (38)	
Foreign born	78 (41)	65 (40)	13 (54)	0.18
Foreign born region				0.01 *
Mexican	27 (35)	20 (31)	7 (54)	
African[3]	6 (7.7)	3 (4.6)	3 (23)	
Other[4]	45 (58)	42 (65)	3 (23)	
Years living in U.S. at second episode				0.30
< 2	1 (1.3)	1 (1.5)	0 (0)	
2-5	15 (19)	11 (17)	4 (31)	
6-10	19 (24)	16 (25)	3 (23)	
11-20	21 (27)	16 (25)	5 (38)	
21 +	22 (28)	21 (32)	1 (7.7)	
median (IQR)	14 (7-24)	14 (8-26)	9 (5-13)	0.08
Risk Factors[5]				0.55
Homeless	39 (21)	34 (21)	5 (21)	1.00
Corrections	17 (9.0)	15 (9.2)	2 (8.3)	1.00
Substance Abuse [6]	90 (48)	79 (48)	11 (46)	0.79
Alcohol[7]	72 (38)	64 (39)	8 (33)	0.53
IDU[6]	14 (7.5)	11 (6.7)	3 (13)	0.40
Non-IDU[8]	52 (28)	48 (29)	4 (17)	0.18

Notes for Table 2a.

*Statistically significant at $p < 0.05$.

[1] Other includes Asian, American Indian/Alaska Native, Hawaiian/Pacific Islander, Multiracial non-Hispanic.

[2] Low-incidence background TB includes 34 states and Puerto Rico.

[3] African includes Ethiopia, Sudan, Somalia, Zimbabwe.

[4] Other includes Bosnia-Herzegovina, Poland, Russia, Ukraine, China, Hong Kong, Republic of Korea, Democratic Peoples Republic of Korea, Taiwan, Ecuador, Guatemala, Honduras, Haiti, El Salvador, Trinidad and Tobago, Indonesia, Cambodia, Laos, Philippines, Thailand, Vietnam.

[5] Yes if ever involved in the risk behavior as recorded at time of either episode.

[6] Missing in two reactivation patients.

[7] Missing in four reactivation patients.

[8] Missing in three reactivation patients.

TB = tuberculosis; IDU = intravenous drug use.

Table 2b. Clinical and treatment characteristics of patients with two episodes of tuberculosis in the United States 1993–2011

		Recurrent TB		
	Total Recurrent	Reactivation	Reinfection	
	(n=188)	(n=164)	(n=24)	
Characteristics	n (%)	n (%)	n (%)	*P* value
Clinical Features				
HIV positive by second episode [1]	29 (15)	22 (13)	7 (29)	0.07
Disease site at first episode				0.13
Both	13 (6.9)	9 (5.5)	4 (17)	
Pulmonary Only	159 (85)	141 (86)	18 (75)	
Extrapulmonary Only	16 (8.5)	14 (8.5)	2 (8.3)	
Cavitary disease at first episode	57 (30)	49 (30)	8 (33)	0.73
Smear positive at first episode	117 (62)	101 (62)	16 (67)	0.63
Treatment Factors				
Treatment type at first episode [2]				0.39
Both	41 (22)	37 (23)	4 (17)	
DOT only	112 (60)	93 (57)	19 (79)	
SAT only	22 (12)	21 (13)	1 (4.2)	
Change in drug resistance between episodes	22 (12)	17 (10)	5 (21)	0.17
Treatment duration (months) at first episode [3]				0.53
Never treated	5 (2.7)	5 (3.1)	0 (0)	
< 2	15 (8.0)	14 (8.5)	1 (4.2)	
2-6	74 (39)	67 (41)	7 (29)	
7-9	49 (26)	40 (24)	9 (38)	
10-12	30 (16)	24 (15)	6 (25)	
13 +	12 (6.4)	11 (6.7)	1 (4.2)	
median (IQR)	7 (6-9)	6 (6-9)	9 (6-10)	0.06
Treatment success at first episode [4]	99 (53)	83 (51)	16 (67)	0.18
Completed therapy[5]	136 (72)	116 (71)	20 (83)	0.24
Sputum conversion[6]	116 (62)	98 (60)	18 (75)	0.16
Other Factors				
Years between episodes				0.34
1-2	84 (45)	74 (45)	10 (42)	
>2 - 3	48 (26)	44 (27)	4 (17)	
>3 - 4	18 (9.6)	16 (9.8)	2 (8.3)	
>4 - 5	24 (13)	20 (12)	4 (17)	
> 5	14 (7.5)	10 (6.1)	4 (17)	
median months (IQR)	26 (19-43)	26 (19-43)	26 (17-60)	0.69
Lineage at second episode [7]				0.10
EuroAmerican	125 (66)	107 (65)	18 (75)	
East Asian	39 (21)	37 (23)	2 (8.3)	
IndoOceanic	7 (3.7)	7 (4.3)	0 (0)	
East African Indian	4 (2.1)	2 (1.2)	2 (8.3)	
Bovis	4 (2.1)	3 (1.8)	1 (4.2)	

Notes for Table 2b.

[1] Records positivity versus other (missing, unknown, not offered, or refused in 39 reactivation and five reinfection patients).

[2] Unknown or missing in 13 reactivation patients.

[3] Unknown or missing in three reactivation patients.

[4] Recorded both treatment completion and sputum conversion; missing one or both qualifying records in 36 reactivation and four reinfection patients.

[5] Missing in eight reactivation patients (including five who never initiated therapy); Lost, moved, refused, or other in 40 reactivation and four reinfection patients.

[6] Missing in 33 reactivation and four reinfection patients.

[7] Unknown in eight reactivation and one reinfection patients.

TB = tuberculosis; HIV = human immunodeficiency virus; DOT = directly observed therapy.

Table 3. Predictors of tuberculosis recurrence among 136 patients completing treatment during the first episode. Bivariate and multivariate analyses. United States 1993–2011.

	Recurrent TB		Outcome (Reinfection)			
	Reactivation (n=116)	Reinfection (n=20)			Adjusted	
Characteristics	n (%)	n (%)	Crude OR	95%CI	OR	95%CI
Demographics						
Female	32 (28)	3 (15)	0.46	0.13-1.69	0.34	0.08-1.51
Black or Hispanic	55 (47)	16 (80)	4.44	1.40-14.08 *	4.35	1.10-17.19 *
Living in State A at diagnosis[1]	15 (31)	7 (64)	3.63	1.25-10.54 *	2.03	0.57-7.26
Mexican birth	12 (10)	7 (35)	4.67	1.56-13.96 *	1.17	0.26-5.26
12 or fewer years livings in U.S. at second episode	21 (18)	9 (45)	3.70	1.36-10.06 *	3.50	1.03-11.86 *
Clinical Features						
HIV positive[2]	14 (12)	4 (20)	1.81	0.53-6.23	0.94	0.21-4.24
Treatment Factors						
Nine or more months of treatment at first episode	42 (36)	12 (60)	2.64	1.00-6.98	2.71	0.87-8.47
Received DOT only at first episode[3]	71 (61)	17 (85)	3.59	1.00-12.95	4.80	1.17-19.78 *

[*] Statistically significant at $p < 0.05$.

[1] State with high-incidence background TB; living in state for both first and second episode diagnoses.

[2] Missing, unknown, not offered, or refused in 26 reactivation and four reinfection patients.

[3] Unknown in one reactivation patient.

TB = tuberculosis; OR = odds ratio; CI = confidence interval; HIV = human immunodeficiency virus; DOT = directly observed

FIGURES

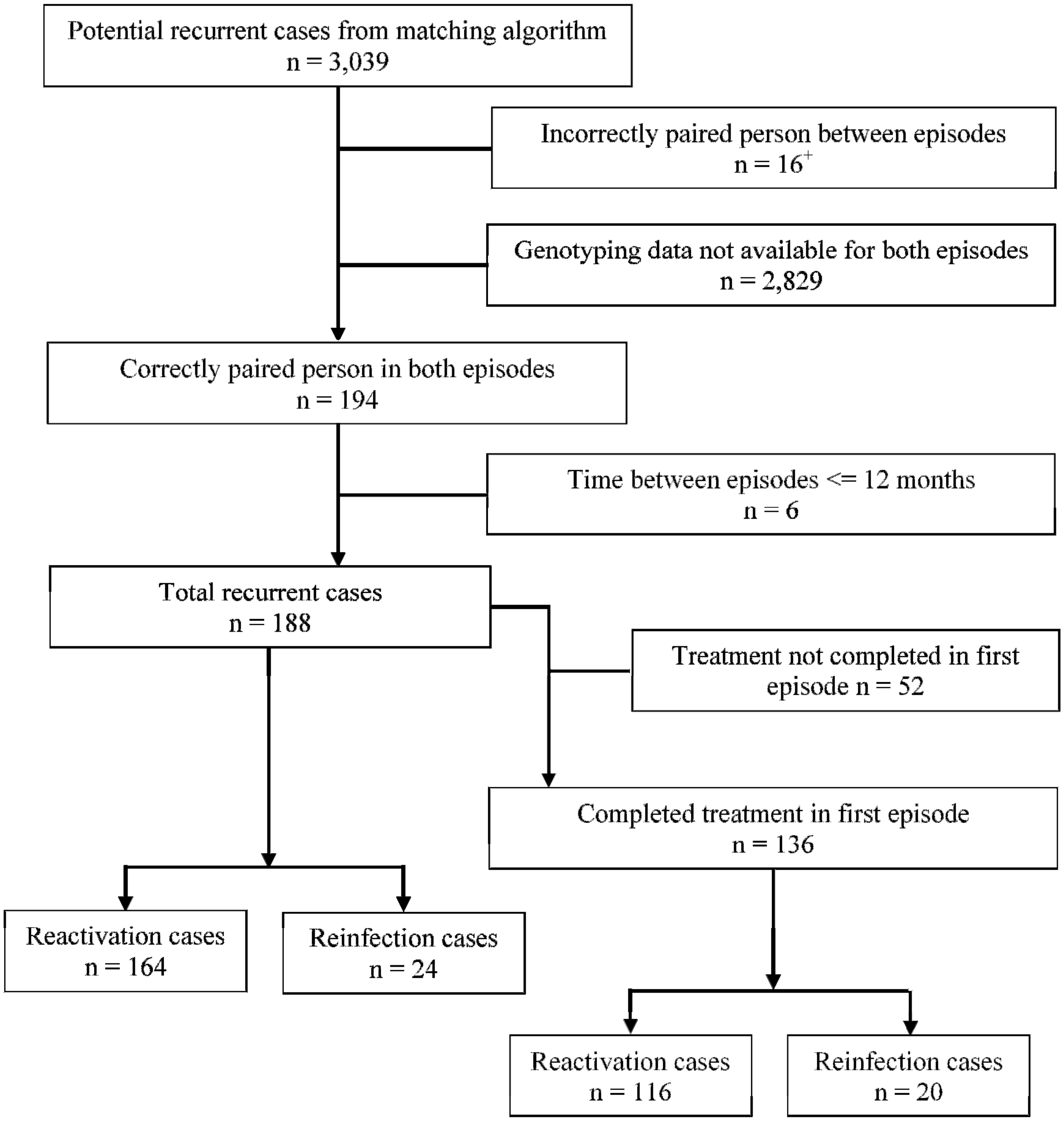

Potential recurrent cases from matching algorithm
n = 3,039
Incorrectly paired person between episodes
n = 16+
Genotyping data not available for both episodes
n = 2,829
Correctly paired person in both episodes
n = 194
Time between episodes <= 12 months
n = 6
Total recurrent cases
n = 188
Treatment not completed in first episode n = 52
Completed treatment in first episode
n = 136
Reactivation cases
n = 164
Reinfection cases
n = 24
Reactivation cases
n = 116
Reinfection cases
n = 20

Figure 1. Description of study cohort. +Twelve patients removed because matching state case numbers between episodes indicated a repeated report of a case, four cases removed because it was unclear if cases were truly a match between first and second episodes.

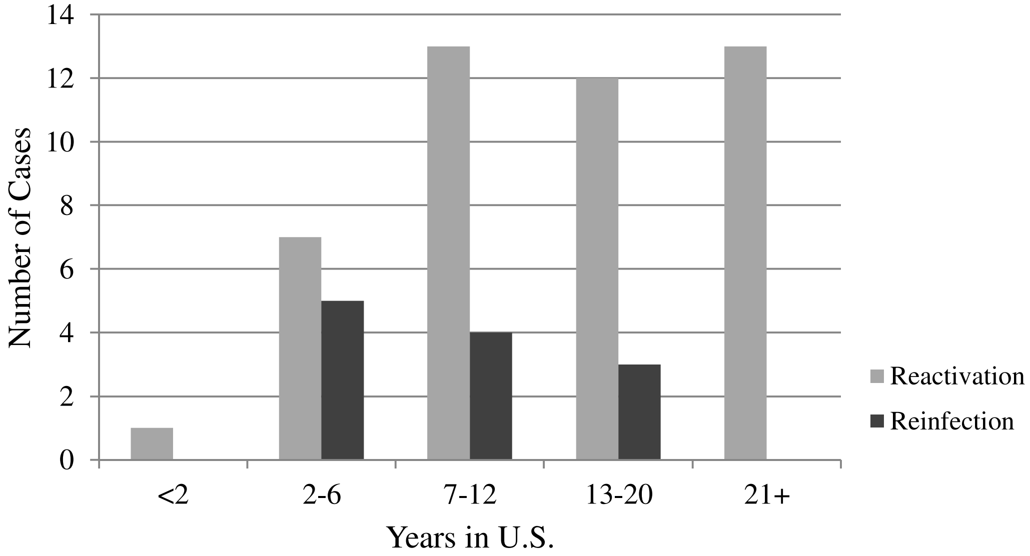


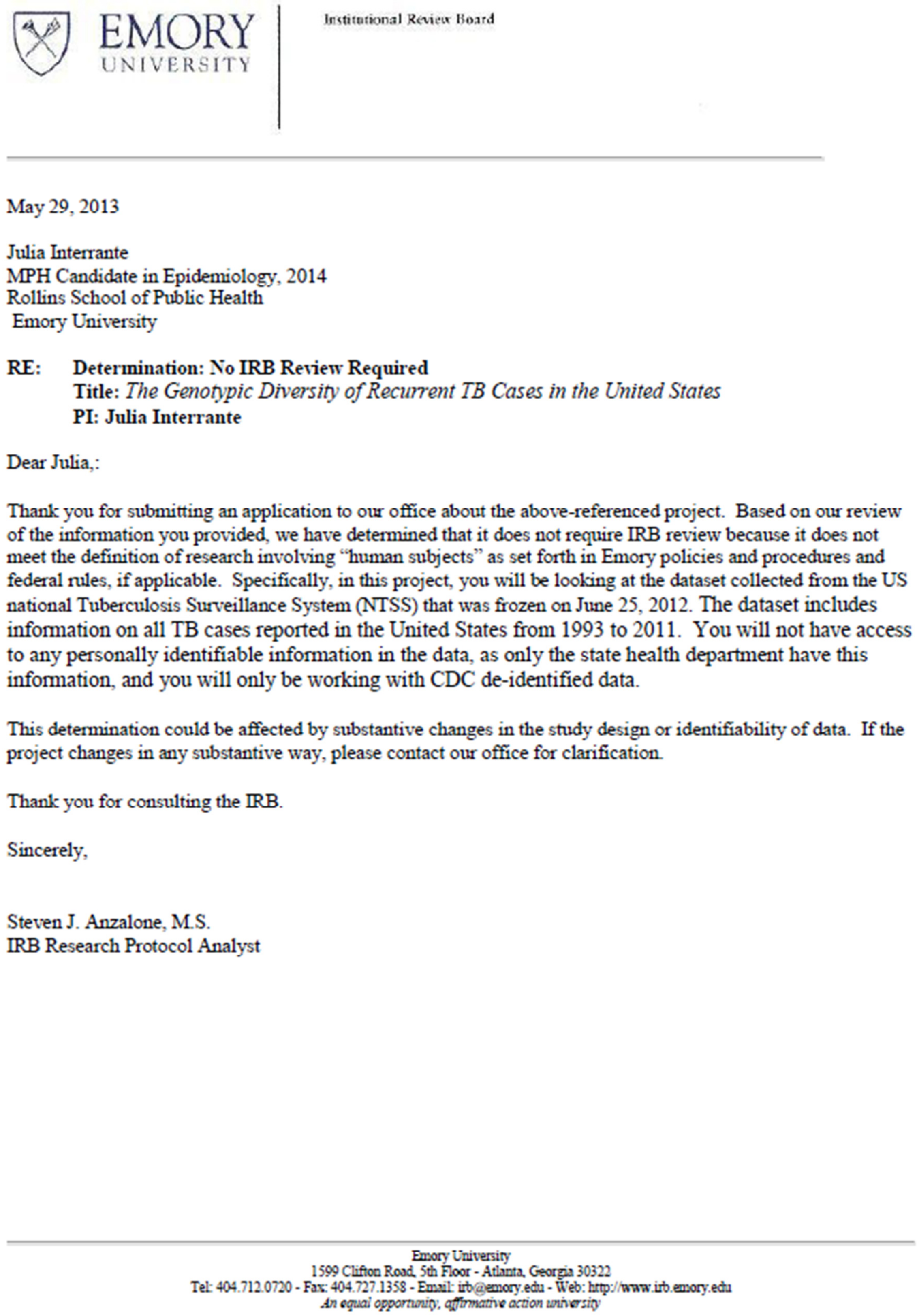
Figure 2. Distribution of years living in the United States at second episode among patients with recurrent TB completing treatment in first episode, 1993–2011

CHAPTER III: PUBLIC HEALTH IMPLICATIONS AND FUTURE DIRECTIONS

This study showed that while endogenous reactivation is the most common form of recurrence for people who have two episodes of tuberculosis (TB) in the United States, exogenous reinfection continues to persist in the United States, suggesting that some groups who recurred were successfully treated during their first episode but remained engaged in risk factors that exposed them to further TB transmission and to eventually become reinfected. Those who were treated exclusively by DOT during their first episode were more likely to experience reinfection if they recurred, meaning that their second episode was not a result of treatment failure during their first episode. Subsequently, this analysis gives support to the use of exclusive treatment by directly observed therapy (DOT) for those at risk for recurrence. Additionally, the fact that minorities and those more recently immigrated to the United States (i.e., populations usually understood as high-risk groups for TB disease) were more likely to become reinfected if they recurred rather developing a reactivation of a previous episode indicates that public health programs are fully and successfully targeting and treating these groups to ensure that they receive cure of their principal TB episodes.

While this study found that public health programs have been successful in treating episodes of TB in some minority populations, the prevalence of exogenous reinfection suggests that some of these patients remain at risk for reinfection with new novel strains of TB due to factors that should be further studied. Future studies should more closely investigate patients who develop reinfection and examine factors related to their areas of work, recreation, and residence that might provide further insight into why these patients continue to be exposed to TB transmission. Since these factors are not aspects of data currently collected as a part of national routine TB surveillance, studies should expand their data collection methods to capture these realities. Additionally, these studies should be developed and conducted with the aim of providing evidence for the development of future public health programs that are able to address this missed area of continued TB transmission.

APPENDIX A: IRB Approval



EMORY UNIVERSITY

Institutional Review Board

May 29, 2013

Julia Interrante
MPH Candidate in Epidemiology, 2014
Rollins School of Public Health
Emory University

RE: **Determination: No IRB Review Required**
Title: *The Genotypic Diversity of Recurrent TB Cases in the United States*
PI: Julia Interrante

Dear Julia,:

Thank you for submitting an application to our office about the above-referenced project. Based on our review of the information you provided, we have determined that it does not require IRB review because it does not meet the definition of research involving "human subjects" as set forth in Emory policies and procedures and federal rules, if applicable. Specifically, in this project, you will be looking at the dataset collected from the US national Tuberculosis Surveillance System (NTSS) that was frozen on June 25, 2012. The dataset includes information on all TB cases reported in the United States from 1993 to 2011. You will not have access to any personally identifiable information in the data, as only the state health department have this information, and you will only be working with CDC de-identified data.

This determination could be affected by substantive changes in the study design or identifiability of data. If the project changes in any substantive way, please contact our office for clarification.

Thank you for consulting the IRB.

Sincerely,

Steven J. Anzalone, M.S.
IRB Research Protocol Analyst

Emory University
1599 Clifton Road, 5th Floor - Atlanta, Georgia 30322
Tel: 404.712.0720 - Fax: 404.727.1358 - Email: irb@emory.edu - Web: http://www.irb.emory.edu
An equal opportunity, affirmative action university

APPENDIX B: Individual Variable Tables

Table 1a. Demographic Characteristics of Recurrent TB Disease among Recurrent Cases Based on the US National Tuberculosis Surveillance System Data Set during 1993-2011

Foreign Born and US Born		
	Recurrent (n=188)	
	No.	%
Age (years)[1]		
< 19	3	1.6
19-29	23	12
30-39	31	16
40-49	41	22
50-59	41	22
60-69	18	9.8
70 +	31	16
Age (binary-years)		
50 +	90	48
Age (binary-years)		
45 +	111	59
Age (biologic importance)		
< 15	1	0.5
15-24	10	5.3
25-49	87	46
50 +	90	48
Age (quartiles)		
< 37	48	26
37-49	50	27
50-60	45	24
61 +	45	24
Age (turtiles)		
< 43	66	35
43-56	65	35
57 +	57	30
Female	47	25
Origin		
US Born	110	59
Foreign Born	78	41

[1] Age at second episode

Table 1a (continued). Demographic Characteristics of Recurrent TB Disease among Recurrent Cases Based on the US National Tuberculosis Surveillance System Data Set during 1993-2011

	Recurrent (n=188)	
	No.	%
Race/ethnicity		
Black, non-Hispanic	52	28
Hispanic	50	27
Other[1]	42	22
White, non-Hispanic	44	23
Black and/or Hispanic	104	55
State		
State B	32	17
State A	32	17
State C_1	10	5.3
State E	9	4.8
State D	7	3.7
State F	7	3.7
State G	7	3.7
State H	6	3.2
State I	5	2.7
State J	5	2.7
State K	4	2.1
State L	4	2.1
State M	4	2.1
State N	4	2.1
State O	4	2.1
State P	4	2.1
State Q	4	2.1
State R	3	1.6
State S	3	1.6
State T	3	1.6
State U	3	1.6
State V	2	1.1
State W	2	1.1
State X	2	1.1
State Y	2	1.1
State Z	2	1.1
State AA	2	1.1
State C_2	2	1.1
State BB	2	1.1
State CC	2	1.1
State DD	1	0.5
State EE	1	0.5
State FF	1	0.5

[1] Asian, American Indian/Alaska Native, Hawaiian/Pacific Islander, Multiracial non-Hispanic

Table 1a (continued). Demographic Characteristics of Recurrent TB Disease among Recurrent Cases Based on the US National Tuberculosis Surveillance System Data Set during 1993-2011

	Recurrent (n=188)	
	No.	%
State (continued)		
State GG	1	0.5
State HH	1	0.5
State II	1	0.5
State JJ	1	0.5
State KK	1	0.5
State LL	1	0.5
State MM	1	0.5
High Incidence State[1]	83	44
State A or State B	64	34

Foreign Born Only

	Recurrent (n=78)	
	No.	%
Nation (5 categories)		
Mexican	27	35
Southeast Asian[2]	20	26
Other[3]	17	22
Latin American/Caribbean[4]	8	10
African[5]	6	7.7
Nation (3 categories)		
Mexican	27	35
African[6]	6	7.7
Other[7]	45	58

[1] Includes State A, State B, State C (including State C_1 and State C_2), and State D

[2] Indonesia, Cambodia, Laos, Philippines, Thailand, Vietnam

[3] Bosnia-Herzegovina, Poland, Russia, Ukraine, China, Hong Kong, Republic of Korea, Democratic Peoples Republic of Korea, Taiwan

[4] Ecuador, Guatemala, Honduras, Haiti, El Salvador, Trinidad and Tobago

[5] Ethiopia, Sudan, Somalia, Zimbabwe

[6] Ethiopia, Sudan, Somalia, Zimbabwe

[7] Bosnia-Herzegovina, Poland, Russia, Ukraine, China, Hong Kong, Republic of Korea, Democratic Peoples Republic of Korea, Taiwan, Ecuador, Guatemala, Honduras, Haiti, El Salvador, Trinidad and Tobago, Indonesia, Cambodia, Laos, Philippines, Thailand, Vietnam

Table 1a (continued). Demographic Characteristics of Recurrent TB Disease among Recurrent Cases Based on the US National Tuberculosis Surveillance System Data Set during 1993-2011

Foreign Born Only

	Recurrent (n=78)	
	No.	%
Decade Moved to US[1]		
Before 1990	23	29
1990-1999	23	29
2000-2009	21	27
Years in US[2]		
< 2	1	1.3
2-5	15	19
6-10	19	24
11-20	21	27
> 20	22	28
13 or More Years in US[11]	40	51

[1] 11 missing

[2] At 2nd episode

Table 1b. Demographic Characteristics of Recurrent TB Disease in Reactivation versus Reinfection Based on the US National Tuberculosis Surveillance System Data Set during 1993-2011

Foreign Born and US Born				
	Reactivation (n=164)		Reinfection (n=24)	
	No.	%	No.	%
Age (years)[1]				
< 19	3	1.8	0	0
19-29	20	12	3	13
30-39	26	16	5	21
40-49	32	20	9	38
50-59	38	23	3	13
60-69	16	9.8	2	8.3
70 +	29	18	2	8.3
Age (binary-years)				
50 +	83	51	7	29
Age (binary-years)				
45 +	101	62	10	42
Age (biologic importance)				
< 15	1	0.6	0	0
15-24	9	5.5	1	4.2
25-49	71	43	16	67
50 +	83	51	7	29
Age (quartiles)				
< 37	40	24	8	33
37-49	41	25	9	38
50-60	41	25	4	17
61 +	42	26	3	13
Age (turtiles)				
< 43	54	33	12	50
43-56	57	35	8	33
57 +	53	32	4	17
Female	43	26	4	17
Origin				
US Born	99	60	11	46
Foreign Born	65	40	13	54

[1] Age at second episode

Table 1b (continued). Demographic Characteristics of Recurrent TB Disease in Reactivation versus Reinfection Based on the US National Tuberculosis Surveillance System Data Set during 1993-2011

	Reactivation (n=164)		Reinfection (n=24)	
	No.	%	No.	%
Race/ethnicity				
Black, non-Hispanic	43	26	9	38
Hispanic	31	25	9	38
Other[1]	39	24	3	13
White, non-Hispanic	41	25	3	13
Black and/or Hispanic	86	52	18	75
State				
State B	26	16	6	25
State A	24	15	8	33
State C_1	10	6.1	0	0
State E	9	5.5	0	0
State D	6	3.7	1	4.2
State F	7	4.3	0	0
State G	7	4.3	0	0
State H	5	3.1	1	4.2
State I	4	2.4	1	4.2
State J	3	1.8	2	8.3
State K	4	2.4	0	0
State L	3	1.8	1	4.2
State M	4	2.4	0	0
State N	3	1.8	1	4.2
State O	4	2.4	0	0
State P	4	2.4	0	0
State Q	4	2.4	0	0
State R	2	1.2	1	4.2
State S	3	1.8	0	0
State T	3	1.8	0	0
State U	3	1.8	0	0
State V	2	1.2	0	0
State W	2	1.2	0	0
State X	1	0.6	1	4.2
State Y	2	1.2	0	0
State Z	1	0.6	1	4.2
State AA	2	1.2	0	0
State C_2	2	1.2	0	0
State BB	2	1.2	0	0
State CC	2	1.2	0	0
State DD	1	0.6	0	0
State EE	1	0.6	0	0
State FF	1	0.6	0	0

[1] Asian, American Indian/Alaska Native, Hawaiian/Pacific Islander, Multiracial non-Hispanic

Table 1b (continued). Demographic Characteristics of Recurrent TB Disease in Reactivation versus Reinfection Based on the US National Tuberculosis Surveillance System Data Set during 1993-2011

	Reactivation (n=164)		Reinfection (n=24)	
	No.	%	No.	%
State (continued)				
State GG	1	0.6	0	0
State HH	1	0.6	0	0
State II	1	0.6	0	0
State JJ	1	0.6	0	0
State KK	1	0.6	0	0
State LL	1	0.6	0	0
State MM	1	0.6	0	0
High Incidence State[1]	68	42	15	63
State A or State B	50	30	14	58

Foreign Born Only

	Reactivation (n=65)		Reinfection (n=13)	
	No.	%	No.	%
Nation (5 categories)				
Mexican	20	31	7	54
Southeast Asian[2]	18	28	2	15
Other[3]	17	26	0	0
Latin American/Caribbean[4]	7	11	1	7.7
African[5]	3	4.6	3	23
Nation (3 categories)				
Mexican	20	31	7	54
African[6]	3	4.6	3	23
Other[7]	42	65	3	23

[1] Includes State A, State B, State C (including State C_1 and State C_2), and State D

[2] Indonesia, Cambodia, Laos, Philippines, Thailand, Vietnam

[3] Bosnia-Herzegovina, Poland, Russia, Ukraine, China, Hong Kong, Republic of Korea, Democratic Peoples Republic of Korea, Taiwan

[4] Ecuador, Guatemala, Honduras, Haiti, El Salvador, Trinidad and Tobago

[5] Ethiopia, Sudan, Somalia, Zimbabwe

[6] Ethiopia, Sudan, Somalia, Zimbabwe

[7] Bosnia-Herzegovina, Poland, Russia, Ukraine, China, Hong Kong, Republic of Korea, Democratic Peoples Republic of Korea, Taiwan, Ecuador, Guatemala, Honduras, Haiti, El Salvador, Trinidad and Tobago, Indonesia, Cambodia, Laos, Philippines, Thailand, Vietnam

Table 1b (continued). Demographic Characteristics of Recurrent TB Disease in Reactivation versus Reinfection Based on the US National Tuberculosis Surveillance System Data Set during 1993-2011

Foreign Born Only				
	Reactivation (n=65)		Reinfection (n=13)	
	No.	%	No.	%
Decade Moved to US[1]				
Before 1990	23	35	0	0
1990-1999	18	28	5	38
2000-2009	17	26	4	31
Years in US[2]				
< 2	1	1.5	0	0
2-5	11	17	4	31
6-10	16	25	3	23
11-20	16	25	5	38
> 20	21	32	1	7.7
13 or More Years in US[11]	36	55	4	31

[1] Missing information: 7 in reactivation, 4 in reinfection
[2] At 2nd episode

Table 2a. Social Risk Factors[1] of Recurrent TB Disease among Recurrent Cases Based on the US National Tuberculosis Surveillance System Data Set during 1993-2011

	Recurrent (n=188)	
	No.	%
Homeless	39	21
Corrections	17	9.0
Long-Term Care	6	3.2
Any substance Abuse[2]	90	48
Substance Abuse Type		
Alcohol[3]	72	38
IDU[4]	14	7.5
Non-IDU[5]	52	28
Unemployed[6]	120	64

[1] Yes if ever involved in the risk behavior as recorded at time of either episode
[2] 2 missing
[3] 4 missing
[4] 2 missing
[5] 3 missing
[6] 21 missing

Table 2b. Social Risk Factors[1] of Recurrent TB Disease in Reactivation versus Reinfection Based on the US National Tuberculosis Surveillance System Data Set during 1993-2011

	Reactivation (n=164)		Reinfection (n=24)	
	No.	%	No.	%
Homeless	34	21	5	21
Corrections	15	9.2	2	8.3
Long-Term Care	5	3.1	1	4.2
Any substance Abuse[2]	79	48	11	46
Substance Abuse Type				
Alcohol[3]	64	39	8	33
IDU[4]	11	6.7	3	13
Non-IDU[5]	48	29	4	17
Unemployed[6]	103	63	17	71

[1] Yes if ever involved in the risk behavior as recorded at time of either episode

[2] Missing: 2 reactivation, 0 reinfection

[3] Missing: 4 reactivation, 0 reinfection

[4] Missing: 2 reactivation, 0 reinfection

[5] Missing: 3 reactivation, 0 reinfection

[6] Missing: 19 reactivation, 2 reinfection

Table 3a. Clinical Features of Recurrent TB Disease among Recurrent Cases Based on the US National Tuberculosis Surveillance System Data Set during 1993-2011

	Recurrent (n=188)	
	No.	%
HIV[1]		
Negative	115	61
Positive	27	14
Change	1	0.5
Refused/Don't Know	24	13
Other[2]	1	0.5
HIV Positive by Region[3]		
US Born	21	11
Southeast Asia	1	0.5
Latin American/Caribbean	2	1.1
African	1	0.5
Mexican	4	2.3
HIV Positive[4]	29	15
Disease Site[5]		
Pulmonary	159	85
Extrapulmonary	16	8.5
Both	13	6.9
Any Pulmonary Disease	172	91
Multiple Disease Sites	13	6.9
Smear Positive[5]		
Negative	63	34
Positive	117	62
Unknown	8	4.3
Cavitary[5]	57	30
Sputum Conversion[6]	116	62

[1] Change from first to second episode; 20 missing
[2] Where 1st episode is unknown but 2nd episode is positive
[3] In first episode; 44 missing and unknown
[4] In 2nd episode, yes versus other
[5] In first episode
[6] In first episode, 37 missing

Table 3b. Clinical Features of Recurrent TB Disease in Reactivation versus Reinfection Based on the US National Tuberculosis Surveillance System Data Set during 1993-2011

	Reactivation (n=164)		Reinfection (n=24)	
	No.	%	No.	%
HIV[1]				
Negative	103	63	12	50
Positive	20	12	7	29
Change	1	0.6	0	0
Refused/Don't Know	24	15	0	0
Other[2]	1	0.6	0	0
HIV Positive by Region[3]				
US Born	17	10	4	17
Southeast Asian	0	0	1	4.2
Latin American/Caribbean	2	1.2	0	0
African	0	0	1	4.2
Mexican	3	1.8	1	4.2
HIV Positive[4]	22	13	7	29
Disease Site[5]				
Pulmonary	141	86	18	75
Extrapulmonary	14	8.5	2	8.3
Both	9	5.5	4	17
Any Pulmonary Disease	150	91	22	92
Multiple Disease Sites	9	5.5	4	17
Smear Positive[5]				
Negative	57	35	6	25
Positive	101	62	16	67
Unknown	6	3.7	2	8.3
Cavitary[5]	49	30	8	33
Sputum Conversion[6]	98	60	18	75

[1] Change from first to second episode; missing: 15 reactivation, 5 reinfection

[2] Where 1st episode is unknown but 2nd episode is positive

[3] In first episode; missing and unknown: 39 reactivation, 5 reinfection

[4] In 2nd episode, yes versus other

[5] In first episode

[6] In first episode; missing: 33 reactivation, 4 reinfection

Table 4a. Treatment Factors of Recurrent TB Disease among Recurrent Cases Based on the US National Tuberculosis Surveillance System Data Set during 1993-2011

	Recurrent (n=188)	
	No.	%
First Episode Initiation Drug[1]		
Isoniazid and Rifampin only	2	1.1
Isoniazid, Rifampin, and Pyrazinamide only	8	4.3
Isoniazid, Rifampin, Pyrazinamide, and Ethambutol Only	154	82
Two or more drugs (not above combinations)	18	9.6
No drugs	5	2.7
Change in Firstline Drug Resistance		
Both susceptible	156	83
Both resistant	10	5.3
Susceptible to resistant	20	11
Resistant to susceptible	2	1.1
Change in Firstline Drug Resistance		
Concordant	166	88
Discordant	22	12
Change in Multidrug Resistance		
Both susceptible	180	96
Both resistant	2	1.1
Susceptible to resistant	2	1.1
Unknown	4	2.1
Treatment Type in First Episode[2]		
DOT	112	60
SAT	22	12
Both	41	22
Unknown	4	2.1
Type of Provider[3]		
Health Department	104	55
Private or Other	33	18
Both	46	24

[1] In first episode; 1 unknown
[2] 9 missing
[3] In first episode; 5 missing

Table 4a (continued). Treatment Factors of Recurrent TB Disease among Recurrent Cases Based on the US National Tuberculosis Surveillance System Data Set during 1993-2011

	Recurrent (n=188)	
	No.	%
Reason for Stopping Therapy in First Episode[1]		
Completed	136	72
Lost	27	14
Moved	5	2.7
Other	5	2.7
Refused	7	3.7
Directly Observed Therapy (DOT)[2]		
Any DOT	153	81
No DOT	31	16
Unknown	4	2.1
DOT Only	112	60
Months of DOT[3]		
< 2 months	15	9.8
2-6 months	44	29
Greater than 6 months	51	33
DOT Site[4]		
Clinic	15	9.8
Field	52	34
Both	66	43
Unknown	5	3.3
Treatment Duration[5]		
Never treated	5	2.7
< 2 months	15	8.0
2-6 months	74	39
Greater than 6 months	91	48
9 or more Months of Treatment[5]	59	31

[1] 8 missing
[2] In first episode
[3] Of those receiving any DOT (153); 43 missing
[4] In first episode; of those receiving any DOT (153) 15 missing
[5] In first episode; 3 missing

Table 4a (continued). Treatment Factors of Recurrent TB Disease among Recurrent Cases Based on the US National Tuberculosis Surveillance System Data Set during 1993-2011

	Recurrent (n=188)	
	No.	%
Completed DOT Only Therapy[5]	88	79
Completed Therapy[1]	136	72
Treatment Success[5]	99	53
Death[2]	21	11

[1] In first episode; 3 missing
[2] In second episode; 4 missing

Table 4b. Treatment Factors of Recurrent TB Disease in Reactivation versus Reinfection Based on the US National Tuberculosis Surveillance System Data Set during 1993-2011

	Reactivation (n=164)		Reinfection (n=24)	
	No.	%	No.	%
First Episode Initiation Drug[1]				
Isoniazid and Rifampin only	2	1.2	0	0
Isoniazid, Rifampin, and Pyrazinamide only	8	4.9	0	0
Isoniazid, Rifampin, Pyrazinamide, and Ethambutol Only	133	81	21	88
Two or more drugs (not above combinations)	15	9.2	3	13
No drugs	5	3.1	0	0
Change in Firstline Drug Resistance				
Both susceptible	138	84	18	75
Both resistant	9	5.5	1	4.2
Susceptible to resistant	16	9.8	4	17
Resistant to susceptible	1	0.6	1	4.2
Change in Firstline Drug Resistance				
Concordant	147	90	19	79
Discordant	17	10	5	21
Change in Multidrug Resistance				
Both susceptible	157	96	23	96
Both resistant	2	1.2	0	0
Susceptible to resistant	1	0.6	1	4.2
Unknown	4	2.4	0	0
Treatment Type in First Episode[2]				
DOT	93	57	19	79
SAT	21	13	1	4.2
Both	37	23	4	17
Unknown	4	2.4	0	0
Type of Provider[3]				
Health Department	90	55	14	58
Private or Other	30	18	3	13
Both	40	24	6	25

[1] Unknown: 1 reactivation, 0 reinfection

[2] Missing: 9 reactivation, 0 reinfection

[3] Missing: 4 reactivation, 1 reinfection

Table 4b (continued). Treatment Factors of Recurrent TB Disease in Reactivation versus Reinfection Based on the US National Tuberculosis Surveillance System Data Set during 1993-2011

	Reactivation (n=164)		Reinfection (n=24)	
	No.	%	No.	%
Reason for Stopping Therapy in First Episode[1]				
Completed	116	71	20	83
Lost	24	15	3	13
Moved	5	3.1	0	0
Other	5	3.1	0	0
Refused	6	3.7	1	4.2
Directly Observed Therapy (DOT)[2]				
Any DOT	130	79	23	96
No DOT	30	18	1	4.2
Unknown	4	2.4	0	0
DOT Only	93	57	19	79
Months of DOT[3]				
< 2 months	15	12	0	0
2-6 months	39	30	5	22
Greater than 6 months	42	32	9	39
DOT Site[4]				
Clinic	13	10	2	8.7
Field	45	35	7	30
Both	56	43	10	43
Unknown	3	2.3	2	8.7
Treatment Duration[5]				
Never treated	5	3.1	0	0
< 2 months	14	8.5	1	4.2
2-6 months	67	41	7	29
Greater than 6 months	75	46	16	67
9 or more Months of Treatment[5]	46	28	13	54

[1] Missing: 8 reactivation, 0 reinfection

[2] In first episode

[3] Of those receiving any DOT (130 reactivation, 23 reinfection); missing: 34 reactivation, 9 reinfection

[4] In first episode; of those any receiving DOT (130 reactivation, 23 reinfection); missing: 13 reactivation, 2 reinfection

[5] In first episode; missing: 3 reactivation, 0 reinfection

Table 4b (continued). Treatment Factors of Recurrent TB Disease in Reactivation versus Reinfection Based on the US National Tuberculosis Surveillance System Data Set during 1993-2011

	Reactivation (n=164)		Reinfection (n=24)	
	No.	%	No.	%
Completed DOT Only Therapy[5]	71	76	17	89
Completed Therapy[1]	116	71	20	83
Treatment Success[5]	83	51	16	67
Death[2]	21	13	0	0

[1] In first episode; missing: 3 reactivation, 0 reinfection

[2] In second episode; missing: 4 reactivation, 0 reinfection

Table 5a. Genotypes of Recurrent TB Disease among Recurrent Cases Based on the US National Tuberculosis Surveillance System Data Set during 1993-2011

	Recurrent (n=188)	
	No.	%
Time between Episodes (months)		
12-16	35	19
17-25	57	30
26-59	79	42
60 +	17	9.0
Time between Episodes (years)		
1-2 years	84	45
-3 years	48	26
-4 years	18	9.6
-5 years	24	13
More than 5 years	14	7.5
Two or More Years between Episodes	110	59
Five or More Years between Episodes	17	9.0
Repeated PCRTypes in Second Episodes[1]		
1 strain	108	57
2 strains	24	13
3 strains	15	8.0
4 or more different strains	18	9.6
Repeated GENTypes in Second Episodes[2]		
1 strain	96	51
2 strains	10	5.3
3 strains	3	1.6
Lineage[3]		
EuroAmerican	125	66
East Asian	39	21
IndoOceanic	7	3.7
East African Indian	4	2.1
Bovis	4	2.1

[1] 23 missing
[2] 79 missing
[3] 9 missing

Table 5b. Genotypes of Recurrent TB Disease in Reactivation versus Reinfection Based on the US National Tuberculosis Surveillance System Data Set during 1993-2011

	Reactivation (n=164)		Reinfection (n=24)	
	No.	%	No.	%
Time between Episodes (months)				
12-16	29	18	6	25
17-25	51	31	6	25
26-59	73	45	6	25
60 +	11	6.7	6	25
Time between Episodes (years)				
1-2 years	74	45	10	42
-3 years	44	27	4	17
-4 years	16	9.8	2	8.3
-5 years	20	12	4	17
More than 5 years	10	6.1	4	17
Two or More Years	96	59	14	58
Five or More Years	11	6.7	6	25
Repeated PCRTypes in Second Episodes[1]				
1 strain	90	55	18	75
2 strains	23	14	1	4.2
3 strains	14	8.5	1	4.2
4 or more different strains	17	10	1	4.2
Repeated GENTypes in Second Episodes[2]				
1 strain	80	49	16	67
2 strains	10	6.1	0	0
3 strains	3	1.8	0	0
Lineage[3]				
EuroAmerican	107	65	18	75
East Asian	37	23	2	8.3
IndoOceanic	7	4.3	0	0
East African Indian	2	1.2	2	8.3
Bovis	3	1.8	1	4.2

[1] Missing: 20 reactivation, 3 reinfection
[2] Missing: 71 reactivation, 8 reinfection
[3] Missing: 8 reactivation, 1 reinfection

APPENDIX C: Models

Table 6a. Characteristics of recurrent tuberculosis cases included in Model 1 (full population, ignoring treatment completion) in the US during 1993-2011

	Recurrent TB					
	Reactivation (n=164)	Reinfection (n=24)	Outcome			
Characteristics	n (%)	n (%)	Crude OR	95%CI	Adjusted OR	95%CI
Demographics						
Female	43 (26)	4 (17)	0.56	0.18-1.74	0.32	0.08-1.27
Black or Hispanic	86 (52)	18 (75)	2.72	1.03-7.20 *	2.24	0.62-8.12
Living in TB high-incidence state,[1] not Mexican	37 (23)	9 (38)	3.25	1.17-9.05 *	5.31	1.52-18.59 *
Living in TB high-incidence state[1] and Mexican	13 (7.9)	5 (21)	5.14	1.46-18.08 *	2.86	0.50-16.35
Not living in a TB high-incidence[1] state and Mexican	7 (4.3)	2 (8.3)	3.82	0.67-21.51	12.27	1.09-137.9 *
12 or fewer years livings in U.S. at second episode	29 (18)	9 (38)	2.79	1.12-7.00 *	2.06	0.59-7.23
Clinical Features						
HIV positive,[2] less than 9 months of treatment at first episode[3]	13 (7.9)	5 (21)	6.73	1.80-25.18 *	30.87	5.11-186.6 *
HIV positive,[2] 9 or more months of treatment at first episode[3]	9 (5.5)	2 (8.3)	3.89	0.68-22.14	3.79	0.52-27.49
Treatment Factors						
Not HIV positive,[2] 9 or more months of treatment at first episode[3]	37 (23)	11 (46)	5.20	1.80-15.06 *	10.05	2.75-36.72 *
Received DOT only at first episode[4]	93 (57)	19 (79)	2.90	1.03-8.15 *	8.34	2.18-31.93 *

[*] Statistically significant at $p < 0.05$.

[1] High-incidence background TB, includes State A and State B.

[2] Missing, unknown, not offered, or refused in 39 reactivation and five reinfection patients.

[3] Treatment never initiated in five reactivation patients and unknown in one reactivation patient.

[4] Unknown in nine reactivation patients.

TB = tuberculosis; OR = odds ratio; CI = confidence interval; HIV = human immunodeficiency virus; DOT = directly observed therapy.

Table 6b. Characteristics of recurrent tuberculosis cases included in Model 2 (full population, accounting for treatment completion and interaction) in the US during 1993-2011

	Recurrent TB					
	Reactivation (n=164)	Reinfection (n=24)	Outcome			
Characteristics	n (%)	n (%)	Crude OR	95%CI	Adjusted OR	95%CI
Demographics						
Female	43 (26)	4 (17)	0.56	0.18-1.74	0.32	0.08-1.28
Black or Hispanic	86 (52)	18 (75)	2.72	1.03-7.20 *	2.21	0.61-8.00
Living in TB high-incidence state,[1] not Mexican	37 (23)	9 (38)	3.25	1.17-9.05 *	5.33	1.52-18.67 *
Living in TB high-incidence state[1] and Mexican	13 (7.9)	5 (21)	5.14	1.46-18.08 *	2.94	0.51-17.09
Not living in a TB high-incidence[1] state and Mexican	7 (4.3)	2 (8.3)	3.82	0.67-21.51	12.54	1.11-141.8 *
12 or fewer years livings in U.S. at second episode	29 (18)	9 (38)	2.79	1.12-7.00 *	2.01	0.57-7.10
Clinical Features						
HIV positive,[2] less than 9 months of treatment at first episode[3]	13 (7.9)	5 (21)	6.73	1.80-25.18 *	31.02	5.14-187.2 *
HIV positive,[2] 9 or more months of treatment at first episode[3]	9 (5.5)	2 (8.3)	3.89	0.68-22.14	3.53	0.47-26.67
Treatment Factors						
Not HIV positive,[2] 9 or more months of treatment at first episode[3]	37 (23)	11 (46)	5.20	1.80-15.06 *	9.39	2.44-36.05 *
Received DOT only at first episode[4]	93 (57)	19 (79)	2.90	1.03-8.15 *	8.07	2.09-31.19 *
Completed treatment at first episode	116 (71)	20 (83)	2.07	0.67-6.37	1.27	0.32-5.09

[*] Statistically significant at $p < 0.05$.

[1] High-incidence background TB, includes State A and State B.

[2] Missing, unknown, not offered, or refused in 39 reactivation and five reinfection patients.

[3] Treatment never initiated in five reactivation patients and unknown in one reactivation patient.

[4] Unknown in nine reactivation patients.

TB = tuberculosis; OR = odds ratio; CI = confidence interval; HIV = human immunodeficiency virus; DOT = directly observed

Table 6c. Characteristics of recurrent tuberculosis cases included in Model 3 (full population, accounting for treatment completion, no interaction) in the US during 1993-2011

	Recurrent TB		Outcome			
	Reactivation (n=164)	Reinfection (n=24)				
Characteristics	n (%)	n (%)	Crude OR	95%CI	Adjusted OR	95%CI
Demographics						
Female	43 (26)	4 (17)	0.56	0.18-1.74	0.50	0.14-1.77
Black or Hispanic	86 (52)	18 (75)	2.72	1.03-7.20 *	2.08	0.65-6.63
Living in TB high-incidence state[1]	50 (30)	14 (58)	3.19	1.33-7.67 *	2.29	0.83-6.30
Mexican foreign birth	20 (12)	7 (29)	2.97	1.09-8.03 *	0.91	0.23-3.61
12 or fewer years livings in U.S. at second episode	29 (18)	9 (38)	2.79	1.12-7.00 *	2.46	0.81-7.44
Clinical Features						
HIV positive[2]	22 (13)	7 (29)	2.66	0.99-7.14	3.33	0.99-11.19
Treatment Factors						
Received DOT only at first episode[3]	93 (57)	19 (79)	2.90	1.03-8.15 *	4.11	1.25-13.51 *
Nine or more months of treatment at first episode[4]	46 (28)	13 (54)	3.03	1.27-7.25 *	2.71	0.97-7.63
Completed treatment at first episode	116 (71)	20 (83)	2.07	0.67-6.37	1.28	0.36-4.54

[*] Statistically significant at $p < 0.05$.

[1] High-incidence background TB, includes State A and State B.

[2] Missing, unknown, not offered, or refused in 39 reactivation and five reinfection patients.

[3] Unknown in nine reactivation patients.

[4] Treatment never initiated in five reactivation patients and unknown in one reactivation patient.

TB = tuberculosis; OR = odds ratio; CI = confidence interval; HIV = human immunodeficiency virus; DOT = directly observed

Table 7a. Characteristics of recurrent tuberculosis cases included in Model 4 (completed treatment population, matching Model 1 variables) in the US during 1993-2011

	Recurrent TB		Outcome			
	Reactivation (n=116)	Reinfection (n=20)				
Characteristics	n (%)	n (%)	Crude OR	95%CI	Adjusted OR	95%CI
Demographics						
Female	32 (28)	3 (15)	0.46	0.13-1.69	0.26	0.06-1.24
Black or Hispanic	55 (47)	16 (80)	4.44	1.40-14.08 *	3.63	0.85-15.43
Living in TB high-incidence state,[1] not Mexican	28 (24)	6 (30)	2.33	0.72-7.52	2.80	0.69-11.41
Living in TB high-incidence state[1] and Mexican	8 (6.9)	5 (25)	6.79	1.74-26.42 *	2.05	0.33-12.82
Not living in a TB high-incidence[1] state and Mexican	4 (3.4)	2 (10)	5.43	0.84-35.07	10.93	0.62-191.4
12 or fewer years livings in U.S. at second episode	21 (18)	9 (45)	3.70	1.36-10.06 *	2.60	0.70-9.68
Clinical Features						
HIV positive,[2] less than 9 months of treatment at first episode	7 (6.0)	2 (10)	3.19	0.54-18.91	6.12	0.66-58.01
HIV positive,[2] 9 or more months of treatment at first episode	7 (6.0)	2 (10)	3.19	0.54-18.91	2.06	0.26-16.20
Treatment Factors						
Not HIV positive,[2] 9 or more months of treatment at first episode	35 (30)	10 (50)	3.19	1.07-9.50 *	5.60	1.43-21.88 *
Received DOT only at first episode[3]	71 (61)	17 (85)	3.59	1.00-12.95	10.61	1.88-59.94 *

*Statistically significant at $p < 0.05$.

[1] High-incidence background TB, includes State A and State B.

[2] Missing, unknown, not offered, or refused in 26 reactivation and four reinfection patients.

[3] Unknown in one reactivation patient.

TB = tuberculosis; OR = odds ratio; CI = confidence interval; HIV = human immunodeficiency virus; DOT = directly observed therapy.

Table 7b. Characteristics of recurrent tuberculosis cases included in Model 5 (completed treatment population, separate test for interaction than Model 1) in the United States during 1993-2011

	Recurrent TB					
	Reactivation (n=116)	Reinfection (n=20)	Outcome			
Characteristics	n (%)	n (%)	Crude OR	95%CI	Adjusted OR	95%CI
Demographics						
Female	32 (28)	3 (15)	0.46	0.13-1.69	0.34	0.08-1.51
Black or Hispanic	55 (47)	16 (80)	4.44	1.40-14.08 *	4.00	1.02-15.64 *
Living in TB high-incidence state[1]	36 (31)	11 (55)	2.72	1.04-7.13 *	1.38	0.43-4.41
Mexican foreign birth	12 (10)	7 (35)	4.67	1.56-13.96 *	1.25	0.28-5.59
12 or fewer years livings in U.S. at second episode	21 (18)	9 (45)	3.70	1.36-10.06 *	3.52	1.05-11.85 *
Clinical Features						
HIV positive[2]	14 (12)	4 (20)	1.81	0.53-6.23	1.09	0.25-4.79
Treatment Factors						
Nine or more months of treatment at first episode	42 (36)	12 (60)	2.64	1.00-6.98	2.77	0.89-8.59
Received DOT only at first episode[3]	71 (61)	17 (85)	3.59	1.00-12.95	5.33	1.31-21.69 *

*Statistically significant at $p < 0.05$.

[1] High-incidence background TB, includes State A and State B.

[2] Missing, unknown, not offered, or refused in 26 reactivation and four reinfection patients.

[3] Unknown in one reactivation patient.

TB = tuberculosis; OR = odds ratio; CI = confidence interval; HIV = human immunodeficiency virus; DOT = directly observed therapy.

Table 7c. Characteristics of recurrent tuberculosis cases included in Model 6 (completed treatment population, no interaction, race as dummy variables) in the US during 1993-2011

	Recurrent TB		Outcome			
	Reactivation (n=116)	Reinfection (n=20)			Adjusted	
Characteristics	n (%)	n (%)	Crude OR	95%CI	OR	95%CI
Demographics						
Female	32 (28)	3 (15)	0.46	0.13-1.69	0.32	0.07-1.47
Race						
Black, non-Hispanic	30 (26)	8 (40)	4.07	1.13-14.59 *	6.71	1.58-28.51 *
Hispanic	25 (22)	8 (40)	4.88	1.34-17.68 *	0.69	0.06-8.66
Living in TB high-incidence state[1]	36 (31)	11 (55)	2.72	1.04-7.13 *	2.02	0.58-7.03
Mexican foreign birth	12 (10)	7 (35)	4.67	1.56-13.96 *	6.23	0.51-75.99
12 or fewer years livings in U.S. at second episode	21 (18)	9 (45)	3.70	1.36-10.06 *	3.99	1.14-14.02 *
Clinical Features						
HIV positive[2]	14 (12)	4 (20)	1.81	0.53-6.23	1.24	0.28-5.57
Treatment Factors						
Nine or more months of treatment at first episode	42 (36)	12 (60)	2.64	1.00-6.98	2.93	0.92-9.31
Received DOT only at first episode[3]	71 (61)	17 (85)	3.59	1.00-12.95	6.59	1.54-28.31 *

[*] Statistically significant at $p < 0.05$.

[1] High-incidence background TB, includes State A and State B.

[2] Missing, unknown, not offered, or refused in 26 reactivation and four reinfection patients.

[3] Unknown in one reactivation patient.

TB = tuberculosis; OR = odds ratio; CI = confidence interval; HIV = human immunodeficiency virus; DOT = directly observed

Table 7d. Characteristics of recurrent tuberculosis cases included in Model 7 (completed treatment population, no interaction, 4 high-incidence states) in the United States during 1993-2011

	Recurrent TB					
	Reactivation (n=116)	Reinfection (n=20)	Outcome			
Characteristics	n (%)	n (%)	Crude OR	95%CI	Adjusted OR	95%CI
Demographics						
Female	32 (28)	3 (15)	0.46	0.13-1.69	0.33	0.08-1.43
Black or Hispanic	55 (47)	16 (80)	4.44	1.40-14.08 *	4.08	1.05-15.94 *
Living in TB high-incidence state[1]	48 (41)	11 (55)	1.73	0.67-4.50	1.11	0.36-3.45
Mexican foreign birth	12 (10)	7 (35)	4.67	1.56-13.96 *	1.36	0.31-5.97
12 or fewer years livings in U.S. at second episode	21 (18)	9 (45)	3.70	1.36-10.06 *	3.47	1.04-11.64 *
Clinical Features						
HIV positive[2]	14 (12)	4 (20)	1.81	0.53-6.23	1.06	0.24-4.58
Treatment Factors						
Nine or more months of treatment at first episode	42 (36)	12 (60)	2.64	1.00-6.98	2.82	0.91-8.70
Received DOT only at first episode[3]	71 (61)	17 (85)	3.59	1.00-12.95	5.47	1.35-22.21 *

[*] Statistically significant at $p < 0.05$.

[1] High-incidence background TB, includes State A, State B, State C, and State D.

[2] Missing, unknown, not offered, or refused in 26 reactivation and four reinfection patients.

[3] Unknown in one reactivation patient.

TB = tuberculosis; OR = odds ratio; CI = confidence interval; HIV = human immunodeficiency virus; DOT = directly observed

Table 8a. Characteristics of recurrent tuberculosis cases included in Model 8 (no treatment completion population, matching Model 1 variables) in the US during 1993-2011

	Recurrent TB					
	Reactivation (n=48)	Reinfection (n=4)	Outcome			
Characteristics	n (%)	n (%)	Crude OR	95%CI	Adjusted OR	95%CI
Demographics						
Female	11 (23)	1 (25)	1.12	0.11-11.89	n/a	n/a
Black or Hispanic	31 (66)	2 (50)	0.55	0.07-4.25	n/a	n/a
Living in TB high-incidence state,[1] not Mexican	9 (19)	3 (75)	10.33	0.95-111.8	n/a	n/a
Living in TB high-incidence state[1] and Mexican	5 (10)	0 (0)	0.00	0->999	n/a	n/a
Not living in a TB high-incidence[1] state and Mexican	3 (6.3)	0 (0)	0.00	0->999	n/a	n/a
12 or fewer years livings in U.S. at second episode	8 (17)	0 (0)	>999	<0.0->999	n/a	n/a
Clinical Features						
HIV positive,[2] less than 9 months of treatment at first episode	6 (13)	3 (75)	150119	<0.0->999	n/a	n/a
HIV positive,[2] 9 or more months of treatment at first episode	2 (4.2)	0 (0)	1.00	<0.0->999	n/a	n/a
Treatment Factors						
Not HIV positive,[2] 9 or more months of treatment at first episode	2 (4.2)	1 (25)	150119	<0.0->999	n/a	n/a
Received DOT only at first episode[3]	22 (46)	2 (50)	1.18	0.15-9.09	n/a	n/a

[*] Statistically significant at $p < 0.05$.

[1] High-incidence background TB, includes State A and State B.

[2] Missing, unknown, not offered, or refused in 13 reactivation and one reinfection patients.

[3] Unknown in eight reactivation patients.

TB = tuberculosis; OR = odds ratio; CI = confidence interval; HIV = human immunodeficiency virus; DOT = directly observed therapy.

Table 8b. Characteristics of recurrent tuberculosis cases included in Model 9 (no treatment completion population, separate test for interaction than Model 1) in the US during 1993-2011

	Recurrent TB		Outcome			
	Reactivation (n=48)	Reinfection (n=4)			Adjusted	
Characteristics	n (%)	n (%)	Crude OR	95%CI	OR	95%CI
Demographics						
Female	11 (23)	1 (25)	1.12	0.11-11.89	0.88	0.06-13.03
Black or Hispanic	31 (66)	2 (50)	0.55	0.07-4.25	0.23	0.02-3.06
Living in TB high-incidence state[1]	14 (29)	3 (75)	7.29	0.70-76.18	10.65	0.82-139
Treatment Factors						
Nine or more months of treatment at first episode	4 (8.3)	1 (25)	3.67	0.31-43.95	5.72	0.32-103.9
Received DOT only[2]	22 (46)	2 (50)	1.18	0.15-9.09	0.57	0.05-6.86

[*] Statistically significant at $p < 0.05$.

[1] High-incidence background TB, includes State A and State B.

[2] Unknown in eight reactivation patients.

TB = tuberculosis; OR = odds ratio; CI = confidence interval; HIV = human immunodeficiency virus; DOT = directly observed therapy.

APPENDIX D: SAS Code

Matching Algorithm

```
*match old cases to new cases;
data work.BEFORE;
set Surv.pc062512;
        BEFDOB=DOB; BEFYEAR=YEAR; BEFSTCASENO=STCASENO;
        BEFCCCASENO=CCCASENO;BEFSTATE=STATE; BEFCITY=CITY;
        BEFCOUNTY=COUNTY; BEFZIPCODE=ZIPCODE; BEFAGE=AGE;
        BEFAGE3=AGE3; BEFSEX=SEX; BEFRACEHISP=RACEHISP;
        BEFORIGIN=ORIGIN; BEFNATION=NATION; BEFYRSIN_US=YRSIN_US;
        BEFUSDATE=USDATE; BEFUSYEAR=USYEAR;BEFPREVTB=PREVTB;
        BEFPREVYR=PREVYR;  BEFPREVMULT=PREVMULT;BEFDIS_SITE=DIS_SITE;
        BEFSPSMEAR=SPSMEAR; BEFSPCULT=SPCULT;
        BEFMICREXAM=MICREXAM; BEFCULTOTHR=CULTOTHR; BEFXRAY=XRAY;
        BEFXRAYCAV=XRAYCAV; BEFXRAYCOND=XRAYCOND;
        BEFTBTEST=TBTEST; BEFTBTESTMM=TBTESTMM;
        BEFPRIMARYOCC=PRIMARYOCC; BEFOCCUPATN=OCCUPATN;
        BEFOCCUHCW=OCCUHCW; BEFOCCUCORR=OCCUCORR;
        BEFOCCUMIGR=OCCUMIGR; BEFOCCUOTH=OCCUOTH;
        BEFOCCUUNEM=OCCUUNEM; BEFOCCUUNK=OCCUUNK;
        BEFHIVSTAT=HIVSTAT; BEFHOMELESS=HOMELESS;
        BEFCORRINST=CORRINST; BEFCORRTYPE=CORRTYPE;
        BEFLONGTERM=LONGTERM; BEFLONGTYPE=LONGTYPE; BEFIDU=IDU;
        BEFNONIDU=NONIDU; BEFALCOHOL=ALCOHOL; BEFISUSDATE=ISUSDATE;
        BEFFIRSTLINE=FIRSTLINE; BEFMDR=MDR; BEFXDR=XDR;
        BEFINITDRG=INITDRG; BEFRXDATE=RXDATE; BEFPROVTYPE=PROVTYPE;
        BEFDOT=DOT; BEFDOTSITE=DOTSITE; BEFDOTWEEKS=DOTWEEKS;
        BEFCONVERT=CONVERT; BEFCPOSDATE=CPOSDATE;
        BEFCNEGDATE=CNEGDATE; BEFSTOPREAS=STOPREAS; BEFSTOPTHER =
        STOPTHER;
BEFSEQ = 'A';MATCHYR=BEFYEAR;
keep    BEFDOB BEFYEAR BEFSTCASENO BEFCCCASENO BEFSTATE BEFCITY
        BEFCOUNTY BEFZIPCODE BEFAGE BEFAGE3 BEFSEX BEFRACEHISP
        BEFORIGIN BEFNATION BEFYRSIN_US BEFUSDATE BEFUSYEAR
        BEFPREVTB BEFPREVYR BEFPREVMULT BEFDIS_SITE BEFSPSMEAR
        BEFSPCULT BEFMICREXAM BEFCULTOTHR BEFXRAY BEFXRAYCAV
        BEFXRAYCOND BEFTBTEST BEFTBTESTMM BEFPRIMARYOCC
        BEFOCCUPATN BEFOCCUHCW BEFOCCUCORR BEFOCCUMIGR
        BEFOCCUOTH BEFOCCUUNEM BEFOCCUUNK BEFHIVSTAT BEFHOMELESS
        BEFCORRINST BEFCORRTYPE BEFLONGTERM BEFLONGTYPE BEFIDU
        BEFNONIDU BEFALCOHOL BEFISUSDATE BEFFIRSTLINE  BEFMDR  BEFXDR
        BEFINITDRG BEFRXDATE BEFPROVTYPE BEFDOT BEFDOTSITE
        BEFDOTWEEKS BEFCONVERT BEFCPOSDATE BEFCNEGDATE BEFSTOPREAS
        BEFSTOPTHER BEFSEQ MATCHYR STATE DOB;
run;

data work.after;
set Surv.pc062512;
```

```
          AFTDOB=DOB; AFTYEAR=YEAR; AFTSTCASENO=STCASENO;
          AFTCCCASENO=CCCASENO; AFTSTATE=STATE; AFTCITY=CITY;
          AFTCOUNTY=COUNTY; AFTZIPCODE=ZIPCODE; AFTAGE=AGE;
          AFTAGE3=AGE3; AFTSEX=SEX; AFTRACEHISP=RACEHISP;
          AFTORIGIN=ORIGIN; AFTNATION=NATION; AFTYRSIN_US=YRSIN_US;
          AFTUSDATE=USDATE; AFTUSYEAR=USYEAR; AFTPREVTB=PREVTB;
          AFTPREVYR=PREVYR;  AFTPREVMULT=PREVMULT;
          AFTDIS_SITE=DIS_SITE; AFTSPSMEAR=SPSMEAR; AFTSPCULT=SPCULT;
          AFTMICREXAM=MICREXAM; AFTCULTOTHR=CULTOTHR;
          AFTXRAY=XRAY; AFTXRAYCAV=XRAYCAV; AFTXRAYCOND=XRAYCOND;
          AFTTBTEST=TBTEST; AFTTBTESTMM=TBTESTMM;
          AFTPRIMARYOCC=PRIMARYOCC; AFTOCCUPATN=OCCUPATN;
          AFTOCCUHCW=OCCUHCW; AFTOCCUCORR=OCCUCORR;
          AFTOCCUMIGR=OCCUMIGR; AFTOCCUOTH=OCCUOTH;
          AFTOCCUUNEM=OCCUUNEM; AFTOCCUUNK=OCCUUNK;
          AFTHIVSTAT=HIVSTAT; AFTHOMELESS=HOMELESS;
          AFTCORRINST=CORRINST; AFTCORRTYPE=CORRTYPE;
          AFTLONGTERM=LONGTERM; AFTLONGTYPE=LONGTYPE;  AFTIDU=IDU;
          AFTNONIDU=NONIDU; AFTALCOHOL=ALCOHOL; AFTISUSDATE=ISUSDATE;
          AFTFIRSTLINE=FIRSTLINE; AFTMDR=MDR; AFTXDR=XDR;
          AFTINITDRG=INITDRG; AFTRXDATE=RXDATE; AFTPROVTYPE=PROVTYPE;
          AFTDOT=DOT; AFTDOTSITE=DOTSITE; AFTDOTWEEKS=DOTWEEKS;
          AFTCONVERT=CONVERT; AFTCPOSDATE=CPOSDATE;
          AFTCNEGDATE=CNEGDATE; AFTSTOPREAS=STOPREAS; AFTSTOPTHER =
          STOPTHER;
AFTSEQ = 'B';MATCHYR=AFTPREVYR;
keep      AFTDOB AFTYEAR AFTSTCASENO AFTCCCASENO AFTSTATE AFTCITY
          AFTCOUNTY AFTZIPCODE AFTAGE AFTAGE3 AFTSEX AFTRACEHISP
          AFTORIGIN AFTNATION AFTYRSIN_US AFTUSDATE AFTUSYEAR
          AFTPREVTB AFTPREVYR AFTPREVMULT AFTDIS_SITE AFTSPSMEAR
          AFTSPCULT AFTMICREXAM AFTCULTOTHR AFTXRAY AFTXRAYCAV
          AFTXRAYCOND AFTTBTEST AFTTBTESTMM AFTPRIMARYOCC
          AFTOCCUPATN AFTOCCUHCW AFTOCCUCORR AFTOCCUMIGR
          AFTOCCUOTH AFTOCCUUNEM AFTOCCUUNK     AFTHIVSTAT
          AFTHOMELESS AFTCORRINST AFTCORRTYPE AFTLONGTERM
          AFTLONGTYPE AFTIDU AFTNONIDU AFTALCOHOL AFTISUSDATE
          AFTFIRSTLINE  AFTMDR  AFTXDR AFTINITDRG AFTRXDATE AFTPROVTYPE
          AFTDOT AFTDOTSITE AFTDOTWEEKS AFTCONVERT AFTCPOSDATE
          AFTCNEGDATE AFTSTOPREAS AFTSTOPTHER AFTSEQ MATCHYR STATE
          DOB;
run;

proc sort data = work.before;by STATE MATCHYR DOB;run;
proc sort data = work.after;by STATE MATCHYR DOB;un;
data work.befaft;merge work.before work.after;by STATE MATCHYR DOB;run;
data work.match1;set work.befaft;if AFTSEQ ne '';if BEFYEAR ne .;run;
data julia.match1;set work.match1;run;
data julia.geno_data;set work.ntgs work.ntgsn;run;

*remerge geno files with recurrence variable;
data matches.geno_data_recur;set work.ntgs work.ntgsn;run;
```

```
********************************************************************;
*                    create final merged document pulling in geno data                    ;
********************************************************************;
*create new file from geno data to seperate first episode;
data before;set matches.geno_data_recur;
*select out all before cases;
          if seq eq 'B' then delete;
*rename all variables before;
          BEFDNAFP=DNAFP;  BEFGENType=GENType; BEFMiru=Miru; BEFMiru2=Miru2;
          BEFNUMBANDS=NUMBANDS; BEFObs=Obs; BEFPCRType=PCRType;
          BEFSPOLIGO=SPOLIGO; BEFSpoligotype=Spoligotype;
          BEFgeno_report_date=geno_report_date; BEFpair_id=pair_id; BEFstateid=stateid;
          BEFSTCASENO=stcaseno; BEFmatch=match;
keep BEFDNAFP BEFGENType BEFMiru BEFMiru2 BEFNUMBANDS BEFObs
          BEFPCRType BEFSPOLIGO BEFSpoligotype BEFgeno_report_date BEFpair_id
          BEFstateid BEFSTCASENO BEFmatch;
run;

*merge of before cases;
proc sort data=matches.match1;by BEFSTCASENO;run;
proc sort data=before;by BEFSTCASENO;run;
data before_merged;merge matches.match1 before;by BEFSTCASENO; run;

*create new file from geno data to seperate second episode;
data after;
          set matches.geno_data_recur;
*select out all before cases;
          if seq eq 'A' then delete;
*rename all variables before;
          AFTDNAFP=DNAFP;  AFTGENType=GENType; AFTMiru=Miru; AFTMiru2=Miru2;
          AFTNUMBANDS=NUMBANDS; AFTObs=Obs; AFTPCRType=PCRType;
          AFTSPOLIGO=SPOLIGO; AFTSpoligotype=Spoligotype;
          AFTgeno_report_date=geno_report_date;  AFTpair_id=pair_id; AFTstateid=stateid;
          AFTSTCASENO=stcaseno; AFTmatch=match;
keep AFTDNAFP AFTGENType AFTMiru AFTMiru2 AFTNUMBANDS AFTObs
          AFTPCRType AFTSPOLIGO AFTSpoligotype  AFTgeno_report_date AFTpair_id
          AFTstateid AFTSTCASENO AFTmatch;
run;

*merge of after cases;
proc sort data=before_merged;by AFTSTCASENO;run;
proc sort data=after;by AFTSTCASENO;run;
data all_merged;merge before_merged after;by AFTSTCASENO;run;

*create the final dataset;
data matches.all_merged2;set all_merged;
          *create recurrent, reactivation, and reinfection variables;
          if BEFmatch=1 then react=1;if AFTmatch=1 then react=1;else react=0;
          if BEFmatch=0 then reinf=1;if AFTmatch=0 then reinf=1;else reinf=0;
          if react=1 or reinf=1 then recurr=1;else recurr=0;run;
```

Data Cleaning and New Variable Creation

```
**************************************************************************************;
* This program writes the code and cleans the data for the recurrent TB      ;
* dataset frozen on 06/25/2012                                                ;
*                                                                             ;
* Uses the merged dataset all_merged2                                         ;
*                                                                             ;
* Written by: Julia Interrante                                                ;
* Created on: 10/12/2013                                                      ;
**************************************************************************************;

*cleans and formats all merged geno and state data for the final dataset;
data matches.merged_AddLab;
set matches.all_merged2;

*adds in genotypes confirmed from the lab (as of 10/3/13 only additions from Lindsay's paper
added);
        *considered a match from Lindsay's analysis (RFLP match);
        if befpair_id=1631 then befmatch=1;
        if aftpair_id=1631 then do;aftmatch=1; react=1; reinf=0; recurr=1; end;
        if befpair_id=1634 then befmatch=1;
        if aftpair_id=1634 then do;aftmatch=1; react=1; reinf=0; recurr=1; end;
        if befpair_id=2752 then befmatch=1;
        if aftpair_id=2752 then do;aftmatch=1; react=1; reinf=0; recurr=1; end;

        *considered a match from Lindsay's (rare spoligo/miru combo);
        if befpair_id=89 then befmatch=1;
        if aftpair_id=89 then do;aftmatch=1; react=1; reinf=0; recurr=1; end;

        *considered a match from Lindsay's (only 1-loci MIRU off);
        if befpair_id=1248 then befmatch=1;
        if aftpair_id=1248 then do;aftmatch=1; react=1; reinf=0; recurr=1; end;

        *considered discordant from Lindsay's analysis;
        if befpair_id=2873 then befmatch=0;
        if aftpair_id=2873 then do;aftmatch=0;react=0;reinf=1;recurr=1;end;

*removes non real pairs and reassigns recurrence type from bef/aft match off;
        *delete where 1st and 2nd episodes were not actually a pair;
        if befstcaseno='XXXXXXXXXXXX' then delete;
        if befstcaseno=' XXXXXXXXXXXX' then delete;
        if aftstcaseno=' XXXXXXXXXXXX' then delete;
        if aftstcaseno=' XXXXXXXXXXXX' then delete;
        if befstcaseno=' XXXXXXXXXXXX' then delete;
        if befstcaseno=' XXXXXXXXXXXX' then delete;
        if befstcaseno=' XXXXXXXXXXXX' then delete;
        if befstcaseno=' XXXXXXXXXXXX' then delete;
        if befstcaseno=' XXXXXXXXXXXX' then delete;
        if befstcaseno=' XXXXXXXXXXXX' then delete;
        if befstcaseno=' XXXXXXXXXXXX' then delete;
```

```
*reassign geno information for 1st and 2nd episodes that are actually pairs;
        if befstcaseno=' XXXXXXXXXXXX' then do;
                aftpair_id=2135;aftmatch=1;react=1;reinf=0;end;
        if befstcaseno=' XXXXXXXXXXXX' then do;
                aftpair_id=2934;aftmatch=9;reinf=0;recurr=0;end;
        if befstcaseno= ' XXXXXXXXXXXX' then befmatch=1;
        if befstcaseno= ' XXXXXXXXXXXX' then do;befmatch=9;aftmatch=9;end;
        if befstcaseno=' XXXXXXXXXXXX'then do;
                befmatch=1;aftmatch=1;recurr=1;react=1;end;
        if befstcaseno=' XXXXXXXXXXXX'then do;
                befmatch=1;aftmatch=1;recurr=1;react=1;end;

*delete case were pairs were reassigned but geno data was not available on the new pair;
        if befstcaseno=' XXXXXXXXXXXX'then delete;
        if befstcaseno=' XXXXXXXXXXXX'then delete;

*temporary controls that will need to be changed when we decide if these are actual pairs;
if aftstcaseno=' XXXXXXXXXXXX'then do; befmatch=6;aftmatch=6;reinf=0;recurr=0;end;
if aftstcaseno=' XXXXXXXXXXXX'then do; befmatch=6;aftmatch=6;react=0;recurr=0;end;

*resets spoligotype for pair_id 773 where numbers were wrong from our records from lab;
if befstcaseno=' XXXXXXXXXXXX'then befspoligotype=700076717763771;
if aftstcaseno=' XXXXXXXXXXXX'then aftspoligotype=700076717760700;

*adds in genotypes confirmed from the lab (as of 1/3/14 only additions from Lab before questions
from Maryam);
        *create temporary variable for added on 1/4/14 indicating obs still need to be cleaned;
        if befpair_id in (2934,2136,2131,765,164,1827,1064,2495,2425) then jan4add=1;
                else jan4add=0;

*considered a match - reactivation;
if befpair_id in (2131,2136,2425,765,164,2495,2934) then befmatch=1;
if aftpair_id in (2131,2136,2425,765,164,2495,2934) then do;
        aftmatch=1;react=1;reinf=0;recurr=1;end;

*considered not a match - reinfection;
if befpair_id in (1064,1827) then befmatch=0;
if aftpair_id in (1064,1827) then do; aftmatch=0;react=0;reinf=1;recurr=1;end;

*adds in genotypes confirmed from the lab (as of 1/10/14 additions after discussion with
Maryam);
        *create temporary variable for added on 1/10/14 indicating obs still need to be cleaned;
        if befpair_id in (2318,2311,704,602,773) then jan10add=1; else jan10add=0;

        *considered a match - reactivation;
        if befpair_id in (2318,2311,704,602,773) then befmatch=1;
        if aftpair_id in (2318,2311,704,602,773) then do;
                aftmatch=1;react=1;reinf=0;recurr=1;end;

*removes non-real pairs from sex, origin, race not matching;
```

```
        *remove non real pairs from sex not matching;
        if befstcaseno=' XXXXXXXXXXXX'then delete;
        if befstcaseno=' XXXXXXXXXXXX'then delete;
        if befstcaseno=' XXXXXXXXXXXX'then delete;
        if befstcaseno=' XXXXXXXXXXXX'then delete;
        if befstcaseno=' XXXXXXXXXXXX'then delete;
        if befstcaseno=' XXXXXXXXXXXX'then delete;
        if befstcaseno=' XXXXXXXXXXXX'then delete;
        if befstcaseno=' XXXXXXXXXXXX'then delete;

        *remove non real pairs from origin not matching;
        if befstcaseno=' XXXXXXXXXXXX'then delete;

        *remove non real pairs from race not matching;
        if befstcaseno=' XXXXXXXXXXXX'then delete;
        if befstcaseno=' XXXXXXXXXXXX'then delete;
        if befstcaseno=' XXXXXXXXXXXX'then delete;
        if befstcaseno=' XXXXXXXXXXXX'then delete;

        *removes other matching and invalid problems;
        *remove where bef and aft state case numbers are the same, not a recurrence;
        *includes 11 cases - all reactivation;
        if befstcaseno=aftstcaseno then delete;

        *remove where probably not a match and treatment stop and start not >12mo;
        if befstcaseno=' XXXXXXXXXXXX'then delete;

        *remove where 1st treatment stop and 2nd treatment start not >12mo;
        if befstcaseno=' XXXXXXXXXXXX'then delete;

        *removed where pair id 717 and 716 are mixed and duplicates;
        *assumption is that 2nd person is duplicate and in error;
        if befstcaseno=' XXXXXXXXXXXX'then delete;
        if befstcaseno=' XXXXXXXXXXXX'then aftpair_id=716;

*cleans matching problems;
        *add a new category in race for when bef and aft race are not equal;
        befracehisp2=befracehisp;
        aftracehisp2=aftracehisp;
        if befracehisp ne aftracehisp and (befracehisp='HISP' or aftracehisp='HISP') then do;
                befracehisp2='HISP';aftracehisp2='HISP';end;
        else if befracehisp ne aftracehisp then do;befracehisp2='MULT';aftracehisp2='MULT';

        *correct befracehisp2 to white when aft is white and origin is bosnia;
        if befstcaseno=' XXXXXXXXXXXX'then do;befracehisp2='WHITE';
                aftracehisp2='WHITE';end;

        *correct bef origin to foreign born, really is russian;
        if befstcaseno=' XXXXXXXXXXXX'then beforigin='FBORN';

        *correct bef origin to foreign born, bef was unknown and aft is foreign;
        if befstcaseno=' XXXXXXXXXXXX'then beforigin='FBORN';

*fixes country codes (nation);
if befnation='USA' then befnation=' ';
if aftnation='USA' then aftnation=' ';
if befnation=' ' and aftnation ne ' ' then befnation=aftnation;
if aftnation=' ' and befnation ne ' ' then aftnation=befnation;

*fixes where asian and ven was input rather than vnm;
if befnation='VNM' and aftnation='VEN' then aftnation=befnation;

*fixes where puerto rico was listed as us born but had foreign born responses;
if befnation='PRI' then do;befusyear=' ';aftusyear=' ';end;

*fixes problems from time between first and second episode;
        *fixes where aftrxdate was reported at 1996 (exact date when 1st episode tx started and
        stopped) but susceptibility testing was in 12/1997 and case was reported in 2/1998;
        if aftstcaseno=' XXXXXXXXXXXX'then aftrxdate=13867;

        *fixes where befstopther was reported at 2010 but that would make it 3+ years of
        treatment, which isn't correct, and conversion date was in 7/05 and followup
        susceptibility testing was 3/07;
        if befstcaseno=' XXXXXXXXXXXX'then befstopther=17466;

*creates new variables for demographic and risk factors;
*create a single variable that is 1 (yes) for ever having experienced the risk factor;
        *occupation;
        *create a variable for all high risk occupations;
        if befoccupatn='HCW' or aftoccupatn='HCW' or befoccupatn='MIGR' or
        aftoccupatn='MIGR'  then highrisk_occu=1;
                else if befoccupatn ne ' ' or aftoccupatn ne ' ' then highrisk_occu=0;
                else highrisk_occu=.;

        *create a variable for employed;
        if befoccupatn='HCW' or aftoccupatn='HCW' or befoccupatn='MIGR' or
        aftoccupatn='MIGR' or befoccupatn='MULT' or aftoccupatn='MULT' or
        befoccupatn='OTH' or aftoccupatn='OTH' then employed=1;
        else if befoccupatn='UNEMP' or aftoccupatn='UNEMP' then employed=0;
        else employed=.;

        *create a variable for unemployed;
        if befoccupatn='UNEMP' or aftoccupatn='UNEMP' then unemployed=1;
        else if befoccupatn='HCW' or aftoccupatn='HCW' or befoccupatn='MIGR' or
        aftoccupatn='MIGR' or befoccupatn='MULT' or aftoccupatn='MULT' or
        befoccupatn='OTH' or aftoccupatn='OTH' then unemployed=0;
        else unemployed=.;

        *ever_homeless;
        if befhomeless='Y' or afthomeless='Y' then ever_homeless=1;
        else if befhomeless='N' or afthomeless='N' then ever_homeless=0;else ever_homeless=.;
```

```
    *ever_correction;
    if befcorrinst='Y' or aftcorrinst='Y' then ever_corr=1;
    else if befcorrinst='N' or aftcorrinst='N' then ever_corr=0;else ever_corr=.;

    *ever_longterm;
    if beflongterm='Y' or aftlongterm='Y' then ever_longterm=1;
    else if beflongterm='N' or aftlongterm='N' then ever_longterm=0;else ever_longterm=.;

    *ever_riskfac;
    if befhomeless='Y' or afthomeless='Y' or befcorrinst='Y' or aftcorrinst='Y' or
    beflongterm='Y' or aftlongterm='Y' then ever_riskfac=1;
    else if (befhomeless='N' or afthomeless='N') and (befcorrinst='N' or aftcorrinst='N') and
    (beflongterm='N' or aftlongterm='N') then ever_riskfac=0; else ever_riskfac=.;

    *ever_subabuse;
    if befalcohol='Y' or aftalcohol='Y' or befidu='Y' or aftidu='Y' or befnonidu='Y' or
    aftnonidu='Y' then ever_subabuse=1;
    else if befalcohol='N' or aftalcohol='N' or befidu='N' or aftidu='N' or befnonidu='N' or
    aftnonidu='N' then ever_subabuse=0; else ever_subabuse=.;

    *ever_alc;
    if befalcohol='Y' or aftalcohol='Y' then ever_alc=1;
    else if befalcohol='N' or aftalcohol='N' then ever_alc=0; else ever_alc=.;

    *ever_idu;
    if befidu='Y' or aftidu='Y' then ever_idu=1;
    else if befidu='N' or aftidu='N' then ever_idu=0;else ever_idu=.;

    *ever_nonidu;
    if befnonidu='Y' or aftnonidu='Y' then ever_nonidu=1;
    else if befnonidu='N' or aftnonidu='N' then ever_nonidu=0;else ever_nonidu=.;

*add calculated values for those who had missing aftyears in US but gave US year to calculate;
    *aftyrsin_us=aftyear-aftusyear;
    if befstcaseno=' XXXXXXXXXXXX'then aftyrsin_us=26;
    if befstcaseno=' XXXXXXXXXXXX'then aftyrsin_us=39;
    if befstcaseno=' XXXXXXXXXXXX'then aftyrsin_us=2;
    if befstcaseno=' XXXXXXXXXXXX'then aftyrsin_us=18;

    *make assumption that years in US is at least as long as the time between episodes and
    that years in US is at least as long as the year of their first episode;
    *where time_btwn_yr gt aftyrsin_us and aftyrsin_us ne .;*if befyrsin_us ne . then use that
    date;*if befyrsin_us eq . then use befyear;
    if befstcaseno=' XXXXXXXXXXXX'then aftyrsin_us=14;
    if befstcaseno=' XXXXXXXXXXXX'then aftyrsin_us=5;
    if befstcaseno=' XXXXXXXXXXXX'then aftyrsin_us=5;
    if befstcaseno=' XXXXXXXXXXXX'then aftyrsin_us=3;
    if befstcaseno=' XXXXXXXXXXXX'then aftyrsin_us=36;

*orders years in US by 1=<2, 2=2-5, 3=6-10, 4=>10 (this only looks at 2nd episode);
if aftyrsin_us ne . and aftyrsin_us lt 2 then yrsin_us4=1;
```

```
else if aftyrsin_us ge 2 and aftyrsin_us le 5 then yrsin_us4=2;
else if aftyrsin_us ge 6 and aftyrsin_us le 10 then yrsin_us4=3;
else if aftyrsin_us gt 10 then yrsin_us4=4;else yrsin_us4=.;

*orders years in US by quantiles: 1=<7, 2=7-11, 3=12-20, 4=>21;
if aftyrsin_us ne . and aftyrsin_us lt 7 then yrsin_us_quant=1;
else if aftyrsin_us ge 7 and aftyrsin_us le 11 then yrsin_us_quant=2;
else if aftyrsin_us ge 12 and aftyrsin_us le 20 then yrsin_us_quant=3;
else if aftyrsin_us ge 21 then yrsin_us_quant=4; else yrsin_us_quant=.;

*order years in US by literature groupings: 1=<2, 2=2-5, 3=6-10, 4=11-20, 5=>20;
if aftyrsin_us ne . and aftyrsin_us lt 2 then yrsin_us5=1;
else if aftyrsin_us ge 2 and aftyrsin_us le 5 then yrsin_us5=2;
else if aftyrsin_us ge 6 and aftyrsin_us le 10 then yrsin_us5=3;
else if aftyrsin_us ge 11 and aftyrsin_us le 20 then yrsin_us5=4;
else if aftyrsin_us gt 20 then yrsin_us5=5;else yrsin_us5=.;

        *adds 0=US born for yrsin_us5;
        yrsin_us5_u=yrsin_us5; if aftorigin='USBORN' then yrsin_us5_u=5;

        *creates new years in US based on the midpoint between median years in US for reinf vs.
        react;
        if aftyrsin_us gt 0 and aftyrsin_us le 12 then yrsin_us13_u=0;
        if aftyrsin_us ge 13 or aftyrsin_us = . then yrsin_us13_u=1;

        *same as above (yrsin_us13_u) but 1=le 12 yrs in US;
        if yrsin_us13_u=0 then yrs_us_mid=1;if yrsin_us13_u=1 then yrs_us_mid=0;

*orders ages by 1=<19, 2=19-29, 3=30-39, 4=40-49, 5=50-59, 6=60-69, 7=70+;
if aftage ne . and aftage lt 19 then aftage7=1;else if aftage ge 19 and aftage le 29 then aftage7=2;
else if aftage ge 30 and aftage le 39 then aftage7=3;
else if aftage ge 40 and aftage le 49 then aftage7=4;
else if aftage ge 50 and aftage le 59 then aftage7=5;
else if aftage ge 60 and aftage le 69 then aftage7=6;
else if aftage ge 70 then aftage7=7;else aftage7=.;

        *orders ages by 1=<50, 2=>=50;
        if aftage ne . and aftage lt 50 then aftage2_50=1;
        else if aftage ge 50 then aftage2_50=0;else aftage2_50=.;

        *orders ages by 1=<45, 0=>=45;
        if aftage ne . and aftage lt 45 then aftage2_45=1;
        else if aftage ge 45 then aftage2_45=0;else aftage2_45=.;

        *orders ages by biological importance 1=<15, 2=15-24, 3=25-49, 4=>49;
        if aftage ne . and aftage lt 15 then aftage_bio=1;
        else if aftage ge 15 and aftage le 24 then aftage_bio=2;
        else if aftage ge 25 and aftage le 49 then aftage_bio=3;
        else if aftage ge 50 then aftage_bio=4;else aftage_bio=.;

        *orders ages by quantiles 1=<37, 2=37-49, 3=50-60, 4=>60;
```

```
    if aftage ne . and aftage lt 37 then aftage_quartile=1;
    else if aftage ge 37 and aftage le 49 then aftage_quartile=2;
    else if aftage ge 50 and aftage le 60 then aftage_quartile=3;
    else if aftage gt 60 then aftage_quartile=4;else aftage_quartile=.;

    *orders ages by turtiles 1=<43, 2=43-56, 3=>56;
    if aftage ne . and aftage lt 43 then aftage_turtile=1;
    else if aftage ge 43 and aftage le 56 then aftage_turtile=2;
    else if aftage gt 56 then aftage_turtile=3;else aftage_turtile=.;

*creates a numeric version of year cases was counted;
befyear_n=befyear+0;aftyear_n=aftyear+0;

*orders year moved to US into decades;
if befusyear ne ' ' and aftusyear ne ' ' and befusyear le 1989 and aftusyear le 1989
    then usyear_decade=1;
else if befusyear ge 1990 and aftusyear ge 1990 and befusyear le 1999 and aftusyear le 1999
    then usyear_decade=2;
else if befusyear ge 2000 and aftusyear ge 2000 and befusyear le 2009 and aftusyear le 2009
    then usyear_decade=3;
else if befusyear ge 2010 and aftusyear ge 2010 then usyear_decade=4;
else if befusyear = ' ' and aftusyear le 1989 and aftusyear ne ' ' then usyear_decade=1;
else if befusyear = ' ' and aftusyear le 1999 and aftusyear ne ' ' then usyear_decade=2;
else if befusyear = ' ' and aftusyear le 2009 and aftusyear ne ' ' then usyear_decade=3;
else usyear_decade=.;

*creates new region variable out of nation;
if befnation='BIH' or befnation='POL' or befnation='RUS' or befnation='UKR' or aftnation='BIH'
or aftnation='POL' or aftnation='RUS' or aftnation='UKR'then region=1;
else if befnation='CHN' or befnation='HKG' or befnation='KOR' or befnation='PRK' or
befnation='TWN' or aftnation='CHN' or aftnation='HKG' or aftnation='KOR' or aftnation='PRK'
or aftnation='TWN' then region=2;
else if befnation='IND' or befnation='KHM' or befnation='LAO' or befnation='PHL' or
befnation='THA' or befnation='VNM' or aftnation='IND' or aftnation='KHM' or aftnation='LAO'
or aftnation='PHL' or aftnation='THA' or aftnation='VNM' then region=3;
else if befnation='ECU' or befnation='GTM' or befnation='HND' or befnation='HTI' or
befnation='TTO' or befnation='SLV' or aftnation='ECU' or aftnation='GTM' or aftnation='HND'
or aftnation='HTI' or aftnation='TTO' or aftnation='SLV' then region=4;
else if befnation='ETH' or befnation='SDN' or befnation='SOM' or befnation='ZWE' or
aftnation='ETH' or aftnation='SDN' or aftnation='SOM' or aftnation='ZWE' then region=5;
else if befnation='MEX' or aftnation='MEX' then region=6;else region=.;

*create region5 variable from region that groups european and asian as other;
if region=3 then region5=1;else if region=4 then region5=2;
else if region=5 then region5=3; else if region=6 then region5=4;
else if region=1 or region=2 then region5=5;else region5=.;

    *adds 0=US born for region5 variable;
    region5_u=region5;if aftorigin='USBORN' then region5_u=0;

    *creates variable for Mexican vs. not Mexican;
```

```
        if region5_u=4 then mexican=1; else if region5_u ne 4 then mexican=0;

        *create variable that groups SE Asian and LA/Car into Other;*because LogReg shows
        quasi-complete separation of data points with Other's group;
        if region5_u in (1, 2, 5) then region3_u=3;else if region5_u=3 then region3_u=1;
        else if region5_u=4 then region3_u=2;    else region3_u=region5_u;

        *region3 above without US born;
        if region5 in (1, 2, 5) then region3=3;else if region5=3 then region3=1;
        else if region5=4 then region3=2;

        *creates numeric foreign born variable;
        if aftorigin='FBORN' then fborn=1;
        else if aftorigin='USBORN' then fborn=0;else fborn=.;

*creates 4 new race categories with other;
if aftracehisp2='WHITE' then race4=1;else if aftracehisp2='BLACK' then race4=2;
else if aftracehisp2='HISP' then race4=3;else if aftracehisp2='AMIND' or aftracehisp2='ASIAN'
or aftracehisp2='MULT' then race4=4;else race4=.;

        *creates race categories comparing black and hispanic together against all others;
        if aftracehisp2='HISP' or aftracehisp2='BLACK' or (befracehisp2='MULT' and
        aftracehisp='BLACK') then race2=1;else if aftracehisp2 ne ' ' then race2=0;else race2=.;

        *creates dummy variables for race as black or hispanic with white as referent;
        if aftracehisp2='BLACK' or (befracehisp2='MULT' and aftracehisp='BLACK') then
        black=1;else black=0;
        if aftracehisp2='HISP' then hisp=1;else hisp=0;

*creates numeric variable for sex;
if befsex='F' then sex=1;else if befsex='M' then sex=0;

*creates binary state variable to compair case coming from a high incidence state or not;
if befstate='CA' or befstate='TX' or befstate='NO' or befstate='NY' or befstate='FL' then
st_highinc=1;else if befstate ne ' ' then st_highinc=0;else st_highinc=.;

*creates binary state variable for just texas and california vs. all others;
if befstate='CA' or befstate='TX' then st_CaTx=1;else if befstate ne ' ' then st_CaTx=0;else
st_CaTx=.;

*creates individual high-incidence state variables;
if befstate='CA' then CA=1;else CA=0;
if befstate='TX' then TX=1;else TX=0;
if befstate='NO' or befstate='NY' then NY=1;else NY=0;
if befstate='FL' then FL=1;else FL=0;

*creates new variables for clinical and treatment factors;
*change from 1st to 2nd episode;
        *creates new variable for HIV and HIV change;
        *fixes hivstat;
        if afthivstat='NEG' and (befhivstat=' ' or befhivstat='NOTOFFRD' or
        befhivstat='REFUSED' or befhivstat='TDUNK' or befhivstat='UNK') then
        befhivstat='NEG';
        if befhivstat='POS' and afthivstat=' ' then afthivstat='POS';

        *creates changes variable for HIV;
        if befhivstat='NEG' and afthivstat='NEG' then HIV=0;
        else if befhivstat='POS' and afthivstat='POS' then HIV=1;
        else if befhivstat='NEG' and afthivstat='POS' then HIV=2;
        else if befhivstat='NOTOFFRD' or befhivstat='REFUSED' or befhivstat='UNK' or
        afthivstat='NOTOFFRD' or afthivstat='REFUSED'  or afthivstat='UNK' then HIV=3;
        else if afthivstat='POS' then HIV=4;       else HIV=.;

        *creates variable for HIV positive;
        if HIV in (1,2,4) then HIVpos=1;else if HIV=0 then HIVpos=0;
        else if HIV in (3) then HIVpos=2;else if HIV=. then HIVpos=.;

        *creates variable for HIV positive w/missing values;
        if HIV in (1,2,4) then HIVpos_m=1;else if HIV=0 then HIVpos_m=0;else HIVpos_m=.;

        *creates variable for HIV positive;
        if HIV in (1,2,4) then HIVpos_nm=1;else if HIV=0 then HIVpos_nm=0;
        else if HIV in (3,.) then HIVpos_nm=2;

        *creates variable for HIV positive vs. other;
        if HIV in (1,2,4) then HIVpos_f=1;else HIVpos_f=0;

*creates variable for change in number firstline drugs administered;
if (befinitdrg='IR' and aftinitdrg='IR') or (befinitdrg='IRZ' and aftinitdrg='IRZ') or
(befinitdrg='IRZE' and aftinitdrg='IRZE') or (befinitdrg='OTHMULT' and
itinitdrg='OTHMULT') then initdrg_chng=0;
else if (befinitdrg='IR' and aftinitdrg='IRZ') or (befinitdrg='IRZ' and aftinitdrg='IRZE') or
(befinitdrg='IR' and aftinitdrg='IRZE') or (befinitdrg='NO DRUGS' and aftinitdrg='IRZ') or
(befinitdrg='NO DRUGS' and aftinitdrg='IRZE') or (befinitdrg='NO DRUGS' and
aftinitdrg='OTHMULT') then initdrg_chng=1;
else if (befinitdrg='IRZE' and aftinitdrg='IRZ') or (befinitdrg='IRZE' and aftinitdrg='NO
DRUGS') then initdrg_chng=2;
else if befinitdrg='OTHMULT' or aftinitdrg='OTHMULT' or befinitdrg='UNK' or
aftinitdrg='UNK' then initdrg_chng=3;else initdrg_chng=.;

*creates binary for no initiation drug;
if befinitdrg='NO DRUGS' then nodrug=1;else if befinitdrg = 'UNK' then nodrug=2;
else if befinitdrg ne ' ' then nodrug=0;else nodrug=.;

*creates variable for change in firstline drug resistance;
if beffirstline='N' and aftfirstline='N' then firstline=0;
else if beffirstline='Y' and aftfirstline='Y' then firstline=1;
else if beffirstline='N' and aftfirstline='Y' then firstline=2;
else if beffirstline='Y' and aftfirstline='N' then firstline=3;
else if beffirstline=' ' or aftfirstline=' ' then firstline=4;else firstline=.;
```

```
*creates concordant/discordant firstline drug resistance;
if firstline in (0,1) then firstline2=0;else if firstline in (2,3) then firstline2=1;else firstline2=.;

        *creates numeric version of firstline drug resistance in 2nd episode;
        if aftfirstline='Y' then aftresist=1;else aftresist=0;

        *creates variable for change in mdr;
        if befmdr='N' and aftmdr='N' then mdr=0;else if befmdr='Y' and aftmdr='Y' then mdr=1;
        else if befmdr='N' and aftmdr='Y' then mdr=2;
        else if befmdr=' ' or aftmdr=' ' or aftmdr='UNK' then mdr=4;else mdr=.;

*importance in the first episode only;
*create clean variable for sputum conversion;
if befconvert='Y' then convert=1;else if befconvert='N' then convert=0;else convert=.;

*create variable for sputum conversion vs. other;
if befconvert='Y' then convert2=1;else convert2=0;

*create variable indicating treatment success;
if befconvert='Y' and befstopreas='COMPLETED' then tx_success=1;
else if befconvert ne ' ' and befstopreas ne ' ' then tx_success=0;else tx_success=.;

*create variable for disease site;
if befdis_site='BOTH' then dis_site=2;
else if befdis_site='PULM ONLY' then dis_site=0;
else if befdis_site='EXTRAPULM ONLY' then dis_site=1;

*create variable for any pulmonary disease site;
if befdis_site='BOTH' then site_anypulm=1;
else if befdis_site='PULM ONLY' then site_anypulm=1;
else if befdis_site='EXTRAPULM ONLY' then site_anypulm=0;

        *create variable for number of disease sites;
        if befdis_site='BOTH' then dis_site2=1;  else dis_site2=0;

*creates variable for smear positive;
if befspsmear='POS' or befmicrexam='POS' then smearpos=1;
else if befspsmear='NEG' or befmicrexam='NEG' then smearpos=0;else smearpos=2;

*creates variable for culture positive;
if befspcult='POS' or befcultothr='POS' then cultpos=1;
else if befspcult='NEG' or befcultothr='NEG' then cultpos=0;else cultpos=2;

*creates variable to indicate cavitary disease on an abnormal xray;
if befxraycav='Y' then cavitary=1;else cavitary=0;

*creates variable for 1st episodes in the 1990's, before the new treatment guidelines;
if befyear lt 2000 then prevguidnc=1;else if befyear ge 2000 then prevguidnc=0;

*creates variable death that show outcome of dealth during treatment;
if aftstopreas='DIED' then death=1;else if aftstopreas ne ' ' then death=0;  else death=.;
```

```
        *fixes death=no if treatment still ongoing;
        if death=. and aftrxdate ne . and aftstopther=. then death=0;

*creates variable that categorizes DOT weeks;
befdotweeks_n=befdotweeks+0;

        *categories are based off of 4.348 weeks per month;
        if befdotweeks_n=0 then dotmonths=.;
        if befinitdrg='NO DRUGS' or befdot='SAT' then dotmonths=0;
        if befdotweeks_n gt 0 and befdotweeks_n lt 8 then dotmonths=1;
        if befdotweeks_n ge 8 and befdotweeks_n le 21 then dotmonths=2;
        if befdotweeks_n gt 21 and befdotweeks_n lt 26 then dotmonths=3;
        if befdotweeks_n ge 26 and befdotweeks_n lt 39 then dotmonths=4;
        if befdotweeks_n ge 39 then dotmonths=5;

        *creates variable that categorizes DOT months into importance in literature;
        if befdotweeks_n ne . and befdotweeks_n le 8 then dotmonths_lit=1;
        else if befdotweeks_n gt 8 and befdotweeks_n le 26 then dotmonths_lit=2;
        else if befdotweeks_n gt 26 then dotmonths_lit=3;
        else if befinitdrg='NO DRUGS' or befdot='SAT' then dotmonths_lit=0;
        else dotmonths_lit=.;

        *creates variable for any DOT to not DOT;
        if befdot='DOT' or befdot='BOTH' then DOT_any=1;
        else if befdot='SAT' or befdot=' ' then DOT_any=0;
        else if befdot='UNK' then DOT_any=2;  else DOT_any=.;

        *creates variable for only DOT to other;
        if befdot='DOT' then DOT_only=1;else DOT_only=0;

        *creates dummy variables for linear trend of amount of DOT;
        if befdot='SAT' or befdot='UNK' or befdot=' ' then dot_lin=0;
        if befdot='BOTH' then dot_lin=1;if befdot='DOT' the dot_lin=2;

                if dot_lin=1 then dot_lin1=1;else dot_lin1=0;
                if dot_lin=2 then dot_lin2=1;else dot_lin2=0;

*creates variable for completed therapy;
if befstopreas='COMPLETED' then compther=1;
else if befstopreas ne ' ' then compther=0;else compther=.;

        *creates all inclusive variable for completed therapy;
        if befstopreas='COMPLETED' then compther2=1;else compther2=0;

        *creates interaction variable for DOT only as well as completed therapy;
        if DOT_only=1 and compther=1 then dotcomp=1;
        else if DOT_only=1 then dotcomp=0;else dotcomp=3;

*change treatment factors in 1st episode for those who never had therapy (no drugs);
if befinitdrg='NO DRUGS' then do;befdot=' ';dot_any=0;dot_only=0;compther=0;end;
```

```
*works with time variables;
        *creates treatment duration from befstopther-befrxdate in months;
        txdur=round((befstopther-befrxdate)/30.436);

        *creates categories for treatment durration time;
        if txdur ne . and txdur lt 2 then txdur_cat=1;
        else if txdur ge 2 and txdur le 6 then txdur_cat=2;
        else if txdur gt 6 then txdur_cat=3;
        else if befinitdrg='NO DRUGS' then txdur_cat=0;else txdur_cat=.;

        *creates 2nd grouping of treatment duration;
        if txdur ne . and txdur lt 2 then txdur_cat2=1;
        else if befinitdrg='NO DRUGS' then txdur_cat2=0;
        else if txdur ge 2 and txdur le 6 then txdur_cat2=2;
        else if txdur gt 6 and txdur le 9 then txdur_cat2=3;
        else if txdur gt 9 and txdur le 12 then txdur_cat2=4;
        else if txdur gt 12 then txdur_cat2=5;else txdur_cat2=.;

        *creates dichotomous variable for txdur based on median difference in reinf vs. react;
        *will include missing as 0 - assuming no treatment(5 were no drug, 1 was unknown, 2
        were started but no data after that);
        if txdur le 8 then txdur9=0;else if txdur ge 9 then txdur9=1;

        *break at 6;if txdur le 5 then txdur6=0;else if txdur ge 6 then txdur6=1;

        *breat at 7;if txdur le 6 then txdur7=0;else if txdur ge 7 then txdur7=1;

        *creates the check variable for report dates in months: aftrxdate-befstopther;
        time_btwn = round ((aftrxdate-befstopther)/30.436);

*fixes problems found from creation of time_btwn first and 2nd episode;
*see above for fixes that involve newly created variables;
*fix where 11 months are listed but just because calculation rounded down;
if befstcaseno=' XXXXXXXXXXXX'or befstcaseno=' XXXXXXXXXXXX'then time_btwn=12;

*fix where . months are listed but just because therapy start date not listed;
if aftstcaseno=' XXXXXXXXXXXX'then time_btwn = round ((aftisusdate-befstopther)/30.436);

*fix where . months are listed but just because therapy never initiated;
if aftstcaseno=' XXXXXXXXXXXX'then time_btwn = round ((aftisusdate-befstopther)/30.436);

*fix where . months are listed but just because treatment unknown;
if aftstcaseno=' XXXXXXXXXXXX'then time_btwn = round ((aftisusdate-befstopther)/30.436);

        *fix where . months are listed but just because therapy never initiated;
        if befstcaseno=' XXXXXXXXXXXX'or befstcaseno=' XXXXXXXXXXXX'or
        befstcaseno=' XXXXXXXXXXXX'or befstcaseno=' XXXXXXXXXXXX'or
        befstcaseno=' XXXXXXXXXXXX'or befstcaseno=' XXXXXXXXXXXX'or
        befstcaseno=' XXXXXXXXXXXX'then time_btwn = round ((aftrxdate-
        befisusdate)/30.436);
        *fix where 9 months are listed but because error in therapy stop date;
*cneg date 8/20/08 and followup susceptibility testing done 11/4/09 - at least 12 months btwn so
use drug susceptibility testing date;
if befstcaseno=' XXXXXXXXXXXX'then time_btwn = round ((aftrxdate-befisusdate)/30.436);

*fix where . months are listed but just because therapy never initiated;
*no susceptibility testing but the closest date to use it the date RVCT submitted 9/7/07;
if befstcaseno=' XXXXXXXXXXXX'then time_btwn = round ((aftrxdate-17416)/30.436);

*fix where . months are listed but just because no treatment information was recorded in the
second episode. case was counted in 2010, treatment completed date (only info available in 2nd
episode) was 5/10/10, and genotyping results on 5/20/10) therefore use 1/1/2010 for default
second episode start date;
if befstcaseno=' XXXXXXXXXXXX'then time_btwn = round ((18263-befstopther)/30.436);

        *will have to remove from dataset because time between cases isn't truely 12 months;
        if befstcaseno=' XXXXXXXXXXXX'or befstcaseno=' XXXXXXXXXXXX'or
        befstcaseno=' XXXXXXXXXXXX'or befstcaseno=' XXXXXXXXXXXX'then delete;

        *create quantile count for time between cases;
        if time_btwn ne . and time_btwn lt 17 then time_btwn_quart=1;
        else if time_btwn ge 17 and time_btwn le 25 then time_btwn_quart=2;
        else if time_btwn ge 26 and time_btwn le 59 then time_btwn_quart=3;
        else if time_btwn gt 59 then time_btwn_quart=4;else time_btwn_quart=.;

        *split above at 2 years;
        if time_btwn ne . and time_btwn lt 24 then time_btwn24=0;
        else if time_btwn ge 24 then time_btwn24=1;else time_btwn24=.;

        *split above at 5 years;
        if time_btwn ne . and time_btwn lt 60 then time_btwn60=0;
        else if time_btwn ge 60 then time_btwn60=1;else time_btwn60=.;

        *grouping by year;
        if time_btwn ne . and time_btwn lt 25 then time_btwn_yr=1;
        else if time_btwn ge 25 and time_btwn le 36 then time_btwn_yr=2;
        else if time_btwn ge 37 and time_btwn le 48 then time_btwn_yr=3;
        else if time_btwn ge 49 and time_btwn le 60 then time_btwn_yr=4;
        else if time_btwn gt 60 then time_btwn_yr=5;else time_btwn_yr=.;

        if time_btwn ne . and time_btwn lt 24 then time_btwn_yr2=1;
        else if time_btwn ge 24 and time_btwn lt 36 then time_btwn_yr2=2;
        else if time_btwn ge 36 and time_btwn lt 48 then time_btwn_yr2=3;
        else if time_btwn ge 48 and time_btwn lt 60 then time_btwn_yr2=4;
        else if time_btwn ge 60 then time_btwn_yr2=5;  else time_btwn_yr2=.;

        *alternative grouping;
        if time_btwn ne . and time_btwn lt 17 then time_btwn3=1;
        else if time_btwn ge 17 and time_btwn le 25 then time_btwn3=2;
        else if time_btwn ge 26 then time_btwn3=3;else time_btwn3=.;
```

```
*creates new variables for genotyping factors;
        *assign values for gentype and pcrtype;
        if aftgentype=' ' then gentype='Missing'; else gentype=aftgentype; pcrtype=aftpcrtype;
        if aftpcrtype=' ' then pcrtype='Missing';

        *add in values for counts for gentypes (from analysis page);
        if gentype='Missing' then gen_count=0;
        else if gentype='G00014' or gentype='G00020' or gentype='G07222' or
        gentype='G15683' or gentype='G00870' then gen_count=2;
        else if gentype='G00018' then gen_count=3;else gen_count=1;

        *adds lineage data to dataset (from Maryam 12/20/13);
        if AFTSpoligo = 1017 then AFTSpoligotype = 717776777760771;  /*Euro Amer*/
        if AFTSpoligo = 1137 then AFTSpoligotype = 377777774020731;  /*Euro Amer*/
        if AFTSpoligo =    2 then AFTSpoligotype = 777777777760771;  /*Euro Amer*/
        if AFTSpoligo =   27 then AFTSpoligotype = 701776777760601;  /*Euro Amer*/
        if AFTSpoligo =   29 then AFTSpoligotype = 700076777760771;  /*Euro Amer*/
        if AFTSpoligo =   30 then AFTSpoligotype = 700036777760731;  /*Euro Amer*/
        if AFTSpoligo =  300 then AFTSpoligotype = 777756777760601;  /*Euro Amer*/
        if AFTSpoligo =   34 then AFTSpoligotype = 000000000003771;  /*East Asian*/
        if AFTSpoligo =  351 then AFTSpoligotype = 777777777360771;  /*Euro Amer*/
        if AFTSpoligo =  680 then AFTSpoligotype = 776000003760771;  /*Euro Amer*/
        if AFTGENType = "G16696" then Lineage = "EuroAmerican (L4)";
        else if AFTGENType in ("G03731", "G03705")  then Lineage = "Bovis";
        else if AFTPCRType in ("PCR00109", "PCR01235") then Lineage = "Bovis";
        else if AFTPCRType in  ("PCR00041", "PCR00388", "PCR00879", "PCR02102",
        "PCR14080", "PCR18115") then Lineage = "IndoOceanic (L1)";
        else if AFTPCRType in  ("PCR00001", "PCR00002", "PCR00036", "PCR00091",
        "PCR00093", "PCR00317", "PCR00803", "PCR00904", "PCR00911",
        "PCR01201","PCR01570", "PCR01571", "PCR01820", "PCR03382", "PCR03456",
        "PCR05894", "PCR08971", "PCR09211", "PCR12363")
                then Lineage="East Asian (L2)";
        else if AFTPCRType in ("PCR00044", "PCR02534", "PCR05792", "PCR08071")
                then Lineage = "East African Indian (L3)";
        else if AFTPCRType in ("PCR00015", "PCR00016", "PCR00017", "PCR00021",
        "PCR00025",  "PCR00039", "PCR00050", "PCR00051", "PCR00062", "PCR00067",
        "PCR00078", "PCR00079", "PCR00172", "PCR00225", "PCR00237",    "PCR00239",
        "PCR00254", "PCR00497", "PCR00556", "PCR00578",   "PCR00617", "PCR00645",
        "PCR00684", "PCR00687", "PCR00719",  "PCR00730", "PCR00743", "PCR00756",
        "PCR00769", "PCR00778",  "PCR00795", "PCR00818", "PCR00874", "PCR00900",
        "PCR00927",  "PCR01024", "PCR01318", "PCR01328", "PCR01332", "PCR01362",
        "PCR01371", "PCR01375", "PCR01379", "PCR01381", "PCR01385",  "PCR01419",
        "PCR01421", "PCR01474", "PCR01556", "PCR01637",   "PCR01669", "PCR01959",
        "PCR02355", "PCR02492", "PCR02587",  "PCR02621", "PCR02753", "PCR03269",
        "PCR03542", "PCR03994",  "PCR04093", "PCR04189", "PCR04200", "PCR04214",
        "PCR04438",    "PCR04626", "PCR04649", "PCR05412", "PCR05708", "PCR05971",
        "PCR06302", "PCR06493", "PCR06594", "PCR06800", "PCR07193",  "PCR07462",
        "PCR07695", "PCR07816", "PCR08159", "PCR08660",   "PCR08842", "PCR09084",
        "PCR09197", "PCR09610", "PCR09864",  "PCR10018", "PCR16176", "PCR17636",
        "PCR18395")
```

```
                then Lineage = "EuroAmerican (L4)";
        else if AFTSpoligotype = 3731 then Lineage = "East Asian (L2)";
        else if AFTSpoligotype = 3771 then Lineage = "East Asian (L2)";
        else if AFTSpoligotype = 777777777760771 then Lineage = "EuroAmerican (L4)";
        else if AFTSpoligotype = 000000004020771 then Lineage = "EuroAmerican (L4)";
        else if AFTSpoligotype = 717776777760771 then Lineage = "EuroAmerican (L4)";
        else if AFTSpoligotype = 377777774020731 then Lineage = "EuroAmerican (L4)";
        else if AFTSpoligotype = 701776777760601 then Lineage = "EuroAmerican (L4)";
        else if AFTSpoligotype = 700076777760771 then Lineage = "EuroAmerican (L4)";
        else if AFTSpoligotype = 700036777760731 then Lineage = "EuroAmerican (L4)";
        else if AFTSpoligotype = 777756777760601 then Lineage = "EuroAmerican (L4)";
        else if AFTSpoligotype = 777777777360771 then Lineage = "EuroAmerican (L4)";
        else if AFTSpoligotype = 776000003760771 then Lineage = "EuroAmerican (L4)";

*creates numeric versions of lineage for model;
if lineage='EuroAmerican (L4)' then lineage_n=1;if lineage='East Asian (L2)' then lineage_n=2;
if lineage='East African Indi' then lineage_n=3;if lineage='IndoOceanic (L1)' then lineage_n=4;
if lineage='Bovis' then lineage_n=5;if lineage=' ' then lineage_n=6;

*creates numeric versions of lineage for model;
if lineage='EuroAmerican (L4)' then lineage_n2=1;
if lineage='East Asian (L2)' then lineage_n2=2;if lineage='East African Indi' then lineage_n2=3;
if lineage='IndoOceanic (L1)' then lineage_n2=5;if lineage='Bovis' then lineage_n2=4;
if lineage=' ' then lineage_n2=5;

*creates numeric versions of lineage for model;
if lineage='EuroAmerican (L4)' then lineage_n3=1;
if lineage='East Asian (L2)' then lineage_n3=2;if lineage='East African Indi' then lineage_n3=3;
if lineage='IndoOceanic (L1)' then lineage_n3=3;if lineage='Bovis' then lineage_n3=3;
if lineage=' ' then lineage_n3=3;

        *create indicator variables for lineage3;
        if lineage_n2=1 then lin1=1;else lin1=0; if lineage_n2=2 then lin2=1;else lin2=0;
        if lineage_n2 in (3,4,5) then lin3=1;else lin3=0;

        *create variable indicating east asian lineage;
        if lineage='East Asian (L2)' then eastasian=1;else eastasian=0;
run;

*create working dataset that removes obs with no geno data - based on all removed from not
matching;
data working_final;
        set matches.merged_AddLab;if befmatch=. and aftmatch=. then delete;
        if befmatch=8 and aftmatch=8 then delete;

        *this is a 1/2/14 addition;
        if recurr ne 1 then delete;
run;
```

Data Analysis

```
*************************************************************************************;
* This program analyzes the data cleaned from the program                          ;
* PREVTB_pairs_20130912                                                            ;
* Uses working dataset working_final                                               ;
* Written by: Julia Interrante                                                     ;
* Created on: 8/25/2013                                                            ;
*************************************************************************************;

*************************************************************************;
*        UNIVARIATE AND BIVARIATE ANALYSIS                               ;
*************************************************************************;
*************************************************************************;
*                         basic demographic analysis by outcome type             ;
*************************************************************************;
*recurrent;
proc freq data=working_final;
        tables aftage7 aftsex race4 race2 aftorigin aftage2_50 aftage2_45 aftage_bio
        aftage_quartile aftage_turtile befstate st_highinc st_CaTx yrsin_us5_u /list missing;
run;
        *get median and IQR;
        proc univariate data=working_final;var aftage;run;
        *foreign born only;
        proc freq data=working_final;
                where aftorigin='FBORN';
                tables region5 region3_u usyear_decade yrsin_us5 yrsin_us13_u /list;run;
        *get median and IQR;
        proc univariate data=working_final;
                where aftorigin='FBORN' and yrsin_us5 ne .;var aftyrsin_us;run;

*reactivation;
proc freq data=working_final;
        where react=1;
        tables aftage7 aftsex race4 race2 aftorigin aftage2_50 aftage2_45 aftage_bio
        aftage_quartile aftage_turtile befstate st_highinc st_CaTx yrsin_us5_u /list missing;
run;
        *get median and IQR;
        proc univariate data=working_final;where react=1;var aftage;run;
        *foreign born only;
        proc freq data=working_final;
                where react=1 and aftorigin='FBORN';
                tables region5 region3_u usyear_decade yrsin_us5 yrsin_us13_u /list;run;
        *get median and IQR;
        proc univariate data=working_final;
                where react=1 and aftorigin='FBORN';var aftyrsin_us;run;

*reinfection;
proc freq data=working_final;
        where reinf=1;
```

```
	tables aftage7 aftsex race4 race2 aftorigin aftage2_50 aftage2_45 aftage_bio
	aftage_quartile aftage_turtile befstate st_highinc st_CaTx yrsin_us5_u /list missing;
run;
	*get median and IQR;
	proc univariate data=working_final;where reinf=1;var aftage;	run;
	*foreign born only;
	proc freq data=working_final;
		where reinf=1 and aftorigin='FBORN';
		tables region5 region3_u usyear_decade yrsin_us5 yrsin_us13_u /list;run;
	*get median and IQR;
	proc univariate data=working_final;
		where reinf=1 and aftorigin='FBORN';var aftyrsin_us;run;

*statistical tests;
proc freq data=working_final;
	tables reinf*race2 reinf*sex reinf*aftorigin reinf*aftage2_50 reinf*aftage2_45
	reinf*aftage_bio reinf*aftage_quartile reinf*aftage_turtile reinf*aftage7 reinf*race4
	reinf*region5 reinf*usyear_decade reinf*yrsin_us5 reinf*st_highinc reinf*st_CaTx
	reinf*region5_u reinf*mexican reinf*yrsin_us5_u reinf*yrsin_us13_u reinf*region3_u
	reinf*region3 /expected chisq fisher cmh;
run;

	proc freq data=working_final;
		where fborn=1; tables reinf*region3/expected chisq fisher;
	run;


*****************************************************************;
*                              social risk factors analysis by outcome type                    ;
*****************************************************************;
*removed highrisk_occu and employed because variables don't tell us anything useful;
*recurrent;
proc freq data=working_final;
	tables ever_homeless ever_corr ever_longterm ever_subabuse ever_alc ever_idu
	ever_nonidu unemployed/list missing nocum;
run;

*reactivation;
proc freq data=working_final;
	where react=1;
	tables ever_homeless ever_corr ever_longterm ever_subabuse ever_alc ever_idu
	ever_nonidu unemployed/list missing nocum;
run;

*reinfection;
proc freq data=working_final;
	where reinf=1;
	tables ever_homeless ever_corr ever_longterm ever_subabuse ever_alc ever_idu
	ever_nonidu unemployed/list missing nocum;
run;
```

```
*statistical tests;
proc freq data=working_final;
        tables reinf*ever_subabuse reinf*ever_riskfac reinf*ever_corr reinf*ever_longterm
        reinf*ever_homeless reinf*ever_alc reinf*ever_idu reinf*ever_nonidu reinf*unemployed
        / expected fisher chisq cmh;
run;

*******************************************************************;
*                        clinical features analysis by outcome type                        ;
*******************************************************************;
*recurrent;
proc freq data=working_final;
        tables hiv convert hivpos*region*beforigin hivpos hivpos_f  dis_site site_anypulm
        dis_site2 smearpos cavitary convert2 tx_success /list missing nocum;
run;

*reactivation;
proc freq data=working_final;
        where react=1;
        tables hiv convert hivpos*region*beforigin hivpos hivpos_f  dis_site site_anypulm
        dis_site2 smearpos cavitary convert2 tx_success /list missing nocum;
run;

*reinfection;
proc freq data=working_final;
        where reinf=1;
        tables hiv convert hivpos*region*beforigin hivpos hivpos_f  dis_site site_anypulm
        dis_site2 smearpos cavitary convert2 tx_success /list missing nocum;
run;

*statistical tests;
proc freq data=working_final;
        tables reinf*hiv reinf*convert reinf*hivpos reinf*hivpos_m reinf*hivpos_nm
        reinf*hivpos_f reinf*dis_site reinf*site_anypulm reinf*dis_site2 reinf*smearpos
        reinf*cavitary reinf*convert2 reinf*tx_success /expected chisq fisher;
run;
        *wilcoxon signed rank test;
        proc freq data=test;tables reinf*react/list missing;run;
                *age;
                PROC NPAR1WAY data=test wilcoxon;Class reinf;Var aftage; Run;
                *years in us;
                PROC NPAR1WAY data=test wilcoxon;Class reinf;Var aftyrsin_us;Run;


*******************************************************************;
*                        treatment factors analysis by outcome type                                ;
*******************************************************************;
*recurrent;
proc freq data=working_final;
        tables befinitdrg initdrg_chng firstline firstline2 mdr befdot befprovtype befstopreas
        death DOT_any DOT_only compther txdur_cat txdur9/list missing nocum;
```

```
run;
        *for DOT receipients;
        proc freq data=working_final;
                where dot_any=1;tables befdotsite dotmonths dotmonths_lit /list missing;run;
        *for DOT only receipients;
        proc freq data=working_final;
                where dot_only=1;tables dotcomp/list missing nocum;run;
        *get median and IQR;
        proc univariate data=working_final;where dot_any=1;var befdotweeks_n;run;
        *get median and IQR;
        proc univariate data=working_final;var txdur;run;

*reactivation;
proc freq data=working_final;
        where react=1;
        tables befinitdrg initdrg_chng firstline firstline2 mdr befdot befprovtype befstopreas
        death DOT_any DOT_only compther txdur_cat txdur9/list missing nocum;
run;
        *for DOT receipients;
        proc freq data=working_final;
                where react=1 and dot_any=1 and dot_any=1;
                tables befdotsite dotmonths dotmonths_lit /list missing;run;
        *for DOT only receipients;
        proc freq data=working_final;
                where react=1 and dot_only=1;  tables dotcomp/list missing nocum;run;
        *get median and IQR;
        proc univariate data=working_final;where react=1 and dot_any=1;
                var befdotweeks_n;run;
        *get median and IQR;
        proc univariate data=working_final;where react=1;var txdur;run;

*reinfection;
proc freq data=working_final;
        where reinf=1;
        tables befinitdrg initdrg_chng firstline firstline2 mdr befdot befprovtype befstopreas
        death DOT_any DOT_only compther txdur_cat txdur9/list missing nocum;
run;
        *for DOT receipients;
        proc freq data=working_final;
                where reinf=1 and dot_any=1;
                tables befdotsite dotmonths dotmonths_lit /list missing nocum;    run;
        *for DOT only receipients;
        proc freq data=working_final;
                where reinf=1 and dot_only=1;  tables dotcomp/list missing nocum;run;
        *get median and IQR;
        proc univariate data=working_final;
                where reinf=1 and dot_any=1;var befdotweeks_n;run;
        *get median and IQR;
        proc univariate data=working_final;where reinf=1;var txdur;run;

*statistical tests;
```

```
proc freq data=working_final;
        tables reinf*befinitdrg reinf*firstline reinf*mdr reinf*DOT_any reinf*dotmonths_lit
        reinf*compther reinf*txdur_cat reinf*death reinf*befprovtype reinf*firstline2
        reinf*DOT_only reinf*prevguidnc reinf*nodrug reinf*initdrg_chng reinf*befdot
        reinf*befstopreas reinf*dotcomp reinf*befdotsite reinf*dotmonths reinf*txdur9
        reinf*aftresist reinf*txdur_cat2 reinf*txdur6 reinf*dot_lin reinf*txdur7 /expected chisq
        fisher cmh;
run;
        proc freq data=working_final;
                where befdot ne ' ' and befdot ne 'UNK'; table reinf*befdot/expected chisq fisher;
        run;

        *wilcoxon signed rank test;
        data test;set working_final;run;
        *dot weeks;
        PROC NPAR1WAY data=test wilcoxon;Class reinf;Var befdotweeks_n;Run;
        *treatment duration;
        PROC NPAR1WAY data=test wilcoxon;Class reinf;Var txdur;  Run;


****************************************************************;
*                          genotyping factors analysis by outcome type                    ;
****************************************************************;
*recurrent;
proc freq data=working_final;
        tables aftgentype aftpcrtype time_btwn_quart time_btwn24 time_btwn60 time_btwn3
        time_btwn_yr gen_count pcr_countg lineage /list missing nocum;
run;
        *get median and IQR;
        proc univariate data=working_final;var time_btwn;run;

*reactivation;
proc freq data=working_final;
        where react=1;
        tables aftgentype aftpcrtype time_btwn_quart time_btwn24 time_btwn60 time_btwn3
        time_btwn_yr gen_count pcr_countg lineage /list missing nocum;
run;
        *get median and IQR;
        proc univariate data=working_final;where react=1;var time_btwn;run;

*reinfection;
proc freq data=working_final;
        where reinf=1;
        tables aftgentype aftpcrtype time_btwn_quart time_btwn24 time_btwn60 time_btwn3
        time_btwn_yr gen_count pcr_countg lineage /list missing nocum;
run;
        *get median and IQR;
        proc univariate data=working_final;where reinf=1;var time_btwn;run;

*statistical tests;
PROC NPAR1WAY data=working_final wilcoxon;Class reinf;  Var time_btwn; Run;
```

```
proc freq data=working_final;
        tables reinf*time_btwn_quart reinf*time_btwn24 reinf*aftgentype reinf*aftpcrtype
        reinf*gen_count reinf*pcr_countg reinf*time_btwn60 reinf*time_btwn3
        reinf*time_btwn_yr reinf*lineage_n reinf*lineage_n2 reinf*eastasian reinf*lineage_n3
        /expected chisq fisher cmh;
run;
        proc freq data=working_final;
                where lineage_n ne .;tables reinf*lineage_n/expected chisq fisher cmh;run;
        *lineage indicator;
        proc freq data=working_final;
                tables reinf*lin1 reinf*lin2 reinf*lineage_n3/expected chisq fisher cmh;run;


*****************************************************************;
*                compute unadjusted ORs                          ;
*****************************************************************;
*for all factors in model;
data unformatted;set working_final;format _all_;run;
proc logistic data=unformatted descending;model reinf = aftage2_50;run;
proc logistic data=unformatted descending;model reinf = sex;run;
proc logistic data=unformatted descending;model reinf = race2;run;
proc logistic data=unformatted descending;model reinf = st_catx;run;
proc logistic data=unformatted descending;model reinf = fborn;run;
proc logistic data=unformatted descending;model reinf = yrsin_us13_u;run;
proc logistic data=unformatted descending;model reinf = yrs_us_mid;run;
proc logistic data=unformatted descending;model reinf = ever_nonidu;run;
proc logistic data=unformatted descending;model reinf = hivpos_f;run;
proc logistic data=unformatted descending;model reinf = dis_site2;run;
proc logistic data=unformatted descending;model reinf = dot_only;run;
proc logistic data=unformatted descending;model reinf = txdur9;run;
proc logistic data=unformatted descending;model reinf = tx_success;run;
proc logistic data=unformatted descending;model reinf = compther2;run;
proc logistic data=unformatted descending;model reinf = time_btwn60;run;
proc logistic data=unformatted descending;model reinf = mexican;run;
proc logistic data=unformatted descending;model reinf = dot_lin1 dot_lin2;run;

*check for linear trends in amount of DOT;
proc logistic data=unformatted descending;class dot_lin/desc;model reinf = dot_lin;run;


*********************************************************************;
*MODEL SELECTIONUSING A 0.1 CUT OFF FOR INDIVIDUAL SIGNIFICANCE;
* w/o TIME_BTWN and dis_site                                        ;
*********************************************************************;
*****************************************************************;
*                collinearity assessment                         ;
*****************************************************************;
*create unformated dataset;
data unformatted;set working_final;format _all_;run;
*run collinearity macro (CIs and VDPs);
```

```
%include '\\cdc\project\NCHHSTP_DTBE_SURV_DATA\Reinfection vs
Reactivation\Datasets\collin_2011.sas';
proc logistic data=unformatted descending covout outest=colin;
        model reinf = aftage2_50 sex race2 st_CaTx mexican yrsin_us13_u HIVpos_f DOT_only
        txdur9 aftage2_50*race2 aftage2_50*Mexican sex*st_CaTx sex*mexican
        sex*yrsin_us13_u sex*HIVpos_f race2*HIVpos_f race2*txdur9  st_CaTx*mexican
        st_CaTx*HIVpos_f st_CaTx*txdur9 mexican*HIVpos_f mexican*DOT_only
        mexican*txdur9  yrsin_us13_u*HIVpos_f yrsin_us13_u*txdur9  HIVpos_f*DOT_only
        HIVpos_f*txdur9  DOT_only*txdur9  /covb;
run;
%collin (covdsn=colin, output=covcheck);run;

*variables dropped from above (1 at a time) because collinearity index above 30 and VDP greater
than 0.5:
        st_CaTx*yrsin_us13_u race2*yrsin_us13_u aftage2_50*sex aftage2_50*st_CaTx
aftage2_50*DOT_only race2*st_CaTx aftage2_50*HIVpos_f mexican*yrsin_us13_u
aftage2_50*yrsin_us13_u aftage2_50*txdur9 sex*DOT_only sex*race2 race2*DOT_only
sex*txdur9 yrsin_us13_u*DOT_only st_CaTx*DOT_only ;

*removed because 0 in one category: race2*mexican;

**************************************************************;
*                interaction assessment                       ;
**************************************************************;

*p-value based backward elimination;
        *drop HIVpos_f*DOT_only: 0.9596; *drop mexican*DOT_only: 0.9499;*drop
        mexican*txdur9: 0.9314;*drop yrsin_us13_u*HIVpos_f: 0.9252;*drop
        mexican*HIVpos_f: 0.8782;*drop aftage2_50*mexican: 0.8398;*drop race2*txdur9:
        0.8374;*drop sex*mexican: 0.7699;*drop st_CaTx*txdur9: 0.6699;*drop sex*st_CaTx:
        0.4070;*drop yrsin_us13_u*txdur9: 0.3767;*drop st_CaTx*HIVpos_f: 0.3870;*drop
        aftage2_50*race2: 0.3541;*drop sex*HIVpos_f: 0.3416;*drop sex*yrsin_us13_u:
        0.2643;*drop race2*HIVpos_f: 0.1978;*drop DOT_only*txdur9: 0.0791;
proc logistic data=unformatted descending;
        model reinf = aftage2_50 sex race2 st_CaTx mexican yrsin_us13_u HIVpos_f DOT_only
        txdur9 st_CaTx*mexican HIVpos_f*txdur9 ;
run;

**************************************************************;
    •   individual variable change-in-estimate assessment     ;
**************************************************************;

*only variables that i could potentially drop that aren't part of interaction or political are: age,
race, yrsinus, dot;

*try using change in estimate backward elimination;
*gold standard model;
proc logistic data=unformatted descending;
        model reinf = aftage2_50 sex race2 st_CaTx mexican yrsin_us13_u HIVpos_f DOT_only
        txdur9 st_CaTx*mexican HIVpos_f*txdur9;
        contrast 'HIV=0 txdur=1' txdur9 1/est=exp;
        contrast 'HIV=1 txdur=0' HIVpos_f 1/est=exp;
        contrast 'HIV=txdur=1' HIVpos_f 1 txdur9 1 HIVpos_f*txdur9 1/est=exp;
        contrast 'st=0 mex=1' mexican 1/est=exp;
        contrast 'st=1 mex=0' st_CaTx 1/est=exp;
        contrast 'st=mex=1' mexican 1 st_CaTx 1 st_CaTx*mexican 1/est=exp;
run;
        *drop age: 0.6205;
        proc logistic data=unformatted descending;
                model reinf = sex race2 st_CaTx mexican yrsin_us13_u HIVpos_f DOT_only
                txdur9 st_CaTx*mexican HIVpos_f*txdur9;
                contrast 'HIV=0 txdur=1' txdur9 1/est=exp;
                contrast 'HIV=1 txdur=0' HIVpos_f 1/est=exp;
                contrast 'HIV=txdur=1' HIVpos_f 1 txdur9 1 HIVpos_f*txdur9 1/est=exp;
                contrast 'st=0 mex=1' mexican 1/est=exp;
                contrast 'st=1 mex=0' st_CaTx 1/est=exp;
                contrast 'st=mex=1' mexican 1 st_CaTx 1 st_CaTx*mexican 1/est=exp;    run;
        *drop race: 0.2957;
        proc logistic data=unformatted descending;
                model reinf = aftage2_50 sex st_CaTx mexican yrsin_us13_u HIVpos_f
                DOT_only txdur9 st_CaTx*mexican HIVpos_f*txdur9;
                contrast 'HIV=0 txdur=1' txdur9 1/est=exp;
                contrast 'HIV=1 txdur=0' HIVpos_f 1/est=exp;
                contrast 'HIV=txdur=1' HIVpos_f 1 txdur9 1 HIVpos_f*txdur9 1/est=exp;
                contrast 'st=0 mex=1' mexican 1/est=exp;
                contrast 'st=1 mex=0' st_CaTx 1/est=exp;
                contrast 'st=mex=1' mexican 1 st_CaTx 1 st_CaTx*mexican 1/est=exp;    run;
        *drop yrsin_us: 0.2638;
        proc logistic data=unformatted descending;
                model reinf = aftage2_50 sex race2 st_CaTx mexican HIVpos_f DOT_only
                txdur9   st_CaTx*mexican HIVpos_f*txdur9;
                contrast 'HIV=0 txdur=1' txdur9 1/est=exp;
                contrast 'HIV=1 txdur=0' HIVpos_f 1/est=exp;
                contrast 'HIV=txdur=1' HIVpos_f 1 txdur9 1 HIVpos_f*txdur9 1/est=exp;
                contrast 'st=0 mex=1' mexican 1/est=exp;
                contrast 'st=1 mex=0' st_CaTx 1/est=exp;
                contrast 'st=mex=1' mexican 1 st_CaTx 1 st_CaTx*mexican 1/est=exp; run;
        *drop age and race;
        proc logistic data=unformatted descending;
                model reinf = sex st_CaTx mexican yrsin_us13_u HIVpos_f DOT_only txdur9
                st_CaTx*mexican HIVpos_f*txdur9;
                contrast 'HIV=0 txdur=1' txdur9 1/est=exp;
                contrast 'HIV=1 txdur=0' HIVpos_f 1/est=exp;
                contrast 'HIV=txdur=1' HIVpos_f 1 txdur9 1 HIVpos_f*txdur9 1/est=exp;
                contrast 'st=0 mex=1' mexican 1/est=exp;
                contrast 'st=1 mex=0' st_CaTx 1/est=exp;
                contrast 'st=mex=1' mexican 1 st_CaTx 1 st_CaTx*mexican 1/est=exp;    run;
        *drop age and yrsin_us;
        proc logistic data=unformatted descending;
                model reinf = sex race2 st_CaTx mexican HIVpos_f DOT_only txdur9
                st_CaTx*mexican HIVpos_f*txdur9;
                contrast 'HIV=0 txdur=1' txdur9 1/est=exp;
                contrast 'HIV=1 txdur=0' HIVpos_f 1/est=exp;
                contrast 'HIV=txdur=1' HIVpos_f 1 txdur9 1 HIVpos_f*txdur9 1/est=exp;
                contrast 'st=0 mex=1' mexican 1/est=exp;
                contrast 'st=1 mex=0' st_CaTx 1/est=exp;
                contrast 'st=mex=1' mexican 1 st_CaTx 1 st_CaTx*mexican 1/est=exp;    run;
        *drop race and years;
        proc logistic data=unformatted descending;
                model reinf = aftage2_50 sex st_CaTx mexican HIVpos_f DOT_only txdur9
                st_CaTx*mexican HIVpos_f*txdur9;
                contrast 'HIV=0 txdur=1' txdur9 1/est=exp;
                contrast 'HIV=1 txdur=0' HIVpos_f 1/est=exp;
                contrast 'HIV=txdur=1' HIVpos_f 1 txdur9 1 HIVpos_f*txdur9 1/est=exp;
                contrast 'st=0 mex=1' mexican 1/est=exp;
                contrast 'st=1 mex=0' st_CaTx 1/est=exp;
                contrast 'st=mex=1' mexican 1 st_CaTx 1 st_CaTx*mexican 1/est=exp;    run;
        *drop age and race and years;
        proc logistic data=unformatted descending;
                model reinf = sex st_CaTx mexican HIVpos_f DOT_only txdur9
                st_CaTx*mexican HIVpos_f*txdur9;
                contrast 'HIV=0 txdur=1' txdur9 1/est=exp;
                contrast 'HIV=1 txdur=0' HIVpos_f 1/est=exp;
                contrast 'HIV=txdur=1' HIVpos_f 1 txdur9 1 HIVpos_f*txdur9 1/est=exp;
                contrast 'st=0 mex=1' mexican 1/est=exp;
                contrast 'st=1 mex=0' st_CaTx 1/est=exp;
                contrast 'st=mex=1' mexican 1 st_CaTx 1 st_CaTx*mexican 1/est=exp;    run;

*final model #1;
proc logistic data=unformatted descending;
        model reinf = sex race2 st_CaTx mexican yrsin_us13_u HIVpos_f DOT_only txdur9
        st_CaTx*mexican HIVpos_f*txdur9;
        contrast 'st=1 mex=0' st_CaTx 1/est=exp;
        contrast 'st=1 mex=1' st_CaTx 1 mexican 1 st_CaTx*mexican 1/est=exp;
        contrast 'st=0 mex=1' mexican 1/est=exp;
        contrast 'HIV=1 txdur=0' HIVpos_f 1/est=exp;
        contrast 'HIV=1 txdur=1' HIVpos_f 1 txdur9 1 HIVpos_f*txdur9 1/est=exp;
        contrast 'HIV=0 txdur=1' txdur9 1/est=exp;
run;

        *get crude ORs for interaction variables;
        proc logistic data=unformatted descending;
                model reinf = st_CaTx mexican st_CaTx*mexican;
                contrast 'st=1 mex=0' st_CaTx 1/est=exp;
                contrast 'st=1 mex=1' st_CaTx 1 mexican 1 st_CaTx*mexican 1/est=exp;
                contrast 'st=0 mex=1' mexican 1/est=exp;
        run;
        proc logistic data=unformatted descending;
                model reinf = HIVpos_f txdur9 HIVpos_f*txdur9;
                contrast 'HIV=1 txdur=0' HIVpos_f 1/est=exp;
                contrast 'HIV=1 txdur=1' HIVpos_f 1 txdur9 1 HIVpos_f*txdur9 1/est=exp;
                contrast 'HIV=0 txdur=1' txdur9 1/est=exp;
        run;
```

```
*final model #1 - using opposite of yrsin_us (yrs_us_mid);
proc logistic data=unformatted descending;
        model reinf = sex race2 st_CaTx mexican yrs_us_mid HIVpos_f DOT_only txdur9
        st_CaTx*mexican HIVpos_f*txdur9;
        contrast 'st=1 mex=0' st_CaTx 1/est=exp;
        contrast 'st=1 mex=1' st_CaTx 1 mexican 1 st_CaTx*mexican 1/est=exp;
        contrast 'st=0 mex=1' mexican 1/est=exp;
        contrast 'HIV=1 txdur=0' HIVpos_f 1/est=exp;
        contrast 'HIV=1 txdur=1' HIVpos_f 1 txdur9 1 HIVpos_f*txdur9 1/est=exp;
        contrast 'HIV=0 txdur=1' txdur9 1/est=exp;
run;

*****************************************************************;
*RERUN ALL USING THE REDUCED POPULATION (compther=1)              ;
*****************************************************************;

*create unformated dataset with reduced population;
data unformatted2;set working_final;where compther=1;format _all_;run;

*run collinearity macro (CIs and VDPs);
%include '\\cdc\project\NCHHSTP_DTBE_SURV_DATA\Reinfection vs
Reactivation\Datasets\collin_2011.sas';
*drop variables: race2*HIVpos_f mexican*DOT_only sex*yrsin_us13_u ;
proc logistic data=unformatted2 descending covout outest=colin;
        model reinf = aftage2_50 sex race2 st_CaTx mexican yrsin_us13_u HIVpos_f DOT_only
txdur9 aftage2_50*race2 aftage2_50*mexican  sex*st_CaTx sex*mexican sex*HIVpos_f
race2*txdur9  st_CaTx*mexican st_CaTx*HIVpos_f st_CaTx*txdur9 mexican*HIVpos_f
mexican*txdur9 yrsin_us13_u*HIVpos_f yrsin_us13_u*txdur9  HIVpos_f*DOT_only
HIVpos_f*txdur9  DOT_only*txdur9  /covb;
run;
%collin (covdsn=colin, output=covcheck);run;

*interaction assessment;
        *drop mexican*HIVpos_f 0.9380;*drop yrsin_us13_u*HIVpos_f 0.9158;*drop
        sex*st_CaTx 0.8546;*drop aftage2_50*mexican 0.7993;*drop race2*txdur9
        0.8001;*drop HIVpos_f*DOT_only 0.7428;*drop st_CaTx*HIVpos_f 0.7312;*drop
        st_CaTx*txdur9 0.6810;*drop mexican*txdur9 0.4850;*drop yrsin_us13_u*txdur9
        0.5144;*drop HIVpos_f*txdur9 0.4218;*drop aftage2_50*race2 0.1944;*drop
        sex*mexican 0.2526;*drop st_CaTx*mexican 0.2212;*drop sex*HIVpos_f 0.0943;*drop
        DOT_only*txdur9 0.1200;
proc logistic data=unformatted2 descending;
        model reinf = sex race2 st_CaTx mexican yrsin_us13_u HIVpos_f DOT_only txdur9;
run;

*get unadjusted values for the reduced model;
proc logistic data=unformatted2 descending;model reinf = sex ;run;
proc logistic data=unformatted2 descending;model reinf = race2 ;run;
proc logistic data=unformatted2 descending;model reinf = st_CaTx ;run;
proc logistic data=unformatted2 descending;model reinf = mexican ;run;
proc logistic data=unformatted2 descending;model reinf = yrsin_us13_u ;run;
proc logistic data=unformatted2 descending;model reinf = yrs_us_mid ;run;
proc logistic data=unformatted2 descending;model reinf = HIVpos_f ;run;
proc logistic data=unformatted2 descending;model reinf = DOT_only ;run;
proc logistic data=unformatted2 descending;model reinf = txdur9 ;run;

*get values for reduced model;
proc freq data=working_final;
        where compther=1;
        tables reinf*sex reinf*race2 reinf*st_CaTx reinf*mexican reinf*yrsin_us13_u
        reinf*yrs_us_mid reinf*HIVpos_f reinf*DOT_only reinf*txdur9;
run;

*run reduced model but including interaction terms from full population model;
proc logistic data=unformatted2 descending;
        model reinf = sex race2 st_CaTx mexican yrs_us_mid HIVpos_f DOT_only txdur9
        st_CaTx*mexican HIVpos_f*txdur9;
        contrast 'sex' sex 1/est=exp;
        contrast 'race2' race2 1/est=exp;
        contrast 'st=1 mex=0' st_CaTx 1/est=exp;
        contrast 'st=1 mex=1' st_CaTx 1 mexican 1 st_CaTx*mexican 1/est=exp;
        contrast 'st=0 mex=1' mexican 1/est=exp;
        contrast 'yrsin_us13_u' yrsin_us13_u 1/est=exp;
        contrast 'HIV=1 txdur=0' HIVpos_f 1/est=exp;
        contrast 'HIV=1 txdur=1' HIVpos_f 1 txdur9 1 HIVpos_f*txdur9 1/est=exp;
        contrast 'HIV=0 txdur=1' txdur9 1/est=exp;
        contrast 'DOT_only' DOT_only 1/est=exp;
run;
        *get crude ORs for interaction terms;
        proc logistic data=unformatted2 descending;
                model reinf = st_CaTx mexican st_CaTx*mexican;
                contrast 'st=1 mex=0' st_CaTx 1/est=exp;
                contrast 'st=1 mex=1' st_CaTx 1 mexican 1 st_CaTx*mexican 1/est=exp;
                contrast 'st=0 mex=1' mexican 1/est=exp;        run;
        proc logistic data=unformatted2 descending;
                model reinf = HIVpos_f txdur9 HIVpos_f*txdur9;
                contrast 'HIV=1 txdur=0' HIVpos_f 1/est=exp;
                contrast 'HIV=1 txdur=1' HIVpos_f 1 txdur9 1 HIVpos_f*txdur9 1/est=exp;
                contrast 'HIV=0 txdur=1' txdur9 1/est=exp;      run;

*run reduced model with race as a dummy variable;
proc logistic data=unformatted2 descending;
        model reinf = sex black hisp st_CaTx mexican yrs_us_mid HIVpos_f DOT_only txdur9;
run;
        *get unadjusted values for the reduced model;
        proc logistic data=unformatted2 descending;model reinf = black hisp;
                contrast 'black' black 1/est=exp; contrast 'hisp' hisp 1/est=exp;run;

*run reduced model with 4 high-incidence states;
proc logistic data=unformatted2 descending;
        model reinf = sex race2 st_highinc mexican yrs_us_mid HIVpos_f txdur9 DOT_only ;
run;
```

```
        *get unadjusted values for the reduced model;
        proc logistic data=unformatted2 descending;model reinf = st_highinc;run;
*run reduced model with only california;
*get updated multivariate estimates;
proc logistic data=unformatted2 descending;
        model reinf = sex race2 ca mexican yrs_us_mid HIVpos_f txdur9 DOT_only ;
run;
        *get unadjusted values for the reduced model;
        proc logistic data=unformatted2 descending;model reinf = ca;run;


****************************************************************************;
* reassess model with completed treatment and race as a dummy variable for                ;
*interaction terms                                                                          ;
******************************************************************************;
*run collinearity macro (CIs and VDPs);
%include '\\cdc\project\NCHHSTP_DTBE_SURV_DATA\Reinfection vs
Reactivation\Datasets\collin_2011.sas';
*drop variables: (black*mexican hisp*mexican)-won't run ;
*drop variables: black*hivpos_f hisp*hivpos_f st_CaTx*mexican black*dot_only hisp*dot_only
black*yrs_us_mid hisp*yrs_us_mid black*sex hisp*sex ;
proc logistic data=unformatted2 descending covout outest=colin;
        model reinf = sex black hisp st_CaTx mexican yrs_us_mid HIVpos_f DOT_only txdur9
black*st_catx black*txdur9 hisp*st_catx hisp*txdur9 HIVpos_f*txdur9   /covb;run;
%collin (covdsn=colin, output=covcheck);run;

*interaction assessment;
        *drop black*st_catx;*drop hisp*txdur9;*drop black*txdur9;*drop hisp*st_catx;*drop
        HIVpos_f*txdur9;
proc logistic data=unformatted2 descending ;
        model reinf = sex black hisp st_CaTx mexican yrs_us_mid HIVpos_f DOT_only txdur9 ;
run;


*********************************************************************;
*RERUN ALL USING THE REDUCED POPULATION (compther=0)               ;
*********************************************************************;
*create unformated dataset with reduced population;
data unformatted3;set working_final;where compther ne 1;format _all_;run;

*get unadjusted values for the reduced model;
proc logistic data=unformatted3 descending;model reinf = sex ;run;
proc logistic data=unformatted3 descending;model reinf = race2 ;run;
proc logistic data=unformatted3 descending;model reinf = st_CaTx ;run;
proc logistic data=unformatted3 descending;model reinf = mexican ;run;
proc logistic data=unformatted3 descending;model reinf = yrsin_us13_u ;run;
proc logistic data=unformatted3 descending;model reinf = HIVpos_f ;run;
proc logistic data=unformatted3 descending;model reinf = DOT_only ;run;
proc logistic data=unformatted3 descending;model reinf = txdur9 ;run;

*get values for reduced model;
```

```
proc freq data=working_final;
        where compther ne 1;
        tables reinf*sex reinf*race2 reinf*st_CaTx reinf*mexican reinf*yrsin_us13_u
        reinf*HIVpos_f reinf*DOT_only reinf*txdur9;
run;

*run reduced model but including interaction terms from full population model;
proc logistic data=unformatted3 descending;
        model reinf = sex race2 st_CaTx mexican yrsin_us13_u HIVpos_f DOT_only txdur9
        st_CaTx*mexican HIVpos_f*txdur9;
        contrast 'sex' sex 1/est=exp;
        contrast 'race2' race2 1/est=exp;
        contrast 'st=1 mex=0' st_CaTx 1/est=exp;
        contrast 'st=1 mex=1' st_CaTx 1 mexican 1 st_CaTx*mexican 1/est=exp;
        contrast 'st=0 mex=1' mexican 1/est=exp;
        contrast 'yrsin_us13_u' yrsin_us13_u 1/est=exp;
        contrast 'HIV=1 txdur=0' HIVpos_f 1/est=exp;
        contrast 'HIV=1 txdur=1' HIVpos_f 1 txdur9 1 HIVpos_f*txdur9 1/est=exp;
        contrast 'HIV=0 txdur=1' txdur9 1/est=exp;
        contrast 'DOT_only' DOT_only 1/est=exp;
run;
        *get crude ORs for interaction terms;
        proc logistic data=unformatted3 descending;
                model reinf = st_CaTx mexican st_CaTx*mexican;
                contrast 'st=1 mex=0' st_CaTx 1/est=exp;
                contrast 'st=1 mex=1' st_CaTx 1 mexican 1 st_CaTx*mexican 1/est=exp;
                contrast 'st=0 mex=1' mexican 1/est=exp;        run;
        proc logistic data=unformatted3 descending;
                model reinf = HIVpos_f txdur9 HIVpos_f*txdur9;
                contrast 'HIV=1 txdur=0' HIVpos_f 1/est=exp;
                contrast 'HIV=1 txdur=1' HIVpos_f 1 txdur9 1 HIVpos_f*txdur9 1/est=exp;
                contrast 'HIV=0 txdur=1' txdur9 1/est=exp;      run;

*run analysis for non Model 1 variable assessment;
*run collinearity macro (CIs and VDPs);
%include '\\cdc\project\NCHHSTP_DTBE_SURV_DATA\Reinfection vs
Reactivation\Datasets\collin_2011.sas';
*drop variables: sex*mexican mexican*HIVpos_f mexican*txdur9 yrsin_us13_u*HIVpos_f
yrsin_us13_u*txdur9 race2*HIVpos_f race2*txdur9 aftage2_50*race2 sex*yrsin_us13_u
HIVpos_f*txdur9 DOT_only*txdur9 st_CaTx*HIVpos_f st_CaTx*txdur9 sex*st_CaTx
sex*HIVpos_f HIVpos_f*DOT_only aftage2_50*mexican ;
*individual dropped from collinearity: yrsin_us13_u HIVpos_f aftage2_50 ;
proc logistic data=unformatted3 descending covout outest=colin;
        model reinf = sex race2 st_CaTx mexican DOT_only txdur9 st_CaTx*mexican
mexican*DOT_only   /covb;run;
%collin (covdsn=colin, output=covcheck);run;

*interaction assessment;
        *drop mexican*DOT_only 0.9990;*drop st_CaTx*mexican 0.9922;*drop mexican - n=0;
proc logistic data=unformatted3 descending;
        model reinf = sex race2 st_CaTx DOT_only txdur9 ;run;
```