# UNDERSTANDING TUBERCULOSIS DYNAMICS IN THE UNITED KINGDOM USING MATHEMATICAL MODELLING

Adrienne Rae Keen
Faculty of Epidemiology and Population Health
London School of Hygiene and Tropical Medicine

Supervisors: Drs. Emilia Vynnycky and Richard White

Submitted to the University of London for the degree of Doctor of Philosophy

I, Adrienne Rae Keen, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signature:

Date: 16 June 2013

## Abstract

In the UK, tuberculosis incidence has risen from the mid-1980s until recently, with the proportion of cases in foreign-born patients increasing to more than 70% of total cases today. Because several features of tuberculosis epidemiology in the UK are unclear, a simulation model was applied to better understand the epidemiology of tuberculosis in the UK. The model was first used to estimate age- and birthplace-dependent risks of disease progression for those infected via the different disease progression pathways—recent infection, reinfection, and latent infection—by fitting the model to incident cases in England and Wales from 1999 – 2009. Results showed that UK-born risks were lower than previous estimates, though foreign-born risks were an estimated 2.5 times higher than UK-born risks. Estimates for the proportion of disease due to recent transmission were higher than previous estimates, at around 46%. Simulations also identified plausible assumptions for the contact rate and the infection status of migrants upon entry to the UK. Results informed a model fitted to Variable Number Tandem Repeat (VNTR) genotyping data from cases in the West Midlands from 2007 – 2011, which was used to estimate the proportion of disease due to recent transmission in the UK and compare estimates to those based on genotyping data. Results showed that an estimated 45 – 63% of cases in the West Midlands were due to recent transmission in the UK, which was underestimated by genotyping data-derived estimates of 35%. Results also identified plausible mutation rates for VNTR profiles and plausible strain type distributions for UK-born and foreign-born individuals. This work suggests there is a large proportion of cases due to recent transmission in the UK, which is underestimated by genotyping data. The study also provides current disease risk estimates and shows a need for better data on migrants to the UK. This work may help focus prevention efforts.

# Acknowledgements

First of all, I would like to thank my supervisors, Emilia Vynnycky and Richard White, immensely for their support and the opportunity to work on this project. I was lucky to have the expertise and strengths of both supervisors. I would also like to extend deep gratitude to Clarence Lehman for numerous contributions to my development as a scientist and for his involvement in this study. I thank committee members Ibrahim Abubakar and Judith Glynn for additional support.

I am grateful for the Health Protection Agency (HPA) Modelling & Economics Unit studentship and the Overseas Research Student Award Scheme scholarship for project funding. I am also grateful to the HPA Department of Statistics, Modelling, and Bioinformatics, the HPA Centre for Infections, and the HPA Tuberculosis section, each of which provided funds for data used in this study.

I thank Jason Evans, Grace Smith and others from the RCM, Birmingham for laboratory data, collaboration, and much helpful advice; the HPA TB section, including Laura Anderson, Jonathan Moore, David Quinn, Nicholas Fulton, Louise Bradshaw and Charlotte Anderson for data extraction and other support; the HPA Bioinformatics unit, especially Steven Platt and Richard Myers, for genotyping database support; John Innes for clinical data expertise; and Andrew Grant for helpful statistical and modelling discussions.

I am also thankful for support and intellectual stimulation from CMMID members at LSHTM, including Ellen Brooks-Pollock, Pete Dodd, Andy Cox, John Edmunds and Ken Eames. I thank Peter White, Albert Jan Van Hoek, Jonathan Read, Rein Houben, Elisabeth Adams and fellow LSTHM students MinHae Park, Giorgio DiGessa, Tazio Vanni, Erin Anastasi, Raphaelle Metras, Inthira Yamabhai, Ngozi Erondu, Ana Llamas-Montoya, Yemi Okwaraji, Lori Miller, Pippa West, Suchi Roy, Esther VanKleef, Tom Peto, Moke Magoma, and all others who have supported me.

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# Abbreviations

| | |
|---|---|
| ANOVA | Analysis of Variance |
| APS | Annual Population Survey |
| ARI | Annual risk of infection |
| BCG | Bacille Calmette-Guérin vaccine |
| DNA | Deoxyribonucleic acid |
| ETR | Exact tandem repeat |
| ETS | *Enhanced Tuberculosis Surveillance* |
| GAD | Government Actuary's Department |
| GOF | Goodness of fit |
| HIV | Human immunodeficiency virus |
| HPA | Health Protection Agency |
| IBM | Individual-based model |
| IGRA | Interferon-gamma release assay |
| IPS | International Passenger Survey |
| LFS | Labour Force Survey |
| LSHTM | London School of Hygiene and Tropical Medicine |
| LTIM | Long-term International Migration |
| MIRU | Mycobacterial Interspersed Repetitive Unit |
| ODD | Overview, Design Concepts, and Details protocol for describing individual-based models |
| ODE | Ordinary differential equation |
| OF-born | Other foreign-born (born in a country outside of the UK or Sub-Saharan Africa) |
| ONS | Office for National Statistics |
| PCR | Polymerase chain reaction |
| PDE | Partial differential equation |
| RFLP | Restriction fragment length polymorphism |
| SSA | Sub-Saharan Africa |
| SSA-born | Sub-Saharan Africa-born |
| TB | Tuberculosis |
| TST | Tuberculin skin test |

| | |
|---|---|
| UK | United Kingdom |
| US | United States |
| VNTR | Variable number tandem repeat |
| WHO | World Health Organization |

# Terms Used Within

Certain terms have ambiguous or conflicting definitions in the tuberculosis literature. For reduced ambiguity, the following terms take specific meanings in this thesis.

| | |
|---|---|
| Clustered isolate | Isolate with a genetic strain type which exactly matches at least one other isolate in the study population |
| Effective contact | Contact that would lead to infection transmission if made between an infectious and an uninfected person |
| Foreign-born | Person born outside of the United Kingdom |
| Infection state | One of the 11 states of *Mycobacterium tuberculosis* infection for classification in the model, analogous to compartments in classical models: 1) *Uninfected*, 2) *Immune*, 3) *Recent Infection*, 4) *Latent Infection*, 5) Reinfection, 6) *Primary Disease* (pulmonary) 7) *Primary Disease* (non-pulmonary), 8) *Reactivation Disease* (pulmonary), 9) *Reactivation Disease* (non-pulmonary), 10) Reinfection Disease (pulmonary), and 11) Reinfection Disease (non-pulmonary) |
| *Latent Infection* | Infection acquired more than five years ago or infection state after recovery from an active disease episode |
| *M. tuberculosis* | *Mycobacterium tuberculosis*, the causative agent of tuberculosis disease, which for simplicity also refers to other mycobacterial species that may cause tuberculosis in humans |
| *Primary Disease* | First disease episode within five years of first infection (*Recent Infection*) |
| *Reactivation Disease* | Disease more than five years after most recent infection or reinfection, or second or subsequent disease episode within five years of most recent infection |

| | |
|---|---|
| *Recent Infection* | First infection with *M. tuberculosis*, which occurred less than five years previously |
| Recent transmission | Transmission which occurred less than five years before disease onset |
| *Reinfection* | Second or subsequent infection with *M. tuberculosis* acquired within the previous five years |
| *Reinfection Disease* | First disease episode within five years of reinfection |
| Unique isolate | Isolate with genetic strain type which does not match any other isolate in the study |

# 1 Introduction

This chapter provides the motivation for this work, outlines the overall aim and specific objectives of the study, and briefly describes the thesis structure.

## 1.1 Study Rationale

Tuberculosis is a leading cause of infectious disease-related mortality in the world, despite effective antibiotic treatment, a vaccine, and many years of research into infection control and prevention. Most of the disease burden lies in developing countries, where infection risks, morbidity, and mortality are higher than in developed countries. However, tuberculosis has re-emerged as a global problem, due to factors such as the Human Immunodeficiency Virus (HIV) epidemic, migration of infected persons, and drug resistance. In some developed countries where incidence had been declining for more than a century, the number of cases has been increasing in recent years.

In the United Kingdom (UK), the tuberculosis notification rate rose from about 9 cases per 100,000 in 1988 to 14 cases per 100,000 in 2007 [1, 2]. During this time, the proportion of cases found in foreign-born individuals also rose. In 1988, 45% of cases in England and Wales occurred in persons born abroad [2] while in 2007, 72% of UK cases, the vast majority of which come from England and Wales, were born abroad [1]. Though many foreign-born persons were probably infected abroad in countries where infection risk is greater than in the UK, it is not clear what proportion of cases are due to recent transmission in the UK versus reactivation of older or imported infections.

Knowledge of the proportion of cases due to a recent infection, versus an older infection or one acquired abroad, is important for evaluating and focusing tuberculosis prevention and control programs. A high or increasing proportion of disease due to recent transmission indicates it may be possible to reduce the incidence of new cases by preventing ongoing transmission. Preventing ongoing transmission in turn requires reducing the infectiousness and infectious period for active cases, which can be achieved by identifying and treating active cases as quickly and effectively as possible. The identification of cases can be improved by actively searching for cases, often by screening high-risk groups in the population for active disease and screening contacts of active cases, rather than waiting for them to present to healthcare services. High-risk populations may include communities of individuals from countries with a high burden of tuberculosis, homeless persons, and individuals living prisons or other

institutionalized housing [3]. For foreign-born individuals, active case finding can also take place at the point of migration or in the migrants' host country, prior to migration [4]. Effective treatment can be achieved by ensuring those with a confirmed diagnosis are started on treatment immediately and, if at risk of defaulting on treatment, are followed carefully by healthcare workers [3]. On the other hand, a high proportion of cases due to the reactivation of older infections would motivate the prioritization of other control measures, including treatment of latent infection with prophylactic antibiotics to prevent active disease from developing, and follow-up of those testing positive for latent infection but not treated prophylactically [5].

Genotyping data is increasingly being used to distinguish between disease due to recent transmission and disease due to reactivation of an infection acquired many years previously, but the interpretation of these data is complex. It is often assumed that isolates with identical strain types—clustered strains—are part of an ongoing chain of transmission. However, the relationship between clustering and recent transmission changes with many factors, including the rate at which the molecular typing marker changes, study duration, study area, case ascertainment, population age structure, and the annual risk of infection (ARI) [6-9]. Some studies have addressed these issues, but it remains unclear what proportion of tuberculosis in the UK is due to recent transmission or how well these data predict the proportion of cases due to recent transmission in the UK for this setting. Little is also known about the proportion of foreign-born persons entering the UK infected and diseased, the current transmission rate—or effective contact rate—in the UK, and the current risks of disease for those with recent infection, recent reinfection, and latent infection.

To reverse the increases in tuberculosis incidence, it is important to better understand the epidemiology of tuberculosis in the UK, especially in foreign-born persons, who represent the majority of cases. This understanding will ensure that control programs are properly focused and prevention efforts, including the national strain typing services, can be evaluated.

## 1.2 Aim of the Thesis

The overall aim of this work is to better understand tuberculosis dynamics and epidemiology in the UK.

## 1.3 Research Objectives

The study aim is to be achieved using modelling approaches to answer a variety of questions related to understanding tuberculosis epidemiology in the UK, with a particular emphasis on the epidemiology of tuberculosis in foreign-born persons, who make up the majority of cases in the UK. The following specific research objectives for the thesis project fulfil the study aim:

1) Construct an individual-based model (IBM) of tuberculosis dynamics in the UK. Design this model to simulate genetic typing data and describe it according to standardized protocols for documenting IBMs, with code made freely available for others to use.
2) Estimate risks of disease for those with a recent infection, recent reinfection, and latent infection in both UK-born and foreign-born individuals using the model applied to notified cases from England and Wales, 1999 – 2009. Also identify plausible assumptions for the effective contact rate and the infection status of migrants entering the UK.
3) Estimate the proportion of cases due to recent transmission in the UK by application of the model to data from England and Wales.
4) Describe the molecular epidemiology of the West Midlands from 2007 – 2011. This description includes the risk factors for genotype clustering and a rough estimate of the proportion of cases due to recent transmission in the UK, using two genotyping methods (15- and 24-locus VNTR) and various clustering definitions.
5) Estimate the proportion of cases due to recent transmission in the UK by application of the model to genotyping data from the West Midlands. Explore the relationship between genotype clustering and recent transmission in the UK.

6) Identify plausible mutation rates for 24-locus VNTR, as well assumptions about the strain type diversity for 24-locus VNTR profiles found in UK-born and foreign-born individuals by application of the model to genotyping data from the West Midlands.

## 1.4 Thesis structure

The remainder of the thesis is organized into seven chapters. Chapter 2 provides a background on tuberculosis natural history and epidemiology and reviews literature relevant to this study. Chapter 2 also introduces the observed notification data used for parameter values and for comparison with model output in subsequent chapters. Chapter 3 describes the individual-based model (IBM) designed in fulfilment of objective one. Chapter 4 presents the data used to form model assumptions and to parameterize the model. Chapter 5 describes the application of the model to notification data from England and Wales to satisfy objectives two and three. Chapter 6 provides a molecular epidemiological analysis of genotyping data from the West Midlands, to fulfil objective four. Chapter 7 describes the application of the model to simulation of the genotyping data from the West Midlands, to satisfy objectives five and six. Finally, Chapter 8 concludes the thesis.

# 2 Background

This chapter provides background on the natural history and epidemiology of tuberculosis, with a particular focus on the UK. Other topics that are covered emphasize areas particularly relevant to this thesis, including the molecular epidemiology of tuberculosis and the mathematical modelling of tuberculosis dynamics. This chapter also includes a discussion of modelling considerations for this thesis. Finally, the chapter describes observed notification data used for parameter values and for comparison with model output in later chapters. More details on several aspects of tuberculosis natural history and epidemiology appear in Chapter 4, which describes specific data and sources of parameter values in the model.

## 2.1 Tuberculosis Natural History

Tuberculosis has caused morbidity and mortality in human populations for millennia. Tuberculosis is primarily a disease of the lungs, though can affect many sites of the body, including the kidney, bones, urogenital tract, and central nervous system. Symptoms of the disease include a chronic cough, fever, weight loss, night sweats, and general malaise. Pulmonary tuberculosis includes any form of disease that includes pulmonary lesions, with or without affecting other sites. Pulmonary disease is sometimes alternatively referred to as 'respiratory' disease, though respiratory disease is slightly broader and also includes disease in pleural effusions and mediastinal nodes [3]. Non-pulmonary disease includes all forms of the disease that do not include pulmonary lesions. Non-pulmonary tuberculosis is also referred to as 'extra-pulmonary' or 'non-respiratory' disease.

### 2.1.1 Aetiological Agent and Transmission

Tuberculosis is caused by *Mycobacterium tuberculosis*, or, less commonly, *M. bovis*, *M. africanum*, *M. canetti*, or *M. microtti*. For simplicity, henceforth *M. tuberculosis* will be considered to include the latter four as well. Coughing, spitting, sneezing, and speaking by a person with infectious pulmonary disease produces droplet nuclei containing *M. tuberculosis* [10]. Inhalation of droplet nuclei may, but does not necessarily, cause infection. Once infected with *M. tuberculosis*, a person may develop disease shortly after infection, may become latently infected and develop disease later, or may never develop disease. A complex interaction between the host's immune system and the pathogen, which is mediated by several factors [11], determines which of these occurs (see Section 2.1.2 below). Infection with *M. tuberculosis* does not provide complete immunity to a subsequent infection. Reinfection is well documented and more common as the annual risk of infection (ARI) increases [12]. Lastly, infection can also be caused by the ingestion of unpasteurised milk from cattle with tuberculosis, though pasteurisation makes this rare today.

Pulmonary tuberculosis is generally the infectious form of disease, though not all pulmonary cases are infectious. Non-pulmonary tuberculosis is generally not

considered infectious. Pulmonary cases with sputum samples that are 'smear-positive' are more likely to be infectious than 'smear negative' cases [13, 14] (see Section 2.2.4 for information on smear status). Although the average number of others infected by an infectious case varies among populations, an untreated infectious person may infect on average around 20 others over the duration of disease [15]. This number is reduced when a person is treated, shortening the duration of infectiousness. It is also reduced when living conditions are improved, including smaller household sizes and better ventilation. In addition, the susceptibility of contacts will influence how many others are infected. Susceptibility may be reduced by vaccination and previous infection, but is increased by poor nutritional status and immune deficiencies, such as infection with HIV. The transmission rate is difficult to measure directly and therefore highly uncertain. Transmission is discussed further in Chapter 4, Section 4.2.2.

### 2.1.2 Progression to Disease

The risk of progression to disease after infection or reinfection with *M. tuberculosis* depends on several factors. Lifetime risks of developing disease following infection vary among individuals and are difficult to measure, though are often thought to be around 10 – 15% [13, 16, 17]. Thus, the majority of infections never cause disease. The risk of disease following infection is greatest within the first year or a few years following infection [13], after which the bacteria lie dormant and active disease is less likely. As with a first infection, persons reinfected are thought to be at greatest risk of developing disease soon after reinfection [18, 19].

In this study, disease occurring due to a 'recent' infection is defined as disease within five years of infection or reinfection, referred to as *'Primary Disease'* and *'Reinfection Disease'*, respectively. Those infected or reinfected for more than five years without developing disease are usually said to have a *'Latent Infection'* and may develop *'Reactivation Disease'* at any point, even many years after infection. The risk of *Reactivation Disease* is thought to be much lower than the risk of *Primary Disease* or *Reinfection Disease*. It should be noted that these definitions are consistent with previous modelling studies [18, 19], but are not universal. A more widely used, clinical definition of latent infection requires a positive test for infection and no sign of clinical

disease [20], though some clinicians and scientists define latent infections as infections more than two years old [21], with the view that the bacteria are usually dormant by then in people who have not developed disease. It is recognized that the definition of latent infection is controversial [22] and the status as defined here could alternatively be thought of as a *'Remote Infection'*.

In addition to the time elapsed since infection, factors that influence disease progression risk include age, sex, problem drug use, and underlying medical conditions such as HIV, other immunosuppressive conditions, and malnourishment [19, 23-25]. Of these, HIV is the most important risk factor, increasing the risk of disease following infection several-fold [26], though this increase depends on the stage of HIV infection and other factors, as discussed further in Chapter 4, Section 4.2.3.1. Age and sex affect risk in several ways. Very young children are at greater risk of tuberculosis before their immune systems mature, though after a few years, children are at lower risk until adolescence when disease risk increases [27]. Risk of disease may be elevated in older individuals [28, 29], though older males experience higher risk of reactivation disease than females [24]. Vynnycky and Fine estimated age-dependent disease risks for white males in England and Wales from 1900 – 1990 by fitting a model to case notifications (this work is discussed further in Section 2.5.2 below). They estimated that cumulative risks of *Primary Disease* and *Reinfection Disease* over five years of infection were 14% and 8%, respectively for males aged 20 years and above. They estimated that the annual risk of *Reactivation Disease* for those with *Latent Infection* was 0.03% for males aged 20 years and above. Since the time period for which these estimates were obtained, it is unknown whether disease risks have changed for UK-born individuals. Disease risks are unknown for foreign-born individuals, who now account for the majority of cases in the UK each year.

Although it is well-established that foreign-born individuals have higher tuberculosis incidence rates than native-born individuals in low-incidence countries, determining what portion of the excess is due to increased risk of disease following infection, versus higher prevalence of infection prior to migration or increased risk of infection in low-incidence countries, is difficult [30]. Some foreign-born individuals have a lower

socioeconomic and health status than native-born individuals [31], including an elevated prevalence of HIV, all of which may contribute to an increased risk of disease following infection [32]. Furthermore, risks of disease depend on time since infection, and this is impossible to deduce for immigrants infected abroad. This means that even evidence of higher reactivation rates in foreign-born individuals [24] does not mean that reactivation disease risk, as defined here, is necessarily higher. Also complicating matters, there may be genetic differences across ethnic groups influencing disease risk [17, 33].

Given the multitude of factors increasing disease risks, they are likely to change among settings and time periods [24]. Especially for foreign-born individuals, estimates of disease risk are important for prioritizing control measures, such as prophylactic treatment of latent infection. Mathematical modelling is therefore used to estimate disease risks in this study. Assumptions about disease risks are discussed further in Chapter 4, Section 4.2.3.1.

The risk of developing pulmonary versus non-pulmonary disease depends on sex, birthplace, and possibly other factors such as age, HIV infection, and probably tuberculous infection status (i.e. *Recent Infection*, *Reinfection*, *Latent Infection*), but this relationship is not well understood. It is likely that risks of non-pulmonary disease are lowest with primary infection [34, 35], though there may be differences across geographic areas and ethnic groups [36]. Females have a higher proportion of disease that is non-pulmonary than males [34, 37-39]. There has been ample evidence in surveillance data that immigrants to low-incidence countries, especially those coming from Africa or South Asia, have a higher proportion of extra-pulmonary disease than those native-born to low-incidence countries [34, 37-41]. The reasons for this are unclear, but may be due to latent infection giving rise to a higher proportion of extra-pulmonary disease than primary infection [34].

The relationship between age and site of disease in the body is not entirely clear. In recent years, the median age for extra-pulmonary cases and pulmonary cases has been similar in the US and the UK [34, 37]. Another recent study showed that age interacts with the birthplace of individuals [42], with no clear pattern for the effect of age

overall. Similarly, a study published in 2005 showed that in France, age was only significantly associated with extra-pulmonary disease in Sub-Saharan Africa (SSA)-born individuals, but not those from other birthplaces [38].

Several studies have shown that HIV-positive patients have an increased likelihood of non-pulmonary tuberculosis [43, 44]. However, a recent, large study in the US showed only a slight correlation between HIV infection and extra-pulmonary disease [37]. Results showed a correlation between certain types of non-pulmonary disease and HIV infection, concluding that having HIV is a risk factor for disseminated tuberculosis or extra-pulmonary and pulmonary tuberculosis, but not for either pulmonary or extra-pulmonary tuberculosis individually [37]. In another study in Canada, high levels of extra-pulmonary disease not attributable to HIV have been observed [36]. A recent study from France showed being HIV-positive was correlated with extra-pulmonary disease for European-born cases, but not for other birthplaces [38].

Once active disease develops, pulmonary tuberculosis and some non-pulmonary forms can be fatal without treatment. The proportion of patients who die as a result of tuberculosis, or 'case fatality rate', depends on many factors including age, site of disease, drug resistance, and co-morbidities such as malnutrition or HIV infection [45-50]. HIV lowers the survival rate of those who develop disease [51]. Increasing age has been shown to increase the likelihood of death from tuberculosis. Generally, pulmonary tuberculosis is more likely to be fatal than non-pulmonary tuberculosis, though this difference depends on the prevalence of highly fatal forms of non-pulmonary tuberculosis. Some forms, such as miliary tuberculosis and tuberculosis meningitis, are highly fatal, while others, such as lymphatic disease, are not likely to be fatal. With treatment, the case fatality rate is drastically reduced. In areas where treatment is widely available, the case-fatality rate has dropped from about 50% before antibiotics [52] to less than 5% more recently, including drug resistant cases [53]. For more detail on the case fatality rate in recent years in the UK, see Section 4.2.3.4.

## 2.2 Tuberculosis Control

### 2.2.1 Historical Context

Formal study of tuberculosis dates to more than 2000 years ago, though it was only in 1882 that Robert Koch discovered the causative agent of the disease, *M. tuberculosis*. Tuberculosis mortality rates were already declining in the US and Western Europe [54] when he made this discovery, but the discovery still led to control programs and increased isolation of infectious cases [55]. Partly because of these interventions, and also because of improved living conditions and nutrition status with industrialization [56], mortality due to tuberculosis continued to decline rapidly in the beginning of the 20th century in Europe and the United States [54]. Declines in mortality and incidence continued over most of the century.

### 2.2.2 Vaccine

A major control milestone was the development of the Bacille Calmette-Guérin (BCG) vaccine, prepared from attenuated *M. bovis* and first used in humans in 1921. It remains the only vaccine for tuberculosis and has been widely used. Unfortunately BCG has not been consistently proven effective. Estimated vaccine efficacy varies from about 0 – 80%, depending on where the vaccine is used and to whom it is given [57]. For example, the United States never adopted a BCG vaccine program because of scepticism over its efficacy, indicated by studies that showed low efficacy in the US [58]. It was later hypothesized that efficacy was low in these trials because they were conducted in a region of the US with a high prevalence of environmental mycobacteria, and thus most of the population had been sensitized to mycobacterial antigens without vaccination. Sensitization by environmental mycobacteria is thought to decrease efficacy of the vaccine, though it is not clear precisely why. In many high incidence countries, BCG has shown little effect in adults. Notably, a large BCG trial in south India found BCG had no protective effect in that setting [59]. However, BCG has consistently shown protective benefits for severe forms of tuberculosis in children, such as miliary tuberculosis and tuberculous meningitis, and is still administered to newborns in many high-incidence countries around the world for this reason [60]. In the UK, a large trial of the BCG vaccine begun in the 1950s reported an average 77%

efficacy of the vaccine after 20 years of follow-up. It should be noted that the protective effect of the vaccine decreased over time and the study could not be continued past 20 years due to low power to detect differences among vaccinated and non-vaccinated individuals. Still, this trial provided evidence in favour of continued vaccine use in the UK [61]. BCG was routinely given to adolescents in the UK from the 1950s to 2005. More details on BCG vaccination practices and efficacy in the UK are found in Chapter 4, Section 4.2.1.

### 2.2.3 Treatment

Also enormously important for control, tuberculosis has been effectively treated using antibiotics since the mid-1940s, although the duration of treatment is long and requires multiple drugs. As mentioned, treatment not only drastically reduces the case fatality rate, saving lives, but also reduces the duration of infectiousness, reducing transmission.

Current treatment guidelines in the UK recommend a standard six-month treatment course consisting of four antibiotics, isoniazid, rifampicin, pyrazinamide, and ethambutol, though the latter two are only required for the first two months [3]. This standard regime is altered for some forms of disease or if a person develops drug resistance. Drug resistance is a concern because it causes lengthier and more expensive treatment and also higher treatment failure rates [62-64]. Drug resistant strains can be transmitted from one person to the other, or resistance can develop spontaneously, usually due to incomplete antibiotic treatment.

In recent years, directly observed treatment (DOT) has provided a solution for improving adherence to tuberculosis therapy. DOT involves the supervision of the ingestion of medication throughout the course of therapy by a healthcare worker or other designated person [65, 66]. DOT has been considered a successful component of tuberculosis control in many settings and remains a major part of the World Health Organization's strategy for tuberculosis control [67, 68]. DOT is not routinely used in the UK. Nonetheless, it could be used for improving treatment success in patients at risk of defaulting on their treatment course [3].

### 2.2.4 Diagnosis of Disease

Especially because of the burden and expense of treatment, a confirmed diagnosis of tuberculosis based on more than clinical symptoms is ideal. However, this is sometimes challenging due to the low sensitivity of available diagnostic tools. Patients with suspected pulmonary tuberculosis are usually given a chest X-ray to confirm active tuberculosis [69], although these tests are not highly sensitive and their interpretation is somewhat subjective. If possible, laboratory tests performed on sputum samples are used to confirm diagnosis. For suspected non-pulmonary tuberculosis, diagnosis can be confirmed through scans of the affected areas of the body or biopsy of the tissues, in combination with laboratory tests on the tissues. Two types of laboratory tests are typical. Firstly, smear microscopy can be used to test the sample for the presence of acid-fast bacilli, designating cases as 'smear-positive' or 'smear negative'. Secondly, the sputum sample can be tested for *M. tuberculosis* growth in culture, designating cases as 'culture positive' or 'culture negative', though culturing can take several weeks. Unfortunately, not all active disease cases result in culture or smear-positive results. Recently, a deoxyribonucleic acid (DNA) test for active disease called 'GeneXpert' was developed, which is fast and very sensitive for detecting tuberculosis, though its cost effectiveness has yet to be determined [70]. A negative test for *M. tuberculosis* infection (see below) can also help to rule out a tuberculosis diagnosis.

### 2.2.5 Tests for Infection

In resource-rich countries such as the UK, contacts of active disease cases are often investigated for active disease or evidence of infection, for treatment or follow-up. Tests for infection are also used for healthcare workers and for immigrants arriving from high-incidence countries (see Section 2.3.2.3.3). The standard test for *M. tuberculosis* infection is the tuberculin skin test (TST). The TST does not actually test for the presence of *M. tuberculosis*, rather it measures immune response to *M. tuberculosis* complex bacteria [71]. Test results are affected by BCG vaccination, exposure to environmental mycobacteria, immune differences among persons, such as from HIV infection, and other factors [71, 72]. The presence of environmental

mycobacteria in some places causes sensitivity to TSTs, making them difficult to interpret. BCG vaccination has the same effect. For these reasons, they are not very specific. However, in children, TSTs are more accurate than in adults due to less exposure to mycobacteria and therefore prevalence and annual risk of *M. tuberculosis* infection in the past has been estimated using TSTs on children. Lastly, the TST results are subject to variation due to test methodology, including both administration and interpretation [73], plus, if repeated, the tests may alter sensitivity in patients [74].

A recent development in testing for infection is the interferon-gamma release assay (IGRA). These assays test blood samples for evidence of immune reactivity to *M. tuberculosis*. The tests are quicker, more accurate, and less subjective than TSTs. Critically, IGRAs are more specific than TSTs since they do not react to BCG vaccination or exposure to environmental mycobacteria [75]. These tests are rapidly being introduced to supplement or replace the use of TSTs. In the US for example, since 2005 IGRA tests have been recommended as replacements for TSTs in all cases [76]. In the UK, 2011 clinical guidelines suggest the use of IGRA tests when TSTs are positive and for a first-line test in some other groups [3]. Although IGRAs have several advantages over TSTs, their cost effectiveness has yet to be determined for general use [75].

In some situations, those who test positive for *M. tuberculosis* infection are treated with isoniazid preventive therapy to reduce the risk of infection developing into disease. In the UK, clinical guidelines suggest this prophylactic treatment for those with a confirmed infection who are at greatest risk of developing disease, including HIV-positive individuals, other immunocompromised persons, and children from high-incidence countries [3]. Patients 35 years and older are not currently recommended to have prophylactic treatment, due to its toxicity to the liver.

### 2.2.6 Control program strategies

Given limited resources for tuberculosis control, tuberculosis management programs must prioritise control measures. Prioritisation of control resources will be setting-specific, depending on factors such as socioeconomic status of the population, the prevalence of risk factors for tuberculosis such as drug and alcohol use, the prevalence of HIV in the population, and the overall burden of tuberculosis in the population or in

subgroups of the population. In addition, and of particular relevance to this thesis, the proportion of disease due to recent transmission can influence the prioritisation of control measures. In a setting with a high proportion of disease due to recent transmission, the focus of control programs will differ from a setting with a high proportion of cases due to older or imported infections.

When a high proportion of cases are due to recent transmission, control programs should be focused on stopping ongoing transmission, achieved by reducing the infectiousness and duration of the infectious period for infectious cases or by vaccination. Given the limitations of the BCG vaccination for *M. tuberculosis*, prevention of ongoing transmission is generally achieved by administering quick and effective antibiotic treatment that reaches a high proportion of infectious cases. The identification of cases can be improved by actively searching for cases, rather than waiting for them to present to healthcare services, by which time they may have already spread infection. The active search for cases, or 'active case finding', can involve screening high-risk populations for active disease, including those from countries with a high burden of tuberculosis, homeless persons, and those living prisons or other institutionalized housing [3]. For foreign-born individuals, active case finding can take place at the point of migration or in the migrants' host country itself, prior to migration [4]. Active case finding can also involve screening contacts of known or suspected infectious cases, an effort which may be expanded when in situations where transmission potential is high [3, 77]. Tools for active case finding have improved with advances in diagnostic tools, such as mobile x-ray units to diagnose disease [78]. Effective treatment can be achieved by ensuring those with a confirmed diagnosis are started on treatment immediately and, if at risk of defaulting on treatment, are followed carefully by healthcare workers, including using DOT to improve adherence if necessary [65, 66].

On the other hand, a high proportion of cases due to the reactivation of older infections would motivate the prioritization of other control measures, including treatment of latent infection with prophylactic antibiotics to prevent active disease from developing or follow-up of those testing positive for latent infection [3, 5]. There

is also a possibility that investment of resources for tuberculosis treatment in high-burden countries may be cost-effective for reducing the tuberculosis burden in countries which receive immigrants from those countries, as shown for the case of the the US and Mexico [79].

## 2.3 Contemporary Tuberculosis Epidemiology

Worldwide, an estimated 8.7 million incident cases of tuberculosis and 1.4 million deaths due to tuberculosis occurred in 2011 [80], the vast majority of which were found in developing countries. The 22 countries designated as 'high-burden' by the World Health Organization (WHO), mostly in Africa and Asia, account for about 80% of total cases each year [68, 80]. India and China alone contribute almost 40% of total cases worldwide [80]. However, SSA has the highest incidence rates per capita in the world, in some countries more than 500 cases per 100,000 population per year. Part of the reason rates are so high in SSA is due to the effect of HIV on tuberculosis dynamics, which severely increases disease risks and mortality rates [26, 81]. WHO estimated that nearly 39% of tuberculosis cases in the African region were co-infected with HIV in 2011 [80]. Still, some progress has been made, as worldwide incidence of tuberculosis has been slowly declining for several years [80], at least partly attributed to prevention and control initiatives brought on by the declaration of tuberculosis as a global emergency in 1993 by WHO.

### 2.3.1 Tuberculosis in Developed Countries

In developed countries, tuberculosis has generally been well controlled. However, recently, the long-term decline in tuberculosis incidence was reversed in many developed countries, including the UK, many Western European countries, and the US [82-86]. The increased incidence rates began around the mid-1980s and lasted until the early 2000s. In other countries, such as the UK, Norway, and Sweden, incidence rates continued to increase until more recently [87]. Immigration from countries with a high prevalence of tuberculosis, homelessness, problem drug use, and HIV infection have been implicated in this rise in tuberculosis incidence [87-93]. As evidenced by these factors, tuberculosis has become less a disease of the general population and more a disease in high-risk groups of the population in developed countries today.

One of the largest risk groups for tuberculosis in developed countries comprises foreign-born persons from high-incidence countries. In many developed countries today, the majority of cases occur in foreign-born persons, though the proportion of

cases in foreign-born persons varies from country to country [94-97]. In Europe, foreign-born persons account for 0 to 82% of total cases [98], generally over-representing their proportion of the total population. It is likely that a mixture of factors cause increased risk of tuberculosis in foreign-born persons.

Foreign-born persons in developed countries may face higher risks of progression to disease [99], as discussed in Section 2.1.2. In addition, more crowded living conditions and greater chance of encountering active cases in foreign-born communities could increase infection risk. Also importantly, migrants from high-incidence countries are more likely to be infected before migration. However, the proportion of foreign-born persons infected abroad is not well known, as testing of migrants for infection is sporadic. Furthermore, even if infection status is known, there is no way to clinically distinguish between disease due to reactivation and disease due to a recent reinfection. The proportion of cases due to a recent infection after migrant entry, versus an older infection or one acquired abroad, is important for evaluating and focusing tuberculosis control programs, as discussed above in Section 2.2.6.

Genotyping data are increasingly being used to help differentiate between cases due to a recent infection and cases due to older infections or infections acquired abroad. Unfortunately, genotyping data have several limitations, as discussed further in Section 2.4.3. In addition, most developed countries have policies regarding screening of migrants for active tuberculosis [100] either before or shortly after migration, though policies vary from country to country. A recent study found about half of developed countries also test for latent infection in migrants [100]. The US is often cited for its successful and comprehensive immigrant screening policy, where all migrants are screened for active disease before entry and tested for latent infection [101]. Persons with latent infection are then followed up by local health departments after settlement in the US [102]. In the UK, screening for active tuberculosis is undertaken for persons from certain high-risk countries at port of entry and migrants are followed up by local health authorities, although this is not consistently undertaken, as discussed in Section 2.3.2.3.3.

### 2.3.2 Tuberculosis in the UK

Tuberculosis trends in the UK have followed patterns similar to those in other developed countries. Declines in tuberculosis mortality that began in the 19th century are partly attributed to isolation of patients in workhouses used for treatment of the poor [103, 104] and are also attributed to an increase in the standard of living, which likely reduced risk of disease progression given infection [103]. Sanatoria were set up in the late 1800s and early 1900s; this isolation of infectious persons probably reduced infection transmission. Nutrition and health status of the population continued to improve over this time. Antibiotic use caused further decline in tuberculosis cases after the late 1940s. Some of the decline in tuberculosis incidence could also be attributed to BCG use as routine vaccination of all 13-year-olds, which began in 1953. The lowest recorded number of tuberculosis notifications since surveillance began in England and Wales was 5,086 cases in 1987. However, since then the number of tuberculosis cases increased until recently, with 8,400 cases in 2011 in England and Wales. Fortunately, the past several years have shown evidence that notifications have stabilized [1, 105].

As in other low-incidence countries, tuberculosis has become less uniformly distributed in the population, with most disease occurring in sub-groups of the population. High-risk sub-groups in the UK include foreign-born individuals from countries of high tuberculosis incidence, HIV-infected persons, problem drug users, homeless persons, alcohol abusers, and those who have spent time in prison. It is also likely that given the decrease in ARI over time, a large fraction of disease in patients born in the UK will be due to reactivation of an older infection, acquired when infection risks in the UK were greater.

It was noticed as early as 1958 that immigrants from countries with high tuberculosis incidence, namely India and Pakistan, experienced high tuberculosis notification rates in the UK [106, 107]. These and other studies provided motivation for a 1965 survey of tuberculosis in immigrants in England and Wales [108]. At this time, 16.5% of tuberculosis cases occurred in immigrants, mostly Irish and South Asian, though foreign-born persons made up about only about 4% of the population [108]. By 1988, 45% of cases were occurred in foreign-born persons, and this increased to 56% in 1998

[2, 109]. In 2001, foreign-born individuals made up only 8% of the population of the UK [110].

Today, more than 70% of incident cases occur in foreign-born persons [105], with the majority of foreign-born cases found in persons born in South Asia or SSA. Foreign-born persons have higher per capita notification rates than those who are UK-born, more than 20-fold higher on average. In the last decade, notification rates have been around 80 – 100 cases per 100,000 population for foreign-born persons and constant at around four cases per 100,000 population for UK-born persons, as shown in Figure 2-1. However, even among the UK-born population, there is a wide disparity among ethnic groups, with non-white persons having elevated tuberculosis notification rates, as shown in Figure 2-2.

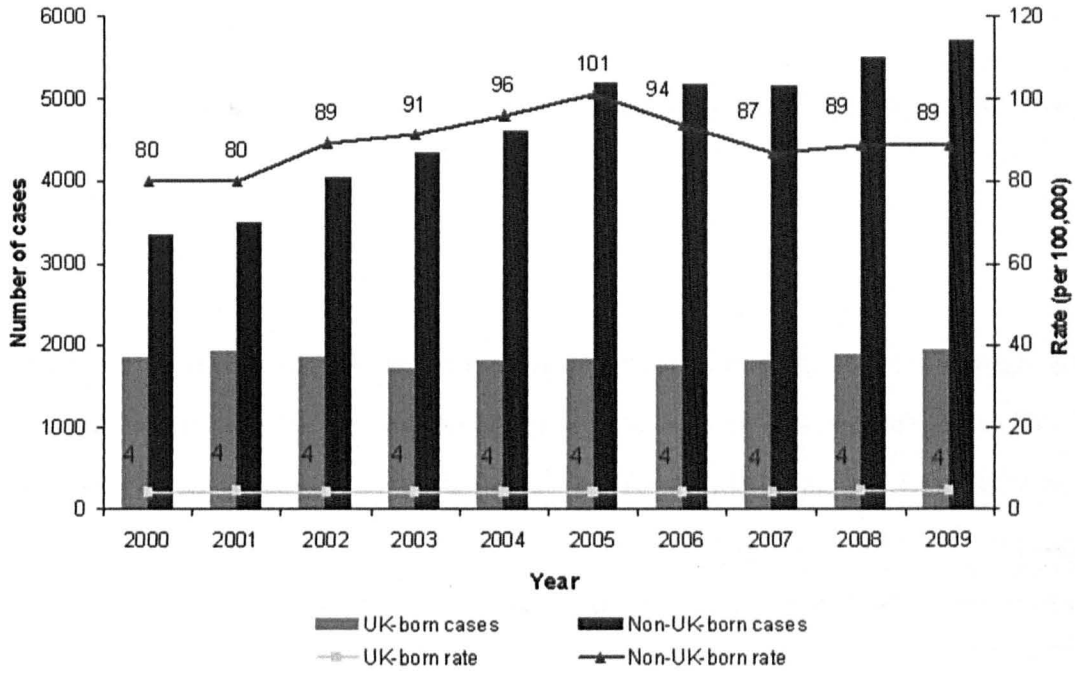


Figure 2-1: Tuberculosis cases and rate per 100,000 population for UK-born and foreign-born individuals in the United Kingdom, 2000-2009.Figure was prepared by the Health Protection Agency Tuberculosis Section. Data sources are Enhanced Tuberculosis Surveillance and the Office for National Statistics (ONS) mid-year population estimates and Labour Force Survey for population sizes.

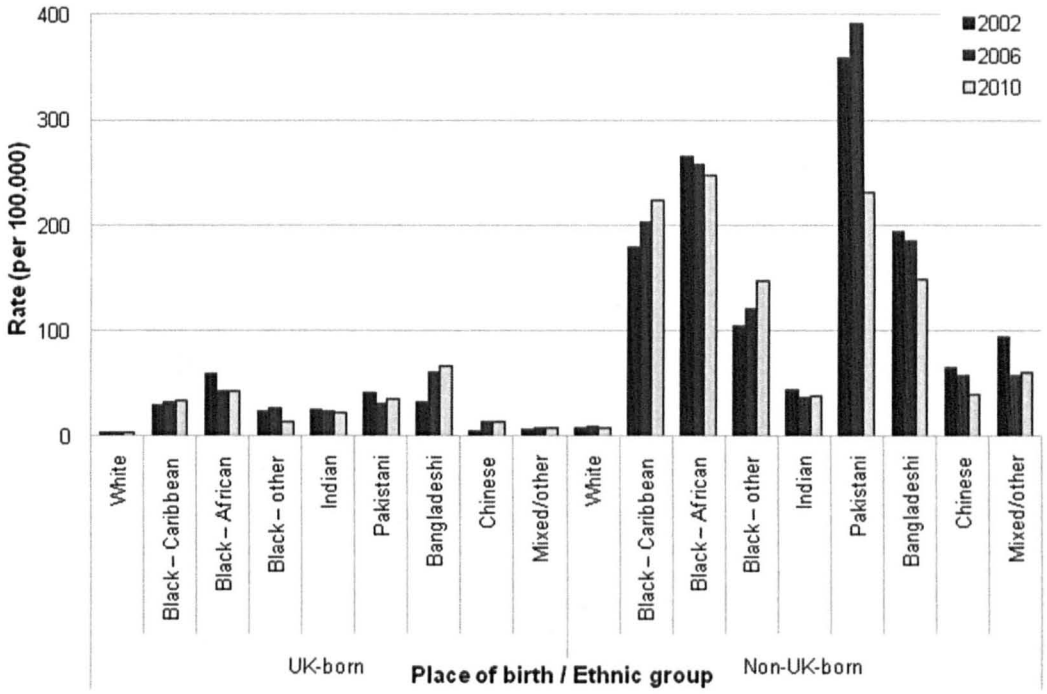


Figure 2-2: Tuberculosis notification rates per 100,000 by birthplace and ethnicity in the United Kingdom, 2011. Figure was prepared by the Health Protection Agency Tuberculosis Section. Data sources are Enhanced Tuberculosis Surveillance and the Office for National Statistics (ONS) mid-year population estimates and Labour Force Survey for population sizes.

Partly due to the uneven distribution of foreign-born and other persons at high risk for tuberculosis, there is also an uneven geographical distribution of cases in the UK. The highest incidence rates are currently found in Leicester, London, and Birmingham, where more than 40 cases of tuberculosis per 100,000 population occur each year. The geographical distribution of tuberculosis in regions of England is illustrated in Figure 2-3.

#### *2.3.2.1 England and Wales*

Trends in the epidemiology of tuberculosis in England and Wales follow those in the UK, since this region comprises the majority of the UK population and majority of the tuberculosis cases in the UK. In 2001, the population size of England and Wales was approximately 52.4 million, while the total UK population was 58.8 million [111]. Also in 2001, 9% of persons in England and Wales were born abroad [111].

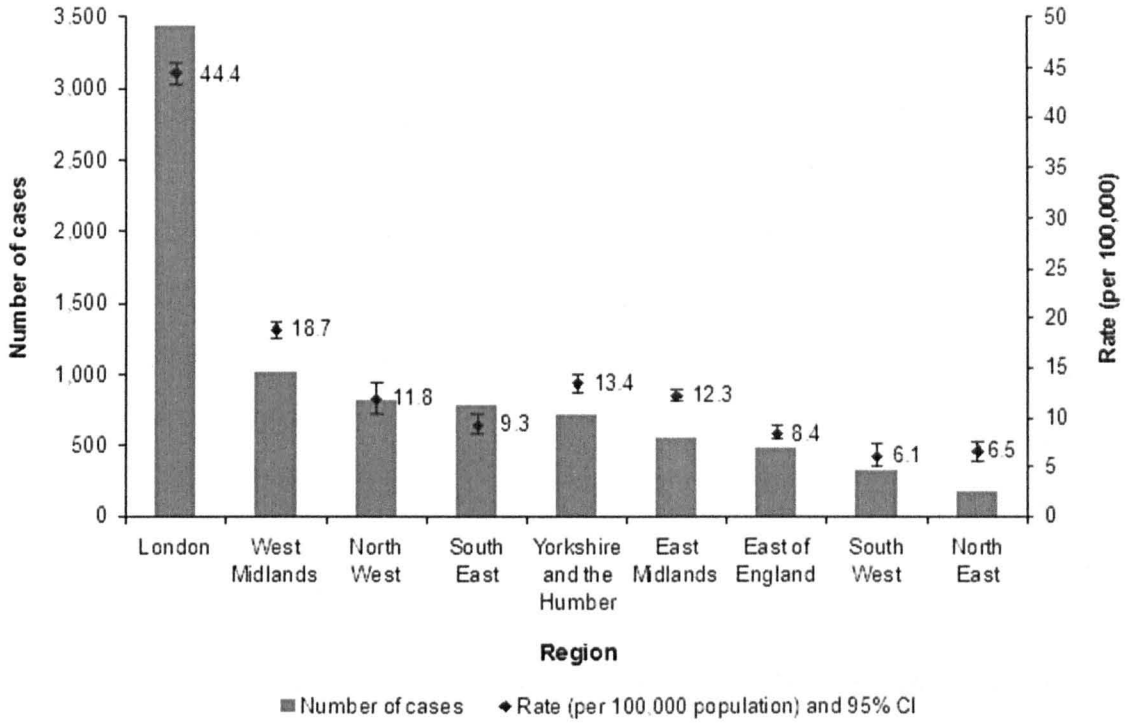


Figure 2-3: Tuberculosis cases and rate per 100,000 population for regions of England, 2011. Figure was prepared by the Health Protection Agency Tuberculosis Section. Data sources are the Enhanced Tuberculosis Surveillance and the Office for National Statistics (ONS) mid-year population estimates.

#### *2.3.2.2 West Midlands*

The West Midlands is a region of the UK which covers the western half of the Midlands, or central England. The population size was 5.3 million in 2001, with about 5.5% of those persons born abroad [110]. The major urban area in the region surrounds Birmingham, the second-largest city in the UK with nearly one million people [110]. The region also includes the cities of Coventry and Wolverhampton. All three cities are noted for relatively high rates of tuberculosis incidence in the UK [112]. In addition, the region as a whole has tuberculosis notification rates above the national average, at approximately 18.5 cases per 100,000 in 2011, compared with 14.4 per 100,000 in the UK [105]. Mirroring national trends, notification rates in the region have increased since the late 1980s, with concurrent increases in the proportion of patients born abroad.

Almost 90% of cases in foreign-born individuals notified from 2007 – 2011 in the West Midlands were found in persons from SSA or South Asia. Also, compared to other regions of the UK, there is a large population of South Asians in the West Midlands, including many who were born in the UK. Over the period 2007 – 2011, approximately 5.5% of UK-born persons in the West Midlands were of South Asian ethnicity, whereas in England and Wales over this period, about 2.7% of UK-born persons were South Asian (proportions obtained from analysis of LFS data, see Chapter 4, Section 4.1.3.1 for information on the dataset). This difference is reflected in the tuberculosis epidemiology of the region. From 2007 – 2011, 39% of UK-born tuberculosis cases were found in ethnic South Asians, while over the same time period, only around 20% of UK-born cases in England and Wales were found in South Asians [105].

#### *2.3.2.3 Contemporary surveillance and control in the UK*

The UK has a comprehensive tuberculosis surveillance and control program. Since 1913, mandated reporting of tuberculosis cases to health authorities has been in place. More recently, the enhanced tuberculosis surveillance system (ETS) was developed (described below). Routine surveillance now includes the results of several laboratory tests, in addition to prospective genotyping of all tuberculosis cases which are laboratory confirmed (also described below). Other elements of tuberculosis control in the UK include contact tracing for each identified case, extensive outbreak investigation for drug resistant cases and those suspected to be highly transmissible, and screening of migrants for tuberculosis upon entry to the UK. For relevance to this study, only the latter of these other elements of tuberculosis control will be further discussed.

##### 2.3.2.3.1 Enhanced tuberculosis surveillance

The ETS system began in 1999 for the collection of additional data on each notified cases of tuberculosis in England, Wales, and Northern Ireland. The enhanced surveillance data help better understanding of tuberculosis epidemiology in the UK and help inform prevention and control. Data collected for each case include place of

residence, date of birth, sex, country of birth, year of entry into UK for foreign-born, date of disease onset, date of notification, and laboratory data. In 2002, treatment outcome data were added to the system. From 2009, other risk factors for tuberculosis, including drug use, alcohol use, homelessness and time in prison have also been collected. All ETS data are held by the HPA Centre for Infections. For each patient notified, ETS data are matched to laboratory data (see below) through a computer algorithm that takes into account various personal identifying information and is supplemented by manual checking where necessary [113]. A similar enhanced surveillance system is used in Scotland, and those data are collated with ETS data to produce annual, UK-wide reports on the epidemiology of tuberculosis in the UK.

#### 2.3.2.3.2 Laboratory data

Clinical specimens are sent to the mycobacterial reference laboratories in the UK for identification of the species of mycobacteria involved in infection, to test for culturing of the organism, to test for antibiotic drug sensitivity of the strain, and to genotype the strain. The UK began universal prospective genotyping of isolates from tuberculosis cases in 2004, using 15-locus VNTR typing. In 2010, the typing system was upgraded to 24-locus VNTR typing. Coverage of universal typing has increased over this period, with an average of 94% of culture-positive isolates typed in 2011 in England [105]. Each of the typing methods is described further in Section 2.4.1.2.

#### 2.3.2.3.3 Screening of migrants

In the UK, it is national policy to screen migrants from high-incidence areas at their port of entry for active tuberculosis with a chest X-ray. New entrants should then be referred to local health authorities for follow-up. Most importantly, they should be referred for active disease screening if the chest X-ray was abnormal or inconclusive, though for some groups, new entrants are referred for latent infection screening [3]. Unfortunately, recent studies have shown that follow-up and latent infection screening by local health authorities was inconsistent [114, 115]. The most recent study did confirm that action was consistently taken on reports an abnormal chest X-ray for an immigrant at port of entry [114].

## 2.4 Molecular Epidemiology of Tuberculosis

Molecular biology tools have become an important part of understanding tuberculosis pathogenesis, drug resistance, infection transmission, and strain distribution worldwide [116, 117]. In particular, genetic strain typing can identify genetically similar strains of *M. tuberculosis*, which has several uses in public health and epidemiology. Genetic strain typing is used to study the phylogeny of *M. tuberculosis*, identify outbreaks of related cases, identify laboratory contamination, detect mixed infections, differentiate between cases of relapse and reinfection in recurrent cases of tuberculosis, and study *M. tuberculosis* transmission on the population level [118, 119].

The use of genotyping data to understand transmission will be the focus of this section, because of its relevance to the thesis. Genotyping data are often used to help distinguish whether disease is due to recent transmission or from reactivation of an older infection because there is no way to clinically distinguish between these forms of disease. To this end, it is often assumed that isolates which 'cluster' together, or have identical genetic strain types, are part of the same recent chain of transmission. On the other hand, isolates which do not match any other strain types in the population are considered 'unique' and assumed to be the result of reactivation of an older infection or one acquired outside the study population. The proportion of isolates clustered is used to estimate the proportion of disease due to recent transmission and identify population-level risk factors for recent transmission. However, interpretation of these data is complicated by several factors, which depend critically on the genotyping method used and other, study-level factors. A description of typing methods follows below. The application of genotyping data to understanding transmission in the population is then discussed. Finally, challenges to interpretation of these data in understanding transmission are presented.

### 2.4.1 Genotyping Methods

The development of genetic typing methods for *M. tuberculosis* has been limited by the small amount of genetic variation in the genome, which is reduced compared to

other bacterial species. Despite this, sequencing has revealed several polymorphic regions, including: Single Nucleotide Polymorphisms (SNPs); insertion elements (e.g. IS*6110*); several regions with variable number tandem repeats (VNTR), including mycobacterium interspersed repetitive units (MIRU) and exact tandem repeats (ETR); Spoligotyping (the direct repeat region); polymorphic GC-rich sequences; and regions of difference polymorphisms. More recently, whole genome sequencing of the bacterium has become a possibility, though routine use in public health has is yet to be seen [120]. Because of their use in UK surveillance, this study focuses on VNTR typing data, with some discussion of insertion element *IS6110*-based RFLP typing because of its importance in other molecular epidemiological studies of tuberculosis. The other genotyping methods are reviewed elsewhere [121].

#### *2.4.1.1 RFLP*

Insertion sequence IS*6110*-based RFLP is the most widely used tool for studying the molecular epidemiology of tuberculosis [122]. This insertion sequence occurs a variable number of times and in different locations throughout the *M. tuberculosis* genome for different strains of the bacterium. RFLP typing requires digestion of the *M. tuberculosis* genome with a restriction endonuclease to cut the DNA at IS*6110* insertion sites and separation of the cut pieces of DNA – called 'restriction fragments' –using gel electrophoresis [123]. After hybridization and labelling, the gels produce a banding pattern which depends on the number and molecular weight of the restriction fragments in a particular *M. tuberculosis* strain. Electrophoresis gels are photographed and then analyzed using specialized software such as BioNumerics® (Applied Maths, Sint-Martens-Latem, Belgium).

Due to a sufficient diversity of strain types in the population and stability of profiles over time periods relevant to infection transmission, it is generally assumed that patients with identical IS*6110* RFLP patterns are part of the same chain of recent transmission. Stability of RFLP patterns over time scales relevant to transmission has been indicated in studies where epidemiologically-linked patients have identical IS*6110* RFLP patterns [124, 125] and where serial isolates from the same patients have shown stability in RFLP patterns over time [124, 126].

The rate of change of RFLP profiles have been estimated to be between 8-25% of profiles each year in disease cases [127-129]. In *Latent Infection*, the mutation rate is thought to be slower due to reduced replication of *M. tuberculosis*, though has only been estimated by one study, to my knowledge. This study compared strains isolated from patients in the 1960s to those isolated from patients in the 1990s to estimate that the mutation rate of strains in latent infection over this time period was approximately 2% per year [130].

Despite widespread use of this method, the validity of using the proportion clustered based on IS*6110* RFLP to estimate recent transmission has been questioned [117, 122, 131-133], largely due to uncertainty over the mutation rate of RFLP profiles [131]. In addition, it is well known that the method is inadequate for isolates with a small number of IS*6110* insertion sites – or low copy number isolates – because of their poor discriminatory power [117, 134-137]. Furthermore, IS*6110* RFLP is labour intensive, is relatively slow because it relies on culturing of *M. tuberculosis* before typing can be performed, and results can be difficult to reproduce [116, 138]. Analogue band patterns must be digitalized, and this process does not always lead to consistent results.

#### *2.4.1.2 VNTR*

A newer technique for studying tuberculosis molecular epidemiology is VNTR typing, which characterizes regions of the genome where short pieces of DNA are repeated in tandem a variable number of times [139, 140]. The regions containing these repeated elements are referred to as VNTR 'loci'. Other similar regions of tandem repeats, MIRU and ETR, will henceforth be referred to as VNTR for simplicity. At each of these loci, a 'repeat number' is assigned, based on how many times the short piece of DNA is repeated at the locus. Thus, a VNTR 'profile' consists of a string of repeat numbers, one for each locus included in the profile. VNTR typing is typically done with 12, 15, or 24 loci, though other combinations are used.

VNTR typing is highly reproducible because of the digital nature of the genotype, and comparisons across laboratories and studies are convenient [116]. VNTR is also

advantageous because it relies on PCR-based determination of repeat numbers, which requires relatively small amounts of DNA and is much quicker than RFLP and other methods which rely on culturing of isolates.

As is the case with any genotyping method, the discriminatory power of VNTR depends on the diversity of strain types in the population and rate of change of profiles. For VNTR, these characteristics depend critically on how many loci are included in the profile and which loci included in the profile, as some loci are more discriminatory than others. Some of the commonly used combinations of VNTR loci are discussed below. The rate of change of VNTR profiles is also discussed below, in Section 2.4.1.3.4.

The disadvantages of VNTR typing include relatively high cost, requirement of specialized laboratory equipment required, and most importantly, the varying and still incompletely known discriminatory ability of the typing system. Furthermore, there is difficulty comparing across studies in the literature to date, since several different sets of loci have been used. It is likely that as VNTR gains in popularity, one standardized set of loci will dominate typing of *M. tuberculosis* isolates, and these problems will subside.

#### *2.4.1.3 VNTR typing systems*

There are many possible combinations of VNTR loci that can be used for *M. tuberculosis* typing. Several of the most important and standardized systems are described below.

##### 2.4.1.3.1 Earliest used loci

In 1998, five variable loci called exact tandem repeats (ETR) were described [141]. These five loci have been used in conjunction with other loci in achieving a high level of discrimination, or as a second-line test, but are not considered discriminatory enough for use on their own [138, 142]. In 2000, Supply et al. described a set of 12 loci called 'MIRU' loci which showed enough discrimination and stability to potentially be used for studying the molecular epidemiology of tuberculosis [140, 143]. Since then, the MIRU loci have been shown to be less discriminatory than originally thought, not

usually comparing well to the discriminatory ability of RFLP [144, 145]. Though some studies have maintained that the 12 MIRU loci compare well with clustering proportions seen in RFLP [146], they are usually only considered as a first-line approach, even when combined with spoligotyping [147]. This combination has been used in the US for large-scale genotyping of tuberculosis isolates [147, 148]. Later, it was proposed that six of the original 12 loci should be discarded from the VNTR typing system due to insufficient variation, and an additional nine be added. This resulted in what is usually considered the standard 15-locus VNTR typing system. Clusters based on these 15 loci have been shown to correlate well with RFLP clusters [144], especially if used together with spoligotyping [149].

#### 2.4.1.3.2 15 loci (UK)

The 15-locus typing system used in the UK from 2004 – 2009 is slightly different than the standard 15-locus typing system found in the literature. This typing system uses all five ETR loci, plus the original 12 MIRU loci, which amounts to 15 distinct loci because two of the loci overlap [150]. Hawkey et al. found this typing system compared well with RFLP in identifying clustered isolates [138], but subsequent studies have shown these 15 loci do not appear to achieve discrimination levels near those of RFLP or the more discriminatory VNTR sets. Gopaul et al. found in three small study populations (n=71, 125, 248) that these fifteen loci were not as discriminatory as RFLP, with more than three times the clustering percentage in one of the studies (n=248) [151]. Hanekom et al. studied Beijing spoligotype isolates in South Africa and found these 15 loci did not discriminate any better than the original 12 MIRU loci and were much less discriminatory than RFLP [122].

#### 2.4.1.3.3 24 loci

In 2006, Supply et al. proposed the use of a set of 24 discriminatory loci for use as genetic typing standard for *M. tuberculosis* [152]. This typing method is gaining popularity, and as of 2010 has been the standard typing method for isolates from tuberculosis cases in the UK. In a Hamburg study in 2007, 24-locus VNTR was found to be slightly more discriminatory than RFLP, though there were some inconsistencies between RFLP and VNTR clusters [153]. In 2008, it was shown that 24-locus VNTR

clusters were highly consistent with RFLP clusters, and gave lower clustering proportions than RFLP, even after excluding low copy number isolates from the analysis [154]. A study set in Venezuela showed 24-locus VNTR was more discriminatory than RFLP, though again, they found some inconsistencies between RFLP and 24-locus VNTR clusters [149]. To date, there have been no population-level studies using 24-locus VNTR typing which analyzed more than a two or three hundred isolates.

#### 2.4.1.3.4 Rate of change of VNTR profiles

The rate of change of strain type profiles has a major impact on the genetic diversity in a population of strains, and the similarity of strains involved in a chain of transmission. Consequently, the mutation rate is a major determinant of the utility of genotyping data for identifying disease cases that are due to recent transmission. Changes to VNTR profiles can result from changes in the number of repeated sequences at any one or more loci in the profile. It is thought that changes to VNTR loci occur during replication due to strand slippage during DNA replication, resulting in increases or decreases in repeat numbers for individual VNTR loci. Repeat numbers are thought to increase or decrease in a stepwise manner, one repeat at a time and loci are also thought to mutate independently.

There have been several studies to estimate the mutation rate of VNTR loci using statistical or mathematical models, often combined with observed VNTR genotyping data. These have resulted in a very wide range of estimates for the per-locus mutation rates, as detailed in Chapter 7, Section 7.1.3.2. Per-locus mutation rates translate into mutation rates for 24-locus profiles ranging from 0.02% per profile per year to 30% per profile per year [155-158]. Although this wide range of mutation rates indicates great uncertainty, the 24-locus VNTR mutation rates may be comparable to those found for RFLP profiles, which are on average around 8-25% per profile per year.

There has been a conventional wisdom that during latent infection, *M. tuberculosis* replication slows or stops [159], making mutation less likely. However, the relative

mutation rate of VNTR profiles between active disease and latent infection has not been studied, to my knowledge.

### 2.4.2 Genotyping Data for Studying Transmission

As mentioned above, one of the main applications of genotyping data is to study transmission in the population. This is typically done by assessing all isolates in the population as clustered or unique, and estimating the proportion of cases due to recent transmission based on the proportion clustered.

The simplest analysis regards the proportion of isolates clustered as a direct proxy for recent transmission, sometimes called the 'n method' [160]. This statistic is often used, though may also be inappropriate for estimating recent transmission in some age groups and settings [161]. Another common statistic based on genotyping data is the 'n-1 method' [162]. This statistic is computed by subtracting the number of clusters from the number of clustered isolates and dividing this figure by the total number of isolates. The idea behind this method is that each cluster has one source case, which is the result of reactivation of a latent infection, while the remaining cases in the cluster have been infected recently. This method may not be appropriate for estimating recent transmission in all age groups, for example in the young where up to 100% of cases in a cluster are due to recent transmission [161] or for studies with reduced case ascertainment [163].

Another method for assessing clustering is the 'retrospective method', which also attempts to distinguish between source cases and recipients of infection. Under this definition, cases with strain type profiles that exactly match one or more other profiles from case(s) in the study population, which were notified previously, within some defined time period, are considered clustered. A case with a strain type profile that does not have a match in the study population from a case notified previously, within the defined time period, is considered unique. Published studies using this method have looked four years and one year prior to define retrospective clustering [164-166]. The retrospective method eliminates some bias in cluster analysis that occurs for other definitions of clustering because cases notified at different times have different follow-up periods for assessing clustering. Using the retrospective method, the same time

period for each isolate is used to look retrospectively for an isolate's match. However, this method results in a loss of data, with increasing losses for longer retrospective time periods considered.

Other methods for estimating the extent of recent transmission include genetic distance-based metrics [117, 132, 167]. In general these metrics estimate higher proportions of disease due to recent transmission due because of more inclusive definitions of clustering, but in the absence of a 'gold standard' it is unknown whether these are more accurate than other methods for estimating recent transmission. These may be appropriate for more discriminatory typing methods, such as whole genome sequencing, when that is more commonly used.

Often clustering is defined by one of these methods, usually the n method, and then risk factors for clustering are assessed, assuming clustering is proxy for recent transmission. The proportions clustered across demographic and other categories of the population are usually reported, along with odds ratios (OR) for clustering among different groups. Commonly identified risk factors for clustering are young age, male sex, pulmonary disease, being native-born, drug use, alcohol use, and homelessness [168-171]. These risk factors for clustering are expected risk factors for transmission and thus provide some confirmation for the utility of genotyping data in detecting recent transmission. To date, risk factors for clustering on the population level have only been studied using RFLP.

### 2.4.3 Limitations

Limitations to the interpretation of genotyping data for assessing recent transmission include uncertainty regarding which methods for defining clustering best correlates with the amount of disease due to recent transmission, but also other factors. These factors include varying discriminatory ability across study settings, uncertainty about mutation rate of genotypes, uncertainty about the effects of ARI in some settings, the effect of low case ascertainment, the effect of a restricted study duration, and the age-dependent utility of genotyping data for predicting the proportion of disease due to recent transmission [122, 168, 172-174].

In 1999, Glynn et al. used simulation modelling to show that increased case ascertainment leads to increased clustering, and populations with small cluster sizes are more susceptible to underestimation of true clustering as case ascertainment is reduced [163]. Two later studies showed clustering underestimates recent transmission as study durations are shortened [7, 9]. Vynnycky et al. used simulation modelling to show how clustering of isolates underestimates the extent of disease due to recent transmission in the young and overestimates this in the older individuals. In 2003, Vynnycky et al. extended this work to show the importance of ARI trend on the relationship between clustering and recent transmission [161]. However, the study excluded immigrants from analysis and so it is uncertain how results translate to tuberculosis epidemiology in countries with a high proportion of cases in immigrants, such as the UK.

In 2002, Murray addressed the influence of latent infection prevalence, population age structure, chemoprophylaxis of contacts, and other factors on the proportion of isolates clustered and cluster size [175]. Among other results, Murray concluded control measures which do not reduce transmission (e.g. screening and treatment of latent infection) may not reduce mean cluster sizes. However, the study was not calibrated to real data, did not take into account evolution of the molecular marker, and did not attempt to explore the effects of immigration [175] so the applicability of results is questionable. In a subsequent paper, Murray also explored bias in the estimation of recent transmission with different sampling, which confirmed results from Glynn et al. [163]

To summarize, there are several factors which influence the relationship between genotype clustering and amount of disease due to recent transmission. It is not clear how to quantify the relationship between clustering and recent transmission, for any particular molecular marker or statistic, in a population of tuberculosis cases as seen in the UK, which is a mixed population of foreign-born and UK-born. These groups will have experienced different ARI over their lifetimes, may currently experience different ARI and disease risks, and may have different risks of developing infectious disease. No study has examined the relationship between clustering and recent transmission in

such a population. In addition, no study has taken into account the likely highly differential mutation rates of bacteria, and therefore genotype profiles, involved in active disease versus infection, when interpreting genotyping data. In the absence of a 'gold standard' it is difficult to evaluate VNTR typing methods and interpret these data reliably across study settings. It may be possible for mathematical modelling to aid interpretation of these data.

## 2.5 Mathematical Modelling of Tuberculosis

Mathematical modelling and simulation of tuberculosis dynamics are helpful in studying the disease, especially because of the unusual complexity in its natural history. Unlike many diseases, infection with the causative agent of tuberculosis is difficult to detect, and an infection is thought to confer incomplete or even little immunity to subsequent infection. Because direct experimentation is not ethical in humans, mathematical models can help shed light on the natural history of the disease [18, 19, 176, 177]. Models can also be used to assess interventions and make projections for disease incidence in the future [178-182]. Models have also been used to better understand the long-term disease dynamics and the effect of major changes to tuberculosis epidemiology, including the HIV epidemic [183-186].

Most tuberculosis models to-date have been deterministic, compartmental models. These models group the population into compartments based on infection states, and possibly other characteristics, with all individuals within a compartment considered equivalent. Some models have further stratified these compartments by age, sex, or other variables. Other models have simulated disease dynamics using Markov processes, which are stochastic but divide the population into states, similar to classical compartmental models [187]. More recently, studies using IBMs of tuberculosis dynamics have been published. These models distinguish each individual in the population, rather than grouping them into compartments. Distinguishing individuals allows for features such as modelling transmission within complex contact networks and describing transmission of individual *M. tuberculosis* strain types.

In the following section, a brief overview of tuberculosis models is presented, with subsequent sections highlighting tuberculosis models in three areas particularly relevant to this work: models based on the UK; models including immigration; models involving genotypes; and IBMs.

### 2.5.1 Chronological Overview of Tuberculosis Models

Since the first tuberculosis modelling paper published in the 1960s [188], many tuberculosis models have been developed, far more than are described here. Because

categorizing these models by common features and assumptions of the models leads to some confusion, since most categories are not mutually exclusive, here I present a roughly chronological description of literature on tuberculosis modelling. I describe studies that comprise a group of diverse and important works relevant to the model used in this study, with a focus on models applied to tuberculosis in developed countries. Reviews of tuberculosis models have been published elsewhere [189-191].

In 1962, Waaler et al. published the first tuberculosis modelling paper, describing disease dynamics using three disease compartments and a system of linear difference equations [188]. They incorporated epidemiological data from South India to solve the equations and concluded that tuberculosis dynamics would remain stable in South India. Waaler extended this work in a number of later papers, most of which focused on evaluating the cost-effectiveness of different interventions [192-196]. By that time, Brogger had already built on Waaler's first model to evaluate different disease control strategies [197]. The transmission function in this model was significantly more complex than in Waaler's. Also in 1967, Revelle introduced a system of nonlinear differential equations to model tuberculosis dynamics with the goal of assessing cost-effectiveness of various interventions in developing countries [198]. Ferebee constructed a model of tuberculosis dynamics in the US using a structure similar to that of Waaler [199], first model studying tuberculosis in a developed country. There were few tuberculosis modelling papers in the 1970s and 1980s, though Sutherland et al. published an age-structure model in 1982 that took into account reinfection [18]. The model was used to estimate risks of developing pulmonary disease for the three disease pathways: recent infection, latent infection and reinfection. Model results showed sex-differences in disease risks.

Many tuberculosis modelling papers appeared in the 1990s. There were several simulation models that offered numerical analyses and projections of tuberculosis burden in various countries or evaluation of control measures [200-203]. In 1995, Blower et al. published a paper on the transmission dynamics of *M. tuberculosis* using two analytically tractable ordinary differential equation (ODE) models. Authors suggest some of the decline in the tuberculosis incidence in developed countries resulted from

a long-term decline in the tuberculosis epidemic [183]. In 1997, Vynnycky and Fine published a partial differential equation (PDE) model of tuberculosis dynamics in the UK, extending the work of Sutherland et al. [18, 19]. Like Sutherland's, this model is age-structured and takes into account reinfection.

In the 2000s, many more models have been published, including some dynamical models of tuberculosis that are analytically tractable. These models offer solutions to threshold conditions and other insights into disease dynamics. The trade-off is they are not always proven realistic with data. In 2000, Aparicio et al. developed a transmission model that accounted for close and casual contacts in what they called a 'generalized households model.' This compartmental model first stratified the population into 'active' and 'inactive' groups, based on whether or not persons were in a cluster—representing households or networks of close contacts—with an active case of tuberculosis. Members of the active clusters were at higher risk of disease, being at risk for both infections from the active cases in their cluster, and also casual contact infection, which can be passed from any infectious case in the population. Inactive cluster members could only be infected by casual contacts, and were at a much lower risk of infection. Aparicio et al. concluded that casual contacts significantly contribute to disease incidence. Other examples may be found in later work by Aparicio and colleagues, as well as work by Song and colleagues [191, 204, 205].

Other important models have been developed but are not described here because of limited relevance to the present study. These include HIV-focused models, models of strain competition, and models of drug resistance [179, 206-213].

### 2.5.2 Tuberculosis Dynamics in the UK

In 1997, Vynnycky and Fine published a model of tuberculosis in white males in England and Wales from 1900 – 1990, extending a model published by Sutherland et al. [18]. Vynnycky and Fine's model was age-structured and had eight infection state compartments, including a reinfection and reinfection compartments. The population modelled was restricted to white ethnic males. Flows between compartments were dictated by a system of PDEs, approximated by ordinary differential equations (ODEs)

and solved by Euler's method. Instead of dynamical transmission modelling, a time-dependent ARI was assumed to apply to the entire population. Fitting the model to notification data allowed estimation of age-specific risks of developing disease via the three disease pathways, recent infection, latent infection, and reinfection for white ethnic males.

Extensions of this work include a 1998 paper that estimated the net reproductive number for tuberculosis in the UK, and discussed implications of the complex natural history of tuberculosis on interpretation of the net reproductive number [214]. In 1999, Vynnycky and Fine used model results and estimates of the ARI in the UK to describe trends in the effective number of contacts for infectious tuberculosis cases over the 20$^{th}$ century [56]. Vynnycky and Fine also used the model to estimate secular trends in age-specific lifetime risks of developing tuberculosis once infected, and serial intervals of the disease in the UK [16].

In 2002 Pitman et al. published a compartmental model of tuberculosis dynamics in the UK to look at the impact of vaccination, chemotherapy, and preventive therapy on disease incidence from 1993 to 1990 [182]. The authors concluded that of these interventions, preventive therapy was most important to infection transmission, as estimates for the duration of infectiousness were long (~2-5 years) and infectiousness begins before symptoms are present. Pitman later extended this compartmental model of tuberculosis dynamics in England and Wales, stratifying the population into five risk groups, based on infection exposure risk (R. Pitman, personal communication). One limitation of both models is that there were many parameters estimated by fitting the models to notification data, many of which were correlated.

### 2.5.3 Tuberculosis and Genotyping Data

Some modelling work has been done to aid interpretation of genetic typing data for tuberculosis isolates, including both dynamic transmission models and statistical models. Most relevant to this thesis, Vynnycky et al. and Murray both used detailed transmission dynamic models to simulate genotyping data [9, 161, 175]. In 2001 Vynnycky et al. used a deterministic, compartmental model to examine the effects of

age and study duration on the relationship between genotype clustering and recent transmission. Analyses excluded immigrants, those clustered with immigrants, and extra-pulmonary cases. This study highlighted the importance of age in determining how well clustering predicts the extent of recent transmission, as discussed above in Section 2.4.2. A later extension of this model showed that the trend in ARI strongly influences the relationship between clustering and recent transmission, also discussed above [161]. The model did not account for reduced rate of change of the molecular marker when infection was latent. However, as there is biological reason for, and now documented evidence of, a much slower rate of change of genotype change in latent infection [130], the relationships between clustering and recent transmission derived from these studies may be altered if more realistic mutation processes are taken into account.

Murray developed an IBM of tuberculosis dynamics to study cluster size distribution and proportion of isolates clustered under different assumptions about control strategies, HIV, prevalence of latent infection, and population age structure [175]. The IBM tracked individual strains of *M. tuberculosis*, though did not allow for strain evolution (the model was only run for four years) and was not calibrated to data. Results show that different control measures can have vastly different effects on proportion clustered and cluster size distribution, depending on whether they reduce infection transmission or reduce reactivation disease. Murray extended this model to examine the effect of the sampling proportion on estimates of recent transmission [8], confirming earlier results, as discussed above in Section 2.4.2. Neither study by Murray took into account migration into the study population or mutation of the genotype profiles.

Some studies have used statistical models to interpret genotyping data. For example, Tanaka et al. used a stochastic model in combination with approximate Bayesian computation to estimate transmission parameters from genotyping data [215]. A paper by Grant et al. modelled the evolution speed of VNTR loci, although their study does not directly address the interpretation of genotyping data for estimating recent transmission [156].

### 2.5.4 IBMs for Tuberculosis

There are relatively few tuberculosis studies using IBMs, though this type of model is increasingly being used. These models are typically used to study contact structure in the population, incorporate non-homogeneous mixing, and study *M. tuberculosis* strain types according to genotype or drug sensitivity [189, 207, 216-218]. There have also been at least two intra-host models of tuberculosis, both of which model individual cellular interactions to understand host response to *M. tuberculosis* infection [219, 220]. Some notable IBMs relevant to this thesis are reviewed below.

In 2007, Cohen et al. published an individual-based network model studying the effects of reinfection on disease dynamics, accounting for non-homogeneous contact patterns among individuals in the model [221]. Authors concluded that non-random mixing leads to increased importance of reinfection on tuberculosis burden, even in low-incidence settings, due to localized clusters of cases. In 2008, Colijn et al. published a related model studying various effects of heterogeneity in contact structure on infection transmission, also in low-incidence settings. Their results suggested that localized outbreaks can result directly from non-random contact structure in the population, even in the absence of variation in host susceptibility or transmissibility of strains. The process for fitting models to data was not specified, and assumed to be *ad hoc*. In 2008, Cohen et al. published an extension of this IBM, which examined the accuracy of estimates of the burden of drug resistant tuberculosis [222] and concluded previous estimates may have been too low. As described above, Murray developed an IBM in 2002 looking at cluster size distribution of M. tuberculosis isolates under many different model assumptions [175]. This model also employed a discrete event simulation algorithm [8].

More recently, Guzzetta et al. published an IBM of tuberculosis dynamics applied to data from the US to study the effects of age-structure and preferential mixing on the ability of a model to simulate observed data [223]. They compared their fully-specified IBM to a deterministic compartmental model and also a simpler IBM, all three of which were based on the structure of Vynnycky and Fine's compartmental model. Results showed that the fully-specified IBM, with spatial structure and contact networks based

on households, schools and workplaces, fit observed data best. Other results include an estimated cumulative risk of 15% for primary disease over the first five years of infection and age-dependent disease risk estimates which showed that individuals over 50 years of age have a four-to-eight-fold increased risk of reactivation disease compared to the general population.

### 2.5.5 Tuberculosis and Immigration

In 2002, Wolleswinkel-van den Bosch et al. used a life-table model of tuberculosis in Dutch natives to assess the impact of immigrants on the Dutch native population [224]. Authors defined the contact rate in their model as 'the average number of infections generated by an infectious case', and this was multiplied by the number of infectious cases to get the ARI. They concluded that by 2030 at least 60% of Dutch tuberculosis cases will have been infected by an immigrant. The model did not explore the impact of the Dutch native cases on the incidence of tuberculosis in immigrants.

McCluskey and van den Driessche used a compartmental, differential equation model solved analytically to study tuberculosis dynamics with immigration [225]. When immigrants enter the population only as susceptibles, the disease-free equilibrium is globally stable when threshold conditions are met– i.e. the reproduction number is below one. However, when immigrants are allowed to enter as infected and diseased, the disease persists endemically—there is no disease-free equilibrium. In 2008, Zhou et al. used a deterministic compartmental model to study tuberculosis in Canada. The population was divided into Canadian-born and immigrants, with transmission occurring within and between the two groups. Their results showed that only about 5% of cases in foreign-born persons were infected in Canada. The model was limited by including very little heterogeneity apart from the two birthplace groupings – for example, no realistic demography, age, or sex differences were taken into account.

Also in 2008, Jia et al. published a study of the impact of immigration on tuberculosis incidence in a compartmental model, also applied to data from Canada [226]. The population was stratified into Canadian-born individuals and immigrants, using three parameters to describe transmission within and between the two groups. This model

was not designed to take into account transmission from natives to immigrants. Authors concluded that immigration will allow the disease to persist in areas where the net reproductive number is below one in the general population, due to importation of infections. Authors advise tuberculosis models should take into account immigration, as leaving immigrants out significantly alters disease dynamics, a conclusion supported by several modelling studies [224, 225, 227].

## 2.6 Summary of Modelling Considerations

### 2.6.1 Tuberculosis Modelling in This Thesis

Important considerations for tuberculosis modelling will be question- and setting-dependent. In this study, one of the most important features of any model used will be its capacity for the simulation of genotyping data. To this end, an IBM would be the most straightforward and flexible approach to handling these data. Strain types can number in the thousands and the population is dynamic, with continual mutation processes, making handling these difficult with a compartmental model. The use of an IBM also allows the flexibility of extending the model in the future to include a more sophisticated contact structure and virtually any type of data. The IBM also simplifies the incorporation of highly stratified parameter values. Lastly, stratification of the population into many groups results in small and variable population sizes for some of these groups, providing another reason for using individual-based modelling. Stochastic effects are easily handled in an IBM.

Another important feature to include in the model is the effect of migration, important because foreign-born persons in the UK now account for the majority of tuberculosis cases. Therefore, migration and explicit division of the population by birthplace is essential for better understanding of tuberculosis dynamics in the UK. Also, allowing age-structure and realistic demography is essential for modelling tuberculosis given the long-term dynamics of the disease [228]. The evidence for the role of reinfection in tuberculosis dynamics necessitates its inclusion [19, 221]. The simulation of large population sizes, specifically England and Wales, makes speed of computation a necessary factor in considering models to use.

### 2.6.2 Other Considerations

New model code was developed to answer my research questions in preference to adapting one of the several models reviewed above. I did this for two main reasons. First, model code is not freely available for any of the above models. Second, the model descriptions, particularly for IBMs, are not sufficient for complete and reliable reproduction of the models. Therefore I have developed my own modelling software, adapted from an existing IBM for HIV dynamics, as described in Chapter 3. This IBM

was available for my use, well described, and allowed me to avoid having to re-design the scheduling architecture of the IBM.

Freely available model code that has been rigorously tested and well documented could serve epidemiologists and the broader modelling community. Especially in some countries with limited budget for designing models themselves, adapting the code of others could potentially save resources and time [229]. Of course, it may not be simple to adapt code, but making code freely available at least makes that option possible. Freely available code can also help ensure reproducibility of results. Finally, it can help identify errors in the code in two ways. Firstly, code that is made freely available would be well-tested and well-documented in preparation for release, which, in itself, reduces errors. Secondly, others reading papers or adapting the model for their own purposes could help discover programming or other errors. These consumers could also suggest improvements to the model and advancements of the science.

These ideas are not new. The immensely popular and successful realm of open source software, which spans many disciplines and applications, serves as a model for these principles. Also, more recently there has been an 'open science' movement which calls for more publishing of data, methods, software, and results in enough detail so that other scientists can reproduce and build on current work [230, 231].

## 2.7 Observed Data Used for the Study

Some of the observed data used in this study are presented here because they are referenced in several chapters. These include notifications from England and Wales, from 1999 – 2009 and notifications from the West Midlands, from 2007 – 2011. Each dataset is described below. Other observed data used in this thesis include genotyping data from the West Midlands, from 2007 – 2011. However, these are not described below because significant processing is required before data can be presented. These data and processing steps are detailed in Chapter 6.

### 2.7.1 Notification Data from England and Wales

The numbers of tuberculosis case notifications recorded in the Enhanced Tuberculosis Surveillance (ETS) system each year from 1999 – 2009 were provided by the HPA Tuberculosis Section. Notifications for each year were stratified by age category, sex, birthplace, and disease site. Age stratification divided notifications into 11 categories for age: under five years, 5 – 9 years, 10 – 14 years, 15 – 19 years, 20 – 24 years, 25 – 34 years, 35 – 44 years, 45 – 54 years, 55 – 64 years, 65 – 74 years, and 75 years and above. Disease site categories are pulmonary and non-pulmonary. Birthplace categories are UK-born, SSA-born and other foreign-born (OF-born). There was no temporal precision beyond the year of the case report. The notifications were reported to HPA and stored in the ETS database, where non-cases were de-notified and cases reported more than once were de-notified (for more details on the ETS database, see Section 2.3.2.3.1).

#### *2.7.1.1 Data processing*

Cases with unknown age, sex, disease site, or birthplace were excluded from stratified notification data provided by HPA, though the total number of cases each year was provided. The ratio of total cases to stratified cases (those stratified by age, sex, birthplace, and disease site) was calculated for each year and used to adjust stratified cases for missing data. It was assumed that cases with missing information did not differ from cases with complete information, so notifications for each stratified category were multiplied by the ratio of total cases to stratified cases for each year.

Table 2-1 gives the total cases and stratified cases for each year, as well as the ratio of total cases to stratified cases used to adjust the numbers of cases in each stratified category. The majority of cases with missing data were missing information on country of birth. After adjustment for cases with missing data, case notifications were grouped into age categories consistent with those used by the HPA Tuberculosis section. These age categories are: 15 years and under, 15 – 44 years, 45 – 64 years, and 65 years and over.

Table 2-1: Tuberculosis case notifications in England and Wales from 1999 – 2009. The number of total cases reported and the number of cases stratified by age, sex, birthplace, and disease site for each year, as provided by the HPA Tuberculosis section, is shown. The ratio of total cases to stratified cases was used to adjust each category (defined by age, sex, birthplace, and disease site) for cases excluded from stratification due to missing information.

| Year | Total cases reported | Cases stratified by age, sex, birthplace and disease site | Ratio of total cases: stratified cases |
|---|---|---|---|
| 1999 | 5701 | 4943 | 1.15 |
| 2000 | 6264 | 5324 | 1.18 |
| 2001 | 6457 | 5520 | 1.17 |
| 2002 | 6783 | 6004 | 1.13 |
| 2003 | 6823 | 6181 | 1.10 |
| 2004 | 7166 | 6531 | 1.10 |
| 2005 | 7879 | 7131 | 1.10 |
| 2006 | 7902 | 7036 | 1.12 |
| 2007 | 7826 | 7090 | 1.10 |
| 2008 | 8109 | 7450 | 1.09 |
| 2009 | 8500 | 7751 | 1.10 |

### *2.7.1.2 Trends in the number of notifications*

From 1999 – 2009, the total number of cases reported in England and Wales increased from 5,701 to 8,500. However, the number of notifications and trends in these numbers were markedly different among the three different birthplace categories, as shown in Figure 2-4—Figure 2-9. UK-born cases made up around 40% of notifications in 1999—over 2300 cases—though this decreased to 25% of total notifications and about 2200 cases in 2009. The number of notifications in UK-born cases for most age groups were roughly constant from 1999 – 2009, though notifications decreased in those aged 65 years and above for both sexes. In males, cases in that age group decreased from over 400 in 1999 to about 270 in 2009. In females the decrease was from about 290 to 200 cases. These trends are shown in Figure 2-4—Figure 2-5.

For OF-born cases, case notifications for most age groups and both sexes increased from 1999 – 2009. Over this time, the proportion of total cases in OF-born individuals rose from approximately 44% in 1999 to 53% in 2009. There was an especially marked increase in those aged 15 – 44 years, with the number of cases in males more than doubling in that age group, from 768 cases in 1999 to 1735 in 2009. In females, cases increased less dramatically, from 315 cases in 1999 to 488 in 2009. These trends are shown in Figure 2-6—Figure 2-7.

In SSA-born cases, the number of case notifications also increased for most age groups and both sexes from 1999 – 2009. The total number of notifications increased from about 870 cases in 1999 to over 1,900 in 2009. The vast majority of cases occured in those aged 15 – 44 years. In this age group, notifications rose over the period 1999 – 2009 from about 670 cases to more than 1400 cases in 2009. There were around 900 cases each for both males and females at the peak incidence in 2006 and 2005, respectively, falling to around 700 cases in each group respectively in 2009. There was a steadier upward trend in those aged 45 – 64. The number of cases more than doubled from 1999 to 2009 for both males and females, with around 150 cases in each group in 2009. There were very few cases in the youngest and oldest age categories,

although the number of cases did rise in these groups. These trends are shown in Figure 2-8—Figure 2-9.

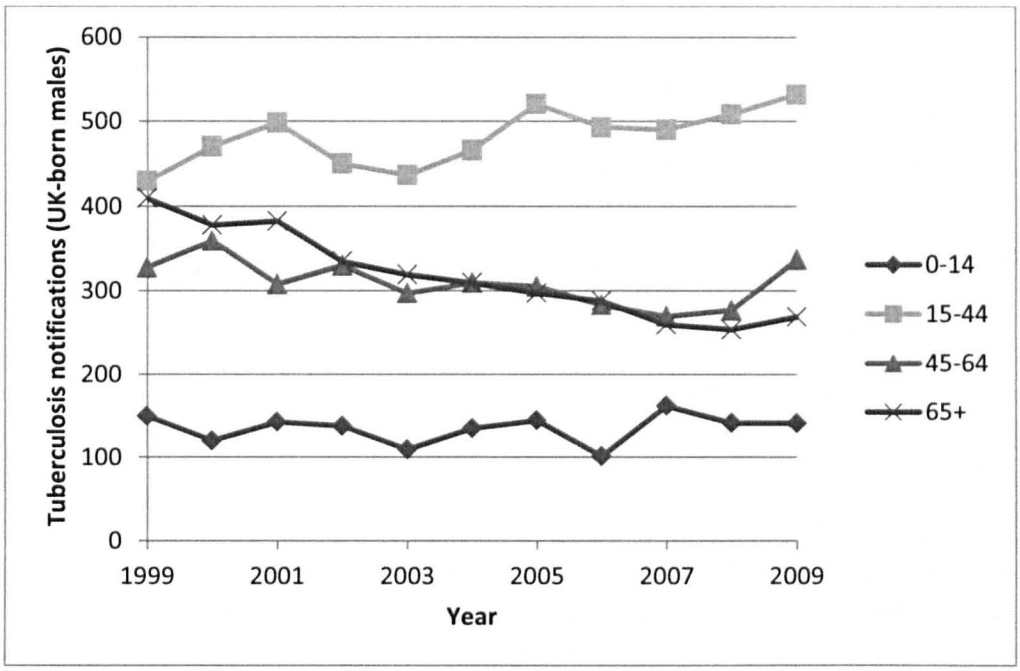


Figure 2-4: Tuberculosis notifications for United Kingdom-born males in England and Wales by age category, 1999 – 2009.

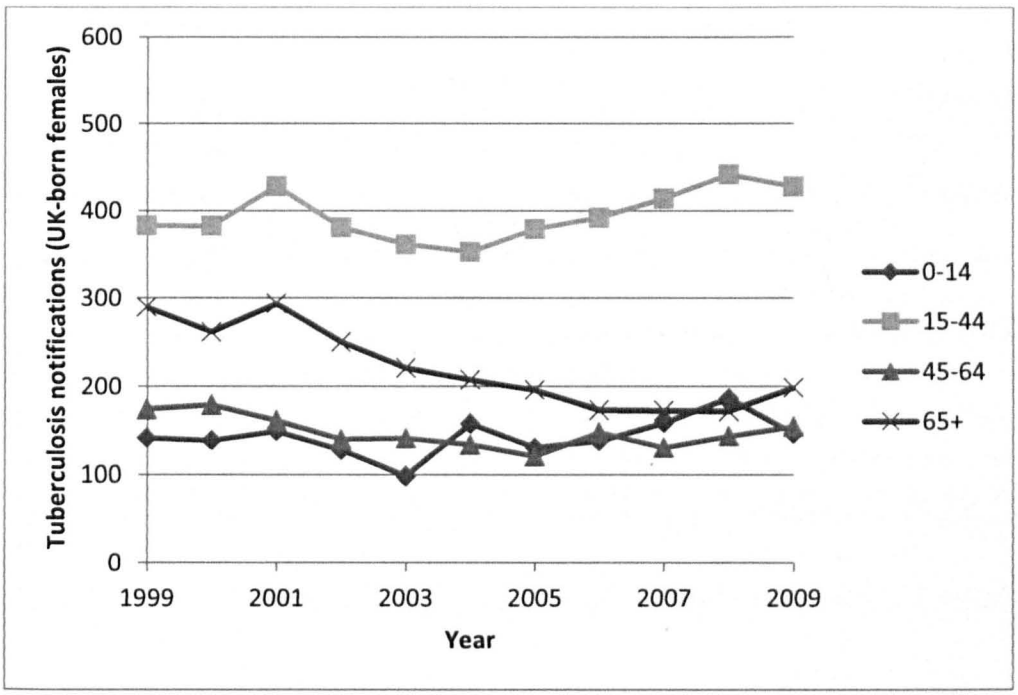


Figure 2-5: Tuberculosis notifications for United Kingdom-born females in England and Wales by age category, 1999 – 2009.

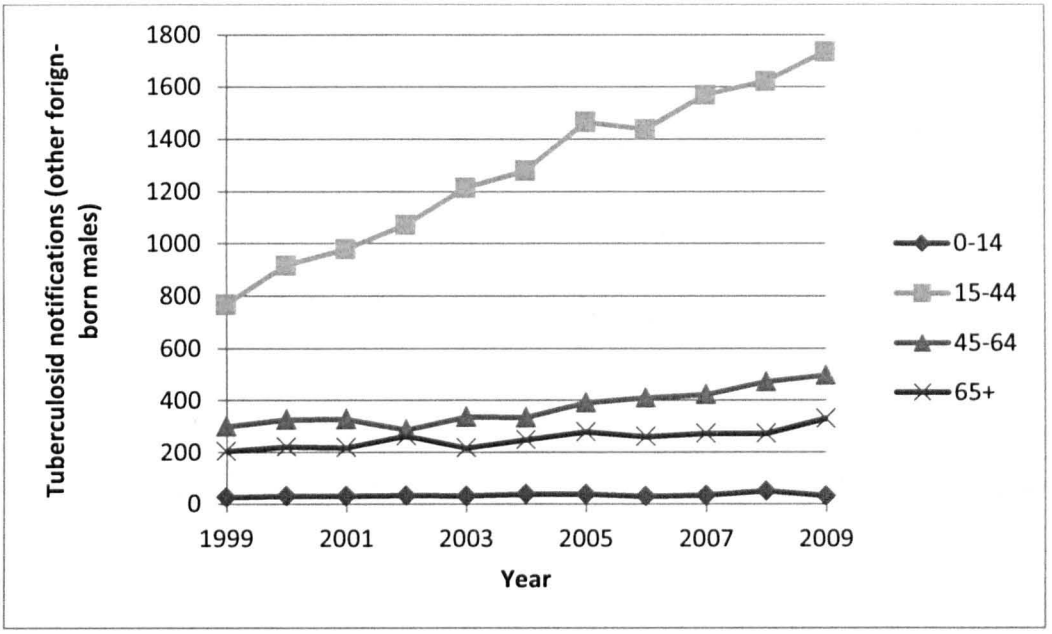


Figure 2-6: Tuberculosis notifications for other foreign-born males in England and Wales by age category, 1999 – 2009.

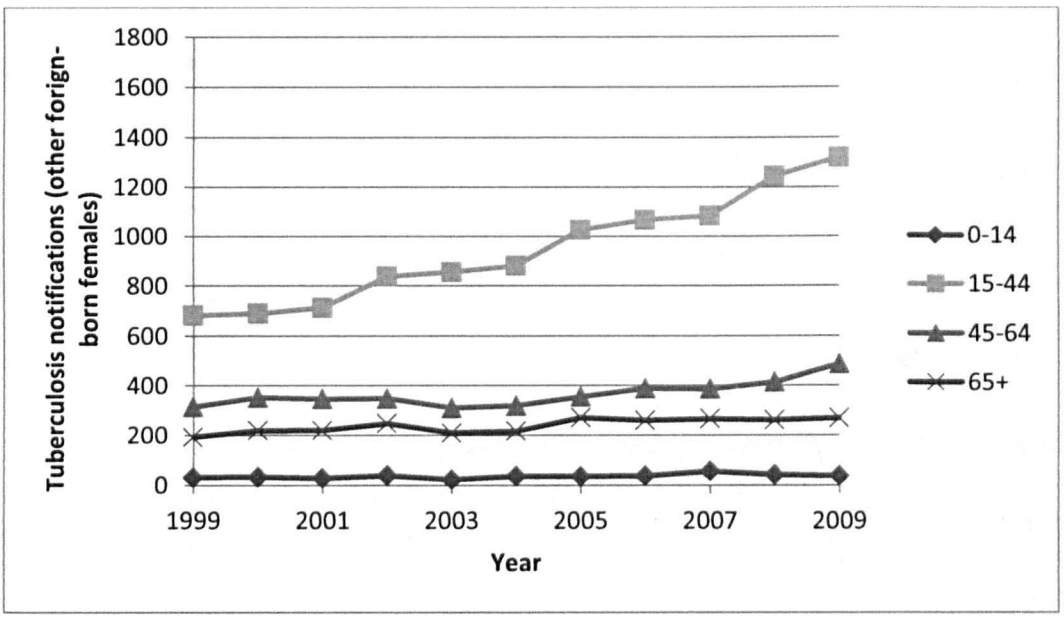


Figure 2-7: Tuberculosis notifications for other foreign-born females in England and Wales by age category, 1999 – 2009.

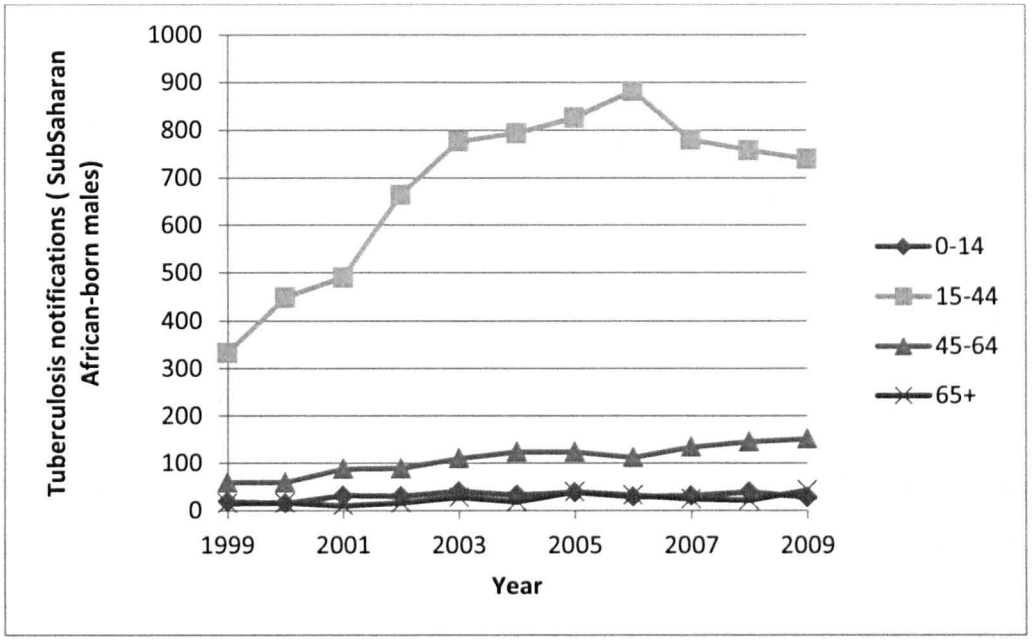


Figure 2-8: Tuberculosis notifications for Sub-Saharan Africa-born males in England and Wales by age category, 1999 – 2009.

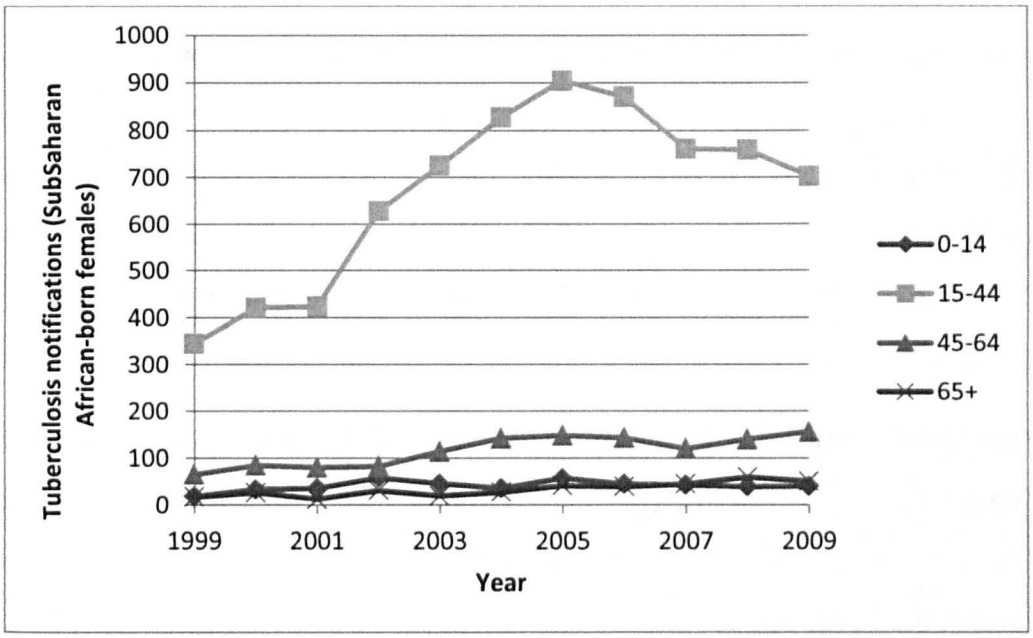


Figure 2-9: Tuberculosis notifications for Sub-Saharan Africa-born females in England and Wales by age category, 1999 – 2009.

### *2.7.1.3 Notification rates*

Notification rates were calculated for each year by dividing the total number of case notifications in each demographic category by the estimated population size for that category. Estimated population sizes were obtained from analysis of the Labour Force Survey (LFS), as described in Chapter 4, Section 4.1.3.1. As per convention for tuberculosis, notification rates were multiplied by 100,000 to obtain rates per 100,000 population per year.

There were vast differences between notification rates among UK, OF, and SSA-born individuals. For UK-born individuals, notification rates were low, ranging from about two cases per 100,000 to 13 cases per 100,000 across all age groups and both sexes from 1999 – 2009, as shown in Figure 2-10 and Figure 2-11. Notification rates were relatively stable from 1999 – 2009, although this varied somewhat with the different age and sex categories. As it was also seen in the number of notifications, there was a downward trend in the case notification rate in those aged 65 years and above. The case notification rate in males fell from about 13 cases per 100,000 in 1999 to eight cases per 100,000 in 2009. In females, there was less of a decrease, as the rate fell from about seven cases per 100,000 in 1999 to about five cases per 100,000 per year in 2009. Generally, notification rates in UK-born males were higher than notification rates in UK-born females.

In OF-born individuals, notification rates were significantly higher than those in UK-born. Rates ranged from a minimum of about 14 cases per 100,000 to more than 130 cases per 100,000 per year across all age groups and both sexes, as shown in Figure 2-12 and Figure 2-13. The trends in notification rates by age showed a different pattern than trends in UK-born cases. In OF-born cases, those aged 0 – 14 years had the lowest notification rates, mostly constant and averaging around 20 cases per 100,000 per year in males and 25 cases per 100,000 per year in females. The highest notification rates were found in those aged 15 – 44 years. Notification rates in this age group increased from 1999 – 2009, though only from 97 cases per 100,000 per year to 112 cases per 100,000 per year in males and from 77 cases per 100,000 per year to 86 cases per 100,000 per year in females. Rates in 45 – 64 year-olds and those 65 years

and older were slightly lower than those aged 15 – 44 years for both males and females. These rates also increased slightly from 1999 – 2009.

In SSA-born individuals, notification rates were significantly higher still. Rates ranged from a minimum of 57 per 100,000 per year to a maximum of over 300 per 100,000 per year across all age groups and both sexes, as shown in Figure 2-14 and Figure 2-15. As with OF-born individuals, notification rates were highest in those aged 15 – 44 years, with the lowest rate being nearly 170 cases per 100,000 and the highest being over 300 cases per 100,00 per year. Also, the notification rate in this age group increased overall, with a peak notification rate in 2003 for males and 2006 for females.

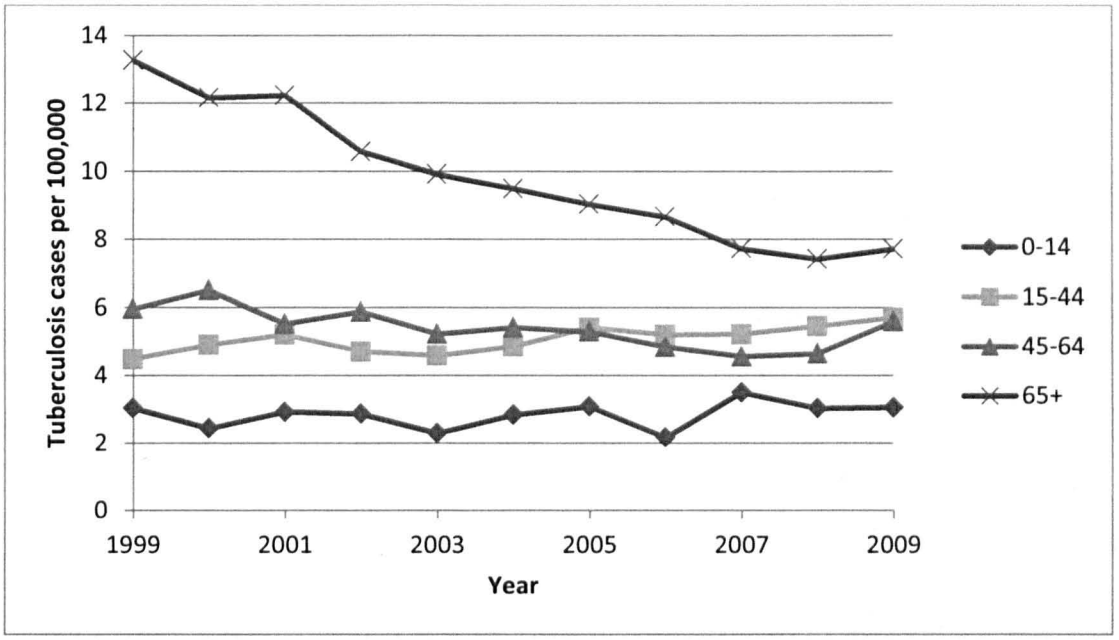


Figure 2-10: Tuberculosis notifications per 100,000 population per year for UK-born males in England and Wales, 1999 – 2009.

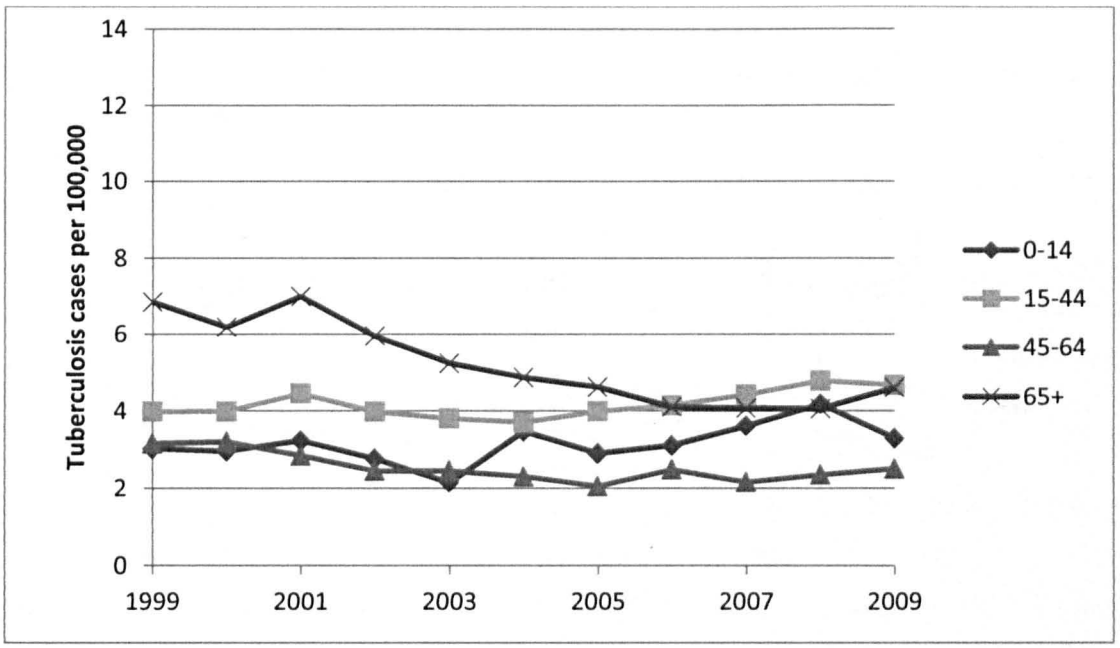


Figure 2-11: Tuberculosis notifications per 100,000 population per year for United Kingdom-born females in England and Wales, 1999 – 2009.

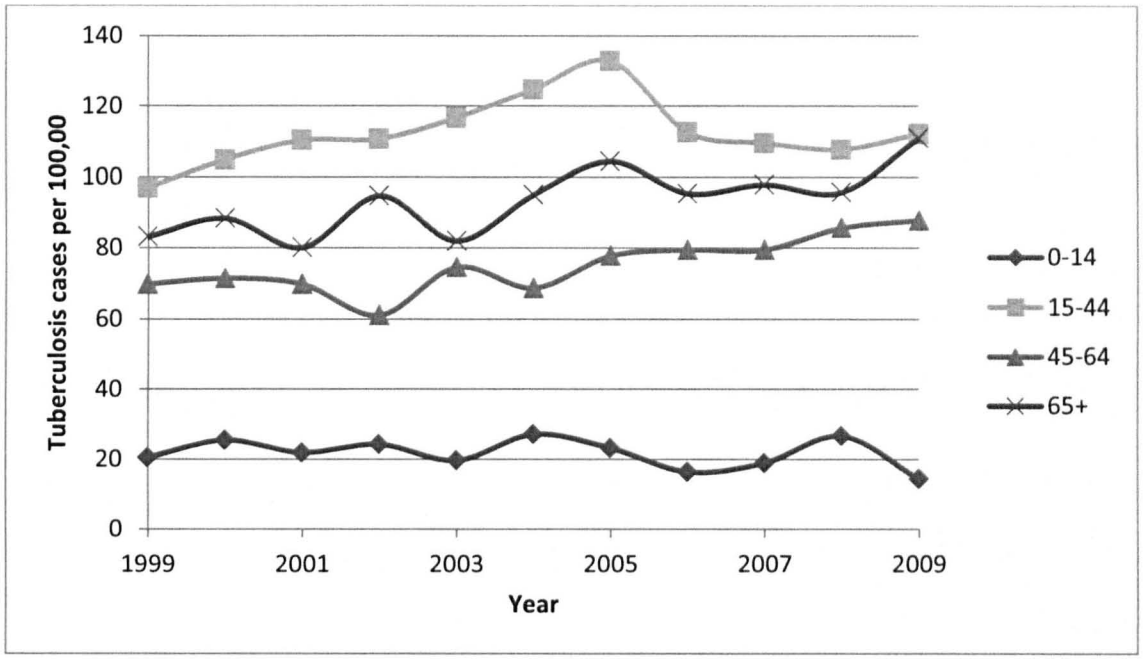


Figure 2-12: Tuberculosis notifications per 100,000 population per year for other foreign-born males in England and Wales, 1999 – 2009.

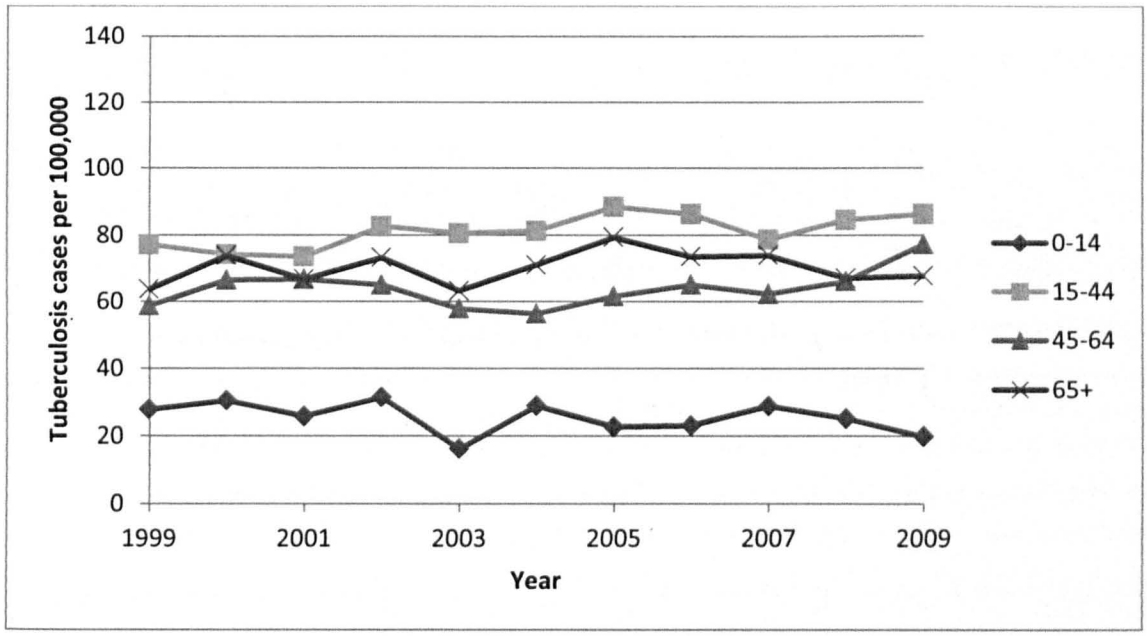


Figure 2-13: Tuberculosis notifications per 100,000 population per year for other foreign-born females in England and Wales, 1999 – 2009.

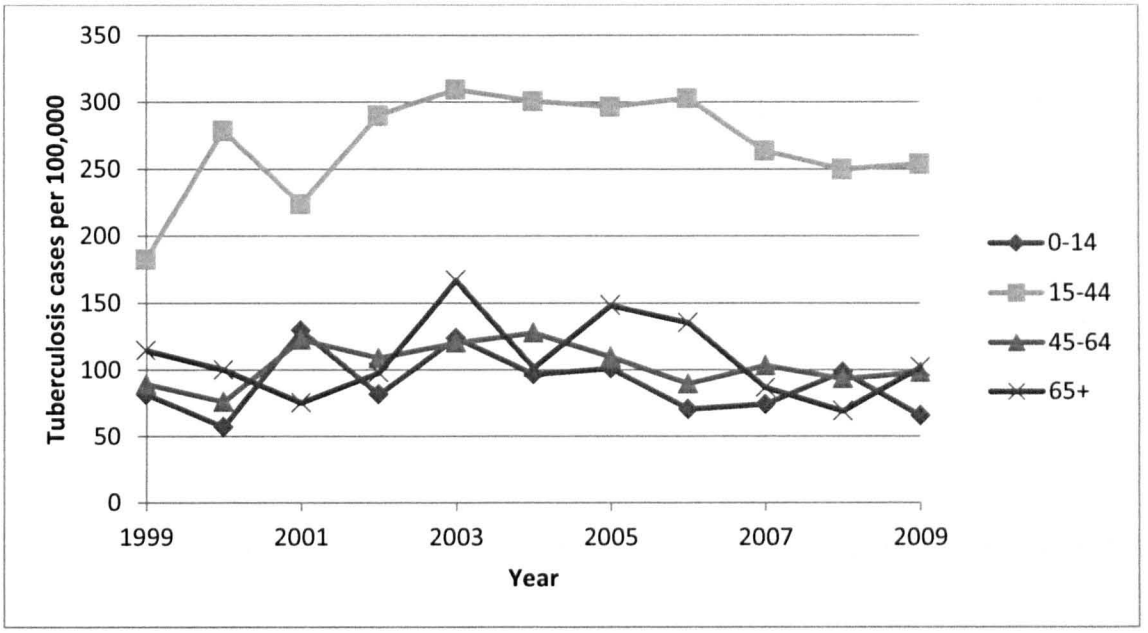


Figure 2-14: Tuberculosis notifications per 100,000 population per year for Sub-Saharan Africa-born males in England and Wales, 1999 – 2009.

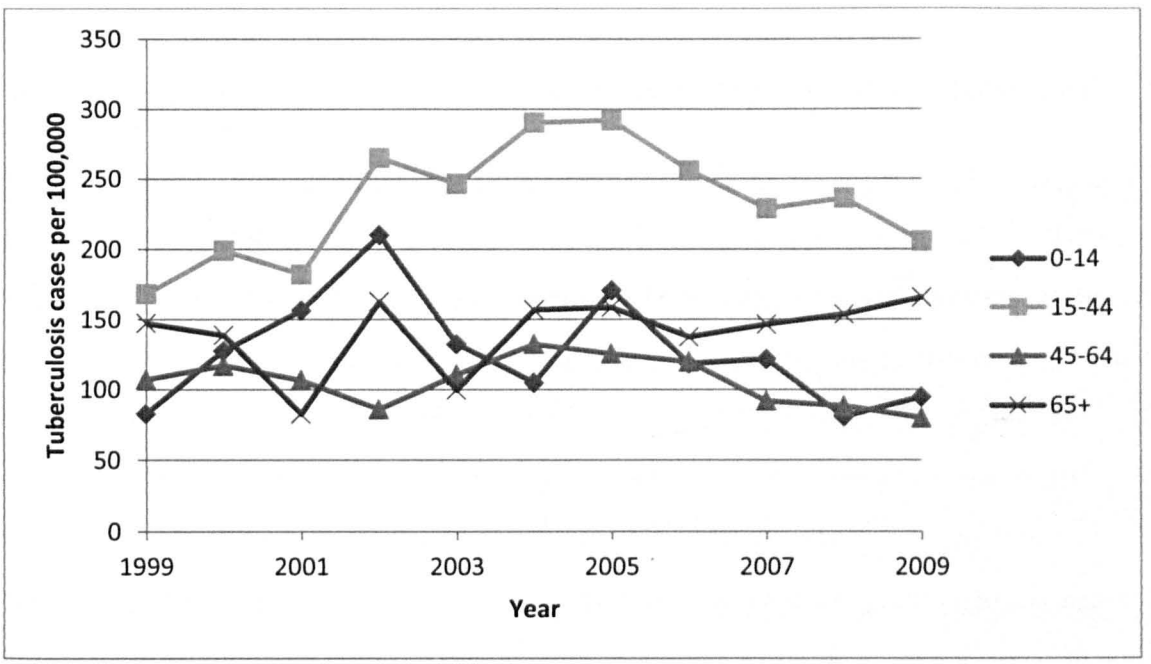


Figure 2-15: Tuberculosis notifications per 100,000 population per year for Sub-Saharan Africa-born females in England and Wales, 1999 – 2009.

### 2.7.2 Notifications from the West Midlands

Tuberculosis notification rates in the West Midlands from 2007 – 2011 were calculated for comparison with model output in Chapter 7. ETS records for each notified case from the West Midlands from 2007-2011 were obtained from the HPA Tuberculosis Section. The records were used to tabulate the number of cases by age, sex, and birthplace. Population denominators for notification rates were obtained by analysis of the LFS data, as described in Section 4.1.3.1.

The overall notification rate in the West Midlands for this time period was about 17 per 100,000 population, though as with the rates in England and Wales, UK-born rates were much lower than foreign-born rates. The UK-born notification rate over this period was 6.5 per 100,000 population while the foreign-born rate was 110 per 100,000 population. The UK-born notification rate in the West Midlands over this time period is higher than that seen in England and Wales from 1999 – 2009, which was about 4.5 per 100,00 overall. For the foreign-born cases in the West Midlands from 2007 – 2011, overall notification rates were also higher than those in England and Wales from 1999 – 2009, where the rate was about 102 per 100,000 population.

The notification rates stratified by age are shown in Figure 7-1 for UK-born males and Figure 7-2 for UK-born females; rates for foreign-born individuals are in Figure 7-3 for males and Figure 7-4 for females. Across age and sex categories, UK-born rates vary between about 1 per 100,000 to about 12 per 100,000, while foreign-born rates vary between about 20 per 100,000 to about 160 per 100,000. Particularly for UK-born cases, notification rates for the West Midlands vary more from year-to-year than England and Wales rates, likely due to the smaller numbers of cases in the West Midlands. This difference could also be partly due to more uncertainty in population size estimates and fluctuation in those estimates.

Some trends in the notification rates for UK-born cases in the West Midlands differ from those seen in the whole of England and Wales. Notably, the rates for UK-born males and females aged 15 – 44 years are higher than in England and Wales over a similar time period. For males in the West Midlands, this ranges from about 8 – 10 per

100,000 from 2007 – 2011. For females, this ranges from about 7 – 11 per 100,000 from 2007 – 2011. In England and Wales, these range from about 4 – 6 per 100,000 for males and 4 – 5 per 100,000 for females from 1999 – 2009. Rates for those aged 15 – 44 in the West Midlands are also higher than the rate for those aged 65 years and above, whereas in England and Wales, the notification rate in those aged 65 years and above are generally higher than that for those aged 15 – 44 years. This trend is always true for males and usually true for females, though the differences in notification rates between the age categories have decreased over time.

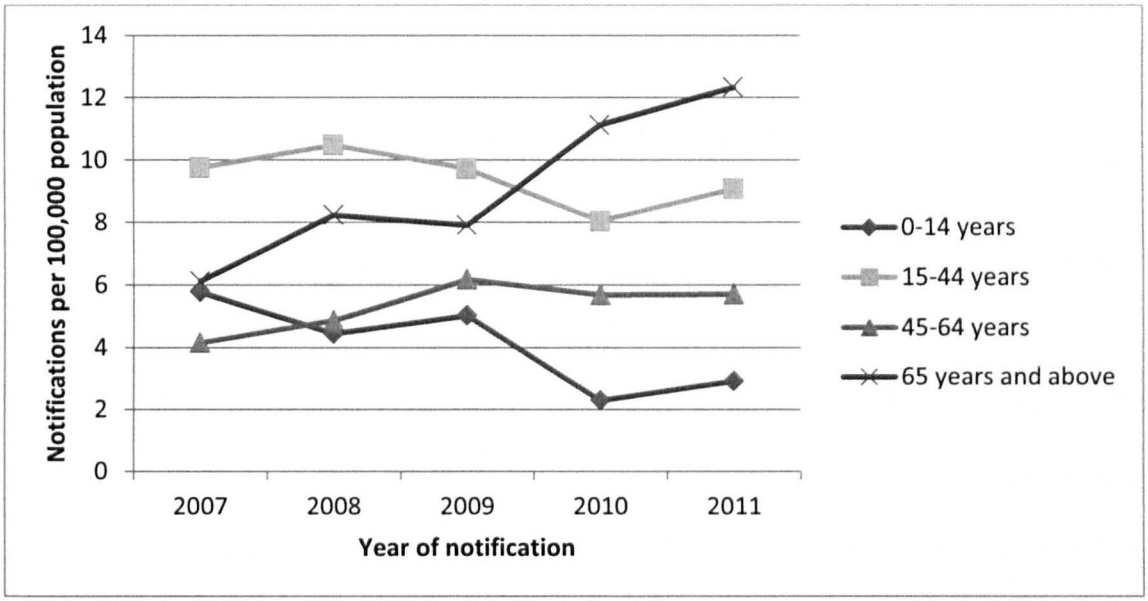


Figure 2-16: Tuberculosis notifications per 100,000 population for United Kingdom-born male cases in the West Midlands, 2007 – 2011.

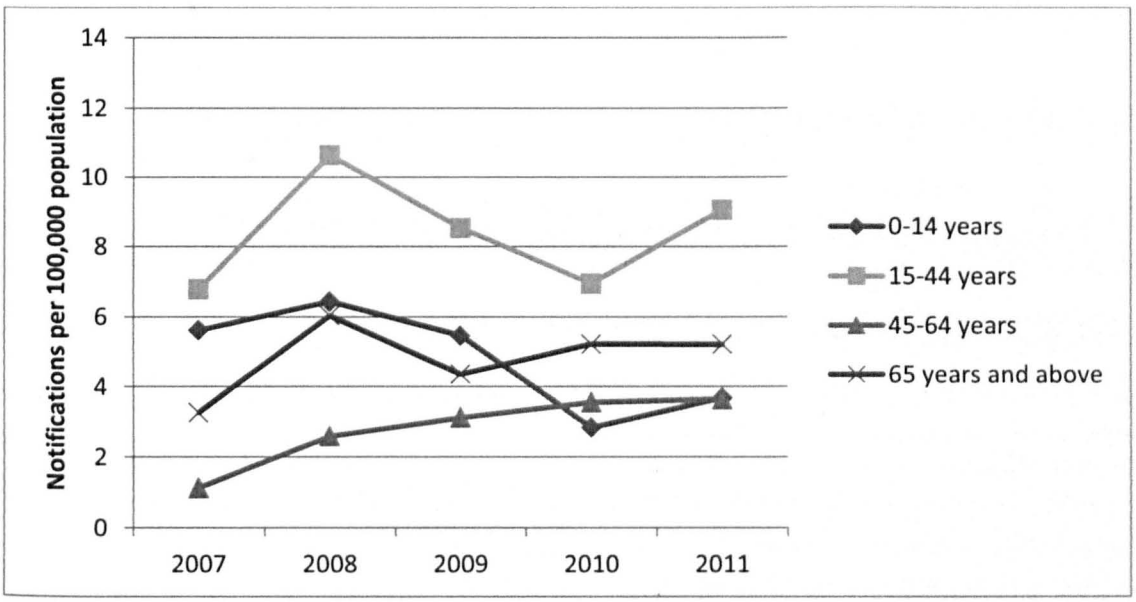


Figure 2-17: Tuberculosis notifications per 100,000 population for United Kingdom-born female cases in the West Midlands, 2007 – 2011.

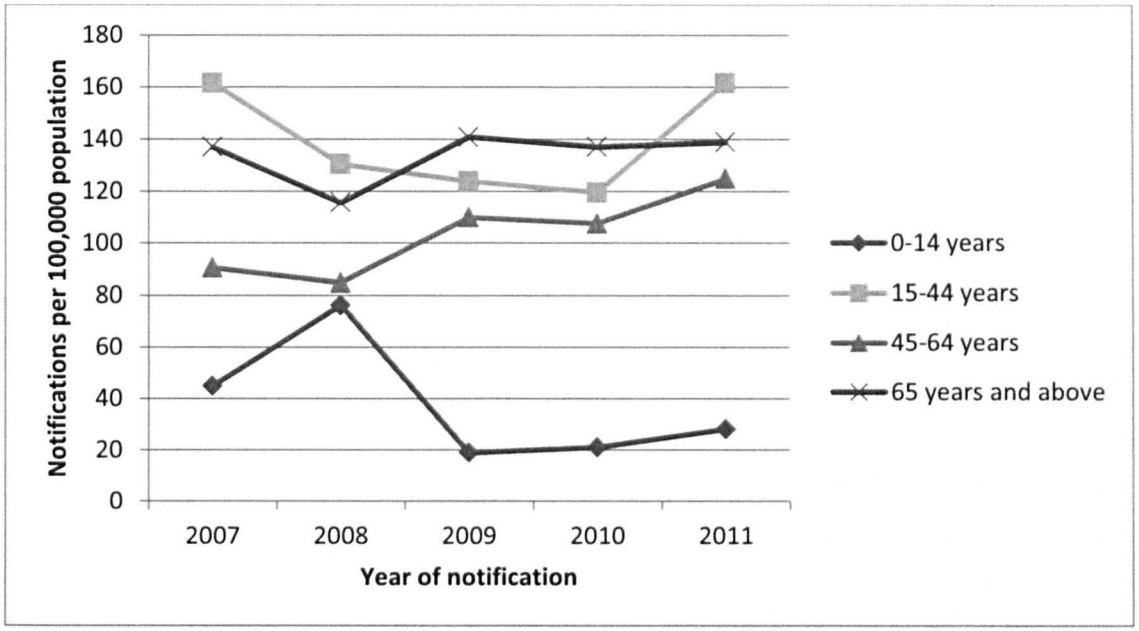


Figure 2-18: Tuberculosis notifications per 100,000 population for foreign-born males in the West Midlands, 2007 – 2011.

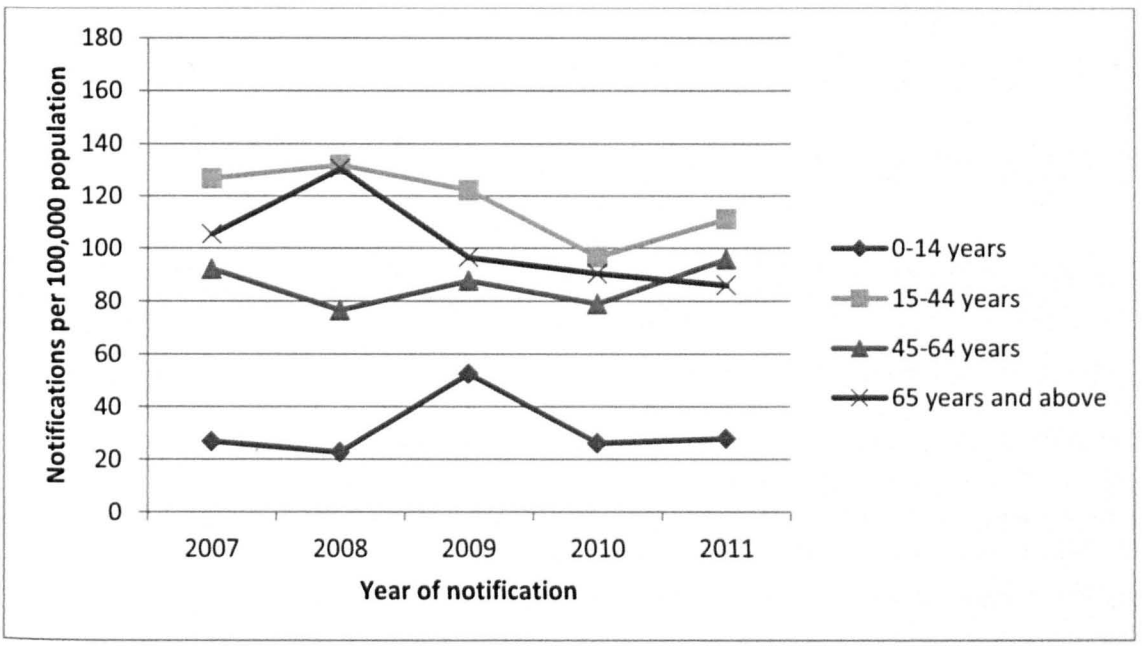


Figure 2-19: Tuberculosis notifications per 100,000 population for foreign-born females in the West Midlands, 2007 – 2011.

# 3 Model Description and Modelling Methods

This chapter describes the model of tuberculosis dynamics in the UK that was developed for answering research questions in this thesis and fulfilling objective one of the study. The model is individual-based so that information can be stored for each person in the simulated population. The model is described below following the Overview, Design concepts, and Details (ODD) protocol for describing IBMs [232, 233]. Following the ODD model description is the account of my own contributions to development of the model, which was adapted from the work of C. Lehman. In addition, I describe my contributions to the individual-based modelling methods that were developed alongside this work, in collaboration with C. Lehman. Lastly, an account of the steps taken for verification of the model and an overview of model validation and fitting methods are also presented.

## 3.1 ODD Model Description

### 3.1.1 Introduction

IBMs can be particularly difficult to describe clearly because of their inherent complexity and because equations and model diagrams are often less applicable to them than to classical models. The ODD protocol aids clarity of description and reproducibility of the model, allowing others to reproduce the model from its description. Furthermore, the standardized and objective approach to description makes it easier to compare different models. Although designed for IBMs in ecology, the protocol is used here because it is appropriate for any type of IBM and because there is currently no standard for describing IBMs in epidemiology.

The elements of the protocol are summarized in Figure 3-1, which is taken from the original publication of the protocol [232]. The guidelines of the ODD protocol were followed as closely as possible. As permitted in the protocol, I omitted some categories of the Design concepts section that were not applicable to this model. Further changes to the protocol included a brief review of the purpose or definition of major sections, which were added so that those sections were clear without prior knowledge of the protocol. Finally, parameter values and data sources, which ODD authors strongly recommend including in the model description, are omitted here. Parameter values and data sources are instead detailed in Chapter 4.

| | |
|---|---|
| **Overview** | **Purpose** |
| | **State variables and scales** |
| | **Process overview and scheduling** |
| **Design concepts** | **Design concepts** |
| **Details** | **Initialization** |
| | **Input** |
| | **Submodels** |

Figure 3-1: Elements of the Overview, Design concepts, and Details protocol for describing individual-based models. Figure was reproduced from the publication describing the protocol [232].

### 3.1.2 Purpose

The purpose of this model is to help understand tuberculosis transmission dynamics and epidemiology in the UK, including how genetic strain typing data can be used to estimate the proportion of tuberculosis cases that are due to recently transmitted infections. However, the full-complexity model as described below, which includes modelling of genotyping data, was not used for both applications of the model in this thesis.

Firstly, a slightly simpler version of the model was applied to tuberculosis dynamics in England and Wales for estimating disease risks used in subsequent applications of the model and for identifying plausible scenarios for transmission parameters and assumptions about the infection status of migrants upon entry to the UK. This version of the model was also used to provide an estimate of the proportion of tuberculosis cases due to recent transmission, independent of genotyping data. Alterations to the model for this application are described in the methods section of Chapter 5.

Secondly, the full-complexity model was used to study tuberculosis dynamics in the West Midlands, where model outputs were fit to molecular epidemiological data to estimate the proportion of tuberculosis cases in the UK caused by recent transmission and to help interpret strain typing data.

### 3.1.3 Entities, State Variables, and Scales

This section describes the structure of the model system, including entities, state variables, and scales. Entities are the low-level units of interest simulated by the model and state variables are the attributes that characterize those entities. Scales of the model include the temporal and spatial scales covered by the simulation.

#### *3.1.3.1 Entities*

The only model entity simulated is the individual person.

#### *3.1.3.2 State variables*

Individuals are characterized by demographic attributes, as well as infection-related attributes. Demographic attributes do not change over the course of the simulation and include birth date, sex, and birthplace. Birthplace has three categories: UK, Sub-Saharan Africa (SSA), and other foreign (OF). Although it would be more realistic to further stratify birthplace, the data needed for a more detailed stratification is lacking. Also, further stratification would result in fewer cases for each demographic category, increasing stochasticity in model output and thereby hindering parameter estimation.

Infection-related attributes include an individual's infection state, date of infection, place infection was acquired (UK or foreign), HIV status (positive or negative), smear status (positive or negative), and an identifier for the strain type involved in infection. Infection states are: *Uninfected*; *Immune*; *Recent Infection*; *Reinfection*; *Latent Infection*; *Primary Disease* (pulmonary and non-pulmonary); *Reactivation Disease* (pulmonary and non-pulmonary); and *Reinfection Disease* (pulmonary and non-pulmonary). HIV status is only considered for SSA-born individuals because the HIV prevalence in that group is higher than in other groups (see Chapter 4, Section 4.2.4).

The strain type profile identifier is simply a different number for each distinct VNTR profile.

Individuals in the *Uninfected* class have never been infected with *M. tuberculosis* and are assumed susceptible to infection. *Immune* individuals have been effectively vaccinated and are immune to infection for the duration of the simulation. Those in the *Recent Infection* state have had a first infection with *M. tuberculosis* that was acquired less than five years previously. Those in the *Reinfection* state have a second or subsequent infection which was acquired less than five years previously. Those in the *Latent Infection* state have an infection acquired more than five years previously or have recovered from disease due to their most recent infection, which may have been acquired less than five years previously. Those with *Latent Infection* are assumed susceptible to a new infection (reinfection). *Primary Disease* is disease caused by a *Recent Infection*. *Reinfection Disease* is disease caused by a *Reinfection*. *Reactivation Disease* is caused by a *Latent Infection*. The three disease types are each split into two categories by disease site, pulmonary and non-pulmonary (see Chapter 2, Section 2.1 for definitions). Pulmonary disease may be infectious, depending on smear status. Non-pulmonary disease is assumed non-infectious. Transition through these states is shown in Figure 3-2 and is discussed further in the next section.

Figure 3-2: Infection state and event diagram for the model. Arrows show transitions among the 11 infection states, with symbols to indicate the events involved in transitions. Symbols in parentheses indicate events that do not change the infection state of the individual. Table 3-1 provides a key to symbols, with a full description of events in Section 3.1.8.

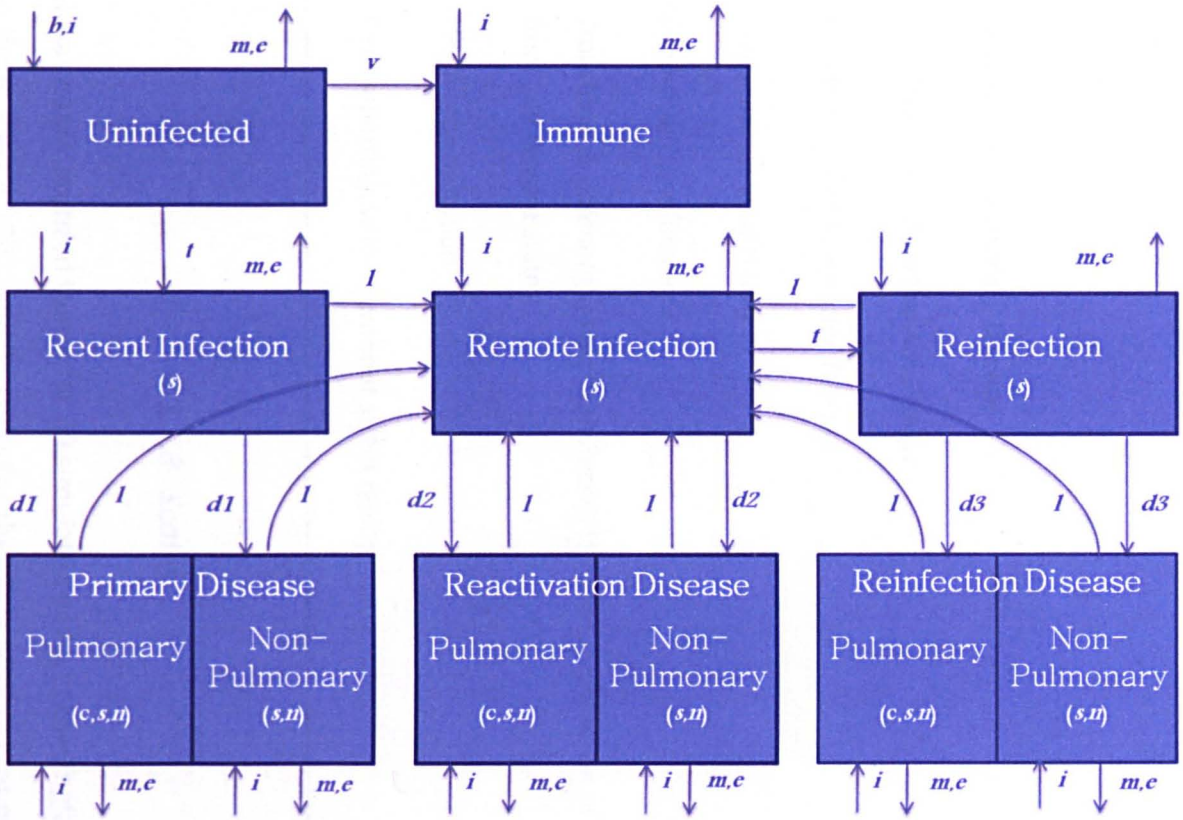

Table 3-1: Model events shown in Figure 3-2. Events are also detailed in Section 3.1.8.

| Symbol | Event |
|---|---|
| *b* | Birth |
| *m* | Death |
| *i* | Immigration |
| *e* | Emigration |
| *v* | Vaccination |
| *d1* | Development of *Primary Disease* |
| *d2* | Development of *Reactivation Disease* |
| *d3* | Development of *Reinfection Disease* |
| *c* | Transmission of infection |
| *t* | Acquisition of infection |
| *l* | Transition to *Latent Infection* from *Recent Infection/Reinfection* or from any disease state |
| *s* | Strain type mutation |
| *n* | Case reporting, with or without strain typing |

#### *3.1.3.3 Scales*

The model was parameterized to run for 29 simulated years in the England and Wales application (from 1981 – 2009) and 31 years for the West Midlands application (from 1981 – 2011). The simulations began before the model output was fit to observed data so that strain type distributions equilibrated and the effects of assumptions about infection and disease prevalence at initialization were minimized. The model was allowed to run for 18 and 26 simulated years before the model output was fit to observed data for the England and Wales and West Midlands simulations, respectively.

The model processes events continuously in time, moving forward from one event in time to the next, instead of by arbitrary time steps. This is often referred to as a 'discrete event simulation.' The closest equivalent to a time step is the average inter-event time, which is the average amount of simulated time, not clock time, that passed between events. The average inter-event time varies throughout the simulation and depends on factors such as the population size simulated. In a typical run for England and Wales, which has over 50 million simulated individuals, events occur on average about once per 15 simulated seconds. In a typical run for the West Midlands, which includes around five million simulated individuals, events occur on average about once per two simulated minutes. Possible events are described below in Section 3.1.4.

The model does not take into account spatial structure explicitly, although in the two applications of the model different geographical regions are covered.

### 3.1.4 Process Overview and Scheduling

This section describes how model entities are processed, including what processes or 'events' can occur and how these are scheduled. The events are briefly described below and detailed in Section 3.1.8. Since this is an event-based simulation, events are processed individually and chronologically from a schedule of pending events. Possible events include: birth; death (due to tuberculosis or causes other than tuberculosis); immigration; emigration; vaccination; acquisition of infection; transition to *Latent Infection* (from *Recent Infection/Reinfection* or recovery from disease); progression from infection to disease; transmission of infection; strain type mutation; and reporting of a disease case (with or without strain typing). The different events possible for an individual will depend only on their infection state, though probabilities of events occurring can change with factors such as the calendar year, age, sex, birthplace, and HIV status. Events are summarized in the paragraph below, with a complete description of each event and the individuals eligible for the event in Section 3.1.8.

The possible events for individuals in each of the 11 infection states are briefly described here and depicted in Figure 3-2, with supporting information in Table 3-1. Any individual can emigrate or die. In addition, *Uninfected* individuals can be vaccinated, moving to the *Immune* state, or become infected, moving to the *Recent*

*Infection* state. *Immune* individuals never change infection states. Those in *Recent Infection* and *Reinfection* states can develop disease, moving to one of the *Primary Disease* or *Reinfection Disease* states, or transition to *Latent Infection*. In addition, their infection strain type can mutate, which does not involve a change of infection state. Those in the *Latent Infection* class can: develop disease, moving to one of the *Reactivation Disease* states; be reinfected, entering the *Reinfection* state; or have their strain type mutate, which does not involve a change of infection states. Individuals in the three pulmonary disease states may be infectious and may infect others. In addition, they can recover from disease, moving to the *Latent Infection* state. They can also be reported as a case or have their strain type mutate, neither of which involves a change of infection state. Those with non-pulmonary disease are assumed to be non-infectious and will not transmit their infection to others, but can recover to the *Latent Infection* state. They can also be reported as a case or have their strain type mutate, neither of which involves a change of infection state. When individuals are born, they enter the *Uninfected* state, and when they immigrate into the population, they may enter any of the 11 infection states.

Events in the schedule of future events are limited to those that can be scheduled ahead of time. This includes all possible events except the acquisition of an infection, which must happen at the time of transmission so that the susceptibility of individuals at that time is known (see Section 3.1.8.6 for more on the transmission event). Future event times are determined using random draws from appropriate cumulative probability distributions for different events, discussed in more detail in Section 3.1.5.3.

At model initialization, detailed in Section 3.1.6, all individuals are placed in the schedule of future events exactly once, for the time of their earliest pending event. This feature of the schedule makes it simple to remove individuals from the population or to cancel or modify events. Once initialization is complete, the simulation proceeds from the simulated start time to end time, processing events chronologically. Event processing generally leads to scheduling of a new event for the individual involved, unless the individual is scheduled for death or emigration from the population. Sometimes, certain scheduled future times for the individual may be modified during event processing. For example, when an individual is scheduled to develop disease,

that event may lead to a new scheduled time of death, since disease brings an additional mortality risk. Or in the case of infection transmission, processing the transmission event for the infectious individual may lead to modifying, adding, or cancelling of events for the recipient of infection. For example, a newly infected individual may be scheduled to develop disease instead of the event which was formerly scheduled for them.

### 3.1.5 Design Concepts

Design concepts are important concepts behind the design of the model. Several of the design concepts listed in the ODD protocol do not apply to this model and are omitted. Sections included here are called: 'basic principles', which describes the fundamentals behind the design of the model; 'interactions', which describes interactions between entities in the model; and 'initialization' which describes how the model is initialized.

#### *3.1.5.1 Basic principles*

Many of the conceptual elements of this model are based on previous work. The infection states and transitions through the states of this model are based on a model of tuberculosis dynamics in England and Wales by Vynnycky and Fine [19]. However, Vynnycky and Fine use a system of partial differential equations solved numerically, while here I use an event-based, IBM that is related to a similar set of partial differential equations. Part of the reason the IBM was used is that when additional complexity is added, particularly when genetic strains are included and arbitrary probability distributions for the infection status of migrants entering the simulation, this model transcends what can be written down in equation form. Furthermore, compared to Vynnycky and Fine, this model considers different time periods, different demographic groups of the population, and makes use of enhanced surveillance data available more recently. The time period is more recent, concerning the last ten years of tuberculosis trends in the UK. Because in this time period the majority of tuberculosis cases in the UK occurred in foreign-born persons, this demographic group is included in the model. In contrast, only white ethnic males were included in Vynnycky and Fine's model. Finally, the infection process is modelled explicitly here, rather than through a force of infection applied to the population, as in Vynnycky and Fine's model.

This model is individual-based so that genetic strain type profiles, of which many hundreds or thousands are possible, can be stored in the model for each infected individual. This allows for comparison of model results to data from strain typing of *M. tuberculosis* isolates. The model is intended to reproduce genotype clustering patterns seen in observed cases to help reveal how these patterns may have been generated and to compare the level of genotype clustering to the proportion of cases due to recent transmission in the simulation.

Beyond the requirements for storing genetic strains, the IBM makes narrowly timed events—such as vaccination at age 13—easy and accurate to accomplish. In addition, early tests indicated that the IBM version ran faster than finite difference approximations to partial differential equations with the same number of individuals and level of accuracy.

#### *3.1.5.2 Interaction*

Interactions between individuals in the model are limited to contacts between individuals which lead to the transmission of infection. The transmission event is detailed in Section 3.1.8.6. There are no indirect interactions between individuals in the model; individuals behave independently of one another apart from transmission of infection.

#### *3.1.5.3 Stochasticity*

Stochasticity is used throughout the model in assigning event times. All event times, apart from the transition to *Latent Infection*, which happens exactly five years after infection by definition (see Section 3.1.8.8), are assigned by random selection of a time to event based on the cumulative probability distribution of event times for that individual. This random element allows for variation in the time to events among individuals rather than fixing times at the average time to event for all individuals. For example, the average duration of disease is six months. Instead of assigning every individual to have a disease duration of six months, an individual's duration of disease is determined by drawing from the inverse cumulative exponential distribution with a mean of six months. Most random draws of event times in the simulation are implemented using a new algorithm developed for the model for choosing random

numbers from any type of distribution, as below in Section 3.3.2 and in a recently published paper [234].

Stochasticity is also used in the model for choosing a random individual to be infected during transmission (Section 3.1.8.6), assigning a disease as pulmonary or non-pulmonary, assigning disease as smear-positive or smear negative (Section 3.1.8.7), assigning demographic and infection-related attributes in model initialization (Section 3.1.6), assigning those attributes for migrants upon entry to the model (Section 3.1.8.3), assigning sex at birth (Section 3.1.8.1), determining whether an individual will be vaccinated and whether the vaccination will be effective (Section 3.1.8.5), and determining whether an individual's disease case will be reported (Section 3.1.8.10).

#### *3.1.5.4 Observation*

The main outputs from the model includes information on reported tuberculosis cases and population sizes. Reported tuberculosis cases include a subset of all tuberculosis cases in the simulation, designed to reflect the reality that not all cases are reported (see Section 3.1.8.10). When an individual develops tuberculosis, they are assigned to be reported or not to be reported. If assigned to be reported, the reporting time precedes recovery, death and emigration. Reported cases are stored by year of reporting and also stratified by age class (0 – 14, 15 – 44, 45 – 64, or 65 years and above), sex, birthplace, and disease site (pulmonary or non-pulmonary).

For the simulation of genetic strain typing data, a subset of reported cases will include genotyping data. Cases are randomly assigned to be genotyped or not, according to a disease site-specific probability that typing occurs. Those typed are included in the genetic data output which is used for cluster analyses.

Population sizes are also reported throughout the simulation both to obtain notification rates (cases per 100,000 population) and to check that population sizes compare well with observed population sizes. Population sizes are totalled twice each year and stored by age class, sex, and birthplace. Simulated notification rates are compared to observed notification rates using automatically generated plots for inspection.

### 3.1.6 Initialization

Model initialization involves constructing the relevant model population, England and Wales or the West Midlands, for the start of the simulation in 1981. All individuals are randomly assigned demographic and infection-related attributes based on data and assumptions which are described in more detail in Chapter 4. Demographic attributes are assigned according to the 1981 census. From the census data, the number of individuals from each birthplace (UK, SSA, OF) is estimated for each five-year age class for both males and females. An exact age within these categories is specified by randomly drawing an age across the age group, resulting in a uniform distribution of ages across the five-year age class. This may be is unrealistic for older age classes. However because there are approximately 25 years of simulation for equilibrations, and more importantly, because tuberculosis parameters will not differ much between small age categories, the uniform distribution of individuals within five-year age categories is unlikely to have any impact on model outcomes.

After demographic attributes are assigned, infection-related attributes are assigned according to probabilities that depend on demographic attributes. Of infection-related attributes, infection state is assigned first, since other attributes will depend on this state. This assignment is made using the probability of each of the possible infection states, which is in turn based on assumptions about the disease prevalence in 1981, vaccination practices in the UK and abroad, and the annual risk of infection experienced over the lifetime of UK-born and foreign-born individuals living in the UK in 1981 (for details see Chapter 4). These probabilities depend on age, sex, and birthplace. After the infection state is assigned, additional characteristics are assigned for those infected or diseased. These include time of infection, place of infection (UK or abroad), HIV status (positive or negative) and genetic strain type of infection. Time of infection is randomly chosen between zero and five years for *Recent Infection* and *Reinfection*. Time of infection has no impact on disease risk for those with *Latent Infection* so this is always set to exactly five years previously. The place of infection is assumed to be abroad for all migrants. All individuals are assumed HIV-negative in 1981.

The strain type is assigned by randomly drawing from a distribution of strain types by birthplace—either UK-born or foreign-born. The strain type distribution for UK-born is

used only at model initialization, while the distribution for foreign-born is used to assign strain types to all migrants to the UK throughout the simulation.

Once their demographic and infection-related attributes are assigned, future event times for each individual are calculated based on their attributes in the same way that event times are calculated throughout the simulation. The earliest event for an individual is scheduled and placed in the schedule of pending events, while any later events for the individual are stored. This constrains the size of the list of future events and speeds operation [235]. When all individuals in the population are assigned attributes and scheduled for exactly one event, initialization is complete and the simulation begins.

### 3.1.7 Input

According to the ODD protocol, 'input data' are external data sources that represent processes that change over time. As these are external to the model, they are not considered parameter values *per se*.

Input data for the model include immigration data and birth data. Immigration is external to the model because of the nature of immigration as coming from outside; no natural process internal to the model catalyses immigration. Immigration data consisted of numbers of migrants by age, sex, and birthplace (UK, SSA and OF) for each year that the model is run. As there was no temporal precision in the data beyond the year of entry, immigrants entered the model population at evenly spaced time intervals throughout the year. See Section 3.1.8.3 for more information on the immigration event.

Birth is external to the model because fecundity of females was not modelled explicitly. Rather, for exact correspondence with reality, the actual numbers of births recorded in England and Wales and the West Midlands were generated in the simulation. Birth data included the numbers of births by sex for each year. Like migration data, there was no additional temporal precision and so births were also evenly spaced throughout the year. See Section 3.1.8.1 for more information on the birth event in the model.

### 3.1.8 Submodels

The 'submodels' below correspond to events in this model. They are described by first explaining how and which individuals arrive at the event and then explaining what the event entails. Figure 3-1 supplements the descriptions with a visual depiction of events in the model, with supporting information found in Table 3-1. Note that these descriptions are general and provide information on the algorithms for events but not specific parameter values and data sources for creating cumulative probability distributions of event times, which are detailed in Chapter 4. To avoid repetition, references to individual sections in Chapter 4 are omitted.

#### *3.1.8.1 Birth*

The number of births per year, $b_y$, is used to schedule a birth every $1/b_y$ years. At birth, all individuals are assigned to the *Uninfected* class and their exact time of birth is recorded. Next, the individual's sex is randomly assigned using sex ratios for births each year. Random times for death, emigration, and vaccination (if applicable) are drawn using appropriate probability distributions and the earliest of these is scheduled.

#### *3.1.8.2 Death*

Individuals in the model can die either from tuberculosis or from other causes. For causes other than tuberculosis, time of death is randomly assigned for each individual using probabilities drawn from cohort mortality data for each year of the simulation. Times of death are assigned at birth for those born in the UK during the simulation, or at model initialization or time of immigration for other individuals.

To generate a random life expectancy from birth, a random number is drawn, compared to the cumulative probabilities of death for the newborn's sex and year of birth, and an age is calculated by linearly interpolating between integer ages given in the cumulative distribution table [234]. For assigning remaining life expectancy for individuals at model initialization or for migrants entering the population at various ages, the cumulative probability table for their sex and birth year is truncated and rescaled to only allow ages from their present age to 121 years. A random life expectancy is then drawn from the rescaled table. In both cases, an exact time of death is chosen by linearly interpolating between integer ages given in the cumulative distribution table. Once a time of death from causes other than tuberculosis is

assigned, either at birth, initialization, or time of immigration into the population, it will not change throughout the simulation.

When a person develops active disease, it is assumed they are at increased risk of death. This assumption is incorporated into the model by assigning a probability of death due to tuberculosis, which is stratified by calendar year, age class, and type of disease (pulmonary or non-pulmonary) at disease onset. If a case is randomly assigned to die from tuberculosis, the individual is assigned a time of death due to tuberculosis *before* the currently stored times for death from other causes, recovery from disease, and emigration.

Death removes an individual from the population and does not catalyse any other event, except a function that transfers the identification number of the highest-numbered individual within the same birthplace to the identification number of the removed individual. This reassignment keeps the array of individuals contiguous, aiding random selection of an infection target during transmission [236].

#### *3.1.8.3 Immigration*

Immigrants are added to the population in fixed intervals, similar to births. The number of immigrants per year is $i_y$ and an immigrant arrives every $1/i_y$ years, assuming the total immigrants for the year are spaced evenly throughout the year.

At immigration, individuals are first assigned a birthplace (UK, SSA, OF), based on the probabilities of each. After this, sex is assigned based on probabilities for each birthplace group. Next, age is assigned conditional on sex and birthplace. Similar to age assignment at population initialization, age is assigned by randomly drawing an age class and then drawing an age from within the age class, assuming individuals are uniformly distributed throughout the age class. Lastly, HIV status is assigned 'negative' to UK-born and OF-born individuals. For SSA-born, HIV status is randomly assigned according to assumptions about the probability of HIV infection by sex and calendar year.

The infection state is then assigned based on probabilities of infection states for migrants, stratified by birthplace, age, sex and year of entry to the UK. To simplify this process, probabilities are assigned for eight infection states, rather than 11, by

combining pulmonary and non-pulmonary disease classes and assigning those later. For those assigned to *Uninfected*, there are no special considerations and they are processed similarly to a birth in the model. For *Immune*, there are also no special considerations.

For those assigned to have infection or disease, some additional steps are taken. Those assigned to *Recent Infection* or *Reinfection* classes are first assigned a time of infection, from zero to five years before the present, chosen randomly from a uniform probability distribution. This assignment allows rescaling of the cumulative distribution for randomly selecting a time to disease progression (See Section 3.1.8.7). Those immigrants assigned to a disease category are randomly designated as pulmonary or non-pulmonary cases. The time of disease onset is assigned to be the present time. Finally, a strain type is assigned by a random draw from the strain type distribution for migrants for all individuals assigned to have infection or disease.

After the infection state has been assigned, future event times based on that infection state and the individual's demographic and infection-related attributes are generated. The earliest of these events is scheduled and any later times are stored.

#### *3.1.8.4 Emigration*

When an individual is born or immigrates into the population, an emigration time is assigned based on estimated annual rates of emigration by birthplace, calculated from emigration data. Emigration times are assumed to be exponentially distributed. The emigration time does not change once assigned.

Emigration removes an individual from the population and does not catalyse any other event, except, like death, it results in the transfer of identification numbers of the highest-numbered individual within the same birthplace to the identification number of the removed individual to keep the array of individuals contiguous [236].

#### *3.1.8.5 Vaccination*

A vaccination time is assigned at birth for those born in the UK and at the time of migration for foreign-born individuals aged 13 and under, if probabilities allow. Vaccination only happens for an individual in the model if a random number drawn is less than the product of vaccine efficacy and vaccine coverage for a given year.

Ineffective vaccinations are insignificant for modelling purposes and never scheduled to happen. For individuals who will be vaccinated effectively, vaccination is scheduled to occur at a random time uniformly distributed between 13 and 14 years of age. If an individual becomes infected before vaccination happens, the vaccination is assumed ineffective and never happens in the simulation.

At vaccination, individuals move to the *Immune* class and retain lifelong immunity, only eligible to die or emigrate from the population. They are scheduled for the earlier of these, at the times previously assigned, since these are not affected by vaccination.

#### *3.1.8.6 Transmission*

Only individuals with smear-positive pulmonary disease are assumed infectious, though in reality, smear-negative cases can also be infectious. When an individual develops smear-positive pulmonary disease, they are assumed to transmit infection to an average of $c$ others each year in a purely susceptible population, or have an average of $c$ 'effective contacts' per year. The time until each infectious case transmits an infection to another person is assigned assuming exponentially distributed times with a mean of $1/c$ per year.

At the time of transmission, a second individual is chosen as the transmission target, either randomly from within the same birthplace as the infectious person or randomly from the whole population. If a random number drawn is less than the probability that the infection target is in the same birthplace as the infectious individual, $pcc$, the infection target will be a randomly chosen individual from within the infectious individual's own birthplace (simplified to UK or foreign). This is meant to loosely represent the proportion of contacts who are from the same household or are other regular contacts. If the random number is greater than or equal to $pcc$, the infection target is chosen from the entire population, including individuals within the same birthplace. The individual targeted for infection is only eligible to be infected if they are in the *Uninfected* or *Latent Infection* states, which are the states assumed to be at risk of (re)infection. Otherwise, the transmission does not occur.

If the infection target is eligible to be infected, that individual's infection state is changed to *Recent Infection* if they were previously *Uninfected* or changed to *Reinfection* if they were previously in the *Latent Infection* class. The time of infection is

recorded and they are assigned to have the same strain type as the individual transmitting the infection. Finally, the individual is assigned future event times and scheduled for the earliest of these. In addition to death and emigration, neither of which is affected by infection, newly infected individuals are eligible for developing disease and for having a strain type mutation.

Whether the transmission actually occurs or not, the infectious individual is assigned a new time for transmission after the transmission event takes place. They are then scheduled for the earliest among the new transmission time and the previously assigned times for recovery from disease, case reporting, strain mutation, death, and emigration, none of which are affected by the transmission event.

#### *3.1.8.7 Disease*

Once infected, an individual may be scheduled to develop disease if probabilities allow. Development of *Primary Disease* and *Reinfection Disease* from *Recent Infection* and *Reinfection* classes happens at a rate which depends on the duration of infection, with the highest probabilities of developing disease in the first year of infection and lower in subsequent years. More specifically, following Vynnycky and Fine [19], for each year in the first five years of infection, the relative risk of developing disease following infection or reinfection in that year is fixed relative to other years [19]. These relative risks are equal for all ages and regions of birth. However, the total probability of developing disease over those five years differs with age and is also allowed to vary with birthplace (UK or foreign) and HIV status. This cumulative probability is used to decide whether or not an individual will develop *Primary Disease* or *Reactivation Disease*. If a random number drawn is less than the cumulative probability, the relative risk of disease over those five years is used to assign a time to disease. Otherwise, a disease time for *Reactivation Disease* is assigned. Note that for migrants entering the *Recent Infection* and *Reinfection* infection states, the time since they were infected is taken into account in assigning time to disease, which is less likely the longer they have had an infection prior to migration.

Development of *Reactivation Disease* from *Latent Infection* is assumed to happen at a constant annual rate, independent of duration of infection, since the infection is older.

Like *Primary Disease* and *Reinfection Disease*, this rate is dependent on age, birthplace (UK or foreign), and HIV status.

For all disease categories, at the scheduled time of disease onset, the individual is first assigned to either pulmonary or non-pulmonary disease. Next, a time for death due to tuberculosis, a time for case reporting, and a time for recovery from disease are calculated for all disease types. Also, a new time for strain mutation is assigned for all disease types, since active disease is assumed to increase the rate of strain type mutation compared to infection without disease. Finally, if disease is pulmonary, the individual is assigned smear-positive or smear-negative disease. For smear-positive individuals, a time to transmit infection is also assigned. These times are compared with the previously calculated times for emigration and for death due to other causes. The earliest of all possible events is scheduled and other times are stored.

#### *3.1.8.8 Transition to Latent Infection*

There are two pathways to *Latent Infection*, one from *Recent Infection* or *Reinfection* and one due to recovery from disease. These are described separately below.

##### 3.1.8.8.1 Transition from *Recent Infection* or *Reinfection* to *Latent Infection*

As defined in Section 3.1.3, individuals with *Recent Infection* or *Reinfection,* by definition, have infections that are less than five years old and that have not yet led to a disease episode. Therefore, from these states, individuals transition to the *Latent Infection* state after exactly five years if they do not die, emigrate, or develop disease before then.

At transition to *Latent Infection*, a new time to develop disease is assigned since those with *Latent Infection* are subject to different risks of disease. In *Latent Infection,* disease risks are constant annual rates of disease progression that are lower than disease development rates in *Recent Infection* and *Reinfection* states. The earliest event among death, emigration, strain mutation, and the new time to disease is scheduled and all other times are stored.

##### 3.1.8.8.2 Recovery from disease to *Latent Infection*

Individuals who do not die from active disease or leave the model through death or emigration may eventually recover from disease and move to the *Latent Infection* state. Time to disease recovery is assigned at disease onset by assuming exponentially distributed times to recovery, with a mean time of six months.

Transition to *Latent Infection* from active disease means an individual is no longer infectious, no longer subject to death from tuberculosis, no longer eligible for case reporting, and no longer subject to the higher strain mutation rate for those with active disease. A new time to strain mutation is calculated to reflect the lowered rate of mutation in infection without disease. Since *Latent Infection* individuals are once again subject to develop disease, a new time to disease is calculated. The earliest of death, emigration, time to disease, and time to strain mutation is scheduled and the other future event times are saved.

#### *3.1.8.9 Strain type mutation*

Strain type mutation is possible for any infected individual, in any of the three infection classes or six infection states. Strain mutation captures only mutations that are found in the region of the genome relevant to the particular strain typing method modelled—in this case 24-locus VNTR—and only those mutations that are successful enough to outcompete the previous strain type of infection. The mutation rate for diseased individuals is assumed to be higher than the rate for infected individuals, as detailed in Chapter 4.

When strain mutation occurs, a new strain type unique to the population replaces the old strain type. Although in reality VNTR profiles could mutate such that they match an existing strain type in the population, this is not taken into account in the model to simplify the mutation process. This decision was made mainly because of data limitations for modelling individual loci of the 24-locus VNTR strain type, described further in Chapter 7, Section 7.1.2.1.

#### *3.1.8.10 Reporting of a tuberculosis case*

To reproduce realistic case reporting in the UK, not all tuberculosis cases in the model are reported. When an individual develops disease, the case is reported if a random number drawn is less than the probability that a case is reported. At the time a case is

reported, it is added to a total of cases for the year, and age class, sex and birthplace for the individual. The case is then assigned for strain typing if a random number drawn is less than the disease site-specific probability that a case is typed. If typed, the strain type is reported for cluster analyses. After reporting, the individual is then scheduled for the earliest of the other events possible for that individual, the times of which are stored and do not change with case reporting.

## 3.2 Model Development

The model was adapted from an IBM written by C. Lehman at the University of Minnesota in 2009, used for simulating HIV transmission in the US (*Lehman et al.*, in revision, manuscript available upon request). The model developed by C. Lehman is a speed-efficient and highly adaptable discrete event simulation. At its core is a collection of algorithms to manage event times, including a list of future events and an event scheduler. These new modelling methods were explained to me by C. Lehman and we worked together in a three-day session in early 2010 to begin converting the HIV model into a basic tuberculosis model. The purpose of this session was to gain experience in adapting the code and to create a template for use in developing the tuberculosis model.

After this working session, the model was a skeleton of what it would become, at around 1000 lines of code, compared with the more than 3,500 lines in the final version, exclusive of scheduling, grouping, fitting, plotting, and utility routines. At that time, the model included nine compartments, a static population size, and no heterogeneity on age, sex, and birthplace. There was also no data input for model parameters or initial conditions. There were no immigration or emigration functions, nor functions for genetic strain type modelling.

From this initial version, the details of the model were expanded and almost completely re-worked, apart from the algorithms for time management and utility algorithms such as random number generation and sorting. The first step in the adaptation of the model was to distinguish UK-born and foreign-born individuals, which meant that the main array of individuals was subdivided and parameters were indexed by birthplace. The management of this subdivided array was the inspiration for a more general method of efficiently handling heterogeneous mixing in simulation models, devised jointly with C. Lehman [236], as described in Section 3.3.1.

Other major changes to the model included a more realistic mixing algorithm, with birthplace contributing to selection of an infection target (see Section 3.1.8.6). Furthermore, the model was adapted to include the entire 11 compartments desired, with appropriate transitions added. Emigration and immigration processes were added, with both UK-born and foreign-born individuals entering and leaving the UK.

Birth and death were de-coupled and the birth function was expanded. Model initialization was made more realistic, with the initial population made to reproduce the UK population in 1981 as closely as possible. The recording of genetic strains and cluster analyses for these data were added. Almost all existing processes in the model, such as vaccination and disease development, were fundamentally changed.

Lastly, empirical data were integrated into the model and parameters were made dependent on calendar year, age, sex, and birthplace. Integration of these data required new methods to be added to handle the creation of and sampling from cumulative distributions for event times, as described in Section 3.3.2 and published in a recent paper [234].

A list of main modules in the model and the associated model fitting routine is found in Table 3-2. For each module, the table provides a brief explanation of the module and the length of the module code including comments, the main author(s), and how the module is used. The model is coded in C and the code is freely available; it is published online at www.cbs.umn.edu/modeling, and in Appendix 10.19.

## 3.3 Contributions to Modelling Methods

In the process of adapting the IBM designed by C. Lehman for use in this thesis, there were several contributions I made to modelling methods. The main contributions were in methods for segregating the main array into collectives, with applications for fast, non-homogeneous mixing, and methods for incorporating real data into the model by drawing random numbers from arbitrary distributions, though both contributions were developed jointly along with C. Lehman. I also contributed to the application and testing of Centinel, a system for handling input and output of data files, as well as documentation and formatting of files. Lastly, I was also part of application and testing of the scheduling software itself, and part of a joint paper on the scheduling algorithms. Each of these contributions is described below.

### 3.3.1 Group Management

One of the first concerns I had for adapting the IBM for my work was segregating individuals based on contact groups. I chose to only divide the population into UK-born and foreign-born. The method for doing this was designed for maximal efficiency in randomly choosing individuals from within these groups, handled by keeping the array of individuals contiguous for each group during the entire simulation. This method was generalized into a group of algorithms for managing groups in simulations, especially helpful for incorporating heterogeneous mixing simulations. The algorithms allow segregation of many, many types of individuals, for example, dividing the population into one-year age classes and by sex. In the end, C. Lehman and I equally contributed to these algorithms. These algorithms are described in a recently published paper, which is reprinted in Appendix 10.19 [236].

### 3.3.2 Choosing Random Numbers From Arbitrary Distributions

With the incorporation of actual data into the model, some major changes in assigning event times occurred. In some cases, data and assumptions only required a change in the annual rate of occurrence for the event, meaning that the method of assigning event time was the same as before, drawing a random event time using the inverse cumulative exponential distribution with a mean value of the annual rate. Such is the

case in assigning duration of disease, which is largely unchanged from the original model.

However, for most transitions in the model, incorporating data required building of arbitrary cumulative distributions of event times. For example, consider the development of disease following *Recent Infection*, which is assigned if the individual is already randomly chosen to develop disease. The cumulative distribution of the probability of disease for these individuals goes from 0 to 1 over the period of 0 to 5 years. A time from the cumulative distribution can be assigned using 'if/else' or 'if' statements to choose a time category (e.g. disease will develop in the first year of infection) and then some interpolation method to get a more specific time, if desired. This technique was sufficient for many processes in the model.

With much of the demographic data however, the sheer volume of data makes this method inefficient and difficult to do correctly. As described in Chapter 4, probabilities of death are indexed by age, sex, and birth cohort. Also, in the case of death times, individuals could enter the population mid-way through the distribution (i.e. at any given age), and so computation of death time from that point forward was very different than assigning a death time at birth for an individual. Therefore, C. Lehman and I jointly devised a method to randomly sample an arbitrary cumulative distribution of event times, entering that distribution at any value (e.g. age) desired. We called this function *RandF* and the code was written by C. Lehman. The algorithm and its description were recently published [234]. The paper is reproduced in Appendix 10.19.

### 3.3.3 Other Contributions

The development of my model led to the first large scale application and testing of algorithms in the original IBM designed by C. Lehman. In some cases, my work led to improvements to the original code, for example, in error handling for the scheduling algorithms. This practical application and large-scale testing of the IBM led to my involvement as co-author in the published paper describing these methods [235], which is also reproduced in Appendix 10.19.

## 3.4 Model Verification

Model verification refers to the act of ensuring the simulation model is programmed according to its specifications [65, 237], in this case, according to the ODD description of the model found above. Methods for verification of computer programs date back to the early days of computer science and had become numerous by the late 1960s [238-241]. Discussion of model verification is now found in most scientific disciplines where simulation models are used (e.g. [65, 242-249]). Model verification is important because the behaviour of many complex systems cannot be known ahead of time, and therefore, it is difficult to know whether the complex models that simulate them are correct. For most complex models there is nothing to compare model output to that proves correctness and it is essential to take steps to ensure correctness of results in other ways.

Unfortunately, model verification steps are rarely reported in epidemiological modelling papers, though the importance of including model verification efforts has been pointed out in a recent article for veterinary epidemiologists [65]. The following sections describe my efforts to ensure correctness of the model. These verification steps are enormously helpful, but not entirely sufficient, as no procedure can prove all algorithms are correct [250].

### 3.4.1 Simple Version of Model Compared to ODE Model

One step in model verification is to compare a simplified version of the complex model to an equation-based model that is solvable using numerical methods, which may be called 'reduction testing' in computing literature. The idea of comparing complex models to analytical models seems to be rarely referenced in epidemiological literature, but sometimes found in other disciplines where simulation modelling is used [237, 251]. The first stage of model testing included matching a simplified version of the model to an ODE model, which could be solved numerically. This method involved first modifying the model, then constructing an ODE model and comparing results from the two.

### *3.4.1.1 Simplified IBM*

In the early stages of model development, before integration of complex data and functions, the model was written with the intention to validate it by comparison to ODEs. To do this, all parameter values and associated event time distributions were made equal for all age, sex, and birthplace categories and made invariant with time. All genetic strain functions were removed. Individuals in the model retained their individual birth dates, sex, and birthplaces, but only to keep the model's structure intact. Individuals were actually homogeneous in this implementation of the model, since parameter values were independent of their attributes. Events times were obtained by a random draw from the inverse cumulative exponential distribution for the IBM. Once all parameter values were constant, annual rates of transition for every function in the model, the simple IBM became merely a differential equation simulator, although implemented very differently than such simulators typically are. These changes allowed the structure of the code to be validated against known results.

#### 3.4.1.1.1 ODE model

An ODE model was written in the R language (R Foundation for Statistical Computing, Vienna, Austria) to correspond to the simplified IBM, which included the testing values for parameters with values matching those used in the IBM. The differential equations were solved using Euler's method with a sufficiently small time step. The numerical solutions from this model provided outputs for the 11 infection states each year for a 29-year period corresponding to the simple IBM output. The equations for the model and its solution are found in Appendix 10.1.

### *3.4.1.2 Comparison*

Although the model used was stochastic, its results were expected to match those of the analogous deterministic ODE with large enough compartments. With large compartments, the variance among runs falls. These comparisons were carried out on an ordinary laptop computer, running about 10 million individuals. Therefore, some infection state categories (e.g. non-pulmonary reinfection disease) remained small and displayed variance between runs, not allowing reliable comparison to the ODE. To remedy this outcome and to ensure variance was as small as possible even in the larger compartments, the contact number (transmission rate) and disease progression

rates were raised to artificially high values. Higher values resulted in enough disease cases throughout the simulation to compare results to the ODE model. After eliminating mistakes from both the simple model and the ODE model, results matched well for the models. The comparisons did not produce an exact numerical match, but one which was very close for all categories—essentially exact for large categories—such that there was virtually no chance of an error in the model.

### 3.4.2 Modular Programming and Unit Tests

Another programming practice I followed was writing the program in 'modular' form as a collection of smaller functions that are easily understood. Functions were then independent of one another, though they could call other functions. Functions were tested individually, often called 'unit testing', helpful because the behaviour of individual functions is usually predictable, or at least the behaviour is more intuitive than that of the entire complex model. Throughout model development, I tested functions individually as they were altered. When possible, function input and output were compared to what was expected, though this process varies from function to function. Also, with some functions is not possible to prove correctness through testing, but merely show that output is what one would reasonably expect or check only a subset of the function output for correctness.

One example of function testing I performed concerns the transition from *Recent Infection* and *Reinfection* to *Latent Infection* in the model. The function is simple, only used to change the state of individual to *Latent Infection* and schedule them for a new event. For testing this function, print statements were first used at the beginning of the function to check that the correct individuals were arriving at the function: those in *Recent Infection* and *Reinfection* states with exactly five years of infection. Print statements also showed the saved event times for this individual, of which only the time to disease should have changed. When the individual exited from the function, print statements were used to check that the new state is correct (*Latent Infection*) and that the individual was scheduled for a new event. The new event should have been one of: death, emigration, disease development, or strain type mutation. Print statements upon exit were also used to check that the time to disease was different than the previously saved time (though in extremely rare cases it could have been the

same exact random time is chosen from two different distributions). Mutation times should not have changed from the previous time saved for the individual. After it was confirmed these conditions were satisfied, the function was considered finished.

### 3.4.3 Pre/Postcondition Programming

Along with modular programming, I adopted a variation of 'pre/postcondition' programming [241], designed to help avoid mistakes in complex computer programs. The method requires for each function a complete and precise description of its requirements for working properly— 'preconditions' —and the resultant state of the system at the end of the function if the preconditions are met — 'postconditions'. This method also requires that all functions are checked rigorously for compliance with pre/postconditions, ideally in joint code review sessions so that they produce the correct results each time. When functions written in pre/postcondition style are reviewed, a function is checked by checking: 1) the pre/postconditions for that function, 2) the code of the function itself, and 3) only the pre/postconditions of any other functions called. This helps restrict the range of attention needed. If all functions are written and checked this way, it is only necessary to verify the pre/postconditions of functions called within other functions, but not the code of the functions called. The called functions are assumed to behave according to those conditions when reviewing or writing code in other functions.

In addition to the pre/postconditions, this method of programming also requires detailed documentation of the code inside the function and a description of the function given in plain language before the preconditions are specified. Although mistakes are still possible using the method, it provides inherent correctness-proving. In addition, the code is kept to small, readable units and is commented well. This brings enormous clarity and structure to a complex model or other program and allows someone other than the writer to understand the code. Examples of this method of programming are found in the model code, reproduced Appendix 10.19.

### 3.4.4 Centinel Input File Handling

Input files specifying population sizes, mortality rates, and other distributions used for initializing the model and assigning times to events for individuals in the model became complex very quickly. For example, mortality rates are specified by birth

cohort, 1870 – 2011, and age 0 – 120 years. It was clear much effort had to be made to avoid errors in input files. Therefore, I began using a module written by C. Lehman for reading formatted input files to help avoid mistakes. The module specifies a documentation structure for input files and allows those files to be both computer- and human-readable. The module is partly described in a published paper I co-authored with C. Lehman regarding using this file format for archiving scientific data, which has many of the same goals in avoiding errors, being computer- and human-readable, and simplistic [252]. An example of a Centinel-formatted input parameter file used for this study is found in Appendix 10.2.

### 3.4.5 Code Review

Lastly, model verification included detailed joint reviews of code with another person, called 'walkthroughs' in the software engineering field. These are used to identify mistakes, improve program documentation and comprehensibility of the code, and re-structure the code for efficiency and clarity. Even the act of explaining the model and code to another person is helpful for improvement and seeing one's own mistakes. C. Lehman volunteered for a detailed walkthrough of the model code twice during model development. The first code review took place part-way through the project, and although primarily useful in correctness-proving and eliminating mistakes, it was also helpful to restructure some code. The second code review took place at the end of the project, before results were obtained.

## 3.5 Model Validation and Calibration

Model validation refers to the process of demonstrating that the model is appropriate for the real-world system it replicates. Model validation has received much attention in modelling literature across various fields, though methods vary from discipline to discipline [243, 244, 247, 250, 251, 253, 254]. In epidemiology, model validation primarily consists of comparing the model to observed data. Methods for model validation used in this thesis varied between applications. In the West Midlands application of the model, the output for several model scenarios was compared to observed genetic and notification data. For the England and Wales application of the model, formal 'model calibration' was undertaken, as described in Section 3.5.2.

Model calibration is the alignment of model output and observed data through variation of some parameters in the model. The best fit(s) of the model output to observed data are found using an optimization algorithm that tests many different parameter sets, each time comparing model output to observed data. For each set of parameters, the fit of model output to observed data is assessed using a measure of goodness-of-fit (GOF) defined *a priori.* The parameter set(s) that result in the best fit of model output to data are taken as best estimates for these unknown parameters, at least for the setting to which they were applied.

This process is easier for deterministic than stochastic models. Because a specific combination of parameter inputs always leads to the same model output, each parameter combination needs to be tested only once. Calibration still requires many model runs, but the computation time is usually reasonable. With stochastic models, such as the model described here, a specific combination of parameter inputs does not produce the same results each time. The variability from run-to-run depends on the number of individuals modelled and the nature of model output, but may require hundreds or even thousands of runs for each combination of parameters to reduce variance. With specialized computer hardware and software, the reduction in variance can be achieved by running many iterations of the program with one set of parameter values in parallel. This reduces the computation time of the optimization algorithm.

### 3.5.1 Model Stochasticity

Because the model is stochastic, there are additional considerations when fitting the model to data. Model output varies from run-to-run, even under the same initial conditions and parameter values, due to chance. Variance in output among model runs is greatest when the population sizes of demographic categories in the model are small or when, even for groups with larger population sizes, there are few disease cases for a category. Variance between model runs was minimized in three ways to allow the model to be fit as if it were a deterministic model.

Firstly, variance between model runs was minimized by simulating the full population size of the study areas, both for England and Wales and the West Midlands, despite the added computer resources required. This simulation required running the model on specialized computer hardware with several software modifications (see Section 3.5.2.6). Secondly, variance between model runs was reduced by eliminating output categories with an unreliable number of case notifications from the fitting process (see methods section of Chapter 5). Thirdly, variance between model runs was reduced by averaging several runs of the model for every evaluation of the model under different parameter values during fitting. For each evaluation of the model, or for every set of parameters tested, there were 30 replicates of the model run with identical initial conditions for the England and Wales application of the model. There were 100 replicates of the model run for the West Midlands application of the model. The only difference in model output was due to different starting seeds for the random number generator in the model.

The number of replicates was chosen based on variation in model output due to chance and keeping computation hours as low as possible. To explore this variation, the model was run with one set of parameter values and initial conditions many times. The model output values averaged over those 500 runs were considered the 'true' values or deterministic values for model output. Then, average values at different numbers of model replicates were taken. Averaging 30 replicates was considered sufficiently close to the minimum number of replicates for which the averaged values matched 'true' model output for even the relatively small output categories for the England and Wales version of the model. More replicates were run for the West Midlands version of the model because of greater stochasticity in those runs, due to smaller population sizes.

Note, in addition to stochasticity in the tuberculosis model, there is stochasticity inherent in the fitting process used for the England and Wales application of the model. This also required consideration, as described further in Section 3.5.2.4.

## 3.5.2 Fitting Process for England and Wales Application

### *3.5.2.1 Optimization algorithm and GOF statistic*

The model applied to England and Wales was fit to observed notification data through an optimization algorithm that combines downhill simplex with simulated annealing [255]. The algorithm was used because the downhill simplex part of the algorithm is relatively efficient in searching parameter space and the simulated annealing algorithm is able to search for optimal parameter sets across a wide range of parameter values, excelling at finding global, as opposed to local, minima. In this context, the global minimum occurs when a set of parameters leads to model output that is closest to observed data, whereas a local minimum occurs at a place where a set of parameters leads to model output that is closest to observed data within a limited part of the parameter space. Code for the optimization routine was taken from the Numerical Recipes collection, published in the C language as the 'amebsa' routine [255, 256].

Downhill simplex with simulated annealing requires specification of an initial 'simplex' or set of $n$+1 points to begin optimization, where $n$ is the number of variable parameters in the model, usually four in this application of the model. Each of these points is first evaluated for the GOF of the model output to data, where a lower GOF statistic implies a better fit of model output to data (see below for details of the GOF statistic used). Then the optimization routine chooses another 'point', or set of values for the variable parameters in the model, based on the GOF of the previous $n$+1 points and partly on random chance. The optimization routine works in this manner, with the $n$+1-vertex shape moving about the parameter space to find a best-fitting set of parameter values. The best-fitting point is stored and reported when the routine converges, or reaches a solution, and little or no change is seen in parameter values or GOF values as the routine searches around that point. Convergence is discussed

further below. As mentioned, the optimization routine requires a function which evaluates the fit of the model output to observed data, also referred to as the GOF statistic. The GOF statistic used was Poisson log likelihood deviance, used in a related studies [257, 258]. The formula for calculating this statistic is

$$-2\sum_{a}\sum_{s}\sum_{r}\sum_{t} o_{a,s,r,t} - e_{a,s,r,t} + o_{a,s,r,t} \times \ln e_{a,s,r,t} - o_{a,s,r,t} \times \ln o_{a,s,r,t}$$

where $a$ is age, $s$ is sex, $r$ is birthplace, $t$ is year, $o_{a,s,r,t}$ is the observed number of notifications by age, sex, birthplace and year, and $e_{a,s,r,t}$ is the simulated number of notifications by age, sex, birthplace and year. This formula is used to calculate the deviance for each set of parameters tested. The best-fitting value for each fitting run is compared with other runs to assess which scenarios fit best. The deviance statistic varies according to a Chi square distribution, with the degrees of freedom equal to the number of data points, less the number of variable parameters.

#### *3.5.2.2 Convergence criteria*

The standard *amebsa* routine for simulated annealing can be run until it has achieved a specified convergence threshold in the GOF statistic [255] or until the rate of change in the GOF statistic has approached zero. The former is part of the standard routine and the latter is a custom extension by C. Lehman (personal communication, code published at http://www.cbs.umn.edu/modeling) to conserved limited computing allocations during fitting runs. The custom routine tracked the relative rate of change in the parameters until that rate repeatedly remained less than 1% per step over the last twenty steps of optimization.

#### *3.5.2.3 Initial conditions*

Initial conditions for a fitting run include the $n$+1 points of the initial simplex discussed above, each with values for the four or six variable parameters in the model. For each of the $n$ variable parameters, a value for each of the $n$+1 points of the simplex was randomly chosen from specified ranges for each parameter (see Chapter 5, Section 5.1.6). Note, the simulated annealing routine searches beyond these input parameters when searching for optimal parameter values and so the ranges, though relatively

large, still do not restrict the optimization routine to find parameter values within these ranges.

#### *3.5.2.4 Replicate fitting runs*

For each fit of the model to data, the fitting routine was run five different times. For each of these replicates, a different set of randomly assigned starting simplex points was used. These initial conditions are the set of $n$+1 parameter sets, each set with one value for each of the $n$ variable parameters. The five replicates were run for two reasons. First, in addition to stochasticity in the tuberculosis model itself, which was addressed in multiple ways as described in Section 3.5.1, the simulated annealing routine is stochastic. Even if the model were deterministic, the fitting routine would take a slightly different path to the optimal parameter set, likely finding a different set of optimal parameter values under runs with the same input parameter values and initial conditions. Secondly, the initial conditions of the fitting run, or points of the initial simplex, may influence the pathway to finding a best fit and thus, the best-fitting values themselves. Running five replicate fitting runs ensures that a range of values for each parameter are tested initially. In summary, replicates help avoid basing conclusions on one run of the fitting routine, which may be an anomaly due to chance or due to choice of initial conditions. Five replicates were the maximum number which could be afforded with the computer hours available.

#### *3.5.2.5 Simulated annealing schedule and temperature*

The 'annealing temperature' is a parameter of the optimization routine that controls the random element which contributes to the algorithm for choosing a new set of variable parameters as the simplex moves about the parameter space, searching for the optimal parameter set. The temperature describes the chance that the routine accepts an 'uphill' step, or a set of parameters that worsen fit to data over current points. This parameter also impacts the amount of random fluctuation on newly chosen simplex points. The temperature is high at the beginning of a fitting run, allowing the optimization routine to accept uphill steps more often. Accepting uphill steps means the simplex may move out of local minima and reduce the chance of finding local minima as an optimal solution. As the optimization routine progresses,

the temperature decreases and the routine becomes less likely to accept uphill points. When the temperature is zero, only 'downhill' steps are taken, or better fits, meaning the optimization routine will not move out of a local minimum.

The 'annealing schedule' is the combination of simulated annealing parameters used in the optimization routine. These include the initial temperature, temperature reduction factor, the number of iterations per cycle and the final temperature convergence tolerance. The initial temperature was set to 2.0. The temperature reduction factor specifies how quickly the temperature decreases, and was set to 0.7. After each annealing cycle, consisting of a specified number of iterations, each of which generates a new point in the simplex, the temperature is multiplied by 0.7. The number of iterations per cycle was set to 25. Lastly, as mentioned, the convergence tolerance and final temperature were not used in this application of amebsa, as the parameter values and GOF values were converging before these were reached.

#### *3.5.2.6 Computer hardware and software*

As described, the fitting routine compared model output under many different parameter sets, searching for the best fit to the observed data. This comparison required many runs of the model, one after the other, in serial. Furthermore, each of the many runs actually required 30 replicate runs, each with the same parameter values but different starting seeds for random number generation, averaged to reduce the effects of stochasticity in model output, as described in Section 3.5.1. One instance of the model used more than 1.7 minutes of computation time on an average processor, for example on the 2.8 GHz Intel "Nehalem" processors used for this work, and an entire fitting run takes on the order of 100 steps, or 100 evaluations of the model, to find the optimal parameter set for the run. This timing meant that approximately 1.7 x 30 x 100 = 5100 minutes or 85 hours of computation time was used per fit. As many fits of the model were needed to explore different input parameter scenarios, this time had to be reduced by running the 30 replicate runs of the model in parallel, or at the same time. Parallel processing required special computer hardware and software.

One additional complication is that this application of the model required more than 50 million individuals in the simulation, using a large amount of computer memory,

approximately six gigabytes of memory for each run. With 30 replicates running in parallel, the memory requirement was 180 gigabytes of memory for each fitting run. Therefore, running the model required computer hardware with sufficient memory and a sufficient number of processors to run 30 replicates of the model in parallel. Running the model in parallel also required special adaptations to the model code, making use of Message Passing Interface (MPI) commands to collate data from multiple processors inside the fitting routine. Lastly, a special compiler which allowed the MPI commands to work, 'mpicc', was used to compile the model after these MPI commands were integrated.

The Minnesota Supercomputing Institute was one place with both the hardware and software available to run multiple, memory-intensive replicates of the fitting routine in parallel. There, an HP Linux cluster called 'Itasca' was used for all fitting runs of the model. Although Itasca has 1,086 compute nodes and 24 gigabytes of memory per node, the computation time for this work was limited by an allocation of computing hours designated for this project.

Table 3-2: Main software modules used for the thesis work. For each module, there is a description of its purpose, the number of lines of code including comments, the author(s) of the modules, their use in the thesis modelling work, and the file in which they exist. The column 'Use' gives the main purpose or place module is used, where 'TIBM' refers to the model of tuberculosis dynamics in the UK; 'IBM' refers to general individual-based model machinery; 'FIT' refers to the fitting routine; and 'UTIL' refers to modules which are utilities, used across many programs and often inside other modules. 'Location' refers to a '.c' file in which the module exists.

| Module | Description | Lines | Author(s) | Use | Location (file) |
|---|---|---|---|---|---|
| *scheduler* | Handles event scheduling and dispatching | 500 | CL | IBM | schedule.c |
| *main* | Sets up the model, runs main loop for simulation | 90 | AK/CL | TIBM | tb36gen.c |
| *dispatch* | Dispatches event to scheduler | 50 | CL | IBM | tb36gen.c |
| *birth* | Adds newborn to the population | 110 | AK | TIBM | tb36gen.c |
| *immigrate* | Adds immigrant to the population | 220 | AK | TIBM | tb36gen.c |
| *immg* | Catalyzes immigration as external event | 40 | AK | TIBM | tb36gen.c |
| *birthg* | Catalyzes birth as external event | 30 | AK | TIBM | tb36gen.c |
| *vaccinate* | Vaccinates an individual | 40 | AK | TIBM | tb36gen.c |
| *infect* | Infects an individual | 120 | AK | TIBM | tb36gen.c |
| *remote* | Moves individual to *Latent Infection* | 70 | AK | TIBM | tb36gen.c |
| *disease* | Moves individual to a disease category | 170 | AK | TIBM | tb36gen.c |
| *transmit* | Transmits (or attempts to transmit) an infection | 90 | AK | TIBM | tb36gen.c |
| *mutate* | Mutates strain type profile | 150 | AK | TIBM | tb36gen.c |
| *death* | Removes individual from population | 60 | AK | TIBM | tb36gen.c |
| *emigrate* | Removes individual from population | 30 | AK | TIBM | tb36gen.c |
| *newstate* | Changes state of individual | 40 | CL | TIBM | tb36gen.c |
| *transfer* | Shifts identification number for contiguous array | 20 | AK | TIBM | tb36gen.c |
| *repcase* | Stores reported case | 120 | AK | TIBM | tb36gen.c |
| *lifespan* | Assigns time of death | 50 | AK/CL | TIBM | tb36gen.c |
| *emtime* | Assigns time of emigration | 50 | AK | TIBM | tb36gen.c |
| *recovery* | Assigns time to disease recovery | 40 | CL | TIBM | tb36gen.c |

| Module | Description | Lines | Author(s) | Use | Location (file) |
|---|---|---|---|---|---|
| *tdisease* | Assigns time to disease | 80 | AK | TIBM | tb36gen.c |
| *init* | Initializes output files | 50 | AK | TIBM | tb36gen.c |
| *data* | Reads data files into arrays | 220 | AK | TIBM | tb36gen.c |
| *param* | Updates cumulative distributions for variable | 190 | AK | TIBM | tb36gen.c |
| *initpop* | Initialize starting population of model | 200 | AK | TIBM | tb36gen.c |
| *final* | Writes final output, plots | 1180 | AK | TIBM | tb36gen.c |
| *(tb30i.c)* | (Entire TIBM code: declarations, modules (including some little-used and not listed here and comments) | 4300 | AK | (TIBM) | tb36gen.c |
| *fit* | Shell program for fitting routine | 160 | AK/CL/NR | FIT | fit5i.c |
| *amebsa* | Downhill simplex with simulated annealing fitting | 120 | NR | FIT | amebsa.c |
| *amotsa* | Selects next point for evaluation in fitting | 50 | NR | FIT | amotsa.c |

# 4 Sources for Model Parameters and Assumptions

The sources used to establish parameter values and other assumptions in the model are described in this chapter. Sources are first described for demographic parameters and assumptions and then for infection-related parameters and assumptions. Sometimes sources differ between the England and Wales and the West Midlands model applications. In these cases, sources are detailed for each application. In cases where sources are the same, specific model applications are not mentioned and the same parameter values and assumptions are used for both applications of the model. Although model output was fitted to data from 1999--2009 for the England and Wales application of the model and from 2007--2011 for the West Midlands, the model was run starting in 1981 for both applications. This reduced the importance of parameter values used at model initialization, such as the prevalence of each infection state in 1981.

## 4.1 Demographic Data

### 4.1.1 Births

For each application of the model, the actual number of births each year was used to ensure that exactly that number of individuals was born in the corresponding simulation year. For England and Wales, the total number of births for each calendar year, as well as the ratio of male births to female births, was obtained from the Office for National Statistics (ONS) in electronic form from their website [111]. Data for 1981 to 2004 were obtained from a download of "Birth Statistics FM1 (Historical Series)" files and data for 2004 – 2009 were downloaded from a 2009 "Birth summary statistics" file.

For the West Midlands, the number of males and females born each year to mothers whose usual residence was the West Midlands was obtained in an electronic file requested from the ONS Vital Statistics Outputs Branch. This file is not available on the ONS webpage (L. Todd, personal communication).

### 4.1.2 Mortality

All-cause mortality rates for England and Wales by year of birth, age and sex were used to assign life expectancies in the model. The ONS Centre for Demography provided an electronic copy of these data on request. Data included tables of $q_x$ rates for males and females for ages 0 – 120 years and birth cohorts from 1870 – 2011. $Q_x$ rates for a given year of birth and age, $x$, represent the proportion of those in the cohort who die between ages $x$ and $x+1$. These rates were based on the number of observed deaths until 2008 and the number of deaths projected thereafter. For cohorts born after 1946 for males and after 1941 for females, $q_x$ rates remain slightly below 1.0 at age 120 years, meaning a small fraction live beyond 120 years. However, for simplicity, it was assumed that no individuals lived beyond 120 years.

All-cause mortality rates were used to calculate the time of death due to causes other than tuberculosis for individuals in the model, although $q_x$ rates included deaths due to tuberculosis. Because the effects of tuberculosis-related mortality were negligible for the general population during the time period of simulations, 1981 – 2011 (see Section

4.2.3.4), this simplifying assumption seems reasonable. It was also assumed that foreign-born individuals were subject to the same average mortality rates as the general population, although some studies have shown migrants to industrialized countries had a lower life expectancy than those who were native-born in industrialized countries [259], while other studies have shown higher life expectancies for migrants in industrialized countries compared to those who were native-born, or were inconclusive [260-262]. In reality, those from some countries may have had increased mortality rates and those from other countries may have had lower mortality rates. Given the conflicting evidence and absence of any other data on life expectancy of those who were foreign-born to the UK, it seemed reasonable to assume the mortality rates among UK-born and foreign-born individuals were the same.

Mortality rates from England and Wales were used for the West Midlands model because data specific to the West Midlands were not available.

### 4.1.3 Population Sizes

Estimates of population sizes by age, sex, and birthplace (UK, SSA, OF) were used to define the initial population of the model in 1981 for England and Wales and the West Midlands, as well as to calculate tuberculosis notification rates and emigration rates for other years. Whilst census data were available for 1981, when simulations started, these data were insufficient for initialising the model population from this year because countries of birth were not distinguished in enough detail to use the data for model initialization. The numbers of SSA-born individuals could not be extracted because birthplace data were aggregated into groupings based on 'Commonwealth' versus 'foreign' states, both of which include countries within SSA. Furthermore, for calculating notification rates and emigration rates in other years, census data were not available for most years because the census occurred only every 10 years. To overcome these issues, population estimates were instead obtained by analysis of Labour Force Survey (LFS) data.

#### *4.1.3.1 Population sizes from LFS data*

Population size estimates were obtained by analysis of individual-level data from the LFS, downloaded from the Economic and Social Data Service [263]. The LFS is a quarterly survey of households in the UK, designed to obtain information on the labour market, but is useful for other purposes due to the comprehensive survey questionnaire used. Survey data were analysed using the HPA Tuberculosis section protocol for obtaining population size estimates (J. Moore, personal communication), to allow comparison with UK national tuberculosis surveillance rates. Under the protocol, population size estimates were obtained using data from the April – June quarter of the LFS, using the 'survey' commands in STATA software version 11.1 (STATACorp, College Station, TX, USA) and individual-level weights provided by ONS. To keep sample sizes reasonably large and age categories consistent with HPA Tuberculosis Section groupings, age was categorized into four classes, 0 – 14 years, 15 – 44 years, 45 – 64 years, and 65 years or above. Methods were identical for producing estimates of population sizes for England and Wales and the West Midlands. Before applying the protocol for survey data analysis, some additional steps were taken to categorize respondents into the three birthplace groupings used in the model, UK-born, SSA-born and OF-born. These steps included distributing those with missing country of birth information for each year, as well as distributing ambiguous birthplace categories for some years. Methods for distributing respondents into birthplace categories are described below.

##### 4.1.3.1.1 Distribution of respondents into birthplace groupings

Redistribution of respondents by birthplace was most complex for 1981. This involved first placing respondents with countries of birth clearly corresponding to one of the three birthplace categories desired into the correct category. These categories included most respondents. Appendix Table 10-1 shows LFS country of birth labels and the birthplace category assigned. All non-bolded categories were assigned directly to one of the three birthplaces. After this, country of birth responses that could not be placed directly into one of the three birthplace categories were processed. These

included six small categories, totalling less than 1% of respondents, labelled 'other new Commonwealth', 'other Africa, foreign', 'rest of the world', 'at sea/in the air', 'not stated' or 'not known' for country of birth. Each of these is shown in bold in the Appendix Table 10-1.

Those assigned to the category 'other new Commonwealth' were randomly assigned to either African or non-African birthplaces based on the observed proportions of each among other Commonwealth countries. All assigned to African birthplaces were then placed in the SSA-born category, since African new Commonwealth countries are all found in SSA. Those assigned as non-African new Commonwealth were categorized as OF-born. Next, those with a birthplace of 'other Africa, foreign' were split into OF-born and SSA-born by random distribution based on the relative proportion of survey respondents from 'foreign, North Africa' and 'foreign, SSA'. Lastly, respondents with birthplace categories of 'rest of the world', 'at sea/in the air', 'not stated' and 'not known', were collectively referred to as 'missing' and were assigned to UK-born, OF-born, or SSA-born according the proportions in each of those categories among all respondents with non-missing country of birth information.

In 1999, LFS categories for country of birth were expanded and the additional detail allowed for straightforward classification of almost all respondents into UK-born, SSA-born, and OF-born. Those who did not fall into these categories were those categorized as 'missing' or 'stateless', both distributed among the three categories, UK-born, SSA-born, and OF-born, proportionally, as before. From 2000 – 2009 the same process was followed.

#### 4.1.3.1.2 Results, limitations and comparisons to other published estimates of population size

Population size estimates obtained from the LFS analysis for 1981 are presented in the Appendix. Table 10-3 and Table 10-5 show population size estimates used for calculating notification rates for England and Wales 1999 – 2009 and the West Midlands 2007 – 2011. Population size estimates for England and Wales were compared against the census and other data sources (see below) to check that analysis

of LFS data provided reasonable population size estimates. This comparison was not undertaken for West Midlands estimates due to a lack of sufficiently detailed data.

Estimates obtained from analysis of LFS data for the initial population in 1981 for England and Wales appeared reasonably close to the census data for that year, however, estimates for SSA-born population sizes were slightly higher than those estimated from the census data. The comparisons are shown in Appendix Table 10-6. Census estimates assumed that no one assigned to the 'remainder of the new Commonwealth' category was SSA-born, while all from 'other Africa' were SSA-born. This assumption is based on LFS classification data, which have similar categories. In LFS classifications, few or none from 'remainder of new Commonwealth' were SSA-born and all or almost all from 'other Africa' were from SSA. Age-structured population sizes from the census, which are stratified by country of birth, did not give detailed enough information on country of birth to estimate the proportion of persons born in SSA by age. Therefore the LSF estimates were used for the 1981 population by age, sex, and birthplace.

Estimates for 2001 from analysis of LFS data for England and Wales were compared to the 2001 census numbers from census Table S015, which provided population estimates by country of birth and sex. The comparison is shown in Appendix Table 10-7. Again, there was an acceptable agreement between the LFS estimates and census numbers.

Estimates from analysis of LFS data from 2004 – 2009 were also compared to the Annual Population Survey (APS) estimates by country of birth for England and Wales, as shown in Appendix Table 10-8. Although the APS was largely based on the LFS and not an independent source of comparison, APS estimates were official ONS estimates, subject to more rigorous analysis than the estimates obtained here through my independent analysis of the LFS data. It appears the LFS estimates obtained were reasonably close to APS estimates. However, since APS estimates were not published by detailed country of birth stratification, only UK-born and foreign-born population sizes could be compared.

Limitations to the estimates obtained from this analysis of LFS data include sampling errors and biases associated with the LFS itself but also potentially inaccuracies due to basing the population size estimates on samples for one quarter of the year. Furthermore, inaccuracies could have resulted from imputation of missing or non-specific sex or country of birth information. Lastly, ONS advises that estimates suggesting that there are fewer than 10,000 individuals are unreliable; there were a small number of these estimates obtained for SSA-born in some age classes, as well as for several demographic categories for the West Midlands.

### 4.1.4 Immigration and Emigration

Data on immigration and emigration were used to specify the numbers of migrants entering the study populations each year and the rate at which individuals left the study populations each year. Immigration and emigration data were derived from the International Passenger Survey (IPS), a sample of passengers entering or leaving principal UK air, sea and Channel Tunnel ports, with approximately 0.2% of all travellers sampled [264]. The IPS's information for inflow and outflow estimates were based on passengers' intended length of stay in or out of the UK. The IPS data informed migration inflow and outflow estimates for 1981 – 1990, produced by ONS. For 1991 – 2011, ONS produced more refined migration estimates, called Long Term International Migration (LTIM) estimates. LTIM estimates were derived from the IPS and supplemented by information from the LFS, Home Office data on asylum seekers and their dependents, and migration data from the Northern Ireland Statistics and Research Agency [265].

Since published IPS and LTIM tables were not stratified by migrants' country of birth in detail nor cross-tabulated by age, sex and region of the UK, customized tables which included inflow and outflow estimates for Sub-Saharan Africans were obtained directly from the ONS Migration Statistics Unit. These tables provided estimates for the number of inward and outward migrants to England and Wales from 1981 – 2009 and the West Midlands from 1981 – 2010 by age category, sex, and birthplace. Age categories in years were under-15, 15 – 24, 25 – 34, 35 – 44, 45 – 59, and 60 and over. Birthplaces included were UK, 'rest of Europe', North Africa, Sub-Saharan Africa, Indian

Subcontinent, 'rest of Asia', America, and Oceania. For both inflow and outflow data, birthplaces were grouped to correspond with the three birthplaces used in the model: UK, OF, and SSA. The numbers of inward migrants stratified by the three birthplaces were used directly in the model (numbers are not reproduced here due to ONS restrictions on publishing these data).

The numbers of outward migrants by birthplace (UK, OF, SSA), sex, and year were divided by population size estimates by birthplace, sex, and year, to obtain emigration rates for each year and birthplace. These rates were averaged over all years to set constant emigration rates per year, because emigration data were based on small numbers of persons and prone to sampling error. Emigration numbers by birthplace are not reproduced here to comply with ONS restrictions on publishing of these data. On average, approximately 0.25% of UK-born, 2.8% of OF-born, and 1.8% of SSA-born in England and Wales emigrated from the UK each year, while approximately 0.17% of UK-born and 1.4% of foreign-born in the West Midlands emigrated from the UK each year.

For the West Midlands model, 2011 migration data were not available. Instead, migration numbers and rates were assumed constant from 2010 – 2011. Provisional data on total migration to and from the UK for 2011 showed that the total numbers of migrants entering and leaving had not changed much from 2010, suggesting this would be an acceptable approximation. Provisional data were accessed in an electronic file from the ONS webpage [266].

One of the major limitations to these estimates is that ONS did not attempt to include the size of the illegally residing population in the UK [111]. In addition, as these estimates were largely based on the sample of migrants who take part in the IPS, they are subject to sampling error, as well as bias due to non-responders having different characteristics than responders and inaccuracy due to misrepresentation of intentions by respondents [111]. Confidence limits on the data indicated a large amount of uncertainty due to sampling error, especially for smaller demographic categories. The most uncertain were estimates for the numbers of SSA-born migrants to England and Wales before 1990 and for estimates for the West Midlands generally. In some years,

for small categories, the standard error of the estimate was more than 50% of the estimate itself.

## 4.2 Infection-Related Parameters

### 4.2.1 Vaccination

BCG vaccination was explicitly included in the model from 1981 to 2005, for both UK-born and foreign-born individuals who were *Uninfected* at the time of vaccination. Although some individuals were vaccinated in the UK after 2005, these were disregarded due to small numbers and limited impact. It was assumed that 75% of children between 13 and 14 years of age were vaccinated from 1981 – 2005, reflecting the policy and estimated coverage in England and Wales over this period [267, 268]. Of those vaccinated, 77% were assumed to be protected from infection for the remainder of their lifetime based on average efficacy estimated from 20 years of follow-up in a trial of BCG given to adolescents in the UK [61]. The trial found that the tuberculosis incidence in the control and vaccinated arms was very similar 20 years after vaccination, which suggests that the assumption that vaccination provides lifelong immunity is overly simplistic. However, the lack of apparent protection in the long-term after vaccination may be attributable to other factors, such as decreasing susceptibility in the unvaccinated group over time, due to exposure to *M. tuberculosis* or environmental mycobacteria [61].

BCG vaccination was also included in the model implicitly for those older than 13 years at the start of the model in 1981. Assumptions about vaccination practices prior to 1981 were used along with other assumptions to estimate the proportion of individuals in different infection states, including *Immune*, at model initialization (see Section 4.2.7). Although BCG vaccination began in 1949 in the UK, coverage was limited and effects are assumed negligible until 1954, when mass vaccination of 13 year-olds began. For those *Uninfected* born 1941 or later, or reached 13 years old in 1954, the proportion *Immune* was obtained from multiplying the proportion *Uninfected* at age 13 by the vaccination coverage and vaccine efficacy. Although the exact numbers vaccinated from 1954 – 1959 are available, population denominators are not readily available (although they could be estimated using data from the 1951 census and age-specific mortality rates). For simplicity, the proportion vaccinated was assumed to increase linearly from 0% in 1953 to 75% in 1960 [267, 268]. From 1960 to

1981, it was assumed a constant 75% of 13-year-olds were vaccinated. Vaccine efficacy is assumed to be 77% [61]; for more on vaccination data, see Section 4.2.1. Vaccination coverage and efficacy were assumed equal for males and females. This assumption may be unrealistic, but sex-specific data were not available.

For foreign-born individuals, it was assumed that none were *Immune* in 1981. This assumption was based on low estimates of average worldwide vaccine coverage, estimated by WHO to be about 16% in 1980 [269]. The average foreign-born individual in the UK in 1981 may have emigrated many years previously, and coverage among these individuals, on average, would have been even lower. Given this and the low vaccine efficacy in many areas of the world [57], the proportion vaccinated *effectively* was likely negligible prior to 1981. Still, this assumption will have underestimated the number of vaccine-protected individuals in the population, because some countries had higher vaccine coverage and some of these vaccinations would have been effective.

### 4.2.2 Infection Transmission

#### *4.2.2.1 Effective contact rate*

The number of new infections per unit time is determined by the 'effective contact rate', defined here as the average number of infections that would be generated by one infectious case each year in a population of uninfected individuals. Henceforth this is referred to as simply the 'contact rate'. It was assumed that only smear-positive pulmonary cases were infectious, as discussed below, in Section 4.2.2.2. Although infectiousness may depend on the time since disease onset, the contact rate in the model was assumed constant for the duration of disease. Due to the difficulty in estimating the contact rate directly, the value of this parameter is highly uncertain.

In estimates of the contact rate for the pre-chemotherapy era, Styblo used the ratio between the ARI and the prevalence of smear-positive disease to estimate that each infectious case infected 20 others for a two-year duration of infectiousness, or 10 per year [13]. More recent calculations using these methods resulted in estimates of a contact rate between 2.6 and 5.8 per year, using data from Korea, China and the Philippines from the 1970s to 1990s [270]. For England and Wales, the effective

contact rate estimated using this method resulted in a contact rate which dropped to about one per year in 1990 [56]. However, the estimated ARI is thought to be uncertain in recent years, making these estimates also uncertain [271-273]. As discussed in Borgdorff et al. [273], age-dependent patterns in contact rates may have led to underestimation of the overall annual risk of infection, and therefore, the effective contact rate, in recent years.

Modelling approaches have also been used to estimate the contact rate. Recently, Zhou et al. [96] used a compartmental model implemented with difference equations to estimate that each infectious case resulted in about 0.6 new infections each month in Canada based on data from 1991 – 2000. This contact rate resulted in 7.2 infections per year, although this estimate was derived from a mixed population of susceptible and non-susceptible individuals, so the corresponding contact rate as defined here would be higher. The contact rate estimated by Zhou et al. may also be incompatible with the rate used in this model due to different assumptions about disease duration. In a compartmental model taking into account households and commuting, Pienaar et al. [274] estimated that on average, each infectious case infects 13 others per year in a hypothetical urban setting with a high incidence of tuberculosis.

Assumptions used by modellers Dye and Williams [185] were derived from older work by Styblo and unpublished data to assume the contact rate ranges from 10 – 18 per year, with a mean of 14 per year for infectious cases, although this work also focused on countries with a higher burden of tuberculosis than that in the UK. Citing the possible underestimation of the ARI in recent years [273], Wolleswinkel-van den Bosch et al. [224] assumed a contact number corresponding to a higher ARI in their model of tuberculosis in the Netherlands. They assumed there were eight effective contacts per infectious case, or a rate of 16 effective contacts per year with the disease duration of six months assumed here.

In summary, effective contact rates have been estimated or assumed to span a range of 1 – 18 effective contacts per infectious case per year in these and related models [56, 96, 179, 185, 224, 274-276]. Given uncertainty over the value of this parameter, a range of contact rates from 4 – 15 per year was used during fitting of model output to

observed data. The methods sections in Chapters 5 and 7 provide details on the values used in each application of the model.

#### *4.2.2.2 Susceptible individuals*

As described in Chapter 3, only individuals in the *Uninfected* or *Latent Infection* states were considered susceptible to infection or reinfection in the model. Furthermore, it was assumed those with *Latent Infection* were equally susceptible to infection as those *Uninfected*, based on analysis from Vynnycky and Fine (1997), which suggested that previous infection imparts little protection against subsequent infection for those with *Latent Infection* [19].

#### *4.2.2.3 Infectious cases*

It was assumed that all non-pulmonary cases were non-infectious and some pulmonary cases were infectious. Of pulmonary cases, those with smear-positive disease (see Chapter 2, Section 2.2.4) are much more likely to be infectious than smear negative cases [13, 14]. For simplicity, it was assumed that all smear-positive pulmonary cases are infectious and all smear negative cases were assumed non-infectious. The proportion of pulmonary cases that were smear-positive was assumed to vary with age.

Data on the age-specific proportion of pulmonary cases that were smear-positive was available from pulmonary tuberculosis notifications in Norway 1951 – 1969 [19, 267]. Following assumptions in Vynnycky and Fine [19] based on these data, it was assumed that 10% of cases in children 10 years and younger were smear-positive, increasing linearly to 65% smear-positive at 20 years of age. From age 20 years to age 90 years, this percentage was assumed to increase from 65% to 85% and it was assumed constant at 85% for those aged 90 years and older. In reality, the likelihood of developing infectious disease depends on age, HIV status and many other factors.

#### *4.2.2.4 Contact patterns*

A simple contact scheme similar to homogeneous mixing was implemented in this model. The contact scheme only considered an individual's birthplace, UK-born or foreign-born, when choosing transmission contacts for an infectious individual. A

proportion of transmission contacts were randomly chosen within the same birthplace group as the infectious individual, meant to roughly correspond to transmission to known contacts, though the model is not spatially explicit and does not model transmission among close contacts or households. The remaining contacts are randomly chosen from the entire population, including both UK-born and foreign-born, meant to represent non-close or 'casual' contacts. In the model it was assumed that 50% of contacts were close and 50% were casual, although it is unknown exactly what proportion of contacts are close versus casual contacts since this is difficult to estimate directly. The assumption agrees qualitatively with conclusions based on modelling studies by Aparicio et al. [177] and Jia et al. [277], who stressed that casual contacts may be as important, or more important, than close contacts in the transmission of infection. Contact patterns in the model do not take into account age, although doing so would have been more realistic [273].

### 4.2.3 Natural History of Infection

#### *4.2.3.1 Disease progression*

It is known that only a portion of those infected with *M. tuberculosis* develop disease and the risks of disease progression vary by age, sex, HIV status and time since infection, as discussed below and in Chapter 2, Section 2.1.2. It is possible, but unknown, whether disease progression risks are different between UK-born and foreign-born individuals after adjusting for these factors. Because disease progression risks are uncertain, these were allowed to vary in both applications of the model, with estimates obtained by fitting model output to notification data from England and Wales data (See methods sections in Chapters 5 and 7). Still, some parameters regarding the development of disease were fixed, including relative disease risks between some groups in the population to decrease the number of parameters varied in the model. The parameters specified mostly followed from assumptions and results from previous modelling work, for lack of appropriate data [19], and are detailed below. The development of *Primary Disease* and *Reinfection Disease* are dictated by similar rules, described together in Section 4.2.3.1.1. The development of *Reactivation Disease* is handled differently, as described in Section 4.2.3.1.4.

#### 4.2.3.1.1 Risks of developing primary disease and reinfection disease

Following assumptions made in Vynnycky and Fine [19], the risk of disease due to *Recent Infection* or *Reinfection*, resulting in *Primary Disease* and *Reinfection Disease* respectively, were characterized by: 1) the cumulative risk of developing disease during the five years of *Recent Infection* or *Reinfection* and 2) the relative risk of disease for each of the five years since infection.

#### 4.2.3.1.2 Cumulative risks of developing *Primary Disease* and *Reinfection Disease*

Cumulative risks of developing disease likely differ between *Recent Infection* and *Reinfection*, as well as by age at infection, sex, birthplace and HIV status. Cumulative disease risks estimated by the model were those for males aged 20 years and above for each infection type. From those risks, disease risks for all other individuals in the model were calculated, using the relationships between disease risks for different groups in the model, described below.

*Risks by age*

Following Vynnycky and Fine [19], dependency of cumulative disease progression risks on age at infection for those in *Recent Infection* and *Reinfection* states was characterized by a constant risk up to 10 years of age and a constant risk from 20 years of age and above, with a linear increase in disease risk assumed between 10 and 20 years of age. For those over 10 and under 20 years of age, cumulative disease progression rate, $D_a$, at age $a$ is given by

$$D_a = D_{10} + (a-10)(D_{20}-D_{10})/10$$

Age-dependency was assumed to follow the same pattern for all individuals in the model, regardless of sex, birthplace, or HIV status.

Because there were few cases in children under 10 years of age, estimation of these parameters was hindered, though few cases also reduced their importance in the model. Therefore, risks for children under 10 years of age were fixed, based on estimates of the risk of respiratory disease by Vynnycky et al. [9, 271] found in Table

4-2 [19]. In particular, risks for *Reinfection Disease* were estimated based on a small number of cases and may be unreliable. Progression risks and rates for respiratory tuberculosis estimated by Vynnycky et al. were divided by *presp*, the proportion of disease that is respiratory in children under 10 years of age. After dividing by *presp*, the new values represent disease risks for all sites of disease, as used in the model. These are found in Table 4-1.

*Presp* was estimated from data on tuberculosis notifications by age and sex for the time period 1982 – 1995. For both males and females, *presp* was calculated by dividing the number of respiratory cases under ten years of age by the total number of tuberculosis notifications for those under ten years of age. For males, this was 0.69 and for females 0.67 on average over this time period. Note, definitions of non-respiratory disease and non-pulmonary disease are different (See Chapter 2, Section 2.1) and notification procedures for the different disease types have changed over time, which is why this value is different from the assumed proportion of disease that is non-pulmonary, described below.

Table 4-1: Assumed risks and rates of developing all forms of tuberculosis for children ages 0 – 10 years. The values in the rows for Recent Infection and Reinfection reflect the cumulative risks of developing tuberculosis over the first five years of infection and reinfection (%) respectively. For Latent Infection these are rates of developing tuberculosis (% per year).

| Infection type | Male | Female |
|---|---|---|
| *Recent Infection* | 5.92 % | 6.16 % |
| *Reinfection* | 10.05 % | 10.45 % |
| *Latent Infection* | $1.43 \times 10^{-7}$ %/year | $1.49 \times 10^{-7}$ %/year |

*Risks by sex*

Also following Vynnycky and Fine [19], disease progression risks for females were assumed lower than those for males for some age categories. The ratios of risk for females to males, $sd_{a,d}$, were calculated for risks of disease for those in *Recent Infection* and *Reinfection* classes and by age using risks estimated for males and females by Vynnycky et al. [9, 19], found in Table 4-2. There is only a difference in risks

by sex for those with *Reinfection* aged 20 years and older. Although Vynnycky et al. only consider respiratory disease, it was assumed here that the risk ratios of disease progression by sex are the same when all forms of disease are considered. These ratios were fixed in the model such that any values of male disease risks could be used to calculate female disease risks.

*Risks by birthplace*

Because disease risks have not been previously estimated for foreign-born individuals in the UK, these risks were assumed to differ from UK-born risks by a factor *df*, such that cumulative risks of *Primary Disease* and *Reinfection Disease* for UK-born were multiplied by *df* to obtain foreign-born risks. As discussed further in the methods section of Chapter 5, this parameter was estimated by fitting the model to England and Wales notification data. The best-fitting values were then used for the West Midlands application of the model, as described in the methods section of Chapter 7.

*Risks by HIV status*

For those with HIV infection, disease progression probabilities are higher than for those without HIV infection [26]. Therefore, cumulative disease risks for *Primary Disease* and *Reinfection Disease* were multiplied by a factor, *ehiv,* for those with HIV infection in a similar manner to *df*. The value of *ehiv* was assumed to be 7 [26, 278], although there is uncertainty over the average magnitude of increased disease risk due to HIV infection. In reality, the factor of increase in disease risk will depend on the length of time a person has been infected with HIV, the state of their immune system, and whether or not they are on antiretroviral treatment. However, for simplicity, the same factor of increase was applied to all HIV-positive individuals, representing the average factor of increased disease risk across all HIV-positive individuals in the model. It is unknown how well this increased risk applies to foreign-born individuals living in the UK where access to HIV treatment, nutrition, and overall health may be better than in the high-burden countries where tuberculosis disease progression in HIV-positive individuals has typically been studied. It is also possible that *ehiv* depends on whether a person has *Recent Infection*, *Reinfection,* or *Latent Infection,* although for

simplification, it was assumed to increase progression for the three disease types equally.

Table 4-2: Risks of developing respiratory tuberculosis by age and sex for the three types of disease, as estimated by Vynnycky and Fine [19, 279]. Risks for *Recent Infection* and *Reinfection* represent cumulative risks of developing disease in the first five years of infection or reinfection. *Latent Infection* risks are annual risks of developing respiratory disease, actually a rate. Risks were used to calculate 'sd', the ratio of disease risk for females to males by age and type of infection. 'sd' was used to derive female disease risks from male risks for other groups, OF-born and SSA-born.

| Infection type | Age (years) | Risk of developing disease | | Female: Male ratio (*sd*) |
|---|---|---|---|---|
| | | Male | Female | |
| *Recent Infection* | 0 – 10 | 4.06% | 4.06% | 1 |
| | 20+ | 13.8% | 13.8% | 1 |
| *Reinfection* | 0 – 10 | 6.89% | 6.89% | 1 |
| | 20+ | 8.25% | 0.01% | 0.001 |
| *Latent Infection* | 0 – 10 | $9.8 \times 10^{-8}$ %/year | $9.8 \times 10^{-8}$ %/year | 1 |
| | 20+ | 0.0299 %/year | 0.0048 %/year | 0.161 |

#### 4.2.3.1.3 Relative risks of disease over the first five years of infection

The relative risk of disease for each year following with *M. tuberculosis* is the factor by which the risk of disease in that year after infection differs from that in the first year after infection. Relative risks of disease were assumed to be highest in the first year following infection and to decrease over time. Relative risks were assumed as in Vynnycky and Fine [19], who based assumptions on results from the UK Medical Research Council BCG trial. The relative risks for each of the five years of *Recent Infection* and *Reinfection* were fixed for all individuals in the model and did not change with age, sex, birthplace, or HIV status. See Table 4-3 for relative risk over the five years of *Recent Infection* or *Reinfection.*

Table 4-3: Relative risks of disease during first five years of infection.

| Years since infection | Relative risk |
|---|---|
| 1 | 1.000 |
| 2 | 0.410 |
| 3 | 0.130 |
| 4 | 0.086 |
| 5 | 0.028 |

#### 4.2.3.1.4 Rate of progression to reactivation disease

Because those in *Latent Infection* have older infections, *Reactivation Disease* risk was assumed to be much lower than the risk of *Primary Disease* or *Reinfection Disease*. Further, it was assumed that *Reactivation Disease* develops at a rate independent of time since infection, again following assumptions made by Vynnycky and Fine [19]. The rate was assumed to be dependent on age the same way *Primary Disease* and *Reinfection Disease* progression was dependent on age, with a constant rate to age 10 years and from age 20 years and over, and a linear increase in the rate from age 10 – 20 years.

As the case for disease progression to *Primary Disease* and *Reinfection Disease, Reactivation* disease progression risks for females were assumed different to risks for males. The ratios of disease progression rates in *Latent Infection* for females to males, $sd_{a,d}$, shown by age, again calculated from disease development rates estimated for males and females by Vynnycky et al. [9, 19], are found in Table 4-2. These relative risks were fixed in the model such that male disease progression rates could be used to calculate female disease progression rates. As with the risk of *Primary Disease* and *Reactivation Disease*, UK-born progression rates were multiplied by the factor *df* to get foreign-born rates. Rates for HIV-positive individuals were further multiplied by the factor *ehiv,* fixed at 7.

As for *Primary Disease* and *Reinfection Disease*, because there were few cases in children under 10 years of age, these disease risks were fixed based on estimates of the risk of respiratory disease by Vynnycky et al. [9, 271] found in Table 4-2 [19]. Again,

progression rates for respiratory tuberculosis estimated by Vynnycky et al. were divided by *presp*, the proportion of disease that is respiratory in children under 10 years of age. After dividing by *presp*, the new values represented disease risks for all sites of disease, as used in the model. These are found in Table 4-1.

#### *4.2.3.2 Proportion of cases that are pulmonary*

For simplicity, it was assumed the proportion of cases that are pulmonary varies only with sex and birthplace (UK, SSA, OF) and is constant over time and for all age groups. These proportions were calculated directly from ETS data from England and Wales, 1999 – 2009 (See Chapter 2, Section 2.7.1) and presented in Table 4-4. Taking the proportions from observed data allowed the model to reproduce the observed proportion but does not help understanding of the mechanism or effect of different disease types on the proportion of disease that was pulmonary. The proportion of cases that are pulmonary may also depend on age, HIV status, and probably infection history, including duration of infection and whether there have been previous infections. However, as discussed in Chapter 2, Section 2.1.2, evidence for these relationships is less clear and does not justify taking them into account.

Table 4-4: Proportion of tuberculosis disease that is pulmonary, by birthplace and sex. Proportions were calculated from ETS data from England and Wales, 1999 – 2009.

| Birthplace | Male | Female |
|---|---|---|
| UK | 0.74 | 0.67 |
| SSA | 0.57 | 0.52 |
| OF | 0.53 | 0.47 |

#### *4.2.3.3 Recovery from disease*

Recovery from active disease to *Latent Infection* was assumed to happen at an annual rate determined by the average disease duration assumed in the model, important mostly for pulmonary disease because it determines the duration of infectiousness. The average disease duration is difficult to observe or measure, but some epidemiological data can be used to inform the assumption. The average delay from

onset of symptoms to beginning of treatment for reported cases in the UK was likely between 1.5 – 3 months [280-282]. Treatment coverage was high in the UK [105] and those on treatment will have lost infectiousness within weeks of beginning treatment [14]. Accounting for unreported cases further lengthens the average duration of the disease across all patients. This is because undetected cases remain infectious for longer than detected cases. If it is assumed that approximately 25% of cases go undetected (see Section 4.2.5), the average duration of disease may increase by several months. Taking into account these factors, an average disease duration of six months was assumed in the model, as assumed in other recent models [191, 283].

#### *4.2.3.4 Death due to tuberculosis*

The overall proportion of cases that die from tuberculosis is referred to as the 'case fatality rate', though it is actually a probability, not a rate. The case fatality rate was assumed to depend on calendar year, age category, and site of disease, pulmonary or non-pulmonary. Since treatment was not modelled explicitly, the case fatality rate was averaged over treated and untreated cases. Sex, birthplace, and other factors potentially influence the case fatality rate but were not taken into account here due to limited availability of data and uncertainty over the impact of these factors.

Estimates of case fatality rates used in the model were derived from 'death to notification ratios' (DNRs) by age category and disease site from two studies conducted in England and Wales [284, 285]. The DNR was obtained by dividing the number of deaths due to tuberculosis by the number of tuberculosis notifications each year. DNRs estimate the case fatality rate as defined here, assuming deaths occur in the same year as the notification, which is reasonable given evidence that most who die of tuberculosis die within a year of diagnosis [286].

DNRs stratified by age category were calculated by Nissar et al. for the years 1974 – 1987 [284] and were used as the basis for case fatality rates in the model for 1981 – 1987. Firstly however, DNRs by age category and calendar year were split into estimated pulmonary and non-pulmonary rates. The overall case fatality rates for any given age group can be calculated as a weighted average of pulmonary and non-pulmonary case fatality rates and can be described by the following equation

$$cfr_{age} = pulm_{age} \times r_{age} \times cfr_{nonpulm,age} + (1\text{-}pulm_{age}) \times cfr_{nonpulm,age}$$

where $cfr_{age}$ is the age-specific case fatality rate, $pulm_{age}$ is the age-specific proportion of cases that are pulmonary, $cfr_{pulm,age}$ is the age-specific case fatality rate for pulmonary cases, $cfr_{non\text{-}pulm,age}$ is the age-specific case fatality rate for non-pulmonary cases, and $r_{age}$ is the ratio of pulmonary to non-pulmonary case fatality rates for a given age category.

Solving the above equation for $cfr_{nonpulm,age}$ results in

$$cfr_{nonpulm,age} = cfr_{age} \div (pulm_{age} \times r_{age} + (1\text{-}pulm_{age}))$$

To use this equation, first the $r_{age}$ ratios were estimated from the ratio of case fatality rates in respiratory cases to non-respiratory cases given by Martineau et al. from data for 1998 – 1999 [285]. Although there are differences in pulmonary versus respiratory disease definitions (see Chapter 2, Section 2.1) and possibly differences in $r_{age}$ over time, these ratios were assumed to approximate the ratios of pulmonary to non-pulmonary case fatality rates over the study period. This simplifying assumption was used for lack of other data. Ratios were calculated by first combining DNRs for all non-respiratory cases, which were divided into eight categories, using a weighted average based on the number of notifications for each category of non-respiratory disease. Next, respiratory rates were divided by these amalgamated non-respiratory rates. Note that the ratio for the age category of 0 – 14 years was set to one since the null rates for non-respiratory disease did not allow division.

Secondly, estimates for $pulm_{age}$ were obtained from ETS data from 1982 – 2008. It was assumed $pulm_{age}$ in 1981 was equal to $pulm_{age}$ in 1982 and also constant from 2008 – 2011.

The above equation was used to calculate $cfr_{nonpulm,age}$ for each age group from the Nissar et al. data from 1981 – 1987, with $cfr_{pulm,age}$ calculated in a similar manner. Results of the calculations of the $cfr_{nonpulm,age}$ and $cfr_{pulm,age}$ are found in Appendix Table 10-9.

For 1988 – 2001, case fatality rates were estimated using age-stratified estimates of the DNR provided by Martineau et al. [285] . Because the rates stratified by calendar

year for this period were not divided into pulmonary and non-pulmonary rates, the same methods as above were used to estimate $cfr_{nonpulm,age}$ and $cfr_{pulm,age}$ for each age group. Again, data from Martineau et al. were used to specify age-specific ratios of case fatality rates in respiratory cases to non-respiratory cases, as well as $pulm_{age}$ estimates from ETS data. Results are also found in Appendix Table 10-9.

For 2002 – 2011, case fatality rates were extrapolated from 2001 values using average rates of change per year for the DNRs as reported by Martineau from 1988 – 2001 [285]. The average decrease in the DNR was assumed to continue from 2002 – 2011. Rates of decrease were zero for the two youngest age categories. For each of the three older age categories, the DNRs fell by about 1.6 – 2.6% per year. The resulting case fatality rate estimates are found in Appendix Table 10-9.

### 4.2.4 HIV Prevalence

HIV prevalence was important to consider in the model because of its impact on tuberculosis disease progression risk, as discussed in Chapter 2, Section 2.1.2. The HIV prevalence in all groups except SSA-born individuals was assumed to be zero, as the effect of HIV on tuberculosis in these groups was likely negligible in England and Wales. In a recent study, the HIV prevalence in non-African, foreign-born tuberculosis cases ranged from 0.6% to 7.1% and was 1.5% for UK-born cases [287]. Cases in Asian-born individuals, representing nearly one-third of the total cases, were found to have an HIV prevalence of only 0.6%. On the other hand, the same study found about 19.8% of African-born and 20.5% of ethnic Black Africans with tuberculosis were co-infected with HIV [287], most or all of whom were likely SSA-born. Because it is much higher than in other groups, SSA-born HIV prevalence was taken into account in the model.

There are no published estimates of HIV prevalence in SSA-born migrants entering the UK, but estimates of HIV prevalence in heterosexual SSA-born individuals living in England and Wales, by sex, for ages 15 – 44 years for 2001 – 2008 were recently published [288]. These were used to estimate HIV prevalence for all SSA migrants entering the UK from 1981 – 2011. HIV prevalence among SSA-born individuals living in the UK in 1981 and SSA-born migrants entering the UK in 1981 was assumed to be zero. It was assumed that from 1981 to 2001 there was a linear increase in HIV

prevalence, reaching the median prevalence levels estimated for England and Wales in 2001 of 1.3% for males and 2.39% for females [288]. The HIV prevalence of SSA-born migrants was assumed equal to the estimated median HIV prevalence in the population of SSA-born individuals in England and Wales from 2001 – 2008, and was assumed constant from 2008 – 2011 at 1.55% for males and 3.39% for females. The resulting assumed HIV prevalence values for SSA-born migrants by year and sex are found in Table 10-10.

These prevalence values may underestimate HIV prevalence in entering migrants since there were SSA-born individuals residing in the UK by 1981 and they were assigned to be HIV negative at model initialisation in 1981. This means that entering migrants in a given year would have a lower HIV prevalence than that needed to result in the prevalence estimated for the overall population residing in the UK at that time. On the other hand, using the same HIV prevalence estimated for 15 – 44-year-olds for all migrants, regardless of age or sexual orientation, may overestimate prevalence in some groups, though there are few migrants in older and younger age groups. It is unknown whether these effects balance.

### 4.2.5 Proportion of Cases That Are Notified

Although the UK requires reporting of tuberculosis cases to health authorities, not all cases are properly reported and some remain undetected altogether. The proportion of cases that are notified includes unreported or otherwise undetected cases, and is important for comparing model output to data for 1999 – 2009 in the England and Wales model application and 2007 – 2011 in the West Midlands model application. Recently, Van Hest et al. estimated under-notification using record-linkage and capture-recapture methods [289]. Based on capture-recapture models, only 56.2% of cases may be notified, though due to inconsistency with other studies and unmet assumptions of the model, they suggested this method may not have been valid. A review by Pillaye et al. found that across several studies, it was estimated that 7 – 27% of cases were not reported. Taking these sources into account, I assumed that 75% of cases were notified.

For assumptions about the infection status of the population in 1981, disease prevalence was adjusted to account for unreported cases. At this time, it was assumed a slightly higher proportion of cases were undetected, 27%, based on a study conducted during that time period [290].

### 4.2.6 Proportion of Cases That Are Typed With 24-Locus VNTR

For the West Midlands application of the model, the proportion of cases typed with 24-locus VNTR was needed to simulate observed typing data. These proportions were taken directly from observed data—the proportion all ETS notifications from the West Midlands that were successfully typed using 24-locus VNTR from 2007 – 2011. Cases typed successfully included no more than two missing loci. These proportions were stratified by disease site and are found in Table 4-5.

Table 4-5: Proportion of cases typed by disease site, for the West Midlands 2007 – 2011. Proportion typed refers to the proportion of notified cases that were successfully typed with 24-locus VNTR, including no more than two missing loci.

| | Proportion typed with 24-locus VNTR | | | | |
|---|---|---|---|---|---|
| Disease site | 2007 | 2008 | 2009 | 2010 | 2011 |
| Pulmonary | 52.5% | 55.1% | 57.9% | 63.6% | 57.5% |
| Non-pulmonary | 35.7% | 35.3% | 37.3% | 38.6% | 32.4% |

### 4.2.7 Infection status of Initial Population, 1981

The proportion of individuals in each of the infection states, by age and birthplace (UK and foreign-born), for all individuals in England and Wales in 1981 was estimated using a simple spreadsheet model. This model took into account disease incidence data, assumptions about vaccination practices in the UK and abroad, the estimated ARIs in England and Wales, and the average estimated ARI experienced by foreign-born individuals living in the UK throughout their time abroad and in the UK. To simplify calculations, individuals were assigned one of eight infection states—instead of 11 states—by combining pulmonary and non-pulmonary disease categories for each of the three disease types, *Primary Disease*, *Reactivation Disease*, and *Reinfection Disease*. Disease was then designated as pulmonary or non-pulmonary when the

simulation model was run, a random assignment based on the same probabilities used for the duration of the simulation (see Section 4.2.3.2). Since the proportions of the population in the different infection states were uncertain and relied on many assumptions, the model was run for several years before model output was fit to notification data, from 1981 – 1998 for the England and Wales application and from 1981 – 2006 for the West Midlands application.

The method for obtaining proportions in each of the infection states was the same for UK-born and foreign-born individuals, although assumptions about ARI, disease incidence, and vaccination practices varied between the two population groups. These assumptions are described below.

#### *4.2.7.1 ARI*

For UK-born individuals, the ARI estimated by Vynnycky and Fine 1997 [291] for each year from 1860 – 1949 for England and Wales was used. A 13% decline in ARI per annum was assumed after 1949 [291]. For foreign-born individuals, the average ARI experienced was unknown and difficult to estimate. This lack of certainty was because the ARI for foreign-born individuals was the average ARI experienced over the average foreign-born individual's time abroad and time in the UK, across all age groups. For lack of data, the ARI for foreign-born was assumed equal to that of UK-born individuals in England and Wales up until 1955 and constant thereafter at 0.84% per year. It was assumed that SSA-born individuals in the UK in 1981 had the same exposure as other foreign-born individuals, though this was altered in some simulations for the England and Wales application of the model, as discussed in Chapter 5, Section 5.1.5. It is unknown how well this assumption reflects reality, however, the assumptions used to initialize the population in 1981 are unlikely to have made a major impact on simulated incidence because the model was run for several years before fitting model output to observed data. A plot of the assumed ARIs experienced by UK-born and foreign-born individuals is found in Figure 4-1.

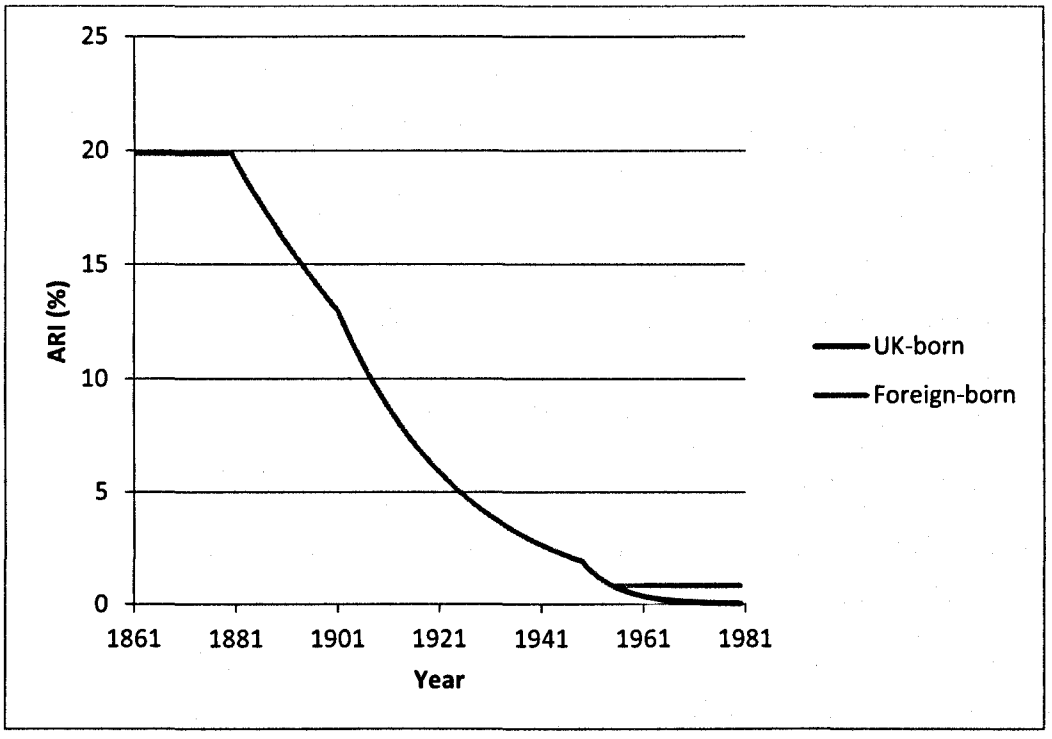


Figure 4-1: Assumed ARIs (%) experienced by those living in the UK in 1981, by country of birth. These ARIs were used to establish the age-specific infection and disease prevalence for the initial model population in 1981.

#### *4.2.7.2 Vaccination practices*

Vaccination practices are detailed in Section 4.2.1.

#### *4.2.7.3 Prevalence of disease*

The number of tuberculosis notifications for 1981 was used to estimate disease prevalence for the start of the simulation at the beginning of 1981. There were 8,128 total incident cases reported in 1981 [292], but this number was adjusted to account for underreporting. Assuming 27% of tuberculosis cases were unreported at this time (See Section 4.2.5 and reference [290]), the total number of cases was calculated as 8,128/(1 – 0.27), for an adjusted total of 11,134 incident cases in 1981.

The birthplace of tuberculosis cases was not reported, so assignment of cases as UK-born and foreign-born was based on trends in the proportion of cases born in the UK in five-year tuberculosis surveys in England and Wales. Surveys reported the birthplace of cases in 1988 and 1993. In 1988, about 55% of cases were UK-born [2], while in 1993, 50% were UK-born [293]. It is assumed that the decreasing trend in the proportion of cases born in the UK from 1988 to 1993 applies to previous years and could be used to

estimate the proportion of cases born in the UK in 1981. Following trends seen from 1988 – 1993, an estimated 62.7% of cases were UK-born in 1981. This proportion resulted in an estimated 6,977 UK-born cases and 4158 foreign-born cases of tuberculosis in 1981.

Next, the number of total *prevalent* cases was estimated. It was assumed the prevalence to annual incidence ratio was 1:2 in 1981, consistent with the six-month duration of infectiousness assumed in the model (See Section 4.2.3.3). The prevalent cases for 1981 was then assumed to be half of the number of incident disease cases, or 3,488 UK-born cases and 2,079 foreign-born cases. The number of prevalent cases for each birthplace was divided by the total population for each group to obtain the proportion of individuals with prevalent disease. Although the age and sex distribution of cases was unrealistic, it should not affect model results because there is no age- or sex-dependent mixing assumed and those cases will have long recovered by the time the model is fit to data after almost 20 years of simulation.

#### *4.2.7.4 Spreadsheet model for calculating infection state prevalences*

The above ARI, vaccination, and disease prevalence assumptions were used for calculations performed using a spreadsheet model used to obtain the proportions in each infection state by birthplace (UK or foreign-born), age, and sex at the end of 1980, to initialize the population starting at the beginning of 1981. The spreadsheet model was implemented in Excel, adapted from work presented in a recent textbook (chapter nine) [294]. First, the age-specific proportion of individuals who avoided infection, or remained *Uninfected*, was calculated up until the age of vaccination, if applicable, for each one-year cohort born from 1861 – 1980. Generally, the proportion of individuals at age $a$ who remain *Uninfected* at age $a$, at the end of year $t$ is

$$U_{a,t}=1 - (1-ARI_{t-a})(1-ARI_{t-a+1})(1-ARI_{t-a+2})....(1-ARI_{t-1})(1-ARI_{t})$$

where $U_{a,t}$ is the proportion *Uninfected* at age $a$, at the end of the year of interest $t$, and $ARI_t$ is the ARI in year t. However, in the simulation model the average individual in

a one-year birth cohort experienced only half the ARI in their year of birth, $t$-$a$, as individuals in a birth cohort were assumed to be born at random times throughout the year of birth. The formula above was modified to take this into account and the proportion $U_{a,t}$ at age $a$ and the end of year $t$ became

$$U_{a,t}=1-(1-0.5xARI_{t-a})(1-ARI_{t-a+1})(1-ARI_{t-a+2})....(1-ARI_{t-1})(1-ARI_t)$$

When vaccination was included in the spreadsheet model for UK-born individuals, the formula above was applied up until the age of vaccination. At the assumed year of vaccination for the cohort, the proportion of *Uninfected* individuals at the end of that year was multiplied by the vaccine coverage for that year, $vc_t$, and efficacy of the vaccine, $ve$, to get the proportion vaccinated effectively, or those assigned to the *Immune* class, $V_{a,t}$. This proportion is given by

$$V_{a,t} = U_{a,t} x\ vc_t x\ ve$$

for age $a$ and year $t$. The proportion $V_{a,t}$ were subtracted from the proportion *Uninfected* left at the end of the year so that $U^*_{a,t}$ represented the corrected number of *Uninfected*, given by

$$U^*_{a,t} = U_{a,t} - V_{a,t}$$

and the new $U^*_{a,t}$ replaced $U_{a,t}$ in subsequent calculations. After vaccination, the ARI was again applied to the remaining *Uninfected* each year, given by

$$U_{a,t} = U_{a-a,t-1} - U_{a-a,t-1} x\ ARI_t$$

This process was continued to the end of 1980 for each birth cohort. The proportions $U_{a,t}$ and $V_{a,t}$ at the end of 1980 for each birth cohort were used for those proportions *Uninfected* and *Immune* for the initial population in 1981. The proportion in *Uninfected* and *Immune* classes was also tabulated and saved for several years before 1981 for use in estimating the proportions in the other infection states.

The remaining individuals, those not assigned *Uninfected* or *Immune*, were then assigned to one of the three infection states preliminarily. The preliminary assignments were used as a starting point for the final proportions in each infection state and also to assign disease cases to one of the three infection states. The first step

in assigning preliminary infection classes was to assign those individuals infected in the five years prior to 1981 to the *Recent Infection* class. This assignment was derived from the tabulation of $U_{a,t}$ such that those preliminarily assigned to *Recent Infection*, $Ip_{a,t}$, at age *a* and at the end of year *t* were given by

$$Ip_{a,t} = U_{a-5,t-5} - U_{a,t}$$

Note that for those under age five, the first term, $U_{a-5,t-5}$, equals one. Also note, this formula did not work for all individuals, as it had to be corrected for those vaccinated, if applicable. In that case the formula became

$$Ip_{a,t} = U_{a-5,t-5} - U_{a,t} - V_{a,t}$$

Next, the proportion of individuals remaining after accounting for the proportion $U_{a,t}$, $V_{a,t}$ and $Ip_{a,t}$ were those who had been infected more than five years previously, preliminarily assigned to the *Latent Infection* class and called $Lp_{a,t}$. The proportion $Lp_{a,t}$ was used as a proxy for those at risk for *Reinfection* and this proportion was calculated for the five years prior to 1981. It should be noted that this was an approximation that over-estimated those at risk for *Reinfection* since some proportion of these would have been reinfected and be in the *Reinfection* or *Reinfection Disease* states. The over-estimation of those at risk led to over-estimation of the proportion in the *Reinfection* class, though since the ARI in the five years prior to 1981 was quite low, this over-estimation was likely small. For each of the five years prior to 1981, the ARI for the year was multiplied by $Lp_{a,t}$ to get the estimated proportion reinfected in that year. The sum over those five years gave the proportion estimated to have *Reinfection* at the end of 1980, calculated by

$$Rp_{a,t} = (Lp_{a,t-5} \times ARI_{t-5}) + (Lp_{a,t-4} \times ARI_{t-4}) + (Lp_{a,t-3} \times ARI_{t-3}) + (Lp_{a,t-2} \times ARI_{t-2}) + (Lp_{a,t-1} \times ARI_{t-1})$$

This proportion, $Rp_{a,t}$, was then subtracted from $Lp_{a,t}$ to approximate the proportion in the *Latent Infection* class, $Lp^*_{a,t}$, calculated by

$$Lp^*_{a,t} = Lp_{a,t} - Rp_{a,t}$$

Lastly, the proportion with prevalent disease in 1981 was divided into *Primary Disease*, *Reactivation Disease* and *Reinfection Disease* classes. The proportion assigned to each

disease type was proportional to the product of the estimated proportions at risk and the estimated risk of disease for each infection class. The proportions at risk are estimated to be those preliminarily assigned to *Recent Infection*, *Latent Infection* and *Reinfection* ($Ip_{a,t}$, $Lp^*_{a,t}$ and $Rp_{a,t}$), respectively. The risk of disease for the three infection classes by age and sex was taken from disease risk estimates for white males in England and Wales from Vynnycky et al. [19]. After the total proportion of prevalent disease was assigned to each disease type, these proportions were subtracted from the corresponding infection class to finalize proportions in each infection state used in the model. The resulting proportions in each infection state by birthplace, age, and sex are found in Appendix 10.7.

### 4.2.8 Infection Status of Migrants Upon Entry to UK

Similar to initialization of the model population in 1981, infection states were assigned to migrants when they entered the population throughout the simulation. Obtaining estimates for the proportion of migrants in each of the infection states was difficult for various reasons. Firstly, even if migrants were screened for infection or disease, there would be no way to clinically distinguish or otherwise ascertain the precise infection states necessary for the model. Secondly, even information on infection and disease prevalence, which could be used for rough groupings for the model, is sparse and difficult to come by. The screening programs that are in place are inconsistently practiced and also not designed for surveillance of infection and disease prevalence in migrants (see Section 2.3.1). Also, up until recently, tests for infection were unreliable (see Section 2.2.5). Therefore, given the uncertainty in estimating infection state proportions for migrants and their importance for model output, two separate methods were used to generate scenarios for the proportions of migrants in the different infection states upon entry to the UK.

The 'screening method' used screening data from new migrants to the UK to base assumptions about the probabilities of different infection states. Two different distributions of infection state probabilities for migrants were generated using this method. The 'ARI method' is similar to the method used to obtain the infection status of the population in 1981 (see above), combining assumptions about the ARI

experienced by migrants before arrival to the UK, disease prevalence in migrants, and vaccination practices abroad to obtain proportions in the different infection states. Three different distributions for the infection state probabilities of migrants were generated using this method. The five total distributions for infection state probabilities for migrants upon entry to the UK were used for model fitting for the England and Wales application of the model in Chapter 5, whereas only the two screening method scenarios were used for the West Midlands application of the model, as described in the methods for Chapter 7.

#### *4.2.8.1 Screening method*

Although there is considerable uncertainty over interpretation of TST data, the screening method used tuberculosis screening studies to estimate disease and infection prevalence in foreign-born migrants to England and Wales. Two related schemes were used to interpret the screening data, *Scr1* and *Scr2*, translating data into infection state probabilities by age, sex, and birthplace (UK, OF, SSA).

##### 4.2.8.1.1 Screening data

Most screening studies conducted in the UK have focused on refugees or asylum seekers, though one study by Ormerod and colleagues looked at all immigrants to a region, and also stratified immigrants by age [295]. Ormerod et al. reported Tine tuberculin test results for new immigrants to the Blackburn, Hyndburn, and Ribble Valley District Health Authority from 1983 – 1988, which consisted mainly of Indian and Pakistani immigrants. They reported the frequency of five test grades, 0 – 4, by ethnic group and age class for all years combined. Age classes reported were: 0 – 4 years; 5 – 14 years; 15 – 29 years; 30 – 44 years; 45 – 64 years; and 65 years and above. Ethnic groups were reported, but they were combined due to small sample sizes. Appendix 10.8 provides a table of the proportions in each test grade by age, estimated from the combined data found in Figure 1 of the Ormerod paper [295]. The test grade data were combined with assumptions about how test grades correspond to infection states to assign proportions in the different infection states for migrants. This was done for OF-born infection state probabilities only. UK-born and SSA-born infection state probabilities were derived using only the OF-born estimates and some further assumptions. This was done because UK-born migrants were not included in

the screening study, so must be handled separately, and because SSA-born migrants were included in the screening study but made up a small proportion of immigrants to the UK (about 11% of all foreign-born individuals in recent years). Another reasons SSA-born were handled separately is because of their likely higher infection and disease prevalence, which was expected to differ from the average values for all immigrants

#### 4.2.8.1.2 Rules for partitioning OF-born individuals into five infection states

In the first scheme, *Scr1*, it was assumed that test grades 0 and 1 corresponded to non-reactions and these individuals were *Uninfected*. It was assumed that test grades 2 and 3 were reactions to inactive infections and these individuals were placed in the *Latent Infection* class. There was a sizeable proportion of individuals with grade 4 reactions, likely more than would have *Recent Infection*, *Reinfection* and active disease. Therefore, it was assumed 10% of grade 4s had active disease, which gave an active disease prevalence roughly comparable to the prevalence found in screening studies. A further 40% of grade 4s were assumed to have a *Recent Infection* or *Reinfection*. The remaining 50% of grade 4s were added to the *Latent Infection* class. It was also assumed that BCG vaccination in migrants was negligible. Although many immigrants had been vaccinated upon arrival to the UK, there is much debate over the efficacy of vaccination in many places in the world. The clearest evidence shows that BCG protects against childhood tuberculosis in many places in the world, but has limited success in prevention of adult disease [60, 296]. Since the protection from childhood tuberculosis in migrants was not important for modelling purposes here, it was assumed that no immigrants entered as *Immune*. Rules and resulting proportions in the different infection states for OF-born are found in Table 4-6.

The second scheme for interpretation of screening data used to devise another distribution for *infimm* is similar to *Scr1*, though differs in how TST results from those aged 15 years and above are interpreted. Table 4-7 details the rules used for translating the TST results into infection states in *Scr2 f*or each age group, along with resulting proportions in a subset of generalized infection states, five in total.

Table 4-6: Rules for translating Tine test grades into infection state categories by age category for OF-born migrants under Scheme *Scr1* and the resulting proportions of OF-born in preliminary infection state categories.

| Infection state | Rule | Proportion assigned to infection state by age class (years) | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0 – 4 | 5 – 14 | 15 – 29 | 30 – 44 | 45 – 64 | 65+ |
| *Uninfected* | Grades 0, 1 | 0.920 | 0.803 | 0.695 | 0.627 | 0.623 | 0.414 |
| *Immune* | None | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| *Latent Infection* | Grades 2, 3 and 50% of grade 4 | 0.070 | 0.180 | 0.263 | 0.321 | 0.294 | 0.496 |
| Recent (re)infection | 40% of grade 4 | 0.008 | 0.013 | 0.033 | 0.041 | 0.067 | 0.072 |
| Active disease | 10% of grade 4 | 0.002 | 0.003 | 0.008 | 0.010 | 0.017 | 0.018 |

Table 4-7: Rules for translating Tine test grades into infection state categories by age category for OF-born migrants under scheme *Scr2* and resulting proportions of OF-born individuals in preliminary infection state categories by age.

| Infection state | Rules (Under 15 yrs) | Proportion assigned to infection state by age class (years) | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 0 – 4 yrs | 5 – 14 yrs | Rules (15 yrs and over) | 15 – 29 yrs | 30 – 44 yrs | 45 – 64 yrs | 65+ yrs |
| *Uninfected* | Grade 0s, 1s | 0.92 | 0.80 | Grade 0s, 1s, 50% 2s | 0.77 | 0.72 | 0.70 | 0.52 |
| *Immune* | None | 0.00 | 0.00 | None | 0.00 | 0.00 | 0.00 | 0.00 |
| *Latent Infection* | Grade 2s, 3s, 50% 4s | 0.00 | 0.00 | Grade 50% 2s, 3s, 50% 4s | 0.01 | 0.01 | 0.02 | 0.02 |
| *Recent Infection* or *Reinfection* | 40% Grade 4s | 0.01 | 0.01 | 40% Grade 4s | 0.03 | 0.04 | 0.07 | 0.07 |
| Active disease | 10% Grade 4s | 0.07 | 0.18 | 10% Grade 4s | 0.19 | 0.23 | 0.22 | 0.39 |

#### 4.2.8.1.3 Rules for subdividing infection and disease categories

The above assumptions split OF-born migrants into five classes, *Uninfected*, *Immune*, *Latent Infection*, *Recent Infection,* or *Reinfection*, and active disease (combination of all six disease classes in the model). The active disease category and the *Recent Infection* or *Reinfection* category were divided by making some additional assumptions. The active disease category was divided according to the age-specific proportions of the three disease types estimated for foreign-born individuals at model initialization in

1981, with values are given in Table 4-8. The *Recent Infection* or *Reinfection* category was also split according to the age-specific estimates obtained for foreign-born individuals at model initialization in 1981, with values given in Table 4-9. After these two categories were split, there were eight disease classes, not the full 11 states found in the model, since pulmonary and non-pulmonary disease were combined for each disease type. Disease site was assigned when the model was run. For *Scr1*, these are found in Table 4-10. For *Scr2*, these are found in Table 4-11.

Table 4-8: Proportion of all active disease assumed for each disease type, *Primary Disease*, *Reactivation Disease*, and *Reinfection Disease* by age category for migrants.

| Infection state | Proportion assigned to infection state | | | | | |
|---|---|---|---|---|---|---|
| | 0 – 4 yrs | 5 – 14 yrs | 15 – 29 yrs | 30 – 44 yrs | 45 – 64 yrs | 65+ yrs |
| *Primary Disease* | 1.00 | 0.97 | 0.99 | 0.98 | 0.93 | 0.28 |
| *Reactivation Disease* | 0.00 | 0.00 | 0.01 | 0.02 | 0.07 | 0.71 |
| *Reinfection Disease* | 0.00 | 0.03 | 0.01 | 0.00 | 0.00 | 0.01 |

Table 4-9: Proportion recent (re)infections assigned to *Recent Infection* and *Reinfection* states by age category for migrants.

| Infection state | Proportion assigned to infection state | | | | | |
|---|---|---|---|---|---|---|
| | 0-4 yrs | 5-14 yrs | 15-29 yrs | 30-44 yrs | 45-64 yrs | 65+ yrs |
| *Recent Infection* | 1.00 | 0.97 | 0.87 | 0.68 | 0.36 | 0.03 |
| *Reinfection* | 0.00 | 0.03 | 0.13 | 0.32 | 0.64 | 0.97 |

Table 4-10: Assumed proportions for each infection state by age category for other foreign-born migrants upon entry to the UK, generated by the *Scr1* scheme of the screening method.

| State | 0 – 4 years | 5 – 14 years | 15 – 29 years | 30 – 44 years | 45 – 64 years | 65+ years |
|---|---|---|---|---|---|---|
| *Uninfected* | 0.919636 | 0.802986 | 0.695497 | 0.627303 | 0.622697 | 0.413556 |
| *Immune* | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| *Recent Infection* | 0.008291 | 0.012976 | 0.028628 | 0.027302 | 0.022967 | 0.002034 |
| *Latent Infection* | 0.070000 | 0.180379 | 0.263103 | 0.321096 | 0.293596 | 0.495889 |
| *Reinfection* | 0.000000 | 0.000332 | 0.004492 | 0.013979 | 0.043999 | 0.070410 |
| *Primary Disease* | 0.002073 | 0.003321 | 0.008225 | 0.010119 | 0.015517 | 0.004729 |
| *Reactivation Disease* | 0.000000 | 0.000006 | 0.000054 | 0.000197 | 0.001200 | 0.007047 |
| *Reinfection Disease* | 0.000000 | 0.000000 | 0.000001 | 0.000004 | 0.000024 | 0.006335 |

Table 4-11: Assumed proportions in each infection state by age category for other foreign-born migrants upon entry to the UK, generated by the *Scr2* scheme of the screening method.

| Infection state | 0-4 years | 5-14 years | 15-29 years | 30-44 years | 45-64 years | 65+ years |
|---|---|---|---|---|---|---|
| *Uninfected* | 0.919636 | 0.802986 | 0.773236 | 0.722191 | 0.696966 | 0.523333 |
| *Immune* | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| *Recent Infection* | 0.008291 | 0.012976 | 0.028628 | 0.027302 | 0.022967 | 0.002034 |
| *Latent Infection* | 0.070000 | 0.180379 | 0.185364 | 0.226208 | 0.219326 | 0.386111 |
| *Reinfection* | 0.000000 | 0.000332 | 0.004492 | 0.013979 | 0.043999 | 0.070410 |
| *Primary Disease* | 0.002073 | 0.003321 | 0.008225 | 0.010119 | 0.015517 | 0.004729 |
| *Reactivation Disease* | 0.000000 | 0.000006 | 0.000054 | 0.000197 | 0.001200 | 0.007047 |
| *Reinfection Disease* | 0.000000 | 0.000000 | 0.000001 | 0.000004 | 0.000024 | 0.006335 |

#### 4.2.8.1.4 Rules for obtaining UK-born infection state proportions

To obtain analogous estimates of infection state proportions for UK-born individuals, the proportion of those in the *Uninfected* class by age category was first estimated using the spreadsheet model described in Section 4.2.7.4. The age specific proportions of *Uninfected* UK-born individuals were calculated for 1995, the average year of migration, to simplify calculations. For these calculations, it was assumed that the ARI experienced was the same as the ARI for England and Wales until the year 1960, after which the ARI was constant at 0.4% until 1995. This ARI was chosen to reflect a higher ARI compared with those living in England and Wales due to potential exposures abroad. The proportion *Uninfected* by age for the average time of migration, 1995, was estimated using the spreadsheet model. These proportions were then averaged into age classes, which corresponded to age classes in the screening data, which were used for the proportion of *Uninfected* UK-born individuals in this method. Next, the proportion *Uninfected* for UK-born was compared the proportion for OF-born individuals for estimating proportions in other infection states. For each age class, the ratio of estimated UK-born *Uninfected* to OF-born *Uninfected* was calculated. These ratios were then used to estimate plausible values for other UK-born infection states by age category. This was done by reducing the *Recent Infection* and *Reinfection* classes and the three disease classes by the inverse of these ratios. The proportion with *Latent Infection* was assigned to be the remainder, after these adjustments. The resulting proportions for UK-born are found in Table 4-12 and Table 4-13.

Table 4-12: Infection state probabilities for UK-born migrants by age category, estimated using the Scr1 method.

| Infection state | 0 – 4 years | 5 – 14 years | 15 – 29 years | 30 – 44 years | 45 – 64 years | 65+ years |
|---|---|---|---|---|---|---|
| *Uninfected* | 0.989655 | 0.959255 | 0.910651 | 0.847613 | 0.625054 | 0.116202 |
| *Immune* | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| *Recent Infection* | 0.007704 | 0.010862 | 0.021864 | 0.020206 | 0.022880 | 0.007239 |
| *Latent Infection* | 0.000715 | 0.026820 | 0.057730 | 0.114198 | 0.291554 | 0.561516 |
| *Reinfection* | 0.000000 | 0.000278 | 0.003431 | 0.010345 | 0.043833 | 0.250587 |
| *Primary Disease* | 0.001926 | 0.002780 | 0.006282 | 0.007489 | 0.015459 | 0.016831 |
| *Reactivation Disease* | 0.000000 | 0.000005 | 0.000042 | 0.000146 | 0.001195 | 0.025079 |
| *Reinfection Disease* | 0.000000 | 0.000000 | 0.000001 | 0.000003 | 0.000024 | 0.022547 |

Table 4-13: Infection state probabilities for UK-born migrants by age category, estimated using the Scr2 method.

| Infection state | 0 – 4 years | 5 – 14 years | 15 – 29 years | 30 – 44 years | 45 – 64 years | 65+ years |
|---|---|---|---|---|---|---|
| *Uninfected* | 0.989655 | 0.959255 | 0.910651 | 0.847613 | 0.625054 | 0.116202 |
| *Immune* | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| *Recent Infection* | 0.007704 | 0.010862 | 0.024308 | 0.023262 | 0.025609 | 0.009161 |
| *Latent Infection* | 0.000715 | 0.026820 | 0.054196 | 0.108422 | 0.281608 | 0.475966 |
| *Reinfection* | 0.000000 | 0.000278 | 0.003814 | 0.011910 | 0.049061 | 0.317105 |
| *Primary Disease* | 0.001926 | 0.002780 | 0.006984 | 0.008622 | 0.017302 | 0.021299 |
| *Reactivation Disease* | 0.000000 | 0.000005 | 0.000046 | 0.000168 | 0.001338 | 0.031736 |
| *Reinfection Disease* | 0.000000 | 0.000000 | 0.000001 | 0.000003 | 0.000027 | 0.028532 |

#### 4.2.8.1.5 Rules for obtaining SSA-born infection state proportions

To obtain estimates for SSA-born infection state proportions, a method similar to the one used for UK-born was employed. The proportion of SSA-born *Uninfected* by age was calculated assuming the same ARIs experienced by those in England and Wales, decreasing until 1947 and constant thereafter at about 1%. The proportion *Uninfected* by age for the average time of migration, 1995, was estimated. These proportions were then averaged into age classes that correspond to age classes in the screening data. For each age class, the ratio between estimated proportions of *Uninfected* SSA-born individuals and *Uninfected* OF-born individuals was calculated and the *Recent Infection* and *Reinfection* classes and the three disease classes were increased by the inverse of this ratio. The proportion with *Latent Infection* was assigned to be the remainder after these adjustments. The resulting proportions for SSA-born individuals can be found in Table 4-14 for *Scr1* and Table 4-15 for *Scr2*.

Table 4-14: Infection state probabilities for SSA-born migrants by age category, estimated using the *Scr1* method.

| Infection state | 0-4 years | 5-14 years | 15-29 years | 30-44 years | 45-64 years | 65+ years |
|---|---|---|---|---|---|---|
| Uninfected | 0.959093 | 0.846127 | 0.686635 | 0.533466 | 0.371400 | 0.069046 |
| *Immune* | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| *Recent Infection* | 0.007950 | 0.012314 | 0.028998 | 0.032105 | 0.038507 | 0.012183 |
| *Latent Infection* | 0.030970 | 0.138086 | 0.271430 | 0.405856 | 0.488254 | 0.388564 |
| *Reinfection* | 0.000000 | 0.000315 | 0.004550 | 0.016438 | 0.073770 | 0.421730 |
| *Primary Disease* | 0.001987 | 0.003152 | 0.008331 | 0.011899 | 0.026016 | 0.028326 |
| *Reactivation Disease* | 0.000000 | 0.000006 | 0.000055 | 0.000232 | 0.002012 | 0.042207 |
| *Reinfection Disease* | 0.000000 | 0.000000 | 0.000001 | 0.000005 | 0.000041 | 0.037946 |

Table 4-15: Infection state probabilities for SSA-born migrants by age category, estimated using the Scr2 method.

| Infection state | 0-4 years | 5-14 years | 15-29 years | 30-44 years | 45-64 years | 65+ years |
|---|---|---|---|---|---|---|
| *Uninfected* | 0.959093 | 0.846127 | 0.686635 | 0.533466 | 0.371400 | 0.069046 |
| *Immune* | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| *Recent Infection* | 0.007950 | 0.012314 | 0.032239 | 0.036961 | 0.043100 | 0.015417 |
| *Latent Infection* | 0.030970 | 0.138086 | 0.266743 | 0.396678 | 0.471514 | 0.244587 |
| *Reinfection* | 0.000000 | 0.000315 | 0.005059 | 0.018924 | 0.082569 | 0.533677 |
| *Primary Disease* | 0.001987 | 0.003152 | 0.009262 | 0.013699 | 0.029119 | 0.035845 |
| *Reactivation Disease* | 0.000000 | 0.000006 | 0.000061 | 0.000267 | 0.002252 | 0.053411 |
| *Reinfection Disease* | 0.000000 | 0.000000 | 0.000001 | 0.000005 | 0.000046 | 0.048018 |

#### *4.2.8.2 ARI method*

The 'ARI method' was used to tabulate age-specific proportions in each infection state using the same method as that used for the initial population in 1981, described above in Section 4.2.7. Briefly, the ARIs for migrants were used along with assumptions about disease incidence to inform a spreadsheet model used to estimate age-dependent proportions in the different infection states for each year and birthplace (UK, OF, SSA). Three schemes were generated this way, one for each of three different assumptions about the ARIs experienced by migrants: 'ARI low', 'ARI mid' and 'ARI high'. ARI and disease prevalence assumptions are described below. This method was only used for fitting scenarios in the England and Wales application of the model.

##### 4.2.8.2.1 ARI estimates

Since older estimates of ARI were not available, it was assumed that ARIs for all countries followed patterns similar to those in the UK, falling to recently estimated values from nearly 20% in 1861. It was assumed that once the ARI fell to the recently estimated value, it was constant thereafter. The recently estimated values were different for each birthplace (UK, OF, SSA) and are outlined below.

Firstly, estimates of the ARI in all regions of the world for recent years were compiled. Recent ARI estimates were divided into UK, 'rest of Europe', North Africa, Sub-Saharan Africa, Indian Subcontinent, 'rest of Asia', America, and Oceania, consistent with migration data provided by ONS. ARIs estimated for each region were obtained from WHO estimates or other literature sources. A range of ARI estimates was established for each region, as well as low, mid and high values for each. These are found in Table 4-16, along with notes on the literature sources for the estimates. Many of these estimates were taken from a 1988 WHO report by Cauthen et al., where ARIs were estimated in many countries of the world using data from 1975 onward [297]. There were also several studies of the ARI in individual countries and regions within countries [298-302].

ARIs for UK-born and SSA-born individuals were assumed constant for recent years, including every year of the simulation, at values listed in Table 4-16. For OF-born migrants, overall ARIs experienced recently for each year of the simulation were estimated by calculating a weighted average of ARIs from the various regions of the world, with weights for each year determined by the number of migrants to England and Wales from each region. These weighted-average ARIs were calculated for each year of the simulation and separately for the low, mid, and high estimates of ARI. Results for the weighted-average ARIs experience in recent years for OF-born for each year of the simulation are found in Figure 4-2.

These ARI estimates for recent years were joined with assumptions about the ARIs experienced in the longer term, using the historic trajectory of ARI described above. For example, if an OF-born migrant entered the UK in 2009, the recently estimated low ARI value was about 0.4%, so it was assumed the ARI for these migrants decreased from nearly 20% in 1861 to about 0.4% in 1961 and remained constant thereafter, as shown in Figure 4-3. This pattern was followed for OF-born individuals in other years and for UK-born and SSA-born individuals. The ARI decreased to the recently estimated level and remained constant thereafter.

Table 4-16: Estimates of the annual risk of infection (ARI) by world region. These values were used to estimate infection and disease prevalence for migrants to the UK in the ARI method. Note, values for the UK reflect assumptions about ARIs experienced by UK-born migrants returning to England and Wales after having lived abroad and reflect higher ARIs than currently estimated for the UK.

| World region | ARI (%) | | | References |
|---|---|---|---|---|
| | Low | Mid | High | |
| UK | 0.400 | 0.800 | 1.700 | To reflect ARIs experienced abroad, ARIs estimated for England and Wales for 1960, 1955, and 1950 [271] were taken as the low, mid, and high estimates. |
| Sub-Saharan Africa | 1.000 | 2.000 | 3.000 | Cauthen et al. African region, Tanzania Tuberculin Survey Collaboration 2001 [297]. |
| Rest of Europe | 0.016 | 0.042 | 0.067 | Sutherland et al. 1983, estimated from data from the Netherlands, where low set at 1979 value, high at 1967 value and mid between the two [303]. |
| North Africa | 0.150 | 0.325 | 0.500 | Cauthen et al.1988, Eastern Mediterranean region data [297]; Ibiary et al. 1999 |
| Indian Subcontinent | 1.000 | 1.500 | 2.000 | Cauthen et al. 1988, South-East Asia region [297]; Gopi et al. 2006 [300]. |
| Rest of Asia | 0.500 | 0.750 | 1.000 | Cauthen et al. 1988, Western Pacific region [297]; Norval et al. 2004 [301]. |
| America | 0.070 | 0.385 | 0.700 | Cauthen et al. 1988, Region of the Americans [297, 300]; Salpeter & Salpeter 1999 [304]. |
| Oceania | 0.070 | 0.385 | 0.700 | Assumed to be the same values as for 'America'. |

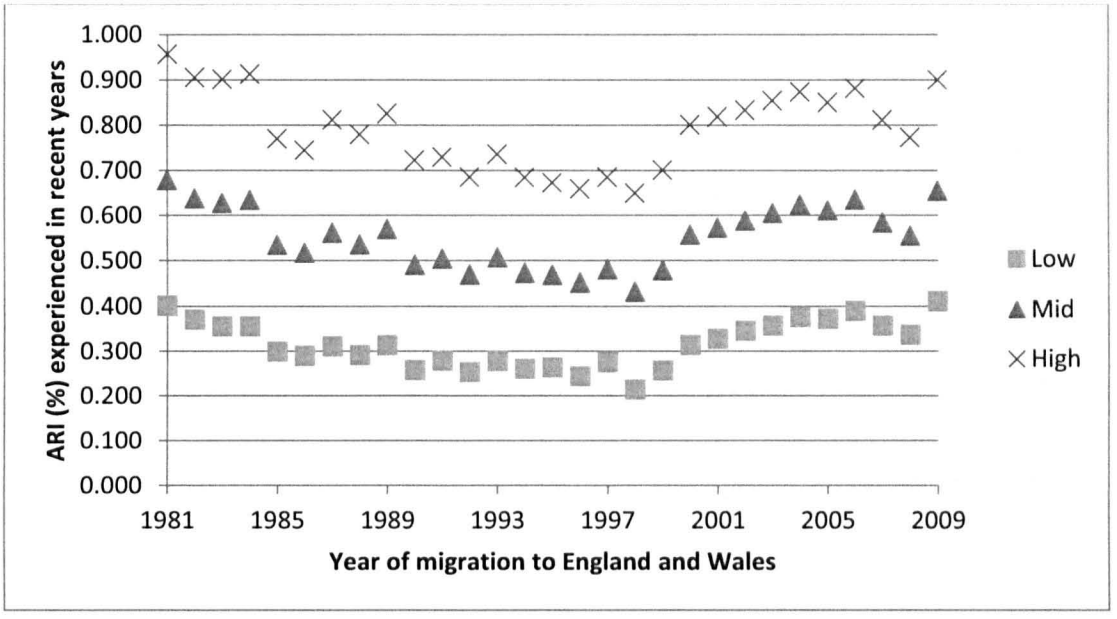


Figure 4-2: Assumed average annual risk of infection (%) experienced in recent years for OF-born migrants by year of entry to England and Wales. These were used along with other assumptions to construct distributions for the probabilities of infection states of OF-born migrants upon entry to England and Wales.

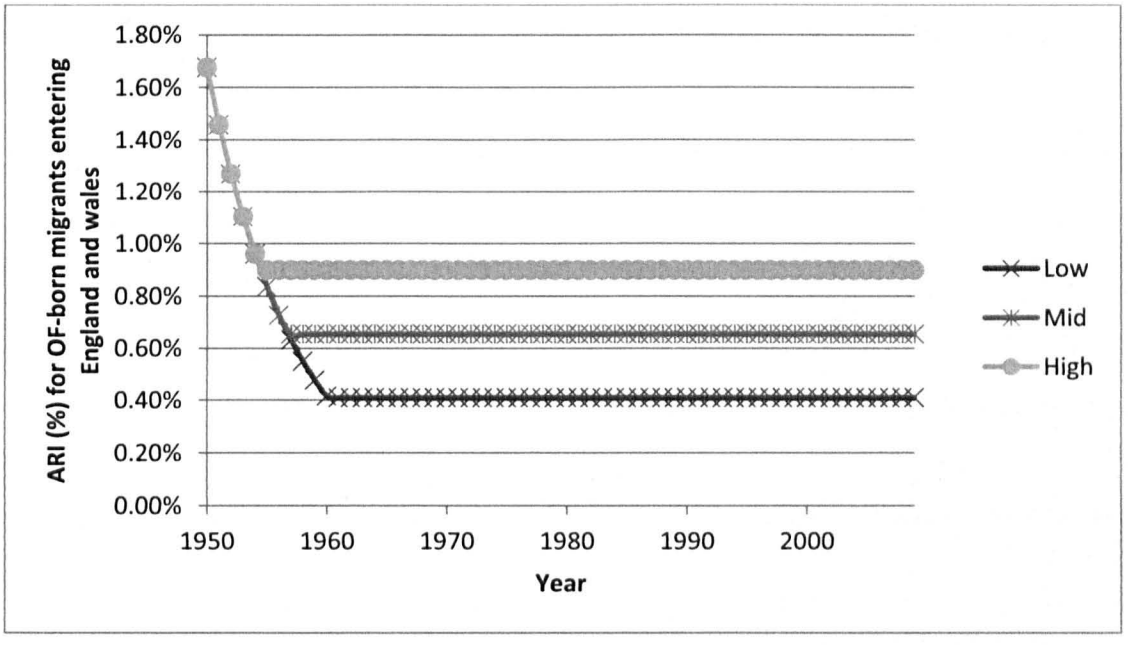


Figure 4-3: Assumed average annual risk of infection (ARI) (%) experienced by OF-born migrants to the UK who entered in 2009. Assumptions for migrants entering in other years, 1981 – 2008, were very similar except the ARI fell to different values in recent years each year of entry, as shown in Figure 4-2.

#### 4.2.8.2.2 Prevalence of disease in migrants

Tuberculosis screening studies were used to make assumptions about the prevalence of disease in foreign-born migrants upon entry to the UK for foreign-born migrants. It

was assumed that 0.55% of OF-born migrants across all age groups entered the UK with active disease, based on two screening studies from the UK [295, 305]. Ormerod reported on immigrants screened from 1983 – 1988 and 1990 – 1994, obtaining a prevalence of all forms of active tuberculosis in regular immigrants to the Blackburn local government area. These two studies gave an average active disease prevalence of 0.55%. Because age-specific data were not available for either study, this prevalence was used for all age groups and was used for all foreign-born migrants. For UK-born migrants returning to the UK, the disease prevalence for 1981 UK-born individuals residing in the UK was used as the disease prevalence for UK-born migrants arriving throughout the simulation, from 1981 – 2009, for lack of other data.

#### 4.2.8.2.3 Results

Three different distributions for the infection state probabilities in UK-born, OF-born, and SSA-born individuals were generated by applying the low, mid, and high ARI assumptions. Results are not shown, as these are cumbersome multi-dimensional tables that include values for eight infection states, 29 years, 120 age classes, both sexes, and three birthplaces for each of three ARI schemes.

# 5 England and Wales Modelling

This chapter describes the application of a simplified version of the model introduced in Chapter 3 to tuberculosis dynamics in England and Wales. The aim was to better understand tuberculosis epidemiology in the region and to inform subsequent modelling work. To satisfy study objective two, the model was fit to tuberculosis notifications from England and Wales from 1999 – 2009 to estimate disease risk parameters and to identify plausible values for effective contact rate and assumptions about the infection status of migrants upon entry to the UK. The model was also used to estimate the proportion of cases due to recent transmission in the UK to satisfy objective three. The quality of model fits to observed data and disease risk estimates are reported for various combinations of effective contact rates and assumptions about the infection status of migrants upon entry to the UK. Results were obtained for two stages of fitting. In stage one, parameter values and input scenarios were outlined *a priori* based primarily on published data. In stage two, parameter values and input scenarios were identical to those used in stage one, apart from four variations: 1) altered parameter distributions for the HIV prevalence of SSA-born immigrants; 2) altered infection status for SSA-born individuals at model initialization in 1981; 3) additional disease risk parameters estimated; and 4) a single foreign-born category combining SSA-born and OF-born during fitting. Reporting focuses on results from the best fits to observed data, which came when combining SSA-born and OF-born during fitting. The input parameters used for subsequent modelling in later chapters are also identified.

## 5.1 Methods

### 5.1.1 Notification Data Used For Fitting Targets

The fitting targets are the observed data to which model output are fitted. In this application of the model, notifications for all forms of tuberculosis, pulmonary and non-pulmonary cases, in England and Wales from 1999 – 2009 were used for fitting targets. These data are fully described in Chapter 2, Section 2.7.1. Since notifications were stratified by age, sex, and birthplace, many demographic categories had a small number of cases observed each year. Some demographic categories also had a small population. In both cases, categories were less reliable for fitting. This is indicated by high random variability in simulations in model runs with identical parameter values and initial conditions, but different random number seeds (see Chapter 3, Section 3.5.1 for discussion of model stochasticity).

As shown in Chapter 2, Section 2.7.1, for each year of notification, sex, and birthplace, the age category of 0 – 14 years resulted in relatively few cases and thus were not well suited as fitting targets. For UK-born and OF-born individuals, the remaining three age classes (15 – 44 years, 45 – 64 years and 65 and over years) for both sexes were used for fitting from 1999 – 2009. These are shown in Figure 5-1 – Figure 5-4. For SSA-born individuals, for both sexes and all years, those aged 45 – 64 years and 65 years and above had small population sizes in the UK and also small numbers of tuberculosis notifications. These age classes were also not used for fitting targets. Figure 5-5 and Figure 5-6 show fitting targets for SSA-born, 1999 – 2009.

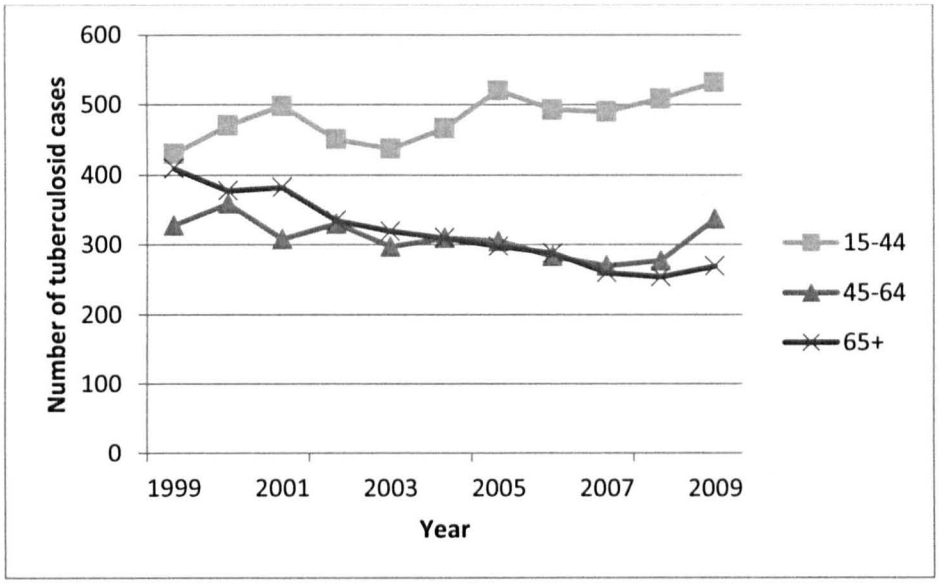


Figure 5-1: Tuberculosis notifications used for model fitting targets for UK-born males by age category (years) for England and Wales, 1999 – 2009.

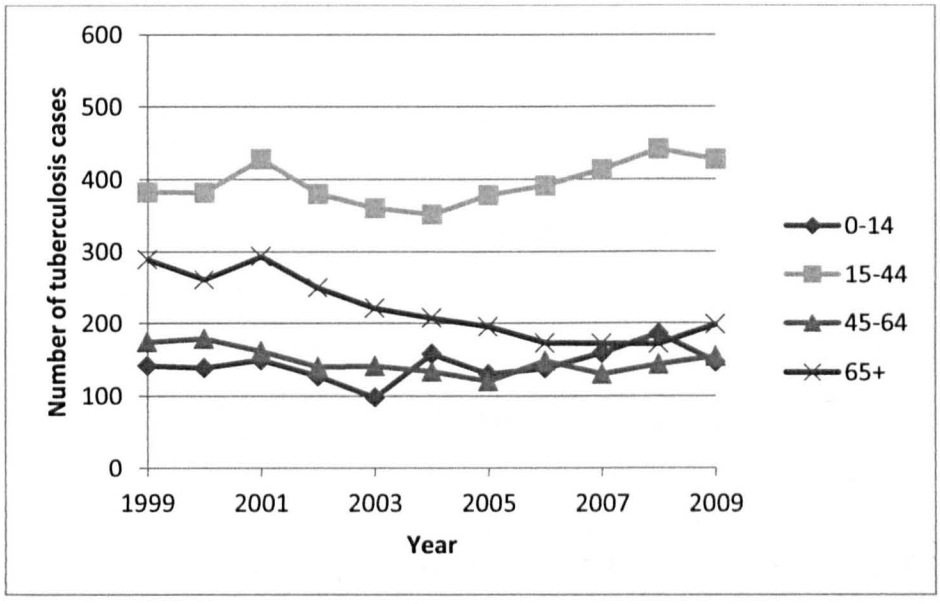


Figure 5-2: Tuberculosis notifications used for model fitting targets for UK-born females by age category (years) for England and Wales, 1999 – 2009.

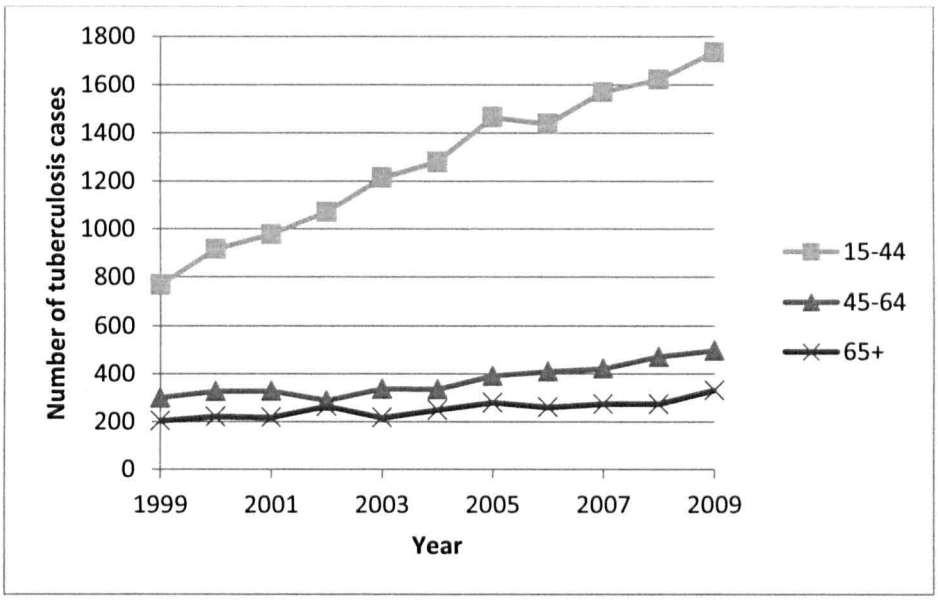


Figure 5-3: Tuberculosis notifications used for model fitting targets for other foreign-born males by age category (years) for England and Wales, 1999 – 2009.

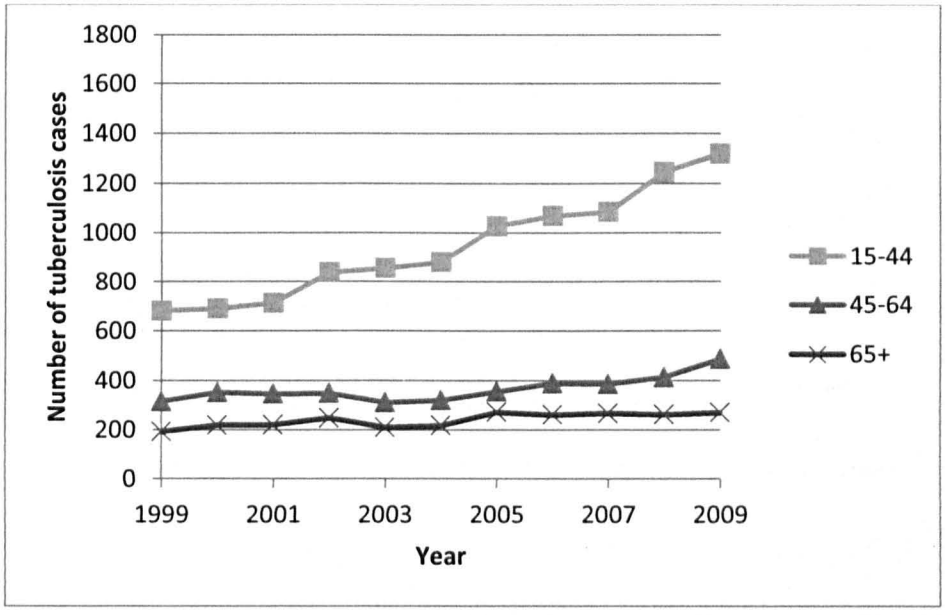


Figure 5-4: Tuberculosis notifications used for model fitting targets for other foreign-born females by age category (years) for England and Wales, 1999 – 2009.

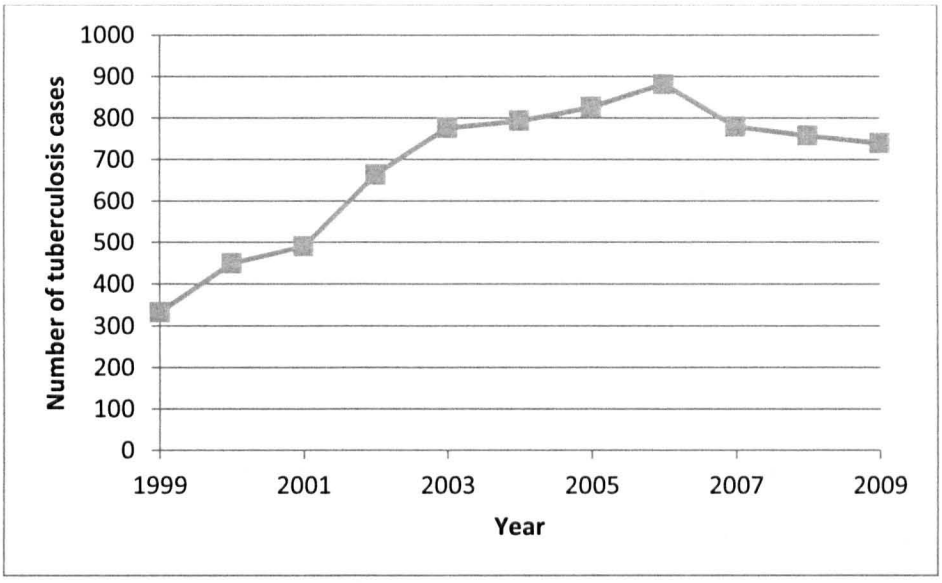


Figure 5-5: Tuberculosis notifications used for model fitting targets for Sub-Saharan African-born males aged 15 – 44 years for England and Wales, 1999 – 2009.

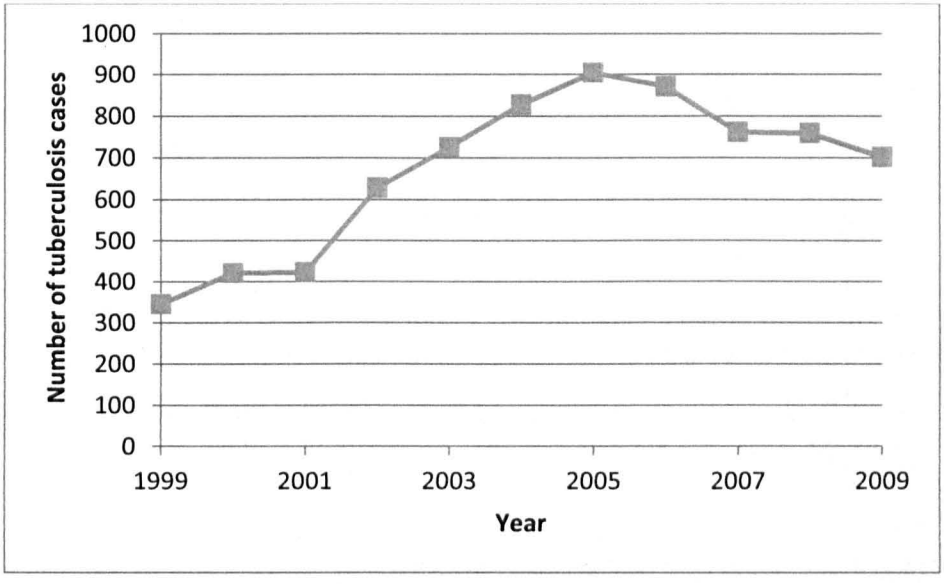


Figure 5-6: Tuberculosis notifications used for model fitting targets for Sub-Saharan African-born females aged 15 – 44 years for England and Wales, 1999 – 2009.

### 5.1.2 Simplified Model

The model of tuberculosis dynamics described in Chapter 3 was simplified for application to notification data from England and Wales. Because genotyping data were not available, records of strain types were left out to save computer memory and computational time. For this application of the model, the model reported numbers of case notifications each year, stratified by age, sex, and birthplace for comparison with the fitting targets described above.

Fitting required many model runs in serial and parallel (see Chapter 3, Section 3.5.2). Therefore, because simulation of the entire population of England and Wales was memory-intensive and because it was important to minimize computation time, only data essential to the fitting process were enabled. For example, data such as the time and place of acquisition of an infection was not needed. After fitting, independent runs using best-fitting disease risk parameters were employed to estimate the proportion of disease due to recent transmission. For these estimates, the model was run on 30 processors in parallel, as done during the earlier fitting runs, and the proportion of cases due to recent transmission was calculated for each demographic category. Results were averaged for the 30 model runs and proportions from each separate run of the model were plotted to depict the variance in simulation results.

### 5.1.3 Model Parameterization

Parameter values and assumptions used in the model are described in Chapter 4 and were based on published data wherever possible. However, as discussed in Chapter 4, some parameters were not well supported in the literature nor found in other available data sources, and thus were estimated or varied during model fitting. Disease risk parameters were particularly uncertain and were estimated during model fitting, as described below. Contact rates and infection status for migrants upon entry to the UK were also both uncertain. These parameters were not estimated during model fitting, but instead they were varied, taking on different values in different fitting scenarios. The contact rates used are described in Section 5.1.3.2, assumptions about the infection status of migrants upon entry to the UK in Section 5.1.3.3, and a summary of the fitting scenarios defined by these parameters in Table 5-3.

### *5.1.3.1 Disease risk parameters*

Parameters for the risk of developing disease following infection with *M. tuberculosis* were estimated by fitting numbers of case notifications in model output to the number of observed notifications in England and Wales for each demographic category from 1999 – 2009. The risk of disease following infection with *M. tuberculosis* was assumed to depend on age, sex, birthplace, HIV-status, and infection type (*Recent Infection, Latent Infection or Reinfection*), as detailed in Chapter 4, Section 4.2.3.1 and summarized below. Recall that risks for *Primary Disease* and *Reinfection Disease* are cumulative risks of developing disease over the five years of *Recent Infection* and *Reinfection* respectively, while the risks for *Reactivation Disease* are proportions of individuals with *Latent Infection* who develop disease each year—actually a rate of disease development.

To reduce the number of variable parameters in the model, several simplifying assumptions were made regarding the effects of age, sex, birthplace, and HIV on tuberculosis progression risks. First, risks for OF-born and HIV-negative SSA-born individuals were related to UK-born disease risks by a factor, *df*, by which UK-born risks were multiplied to get OF-born and HIV-negative SSA-born risks. This factor implies the relative differences between risks of disease for the different infection types and between the sexes was the same for UK-born and foreign-born individuals. It is unknown if this assumption is valid. The value of *df* was estimated in model fitting, as discussed below.

Risks of developing tuberculosis for HIV-positive SSA-born individuals were assumed to be related to risks for OF-born and HIV-negative SSA-born individuals of the same age and sex by a factor of increase in disease risk due to HIV, *ehiv*. The value of *ehiv* was fixed at 7.0 in the model for all infection types, as discussed in Chapter 4, Section 4.2.3.1. In addition to defining *df* and *ehiv*, the number of variable parameters was reduced by fixing disease risk ratios between males and females. These were fixed based on work by Vynnycky et al. [9, 19] as discussed in Section 4.2.3.1; values are given in Table 5-1.

For each birthplace and sex, age-dependency of disease risk parameters was simplified to take into account only two age classes, those aged 10 years and under and those 20 years and over, following Vynnycky and Fine [19]. A linear increase in risk was assumed between ages 10 and 20 years. The number of variable parameters was reduced by fixing the disease progression risks for UK-born individuals under 10 years of age for both males and females. These were fixed based on estimates of the risk of respiratory disease by Vynnycky et al., as discussed in Section 4.2.3.1 and shown in Table 5-1 [19]. Progression risks and rates for respiratory tuberculosis estimated by Vynnycky et al. were divided by the proportion of disease that is respiratory in children under 10 years of age. This division converted respiratory disease risks into disease risks for all sites of disease combined, as used herein. These risks were fixed because there were few cases on which to base estimates and because these cases were excluded from the data used for model fitting. Furthermore, these risks were not relevant to a large segment of the population, as relatively few children in England and Wales were infected with *M. tuberculosis*.

For stage one fits, four disease risk parameters were estimated during fitting of model output to observed data. These parameters and their relationships to other parameters in the model are shown in Table 5-2, with fitted parameters marked in bold. For *Recent Infection* and *Reinfection,* cumulative disease risks for UK-born males aged 20 years and above were estimated in fitting. For *Latent Infection,* annual rates of disease progression for UK-born males aged 20 years and above were estimated in fitting. Lastly, foreign-born risks were estimated by the model using a parameter for the factor by which UK-born disease risks and rates were multiplied to get foreign-born disease risks and rates, *df*. For each fit of the model to observed case notifications, estimates for these four parameters were obtained. Different fits of the model were obtained for 25 different input scenarios under which the effective contact rate and assumptions about the infection status of migrants upon entry to the UK were varied, as discussed below.

Table 5-1: Key to parameter names with values for fixed parameters where applicable.

| Description | Parameter | Value | Units |
|---|---|---|---|
| Risk of Primary Disease for those aged <10 years | *d1uk10* | 4.06 | Cumulative % |
| Risk of Reactivation Disease for those aged <10 years | *d2uk10* | $9.82 \times 10^{-8}$ | Cumulative % |
| Risk of Reinfection Disease for those aged <10 years | *d3uk10* | 6.89 | % per year |
| Risk of Primary Disease for those aged 20 years and above | *d1uk20* | Estimated in fitting | Cumulative % |
| Risk of Reactivation Disease for those aged 20 years and above | *d2uk20* | Estimated in fitting | % per year |
| Risk of Reinfection Disease for those aged 20 years and above | *d3uk20* | Estimated in fitting | Cumulative % |
| Ratios of female:male disease risks by infection type and age specified below. | *sd[]* | | |
| *Recent Infection*, age <10 | *sd[d1a10]* | 1 | None |
| *Latent Infection*, age <10 | *sd[d2a10]* | 1 | None |
| *Reinfection*, age <10 | *sd[d3a10]* | 1 | None |
| *Recent Infection*, age 20+ | *sd[d1a20]* | 1 | None |
| *Latent Infection*, age 20+ | *sd[d2a20]* | 0.16054 | None |
| *Reinfection*, age 20+ | *sd[d3a20]* | 0.00121 | None |
| Ratio of foreign-born: UK-born disease risks | *df* | Estimated in fitting | None |
| Ratio of HIV+:HIV- disease risks for SSA-born individuals | *ehiv* | 7 | None |

Table 5-2: Table of disease risk parameters by infection type, age, sex, birthplace, and Human Immunodeficiency Virus (HIV) status. Age is divided into those under 10 years of age and those aged 20 years and over. Birthplaces are the United Kingdom (UK), Sub-Saharan Africa (SSA) and other foreign countries (OF). Parameters estimated in the model are printed in bold. Key to parameter names and values for fixed parameters are given in

**UK-born**

| | Ages 10 years and under | | Ages 20 years and over | | |
|---|---|---|---|---|---|
| Infection type | Male | Female | Male | Female | Units of disease risk |
| *Recent Infection* | *d1uk10* | *d1uk10*sd[1]* | ***d1uk20*** | ***d1uk20*sd[2]*** | Cumulative % developing disease over 5 years |
| *Reinfection* | *d2uk10* | *d2uk10*sd[3]* | ***d2uk20*** | ***d2uk20*sd[4]*** | Cumulative % developing disease over 5 years |
| *Latent Infection* | *d3uk10* | *d3uk10*sd[5]* | ***d3uk20*** | ***d3uk20*sd[6]*** | % developing disease each year |

**OF-born and HIV-negative SSA-born**

| | Ages 10 years and under | | Ages 20 years and over | | |
|---|---|---|---|---|---|
| Infection type | Male | Female | Male | Female | Units of disease risk |
| *Recent Infection* | ***df*d1uk10*** | ***df*d1uk10*sd[1]*** | ***df*d1uk20*** | ***df*d1uk20*sd[2]*** | Cumulative % developing disease over 5 years |
| *Reinfection* | ***df*d2uk10*** | ***df*d2uk10*sd[3]*** | ***df*d2uk20*** | ***df*d2uk20*sd[4]*** | Cumulative % developing disease over 5 years |
| *Latent Infection* | ***df*d3uk10*** | ***df*d3uk10*sd[5]*** | ***df*d3uk20*** | ***df*d3uk20*sd[6]*** | % developing disease each year |

**HIV-positive SSA-born**

| | Ages 10 years and under | | Ages 20 years and over | | |
|---|---|---|---|---|---|
| Infection type | Male | Female | Male | Female | Units of disease risk |
| *Recent Infection* | *ehiv*df*d1uk10* | *ehiv*df*d1uk10*sd[1]* | *ehiv*df*d1uk20* | *ehiv*df*d1uk20*sd[2]* | Cumulative % developing disease over 5 years |
| *Reinfection* | *ehiv*df*d2uk10* | *ehiv*df*d2uk10*sd[3]* | *ehiv*df*d2uk20* | *ehiv*df*d2uk20*sd[4]* | Cumulative % developing disease over 5 years |
| *Latent Infection* | *ehiv*df*d3uk10* | *ehiv*df*d3uk10*sd[5]* | *ehiv*df*d3uk20* | *ehiv*df*d3uk20*sd[6]* | % developing disease each year |

#### 5.1.3.2 *Contact rate*

Although only disease risk parameters were estimated during model fitting, there was also uncertainty in the contact rate parameter in the model, as detailed in Section 4.2.2.1. The contact rate was not estimated during fitting due to the expected inverse correlation between this parameter and disease risk parameters. Increasing the transmission parameters should result in decreasing the disease risk, and vice versa. Instead of estimating the contact rate during model fitting, five plausible schemes for the contact rate were tested in separate model fits. Given uncertainty in the value of this parameter, a range from two per year to 10 per year was explored, representing a range of published estimates of the contact rate, reviewed in Section 4.2.2.1. Contact rate assumptions include schemes where all individuals have the same contact rate and also where the contact rate differs by birthplace (UK or foreign). The five different schemes for contact rates are: 1) all = 4; 2) all = 6; 3) all = 8; 4) all = 10; and 5) UK-born = 4, Foreign-born = 8.

#### *5.1.3.3 Infection status of migrants upon entry to the UK*

As with the effective contact rate, there was uncertainty about the probabilities of the different infection states for migrants entering the UK, as discussed in Chapter 4, Section 4.2.8. The probabilities of infection states for migrants entering the UK can be thought of as probability distributions for randomly selecting the infection status of migrants upon entry to the UK, hereafter abbreviated as the 'infection status of migrants'. These parameters were not estimated using model fitting for two reasons. First, the parameters would likely be highly correlated with the disease risk parameters. This is because increasing the proportion infected or diseased entering the simulation would increase the number of disease cases, and vice versa, which is also true for disease risks. Second, there would be a very large number of variable parameters related to these distributions and this could be difficult for fitting and make several parameter estimates dependent on too little data. Therefore, two separate methods for arriving at assumptions about the infection status of migrants upon entry to the UK were used to produce five schemes for the infection status of migrants. These are detailed in Chapter 4, Section 4.2.8. Briefly, the ARI method used a

spreadsheet model to produce three schemes for the infection status of migrants, based on low, medium and high assumptions for ARI experienced by migrants before entry. The screening method was used to produce two schemes for the infection status of migrants, based on data from studies which screened migrants to the UK for TST reactions and active disease upon arrival [295].

### 5.1.4 Stage One Fitting

The five contact rate schemes and the five schemes for the infection status of migrants are combined into 25 input parameter scenarios, each of which is fit to observed data. These scenarios are enumerated in Table 5-3.

### 5.1.5 Stage Two Fitting

After stage one fitting of the 25 scenarios, four separate variations were tested for improved fits, as described in the sections below. For each variation, 10 of the 25 fitting scenarios were used to reduce computational time and to eliminate scenarios that were unlikely to result in good fits of the model to data. These 10 scenarios included the two best-fitting contact parameter schemes for each of the five schemes for the infection status of migrants, marked in bold in Table 5-3. All parameter values and assumptions for stage two fits are the same as for stage one fits, unless specified otherwise.

#### *5.1.5.1 Variation One*

The first of the four variations tested a new distribution of HIV prevalence for SSA-born immigrants entering the study population each year. This distribution was identified as potentially problematic because it was based on HIV prevalence in the UK for all SSA-born individuals living in the UK (see Section 4.2.4), but used in the model to assign HIV status only to those SSA-born individuals *arriving* in a given year. These data were used because they were the only data available. However, these may have underestimated the HIV prevalence in new SSA-born migrants each year, since many SSA-born adults living in the UK arrived before the recent HIV epidemic. These individuals would decrease the overall average HIV prevalence compared with new migrants from SSA, among whom HIV prevalence is higher. Again, due to lack of data, the factor by which HIV prevalence values should have been increased was unknown. As a starting point, HIV prevalence was increased by 50% for both sexes each year.

### *5.1.5.2 Variation Two*

The second variation tested a new parameter distribution for the *M. tuberculosis* infection status of SSA-born individuals in 1981, to allow for increased prevalence of *M. tuberculosis* infection and disease compared with OF-born. In stage one fits, the infection and disease prevalences of SSA-born individuals was assumed equal to those in OF-born individuals in 1981 (see Chapter 4, Section 4.2.7). This assumption was for simplicity, and also because data to support more complex assumptions was lacking. Here a higher *M. tuberculosis* infection and disease prevalence in SSA-born was tested. The higher infection and disease prevalence assumptions were established using the same methods for estimating the infection status of UK-born and OF-born individuals in 1981 (again, see Chapter 4, Section 4.2.7), but assuming SSA-born individuals in 1981 had experienced a higher ARI than those who were UK-born or OF-born. The ARI for *M. tuberculosis* in SSA-born individuals living in England and Wales in 1981 was assumed to decline as it did in England and Wales from around 1900, but remain constant from 1952 to 1980 at the estimated value for the ARI in England and Wales for the time, about 0.6% [271].

### *5.1.5.3 Variation Three*

In the third variation, six disease risk parameters in the fitting routine were estimated. Instead of fixing the ratio between foreign-born and UK-born disease risks, *df,* for each type of disease, three additional disease risk parameters applied to foreign-born adult males, one for each disease type. These three disease risk parameters were analogous to the three risks for UK-born individuals, all specified for males aged 20 years and above. To clarify how these new parameters fit into parameters outlined in Table 5-2, recall that foreign-born *Primary Disease* risk for males aged 20 years and above is given by *df*d1uk20.* Under this fitting scheme, this equation became a new parameter, *d1nuk20.* The new parameter also took the place of *df*d1uk20* for other disease risks which are derived from this—for example, foreign-born females aged 20 years and above, given by *d1nuk20*sd*[1]. The situation is analogous for *Reactivation* and *Reinfection Disease,* and also applies to risks in Table 5-2.

#### *5.1.5.4 Variation Four*

The last of the four variations used a single foreign-born category during fitting, combining SSA-born and OF-born. Although the groups were combined in fitting and plotting of results, for expediency, the model itself was not changed and they were handled separately. The groups were combined in fitting because SSA-born population sizes are small and model output is uncertain for the group, carrying a high degree of stochasticity. Even estimates of SSA-born population sizes are uncertain. Furthermore, also due to data that were uncertain or lacking, the model did not differentiate between SSA-born and OF-born for some parameters, for example the contact rate and disease risk parameters for HIV-negative individuals, which caused conflicts for fitting. The model was not fully specified when fitting the two groups separately, hence the logic for combining them in this analysis.

Table 5-3: Input parameter scenarios, as defined by assumptions about the infection status of migrants upon entry to the UK and contact rate parameters. All 25 input scenarios were used in stage one fits, whereas stage two fits used only the 10 scenarios in bold. For the infection status of migrants upon entry to the UK, the ARI method was used to produce three the infection status of migrants schemes, 'ARI low', 'ARI med', and 'ARI high', based on low, medium and high assumed ARIs experienced by migrants before entering the UK. The screening method was used to produce two the infection status of migrants schemes based on a study which screened migrants to the UK upon arrival, resulting in schemes 'Scr1' and 'Scr2'.

| Scenario | Infection status of migrants | Contact rate (number per year) |
|---|---|---|
| 1 | Scr1 | all = 4 |
| 2 | Scr1 | all = 6 |
| **3** | **Scr1** | **all = 8** |
| **4** | **Scr1** | **all = 10** |
| 5 | Scr1 | UK-born=4, Foreign-born=8 |
| 6 | Scr2 | all = 4 |
| 7 | Scr2 | all =6 |
| **8** | **Scr2** | **all = 8** |
| **9** | **Scr2** | **all = 10** |
| 10 | Scr2 | UK-born=4, Foreign-born=8 |
| 11 | ARI low | all = 4 |
| 12 | ARI low | all = 6 |
| **13** | **ARI low** | **all = 8** |
| **14** | **ARI low** | **all = 10** |
| 15 | ARI low | UK-born=4, Foreign-born=8 |
| **16** | **ARI med** | **all = 4** |
| **17** | **ARI med** | **all = 6** |
| 18 | ARI med | all = 8 |
| 19 | ARI med | all = 10 |
| 20 | ARI med | UK-born=4, Foreign-born=8 |
| 21 | ARI high | all = 4 |
| **22** | **ARI high** | **all = 6** |
| **23** | **ARI high** | **all = 8** |
| 24 | ARI high | all = 10 |
| 25 | ARI high | UK-born= 4, Foreign-born= 8 |

### 5.1.6 Fitting Procedure

Methods for fitting the model to observed data are detailed in Chapter 3, Section 3.5.2. Briefly, disease risk parameters for UK-born males aged 20 years and above for each infection type, *Recent Infection*, *Latent Infection*, and *Reinfection*, plus the ratio of disease risk in foreign-born individuals to disease risks in UK-born individuals, were varied to fit the model output to observed tuberculosis notifications. For each set of parameter values tested during fitting, the model output was averaged over 30 replicates. The averaged model output was then fit using a simulated annealing with downhill simplex optimization routine [255]. The optimization routine, or 'fitting routine', returned the best-fitting estimates for each of the four variable parameters and the value of the Poisson log likelihood deviance GOF statistic, hereafter called the 'GOF statistic', which is a measure of the quality of fit of model output to observed data. This GOF statistic is lowest for the best fits, indicating less deviance from the observed data. Fitting terminated when the rate of change in the GOF statistic and estimated parameters converged to less than 1% change on average over the previous 20 runs, repeatedly. See Chapter 3, Section 3.5.2.2 for more information.

For each input parameter scenario, the fitting routine was run five times in replicate with different initial conditions, as discussed below and in Chapter 3, Section 3.5.2. Each replicate provided estimates for the four disease risk parameters and the GOF statistic measuring the fit of model output to data. Initial conditions were defined by starting values for the variable parameters, including five sets of the four variable parameters necessary to initialize the optimization routine. These were randomly drawn for each run of the fitting routine from the means and standard deviations shown in Table 5-4. Replicates are not always necessary with simulated annealing, since the method is designed to explore a wide parameter space to find the best-fitting set. Nonetheless, replicates were used here to help assure the best fit would be found. Plausible means and standard deviations of parameters to be estimated were derived from ranges extracted from literature values and values assumed in related models when possible, capturing broad ranges intentionally. Ranges for the three disease risks for UK-born males aged 20 years and above were derived from risks estimated and assumed in several tuberculosis modelling studies [19, 179, 182, 224]. For the ratio between disease risks in foreign-born and UK-born individuals, estimates were not

available, so a range of 0.5 – 3.0 was used. Ranges were converted into means and standard deviations by setting the midpoint of each range as the mean of the parameter distribution and assuming the range covered the 95% confidence bounds of the distribution. The variance was calculated assuming a normal distribution. Note that the initial values for parameters drawn from the distributions in Table 5-4 do not restrict parameter estimates to these ranges.

Table 5-4: Means and standard deviations used to randomly assign initial values for the four variable disease risk parameters in the fitting routine.

| Parameter | Description | Mean | SD |
|---|---|---|---|
| *df* | Ratio of disease risk foreign-born to disease risk in UK-born | 1.75 | 0.638 |
| *d1uk20* | Cumulative risk (%) of disease resulting from *Recent Infection*, males aged 20 years and above | 0.077 | 0.037 |
| *d2uk20* | Annual risk (%) of disease resulting from for *Latent Infection*, males aged 20 years and above | 0.000807 | 0.0003538 |
| *d3uk20* | Cumulative risk (%) of disease resulting from *Reinfection*, males aged 20 years and above | 0.077 | 0.037 |

### 5.1.7 Statistical Analysis of Model Results

Although results generally focused on the best-fitting replicate for each scenario to report and compare fits across input scenarios, to supplement this approach, an analysis of variance (ANOVA) was performed to compare variance in the GOF statistic among replicates within a scenario to variance across scenarios. ANOVA was used to help identify whether there were true differences in fit quality across scenarios. In addition to helping quantify the effect of the contact rate and the infection status of migrants upon entry to the UK, a two-way ANOVA was used to test the effects of these parameters on GOF values and disease risk estimates. For parameters significantly associated with GOF or disease risk, simple linear regression was also used to estimate coefficients of variation. All ANOVA and regression tests were performed using the R: A language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria).

## 5.2 Results

Results of fitting model output to observed tuberculosis notifications in England and Wales from 1999 – 2009 are divided into two sections. Firstly, results from stage one fits are reported. Next, results from stage two fits are reported. Since the best fits of model output to data were found in stage two fitting when a single foreign-born category was used, only those stage two results are detailed. For both stage one and two fits, the quality of fits to observed data, disease risk estimates, the estimated proportion of cases due to recent transmission in the UK, and the effect of parameter values for the contact rate and infection status of migrants upon entry to the UK are reported.

### 5.2.1 Stage One

A summary of the results of fitting the 25 scenarios to case notifications from England and Wales is shown below in Table 5-5. Best-fitting disease risk estimates for *Primary*, *Reactivation,* and *Reinfection Disease* in UK-born males aged 20 years and above and the relative risk of disease between foreign-born and UK-born individuals are shown for each input scenario. The GOF statistic and the GOF rank for each of the scenarios from the best-fitting (1) to the worst-fitting (25), or equivalently from the lowest GOF statistic to the highest are given. To simplify results, only the best-fitting of the five replicates run for each input scenario is reported, regarding it as the true best fit for that scenario. However, the five replicates within a scenario did generally result in differing parameter estimates and GOF statistics, as discussed further in Section 5.2.1.1.2.

Table 5-5: Results of stage one model fitting to the 25 input scenarios. Scenarios are defined by the contact rate and five schemes for the infection status of migrants upon entry to the UK, including Scr1 and Scr2, both based on immigrant screening data, and ARIlow, ARImed, and ARIhigh, based on assumptions about the ARI experienced by migrants prior to arrival in the UK. The best-fitting of the five replicate fitting runs is shown for each scenario. Asterisks denote scenarios used for stage two fitting. Results of model fitting to observed case notifications include disease risk estimates and the value of the goodness of fit (GOF) statistic measuring the quality of fit of the model to data. The GOF rank orders fits from the best (1) to the worst (25). Only the best-fitting replicate for each scenario is shown.

| | | | Estimated disease risks for UK-born adult males by disease type | | | | | |
|---|---|---|---|---|---|---|---|---|
| Scenario | Infection status of migrants | Contact rate (number per year) | Primary (%) | Reactivation (% per year) | Reinfection (%) | Foreign:UK-born risk ratio (*df*) | GOF | GOF rank |
| 1 | Scr1 | 4 | 23.4 | 0.010 | 18.1 | 1.68 | 10452 | 23 |
| 2 | Scr1 | 6 | 17.4 | 0.012 | 3.1 | 1.81 | 10235 | 21 |
| 3* | Scr1 | 8 | 12.4 | 0.009 | 7.6 | 1.90 | 10052 | 17 |
| 4* | Scr1 | 10 | 10.3 | 0.009 | 6.2 | 1.85 | 9925 | 16 |
| 5 | Scr1 | UK-born=4, Foreign-born=8 | 14.1 | 0.015 | 3.7 | 1.76 | 10361 | 22 |
| 6 | Scr2 | 4 | 23.9 | 0.003 | 27.4 | 1.56 | 9678 | 12 |
| 7 | Scr2 | 6 | 15.5 | 0.008 | 12.3 | 1.77 | 9437 | 6 |
| 8* | Scr2 | 8 | 10.9 | 0.011 | 5.9 | 1.97 | 9387 | 5 |
| 9* | Scr2 | 10 | 9.3 | 0.010 | 5.9 | 1.88 | 9274 | 3 |
| 10 | Scr2 | UK-born=4, Foreign-born=8 | 15.2 | 0.012 | 6.5 | 1.50 | 9537 | 9 |
| 11 | ARIlow | 4 | 20.1 | 0.018 | 13.2 | 3.28 | 9920 | 15 |
| 12 | ARIlow | 6 | 15.0 | 0.016 | 7.3 | 3.17 | 9686 | 13 |

| Scenario | Infection status of migrants | Contact rate (number per year) | Estimated disease risks for UK-born adult males by disease type: Primary (%) | Reactivation (% per year) | Reinfection (%) | Foreign:UK-born risk ratio (*df*) | GOF | GOF rank |
|---|---|---|---|---|---|---|---|---|
| 13* | ARIlow | 8 | 10.0 | 0.021 | 0.4 | 3.79 | 9368 | 4 |
| 14* | ARIlow | 10 | 8.5 | 0.020 | 0.4 | 3.69 | 9201 | 2 |
| 15 | ARIlow | UK-born=4, Foreign-born=8 | 14.4 | 0.021 | 1.7 | 2.74 | 9531 | 8 |
| 16* | ARImed | 4 | 32.6 | 0.022 | 0.1 | 1.69 | 9595 | 10 |
| 17* | ARImed | 6 | 23.1 | 0.017 | 8.0 | 1.65 | 9639 | 11 |
| 18 | ARImed | 8 | 16.4 | 0.012 | 16.5 | 1.64 | 10097 | 18 |
| 19 | ARImed | 10 | 14.5 | 0.013 | 9.2 | 1.60 | 10201 | 20 |
| 20 | ARImed | UK-born=4, Foreign-born=8 | 23.1 | 0.019 | 4.0 | 1.35 | 10858 | 24 |
| 21 | ARIhigh | 4 | 20.8 | 0.025 | 7.8 | 2.37 | 9759 | 14 |
| 22* | ARIhigh | 6 | 18.2 | 0.020 | 7.9 | 2.02 | 9171 | 1 |
| 23* | ARIhigh | 8 | 14.8 | 0.022 | 7.6 | 1.89 | 9505 | 7 |
| 24 | ARIhigh | 10 | 11.1 | 0.018 | 8.0 | 2.06 | 10126 | 19 |
| 25 | ARIhigh | UK-born=4, Foreign-born=8 | 17.1 | 0.020 | 7.2 | 1.71 | 11331 | 25 |
| Mean | | | 16.5 | 0.015 | 7.8 | 2.09 | | |

### *5.2.1.1 Quality of fits*

Across input scenarios, the model output captured most, but not all, trends found in observed case notifications in England and Wales from 1999 – 2009. To illustrate the quality of model fit to observed data, plots of model output versus observed notification data are shown for the two best-fitting scenarios of the 25 stage one fits, scenarios 14 and 21, though GOF statistics and plots suggest that several scenarios fit the data roughly equally well. Plots of simulated numbers of case notifications versus observed case notifications are found in Figure 5-7 – Figure 5-12 and described below in Section 5.2.1.1.1 while plots of simulated versus observed notification rates per 100,000 population are described in the Appendix 10.9. The variance in GOF statistics among the five replicates for each scenario is discussed in Section 5.2.1.1.2. These GOF statistics and plots for each of the five replicates for each scenario indicate that several of the scenarios resulted in the model output fitting the observed data approximately equally well (see Section 5.2.1.1.2).

#### 5.2.1.1.1 Plots of model fit to case notifications

Trends in the observed number of case notifications differed markedly among the three birthplace groups, and major differences were largely reflected in model output. Plots showing model output versus observed case notifications for best-fitting scenarios 14 and 22 are found in Figure 5-7 – Figure 5-12, with a plot for each birthplace and sex combination.

OF-born individuals had the highest number of cases, which generally increased from 1999 – 2009, with the trend captured well by the model in both scenarios 14 and 22 as shown in Figure 5-8 and Figure 5-11. Age-specific trends in this group were also captured well by the model. The age group of 15 – 44 years for both males and females accounted for the majority of cases, increasing to over 1700 cases in males and over 1300 cases in females by 2009. Under scenario 14, in this age group, the number of cases predicted by the model was fairly consistent with observed notifications although on average, underestimated for males and overestimated for females. The number of notifications in males and females aged 45 – 64 years were generally overestimated by model predictions for both males and females under

scenario 14. There were few cases in children and those aged 65 years and above, but trends and numbers of cases were reproduced well by the model.

Under scenario 22 for OF-born individuals, cases in the 15 – 44 age group were reproduced well for males from 1999 – 2005, though overestimated by the model from 2005 onward. The number of cases in females was consistently overestimated, but the model output matched well to qualitative trends in the data. For the other three age groups, the model predicted the number of cases and reproduced qualitative trends well, albeit slightly overestimated the number of cases in the 45 – 64 age group for both males and females.

UK-born individuals also accounted for a substantial number of cases, which were more evenly distributed among age classes than for OF-born individuals. These age-specific trends for UK-born individuals were not captured as well for as those for OF-born, as shown in Figure 5-7 and Figure 5-10 for Scenarios 14 and 22. The model fits by age group varied between the two scenarios and between the sexes. For males under scenario 14, trends in notifications for those aged 45 – 64 years, 0 – 14 years, and 65 years and above were reproduced well by the model (Figure 5-7-A). However, the number of notifications in those aged 15 – 44 years—the group with the largest number of cases for UK-born males—was underestimated by the model. The qualitative trend of stable or slightly increasing numbers of notifications observed was captured by the simulation. For UK-born females under scenario 14, the model predicted the number of notifications in the 15 – 44 years age group well (Figure 5-7-B). However, in older females, the model failed to show the observed trends of decreasing numbers of notifications. The model also predicted increasing notifications in the 45 – 64 age group when observed cases for females aged 45 – 64 years were falling. This problem was also present for females aged 65 years and above, where observed notifications decreased but the model predicted a stable or slightly increasing number of cases each year.

Under scenario 22, notifications in the 15 – 44 age group for UK-born individuals were underestimated for both males and females, although the qualitative trend of stable to slightly increasing incidence was reproduced well by the model, as shown in Figure 5-10. Also, for both males and females, observed trends and case notifications in those

aged 45 – 64 years were reproduced successfully by the model. Most problematic for both males and females was the age group of 65 years and above, in which the observed number of case notifications decreased from 1999 – 2009. The model failed to reproduce the decreasing trend. In males, cases in those aged 65 years and above were stable in the model output while in females, cases were increasing in the model.

For SSA-born individuals, there were fewer total cases than among the OF-born or UK-born individuals, however there were a substantial number of cases in the 15 – 44 age group, the only age group for SSA-born individuals used in model fitting. There was a noticeable time trend of rise and fall in the number of SSA–born cases for both males and females from 1999 – 2009. As shown in Figure 5-9 and Figure 5-12, this qualitative trend in SSA-born notifications was captured by the model especially well in scenario 22, although the model underestimated the number of cases seen in that group. For both scenarios 14 and 22, and for males and females, the model predicted far fewer cases than were observed. Despite that other age groups were not used in fitting, the model reproduced the case numbers well in the three other age groups for SSA-born individuals.

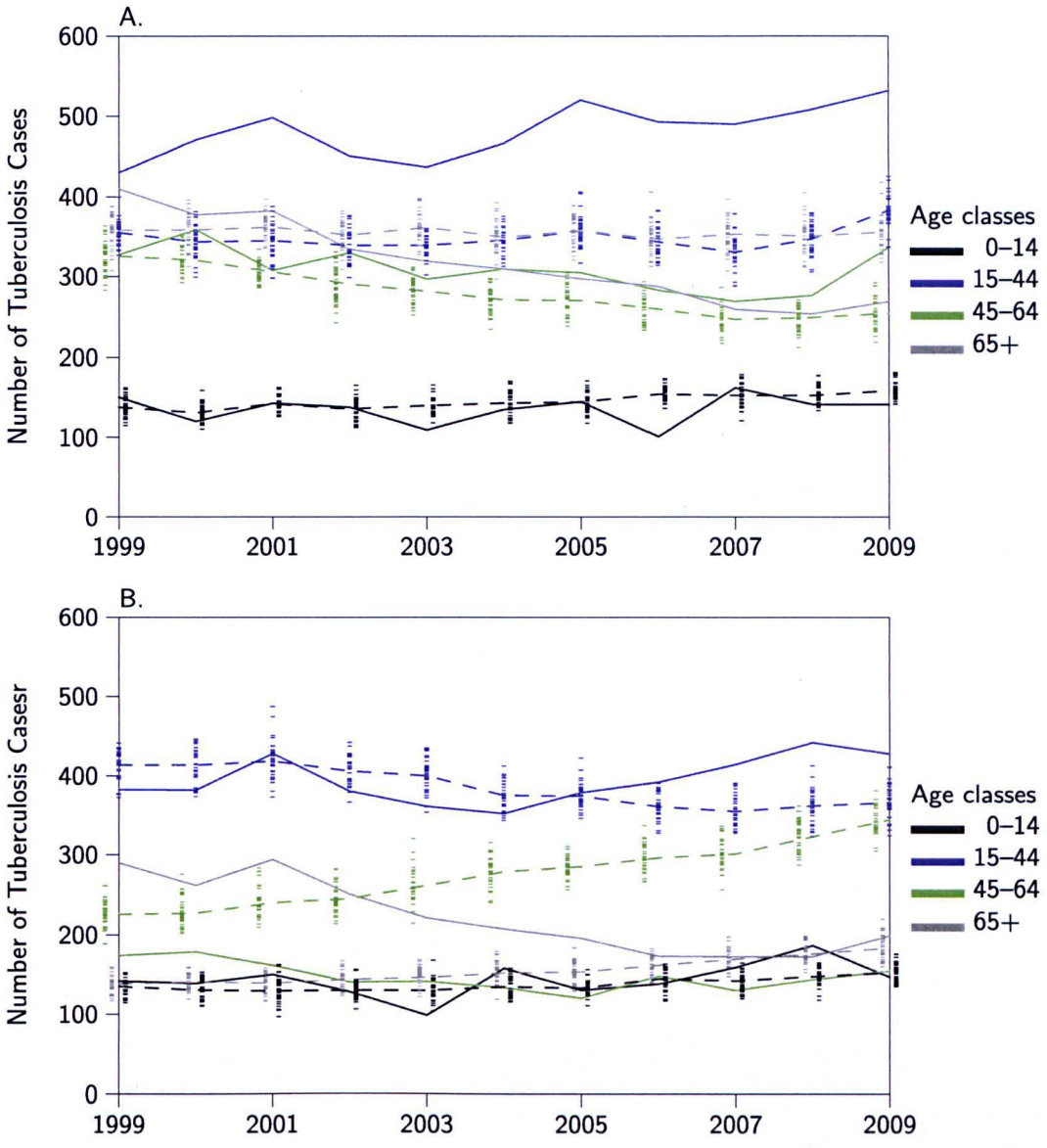


Figure 5-7: Simulated and observed cases notified in England and Wales for UK-born males (A) and females (B) by age category for 1999 – 2009, under fitting Scenario 14. Average values of model output follow the dashed line and individual runs of the model are denoted with a '-' (there are 30 for each data point). Observed numbers of case notifications follow the solid line. Age categories are as follows: ages 0 – 14 years are in black; ages 15 – 44 years are in blue; ages 45 – 64 are in green; and ages 65 years and over are in grey.

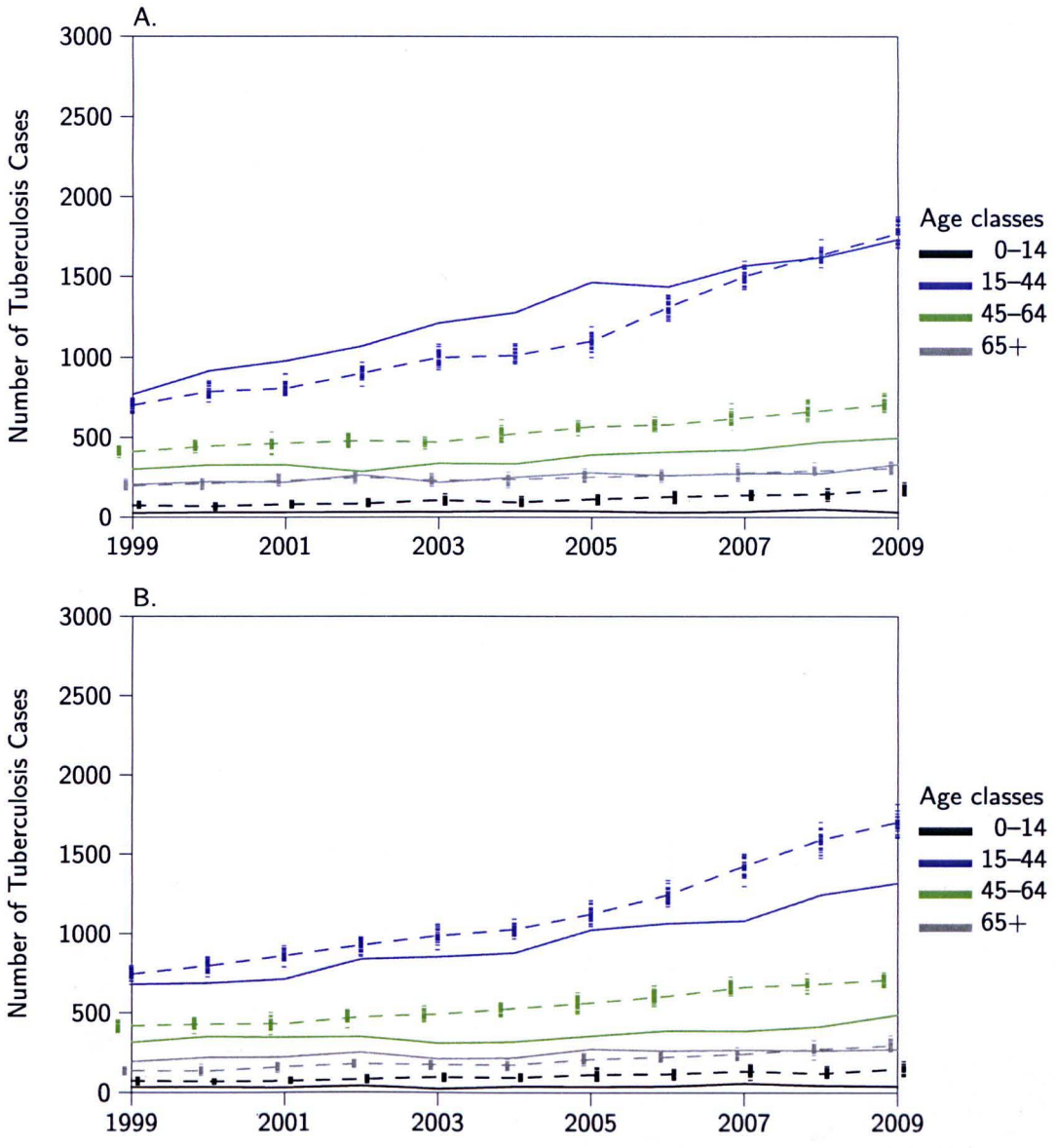


Figure 5-8: Simulated and observed cases notified in England and Wales for OF-born males (A) and females (B) by age category for 1999 – 2009, under fitting Scenario 14. Average values of model output follow the dashed line and individual runs of the model are denoted with a '-' (there are 30 for each data point). Observed numbers of case notifications follow the solid line. Age categories are as follows: ages 0 – 14 years are in black; ages 15 – 44 years are in blue; ages 45 – 64 are in green; and ages 65 years and over are in grey.

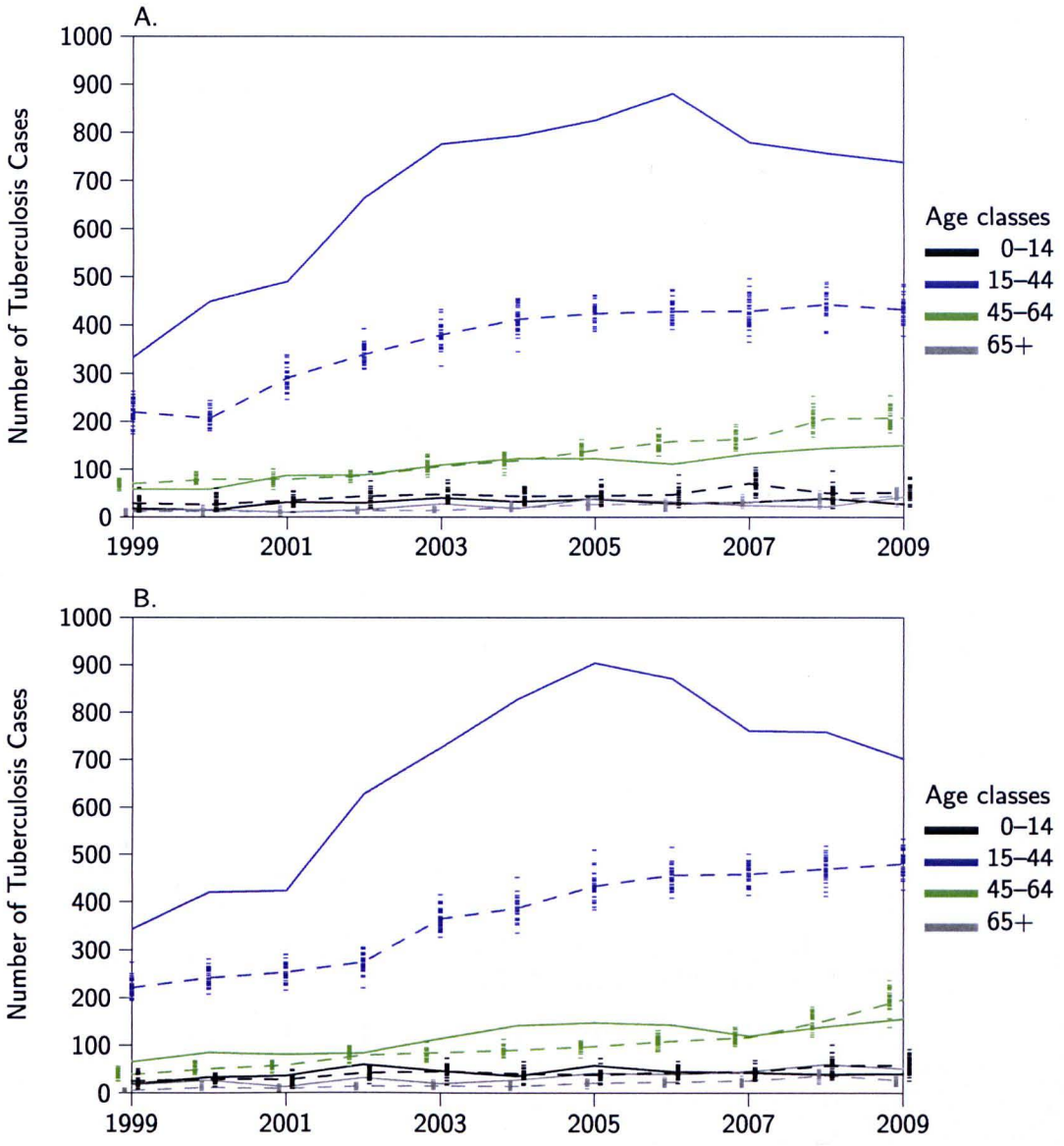


Figure 5-9: Simulated and observed cases notified in England and Wales for SSA-born males (A) and females (B) by age category for 1999 – 2009, under fitting Scenario 14. Average values of model output follow the dashed line and individual runs of the model are denoted with a '-' (there are 30 for each data point). Observed numbers of case notifications follow the solid line. Age categories are as follows: ages 0 – 14 years are in black; ages 15 – 44 years are in blue; ages 45 – 64 are in green; and ages 65 years and over are in grey.

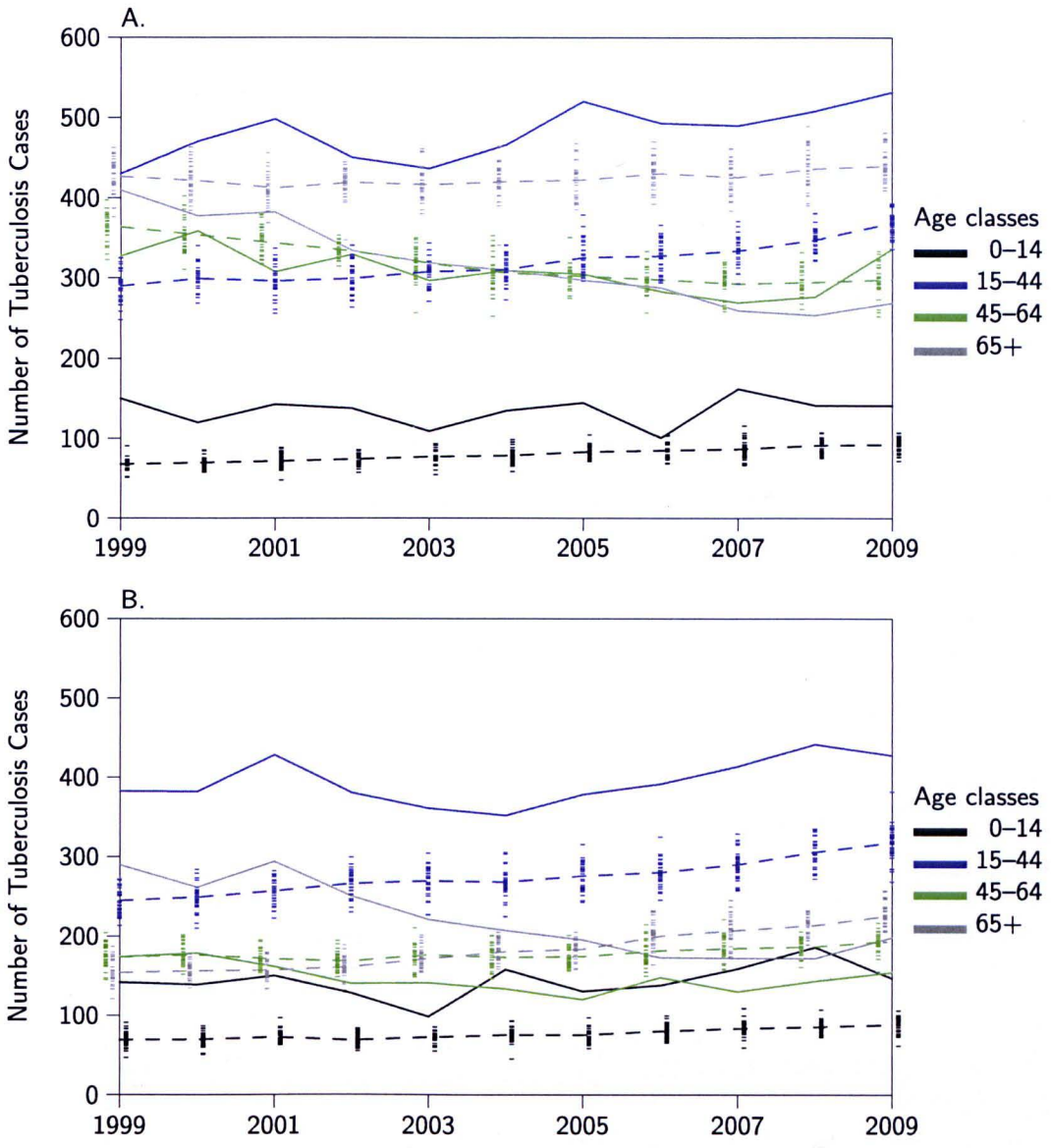


Figure 5-10: Simulated and observed cases notified in England and Wales for UK-born males (A) and females (B) by age category for 1999 – 2009, under fitting Scenario 22. Average values of model output follow the dashed line and individual runs of the model are denoted with a '-' (there are 30 for each data point). Observed numbers of case notifications follow the solid line. Age categories are as follows: ages 0 – 14 years are in black; ages 15 – 44 years are in blue; ages 45 – 64 are in green; and ages 65 years and over are in grey.

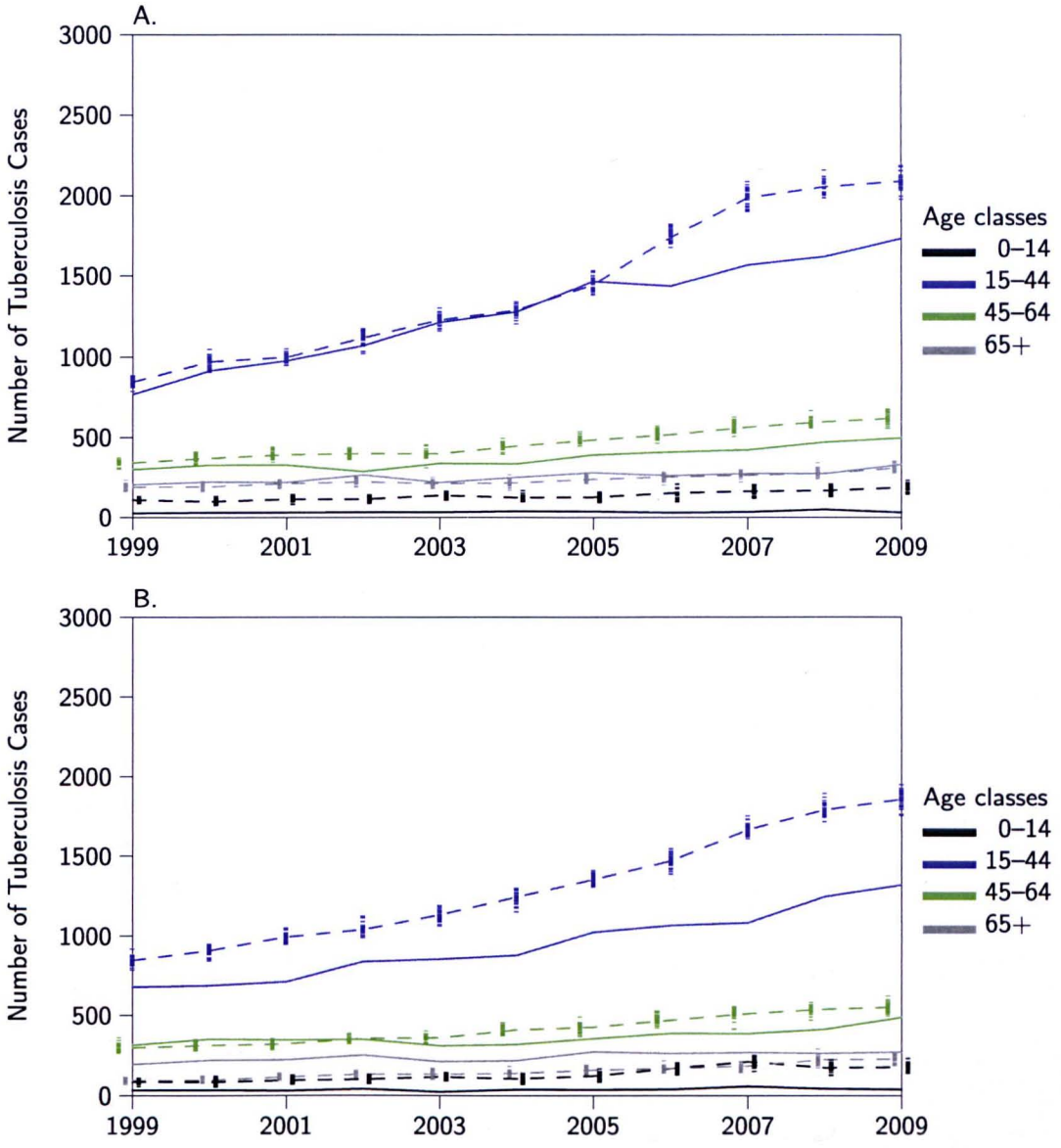


Figure 5-11: Simulated and observed cases notified in England and Wales for OF-born males (A) and females (B) by age category for 1999 – 2009, under fitting Scenario 22. Average values of model output follow the dashed line and individual runs of the model are denoted with a '-' (there are 30 for each data point). Observed numbers of case notifications follow the solid line. Age categories are as follows: ages 0 – 14 years are in black; ages 15 – 44 years are in blue; ages 45 – 64 are in green; and ages 65 years and over are in grey.

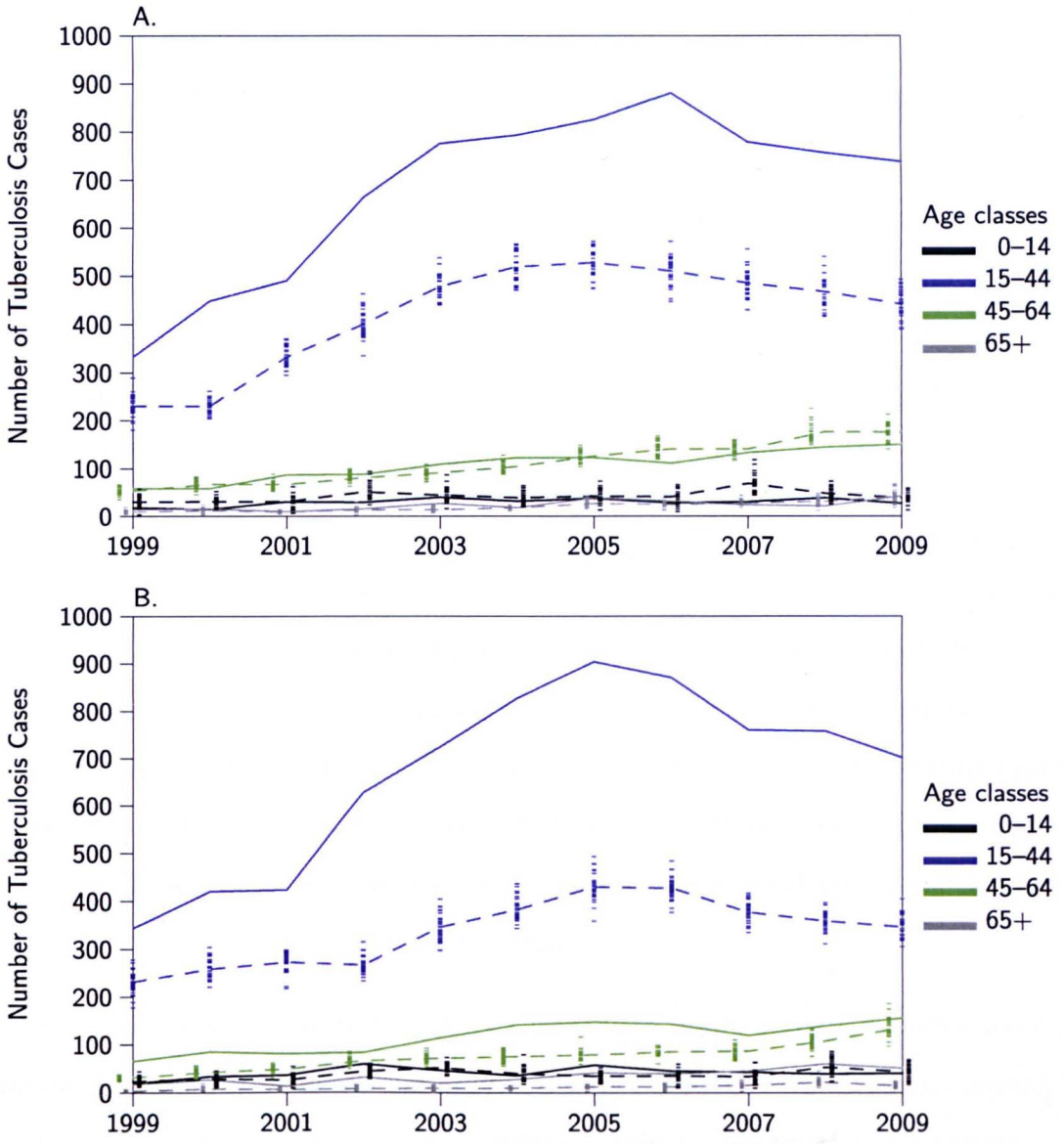


Figure 5-12: Simulated and observed cases notified in England and Wales for SSA-born males (A) and females (B) by age category for 1999 – 2009, under fitting Scenario 22. Average values of model output follow the dashed line and individual runs of the model are denoted with a '-' (there are 30 for each data point). Observed numbers of case notifications follow the solid line. Age categories are as follows: ages 0 – 14 years are in black; ages 15 – 44 years are in blue; ages 45 – 64 are in green; and ages 65 years and over are in grey.

#### 5.2.1.1.2 Replicates

The five replicates run for each scenario are relevant to assessing the quality of fit of the model to data and choice of best-fitting scenarios in two ways. First, replicates show the kind of variation in fits that occurred within a scenario and give an indication of the acceptability of using only the best-fitting of the replicates. Secondly, the five replicates allow comparison of variation within a scenario to variation among the 25 scenarios. This comparison can help determine whether there were significant differences in the quality of fit among scenarios.

To illustrate variation across replicates, the five replicates from scenarios 14 and 22 are shown below in Table 5-6 and Table 5-7. The patterns of variation among replicates were typical. The patterns support considering the best-fitting among the five replicates as the true best fit, although more replicates would have increased the chance of improving this fit. For example, in scenario 14, the two best-fitting replicates were almost identical in GOF statistics and best-fitting parameter estimates. For scenario 22 parameter estimates were similar for the two best fits, though the GOF statistics varied more between the two best fits.

As illustrated with scenarios 14 and 22, GOF values among replicates within a scenario varied substantially. However, an ANOVA on the variance in GOF statistics among scenarios showed that, despite this variation within a scenario, the GOF results differed significantly among the 25 scenarios (see Figure 5-8, p-value 0.00003). When including only the 10 scenarios which were used for stage two modelling, ANOVA results showed GOF statistics still differed significantly among the 10 scenarios at the p=0.05 level (see Figure 5-9, p-value 0.014), though the p-value was much larger. After removing the two scenarios with the highest mean GOF statistics, scenarios 17 and 23, ANOVA results on the eight remaining best scenarios showed that GOF statistics were not significantly different among scenarios (p-value 0.127).

Table 5-6: Results for the five replicates of Scenario 14 in stage one fitting. Results include estimated disease risks and the value of the goodness of fit (GOF) statistic measuring the quality of fit of the model to data. The GOF rank orders fits from the best (1) to the worst (5).

| Primary (%) | Reactivation (% per year) | Reinfection (%) | Foreign:UK-born risk ratio (*df*) | GOF | GOF rank |
|---|---|---|---|---|---|
| 8.5% | 0.0195% | 0.4% | 3.7 | 9201 | 1 |
| 8.8% | 0.0198% | 0.3% | 3.6 | 9235 | 2 |
| 11.0% | 0.0106% | 5.7% | 2.8 | 10019 | 3 |
| 10.6% | 0.0066% | 10.7% | 2.7 | 10508 | 4 |
| 12.5% | 0.0002% | 15.7% | 2.3 | 11735 | 5 |

Table 5-7: Results for the five replicates of Scenario 22 in stage one fitting. Results include estimated disease risks and the value of the goodness of fit (GOF) statistic measuring the quality of fit of the model to data. The GOF rank orders fits from the best (1) to the worst (5).

| Primary (%) | Reactivation (% per year) | Reinfection (%) | Foreign:UK-born risk ratio (*df*) | GOF | GOF rank |
|---|---|---|---|---|---|
| 18.2% | 0.0204% | 7.9% | 2.0 | 9171 | 1 |
| 15.8% | 0.0207% | 8.2% | 2.2 | 10288 | 2 |
| 14.3% | 0.0237% | 8.4% | 2.4 | 10477 | 3 |
| 18.3% | 0.0253% | 1.2% | 2.0 | 11571 | 4 |
| 10.2% | 0.0286% | 14.2% | 2.9 | 12515 | 5 |

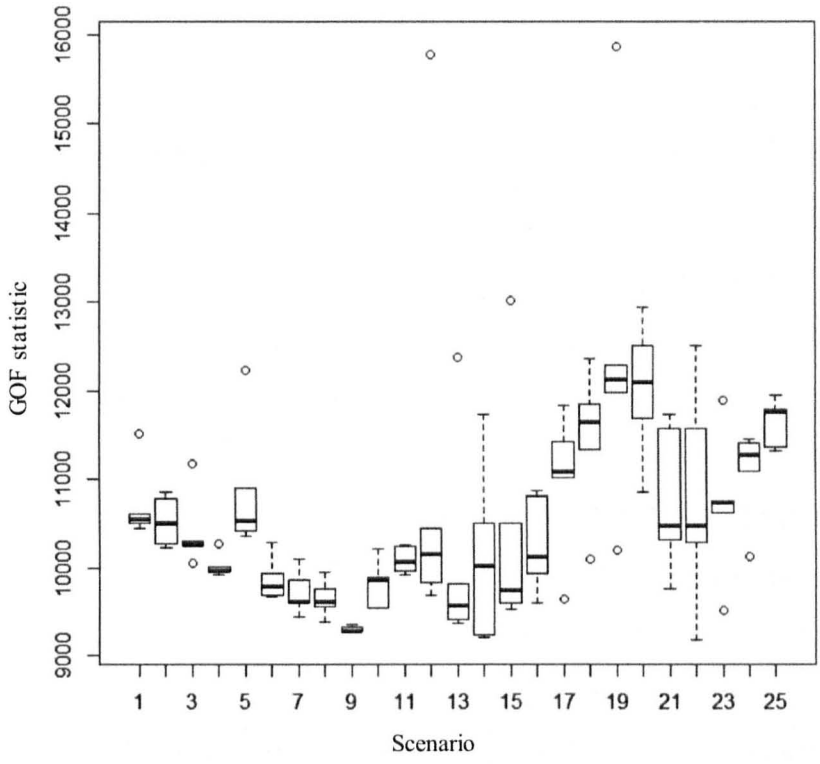


Figure 5-13: Box plot of goodness-of-fit (GOF) statistics for each of the 25 fitting scenarios, with five replicates each. GOF statistics differ significantly among scenarios according to a one-way analysis of variance (p-value 0.00003).

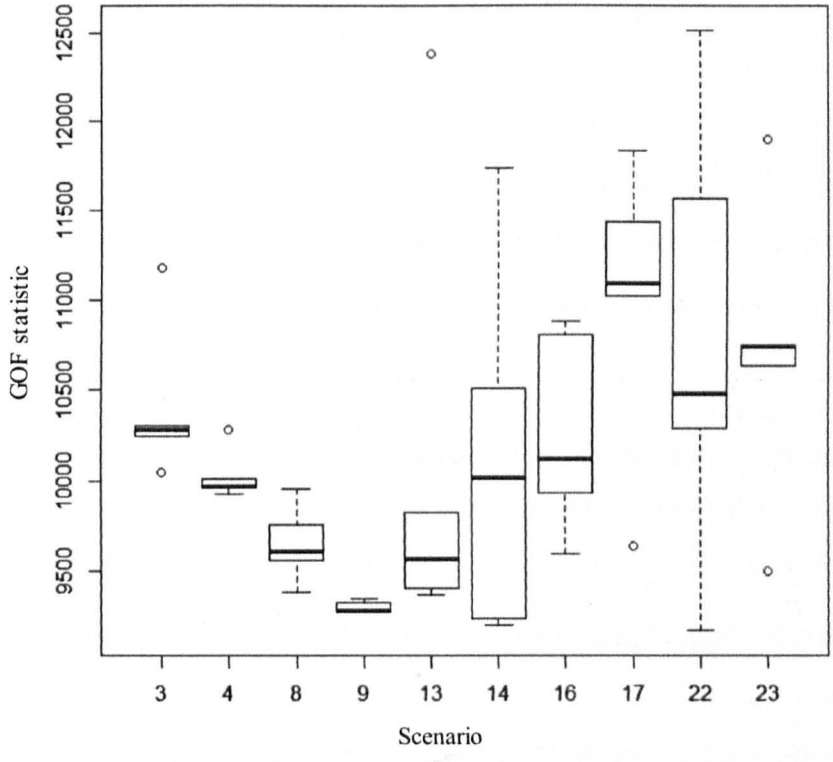


Figure 5-14: Box plot of goodness-of-fit (GOF) statistics among the 10 scenarios used for stage two fitting. GOF statistics differ significantly according to the one-way analysis of variance (p-value 0.0141). Removing scenarios 17 and 23 resulted in no significant differences in GOF statistics for the remaining eight scenarios (p-value 0.127).

### *5.2.1.2 Disease risk estimates*

For each of the 25 scenarios, the best-fitting values for the four disease-risk parameters and the GOF statistics are shown in Table 5-5, for the best-fitting replicate of each scenario. Although the quality of fits of the model to observed data were similar for many scenarios, estimates of the risk disease risk varied across scenarios. The cumulative risk of *Primary Disease* over the five years of *Recent Infection* ranged from about 8% to over 30% across the 25 scenarios, as estimated for UK-born adult males. Estimates of the risk of *Reactivation Disease*, or the annual percentage with *Latent Infection* who developed disease, ranged from about 0.003% to 0.025% per year, on average 0.015% per year for UK-born adult males. Estimates for the cumulative risk of *Reinfection Disease* over the first five years of reinfection had a relatively larger range across scenarios, with estimates from about 0.01% up to over 27% for UK-born adult males. Across all scenarios, the estimates averaged at about 8%. The estimated ratios of disease risk between foreign-born and UK-born adult males, *df,* ranged from 1.35 – 3.79. On average, *df* was 2.1 across all scenarios.

For simplification of reporting and because for stage two fitting only 10 of 25 scenarios were tested, the remainder of this chapter will focus on results from those 10 scenarios. These are summarized in Table 5-8. Comparison of this table with Table 5-5 shows that average estimates for the 10 best scenarios are similar to those for all 25 stage one fitting scenarios. When considering just the 10 scenarios, the estimated risk for *Reinfection Disease* was slightly lower, about 5%, and *df* was slightly higher, around 2.2.

For illustration of the differences in disease risk among demographic groups, Table 5-9 shows foreign-born and female disease risks calculated from the UK-born male risks estimated for the 10 best-fitting scenarios. Risks for foreign-born individuals, including both those who were OF-born and those who were HIV-negative SSA-born, were calculated by multiplying *df* by the risks for each disease type. Consistent with model assumptions, risk ratios for foreign-born to UK-born disease risk were equal for each disease type. Disease risks for females were obtained by multiplying male risks for a given category by the ratio between female and male disease risks, as described in Section 5.1.3.1. These values highlight the differences in risks between males and

females, particularly for *Reinfection Disease* risk. For simplification of the presentation of results, only the disease risk parameters for UK-born males aged 20 years and older will be reported for the remainder of fits.

The estimates for cumulative risk of developing *Primary Disease* for adults were greater than the cumulative risk of developing *Reinfection Disease* for each of the 10 best-fitting scenarios, as shown in Table 5-10. The average risk ratio for *Reinfection Disease* to *Primary Disease* was 0.39, corresponding to a 61% reduction in the risk of *Reinfection Disease* compared to *Primary Disease*.

Table 5-8: Disease risk estimates for each of the 10 of the best-fitting scenarios of stage one fits.

| Scenario | Primary (%) | Reactivation (% per year) | Reinfection (%) | Foreign:UK-born |
|---|---|---|---|---|
| 3 | 12.4% | 0.009% | 7.6% | 1.90 |
| 4 | 10.3% | 0.009% | 6.2% | 1.85 |
| 8 | 10.9% | 0.011% | 5.9% | 1.97 |
| 9 | 9.3% | 0.010% | 5.9% | 1.88 |
| 13 | 10.0% | 0.021% | 0.4% | 3.79 |
| 14 | 8.5% | 0.019% | 0.4% | 3.69 |
| 16 | 32.6% | 0.022% | 0.1% | 1.69 |
| 17 | 23.1% | 0.017% | 8.0% | 1.65 |
| 22 | 18.2% | 0.020% | 7.9% | 2.02 |
| 23 | 14.8% | 0.022% | 7.6% | 1.89 |
| Mean | 15.0% | 0.016% | 5.0% | 2.23 |

Table 5-9: Disease risk estimates by sex and birthplace for adults 20 years and older for the 10 best-fitting scenarios of stage one fitting. Risks for UK-born adult males and the disease risk ratio between foreign-born and UK-born individuals were estimated by the model. The disease risk ratio was used to calculate foreign-born risks. Female risks were calculated from parameter risk ratios between females and males, which were fixed in the model (see Table 5-1).

| | Disease risks for UK-born males | | | Disease risks for foreign-born males | | | Disease risks for UK-born females | | | Disease risks for foreign-born females | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | Primary (%) | React. (%/year) | Reinf. (%) | Primary (%) | React. (%/year) | Reinf. (%) | Primary (%) | React. (%/year) | Reinf. (%) | Primary (%) | React. (%/year) | Reinf. (%) |
| 3 | 12.4 | 0.009 | 7.6 | 23.6 | 0.017 | 14.4 | 12.4 | 0.001 | 0.01 | 23.6 | 0.003 | 0.02 |
| 4 | 10.3 | 0.009 | 6.2 | 19.1 | 0.016 | 11.4 | 10.3 | 0.001 | 0.01 | 19.1 | 0.003 | 0.01 |
| 8 | 10.9 | 0.011 | 5.9 | 21.4 | 0.021 | 11.6 | 10.9 | 0.002 | 0.01 | 21.4 | 0.003 | 0.01 |
| 9 | 9.3 | 0.010 | 5.9 | 17.5 | 0.019 | 11.0 | 9.3 | 0.002 | 0.01 | 17.5 | 0.003 | 0.01 |
| 13 | 10.0 | 0.021 | 0.4 | 38.0 | 0.078 | 1.6 | 10.0 | 0.003 | 0.00 | 38.0 | 0.012 | 0.00 |
| 14 | 8.5 | 0.019 | 0.4 | 31.4 | 0.072 | 1.4 | 8.5 | 0.003 | 0.00 | 31.4 | 0.012 | 0.00 |
| 16 | 32.6 | 0.022 | 0.1 | 55.1 | 0.036 | 0.2 | 32.6 | 0.003 | 0.00 | 55.1 | 0.006 | 0.00 |
| 17 | 23.1 | 0.017 | 8.0 | 38.1 | 0.028 | 13.2 | 23.1 | 0.003 | 0.01 | 38.1 | 0.004 | 0.02 |
| 22 | 18.2 | 0.020 | 7.9 | 36.7 | 0.041 | 16.0 | 18.2 | 0.003 | 0.01 | 36.7 | 0.007 | 0.02 |
| 23 | 14.8 | 0.022 | 7.6 | 27.8 | 0.041 | 14.4 | 14.8 | 0.003 | 0.01 | 27.8 | 0.007 | 0.02 |
| Mean | 15.0% | 0.016% | 5.0% | 30.9% | 0.037% | 9.5% | 15.0% | 0.003% | 0.0% | 30.9% | 0.006% | 0.0% |

Table 5-10: Estimated risks of *Primary Disease*, *Reinfection Disease*, and the risk ratio between *Reinfection Disease* and *Primary Disease* for 10 of the best-fitting scenarios from stage one.

| Scenario | Primary Disease (%) | Reinfection Disease (%) | Risk ratio |
|---|---|---|---|
| 3 | 12.4% | 7.6% | 0.61 |
| 4 | 10.3% | 6.2% | 0.60 |
| 8 | 10.9% | 5.9% | 0.54 |
| 9 | 9.3% | 5.9% | 0.63 |
| 10 | 15.2% | 6.5% | 0.43 |
| 13 | 10.0% | 0.4% | 0.04 |
| 14 | 8.5% | 0.4% | 0.04 |
| 16 | 32.6% | 0.1% | 0.00 |
| 22 | 18.2% | 7.9% | 0.44 |
| 23 | 14.8% | 7.6% | 0.52 |
| Mean | | | 0.39 |

#### *5.2.1.3 Proportion of disease due to recent transmission in the UK*

The proportion of cases due to recent transmission in the UK across the 10 best-fitting input scenarios is summarized in Table 5-11. The table presents average values for each age and birthplace category, as well as the interquartile range across years, input scenarios, and model replicates, as an indicator of variability for each estimate. Age-specific trends vary by birthplace. For UK-born individuals, the highest proportion of disease due to recent transmission was estimated for UK-born children 0 – 14 years of age, at 95% on average. This percentage decreased as age increased, with the proportion of cases due to recent transmission in the UK estimated at 48% on average in those aged 65 years and above. For OF-born and SSA-born individuals, the proportion of cases due to recent transmission in the UK increased with increasing age. Those aged 0 – 14 years had the lowest estimated proportions of disease due to recent transmission in the UK, at 46% and 37% on average respectively, whereas those aged 65 years and above had the highest proportion of disease due to recent transmission, at 78% and 53% on average. Across all age, sex, and birthplace categories from 1999 – 2009 an estimated 59.8% of cases were due to recent transmission in the UK. For UK-born individuals, this percentage was estimated to be 67% and for foreign-born individuals (OF-born and SSA-born combined) this percentage was estimated to be 56%.

This proportion of cases due to recent transmission in the UK differs among input scenarios, as shown in Table 5-12 and illustrated in plots of the estimates for scenarios 14 and 22, shown in Figure 5-10 – Figure 5-20. As in the case over all 10 best-fitting scenarios, both scenarios showed those aged 0 – 14 years, both males and females, were mostly due to recent transmission in the UK. Under scenario 22, the proportion was close to 100%. On the other hand, those aged 65 years and above had the lowest proportion of cases due to recent transmission in the UK, around 20% on average for males under Scenario 14, though increased over time. Under scenario 22, the proportion was higher, increasing over time from about 30 to 40% for males. Similar trends were found females, although the proportion of cases due to recent transmission was higher than in males.

For OF-born individuals, estimates of the proportion of cases due to recent transmission in the UK were generally very high under Scenario 14, usually between 80 – 90% for all age groups apart from males aged 65 years and above, where the proportion was closer to an average of about 60%. These proportions were lower under Scenario 22, though were still fairly high for older individuals, aged 45 – 64 years and 65 years and above, which were estimated to have 70 – 90% of cases due to recent transmission in the UK.

For SSA-born individuals, estimates of the proportion of disease due to recent transmission in the UK showed more stochasticity than estimates for UK-born and OF-born. Age-specific trends were also less clear. Generally, for both illustration scenarios, those aged 0 – 14 years and 15 – 44 years had a lower proportion of disease due to recent transmission in the UK. Those aged 45 – 64 years and 65 years and above had a higher proportion of disease due to recent transmission in the UK.

Table 5-11: Estimated proportion of cases due to recent transmission in the UK by age and birthplace, across the 10 best-fitting scenarios of stage one, 1999 – 2009 for England and Wales. Estimates were averaged for 30 replicates of the model per scenario.

| Birthplace | Age category | Mean | 0.25 | 0.75 |
|---|---|---|---|---|
| UK | 0 – 14 years | 0.95 | 0.93 | 0.98 |
| | 15 – 44 years | 0.72 | 0.63 | 0.83 |
| | 45 – 64 years | 0.68 | 0.55 | 0.8 |
| | 65+ years | 0.48 | 0.37 | 0.59 |
| OF | 0 – 14 years | 0.46 | 0.26 | 0.59 |
| | 15 – 44 years | 0.52 | 0.38 | 0.55 |
| | 45 – 64 years | 0.8 | 0.74 | 0.87 |
| | 65+ years | 0.78 | 0.73 | 0.86 |
| SSA | 0 – 14 years | 0.37 | 0.17 | 0.54 |
| | 15 – 44 years | 0.39 | 0.28 | 0.49 |
| | 45 – 64 years | 0.59 | 0.46 | 0.72 |
| | 65+ years | 0.53 | 0.29 | 0.75 |

Table 5-12: Proportion of cases due to recent transmission in the UK by input scenario, averaged over 1999 – 2009 and over all demographic categories.

| Scenario | Mean | 0.25 | 0.75 |
|---|---|---|---|
| 3 | 0.56 | 0.39 | 0.74 |
| 4 | 0.58 | 0.4 | 0.75 |
| 8 | 0.52 | 0.35 | 0.7 |
| 9 | 0.55 | 0.37 | 0.72 |
| 13 | 0.72 | 0.59 | 0.88 |
| 16 | 0.59 | 0.35 | 0.85 |
| 17 | 0.65 | 0.46 | 0.88 |
| 22 | 0.59 | 0.38 | 0.81 |
| 23 | 0.62 | 0.42 | 0.84 |

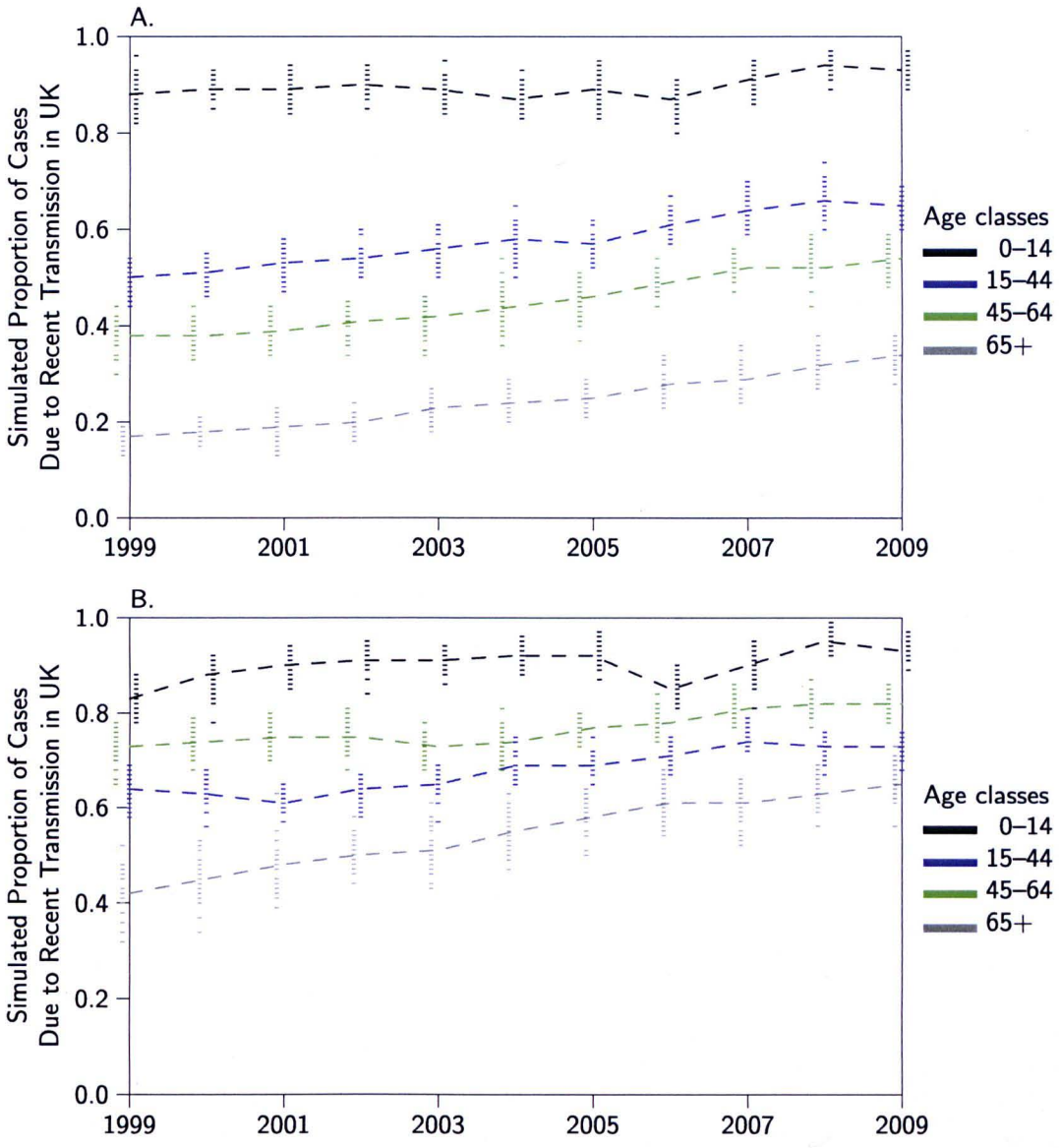


Figure 5-15: Proportion of cases due to recent transmission in simulation results for UK-born males (A) and females (B) by age category for 1999 – 2009, under fitting Scenario 14. The averaged model output follows the dashed line and individual runs of the model are denoted with a '-' (there are 30 for each data point). Age categories are as follows: ages 0 – 14 years are in black; ages 15 – 44 years are in blue; ages 45 – 64 are in green; and ages 65 years and over are in grey.

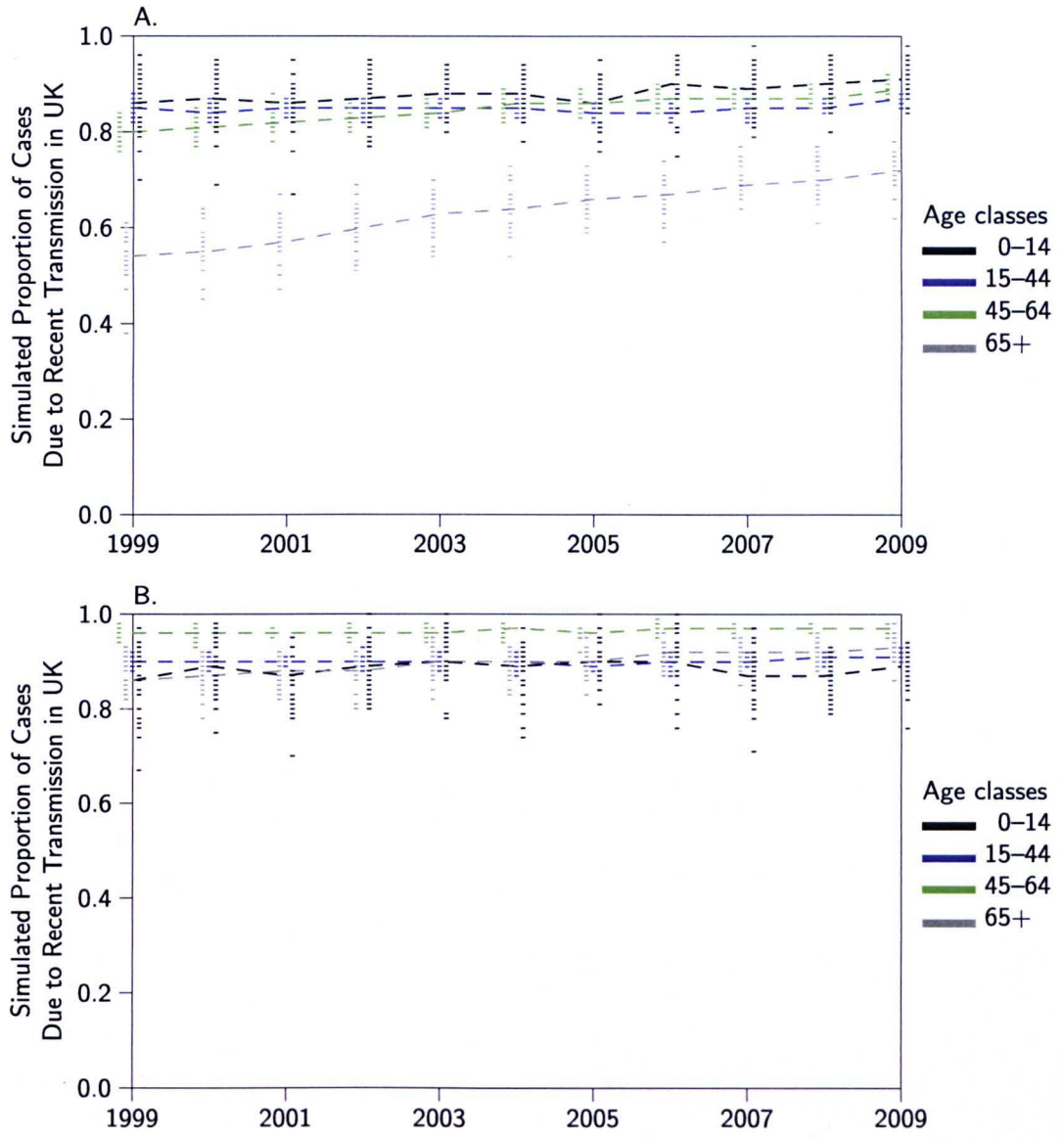


Figure 5-16: Proportion of cases due to recent transmission in simulation results for OF-born males (A) and females (B) by age category for 1999 – 2009, under fitting Scenario 14. The averaged model output follows the dashed line and individual runs of the model are denoted with a '-' (30 for each data point). Age categories are as follows: ages 0 – 14 years are in black; ages 15 – 44 years are in blue; ages 45 – 64 are in green; and ages 65 years and over are in grey.

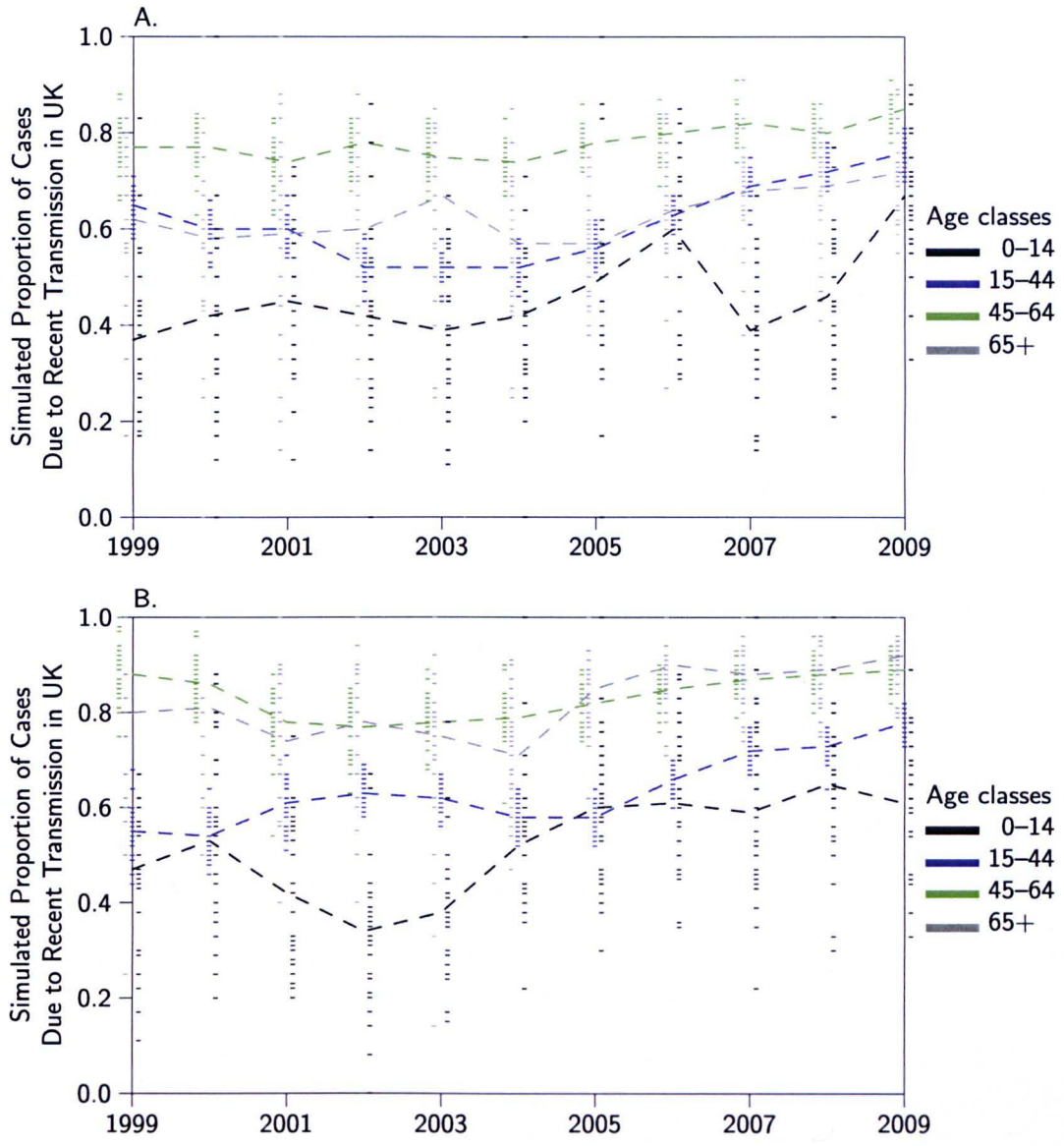


Figure 5-17: Proportion of cases due to recent transmission in simulation results for SSA-born males (A) and females (B) by age category for 1999 – 2009, under fitting Scenario 14. The averaged model output follows the dashed line and individual runs of the model are denoted with a '-' (30 for each data point). Age categories are as follows: ages 0 – 14 years are in black; ages 15 – 44 years are in blue; ages 45 – 64 are in green; and ages 65 years and over are in grey.

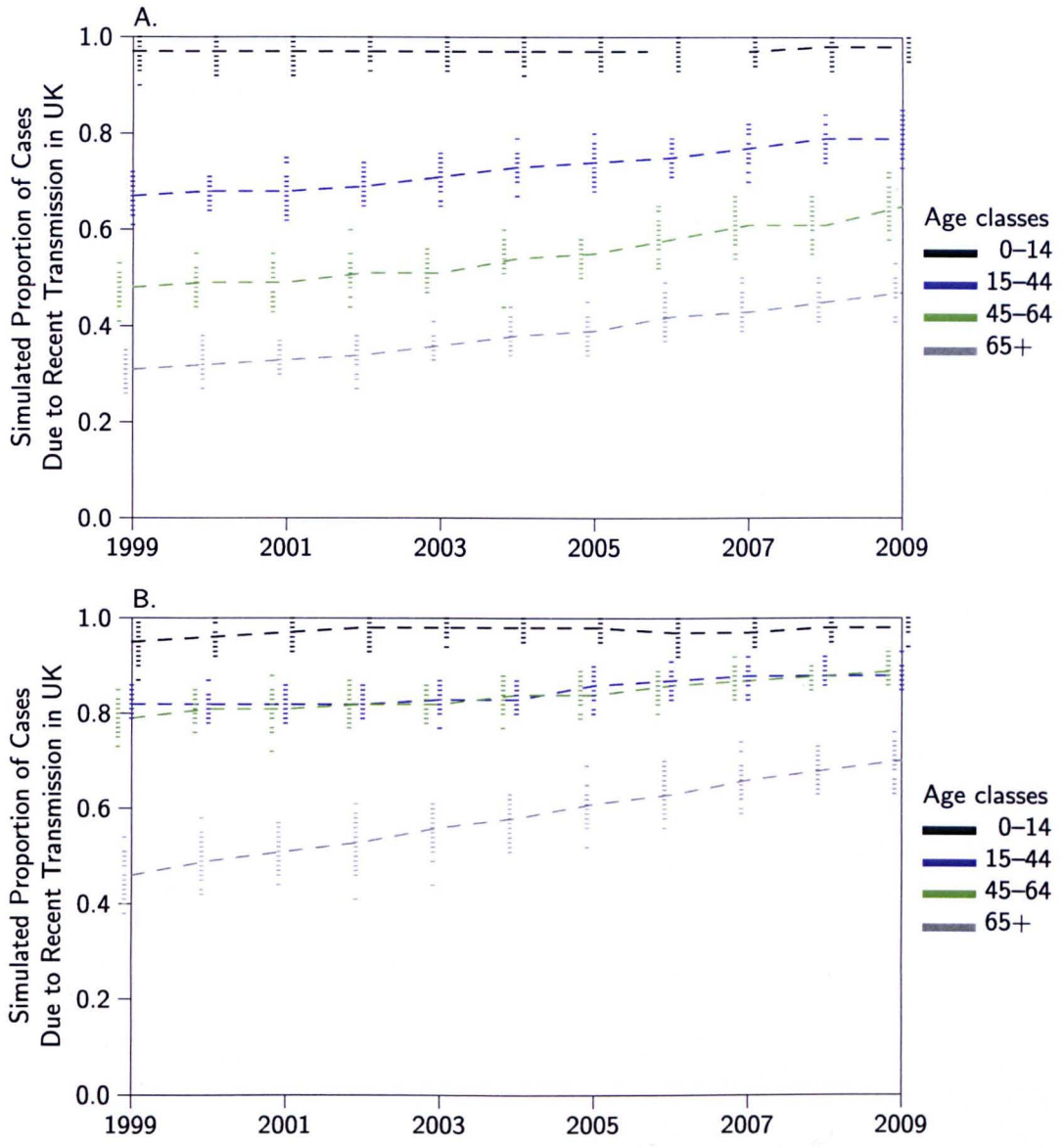


Figure 5-18: Proportion of cases due to recent transmission in simulation results for UK-born males (A) and females (B) by age category for 1999 – 2009, under fitting Scenario 22. The averaged model output follows the dashed line and individual runs of the model are denoted with a '-' (there are 30 for each data point). Age categories are as follows: ages 0 – 14 years are in black; ages 15 – 44 years are in blue; ages 45 – 64 are in green; and ages 65 years and over are in grey.

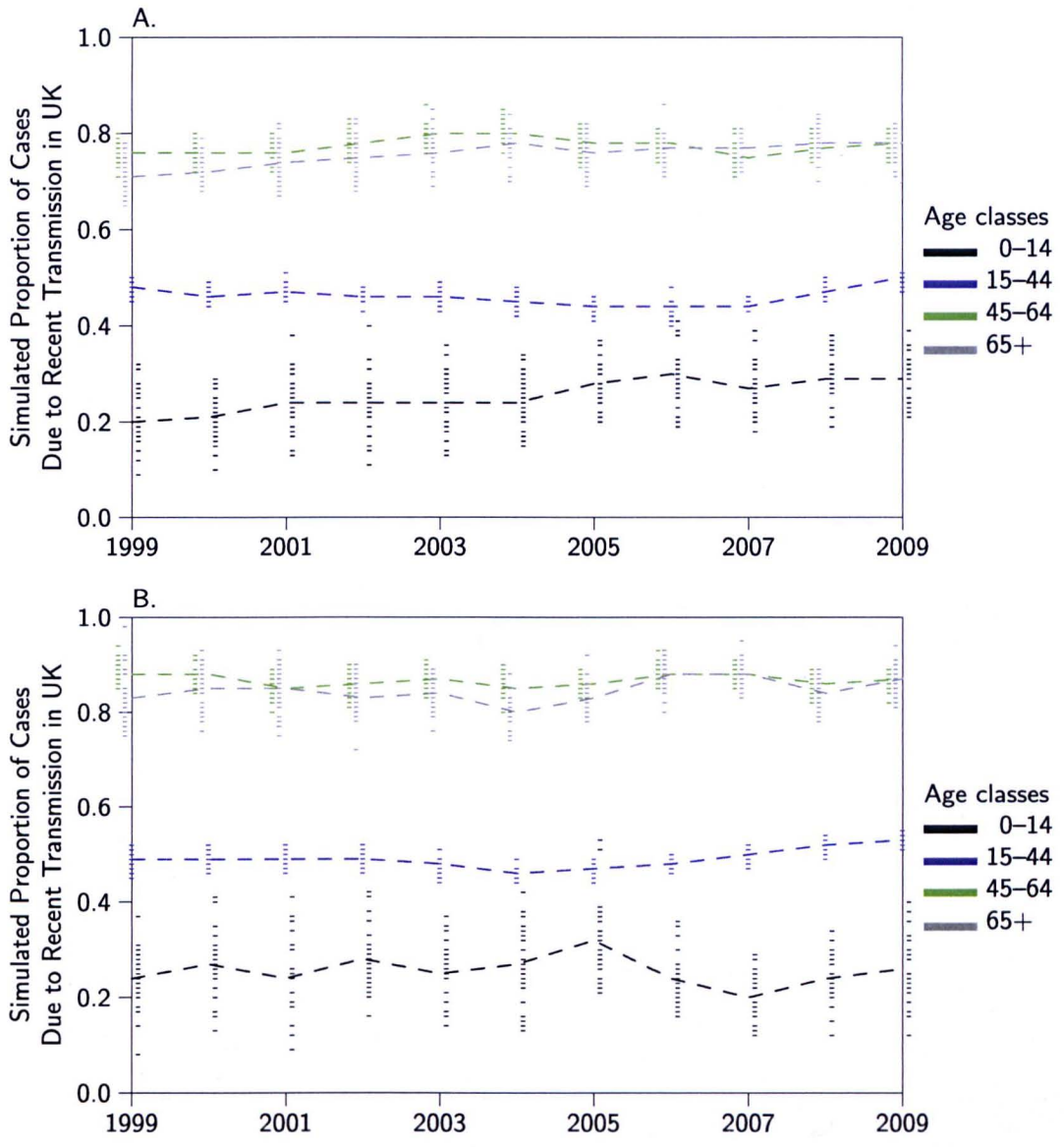


Figure 5-19: Proportion of cases due to recent transmission in simulation results for OF-born males (A) and females (B) by age category for 1999 – 2009, under fitting Scenario 22. The averaged model output follows the dashed line and individual runs of the model are denoted with a '-' (30 for each data point). Age categories are as follows: ages 0 – 14 years are in black; ages 15 – 44 years are in blue; ages 45 – 64 are in green; and ages 65 years and over are in grey.

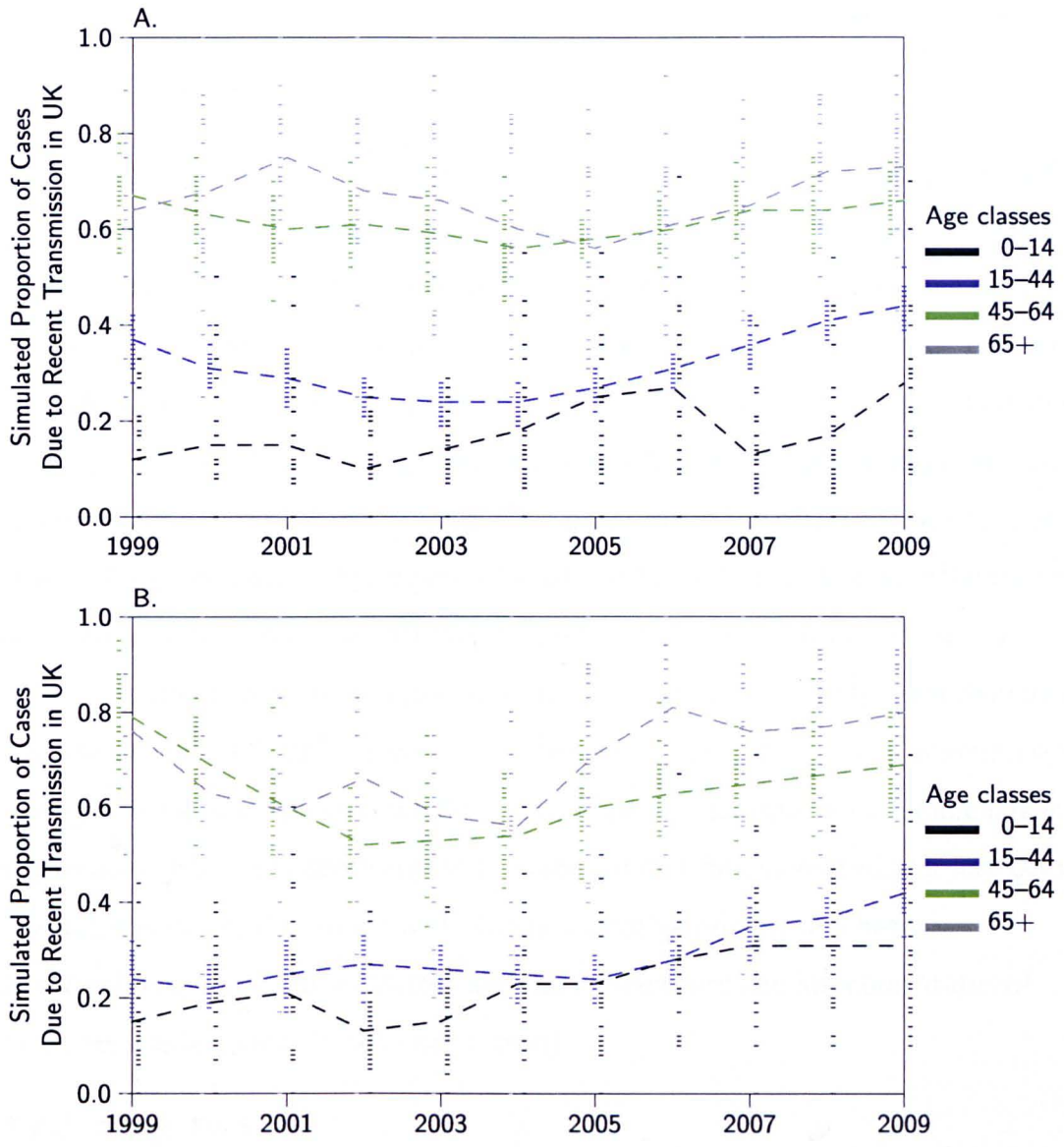


Figure 5-20: Proportion of cases due to recent transmission in simulation results for SSA-born males (A) and females (B) by age category for 1999 – 2009, under fitting Scenario 22. The averaged model output follows the dashed line and individual runs of the model are denoted with a '-' (30 for each data point). Age categories are as follows: ages 0 – 14 years are in black; ages 15 – 44 years are in blue; ages 45 – 64 are in green; and ages 65 years and over are in grey.

#### *5.2.1.4 Effect of contact rate and infection status of migrants*

As detailed in Section 5.2.1.1, results showed several scenarios fit the data roughly equally well. Since scenarios were defined by the contact rate and the infection status of migrants, the effects of these parameters on GOF statistics and disease risks estimates were explored quantitatively. Two-way ANOVA was performed for testing the effects of these parameters on GOF statistics and disease risks, restricting analysis to the 10 scenarios used in stage two fitting. ANOVA showed the contact rate had a significant effect on GOF statistics (p-value 0.02), while the effect of the infection status of migrants was not significant (p-value 0.12). When looking at estimated risks of *Primary Disease*, two-way ANOVA showed that only the contact rate had a significant effect on estimates (p-value $<2x10^{-16}$), which was also true for *Reactivation Disease* (p-value $5.5x10^{-7}$). However, neither the contact rate nor the infection status of migrants significantly affected *Reinfection Disease* risk estimates. Simple linear regression coefficients confirmed that as the contact rate increased, *Primary Disease* risk decreased (results not shown). The quantitative relationship between *Reactivation Disease* risk estimates and the contact rate and infection status of migrants was less clear (results not shown).

### 5.2.2 Stage Two

Stage two fits included four variations on stage one fits, targeting parameters and assumptions that were based on the least reliable data. These were: 1) altered parameter distributions for the HIV prevalence of SSA-born immigrants; 2) altered infection status for SSA-born individuals at model initialization in 1981; 3) additional disease risk parameters estimated; and 4) fitting of a single foreign-born category, comprised of both SSA-born and OF-born individuals. Results from the first three of the four variations showed that fits of model output to observed data were not improved. These are reported in Appendices 10.10, 10.11, and 10.12. Improved fits were obtained with the fourth variation, in which a single foreign-born category was fit to data, as described in 5.1.5.4. The remainder of this section will report on those results.

### *5.2.2.1 Quality of fits*

Fitting of the single foreign-born group resulted in noticeably improved fits, especially for the foreign-born group, both when looking qualitatively at plots of observed data versus model output and when comparing the GOF statistics, given in Table 5-13, to those of previous fits. Fits of the model to observed data are described for both the numbers of case notifications that were used to fit the model and for the observed notification rates. Variance in GOF statistics among and between the 10 scenarios is also discussed.

#### 5.2.2.1.1 Fits to case notifications

It should be noted that the GOF statistics for stage two fits with a single foreign-born category were lower than those for stage one fits, indicating better fits to data, but this difference is at least partly due to the fact that there were fewer data points used to measure the GOF for this fitting scheme. In stage one fits, the GOF statistic was calculated using 154 data points, with 66 data points each for UK-born and OF-born cases and 22 data points for SSA-born cases. The 66 data points for UK-born and OF-born cases were derived from 11 years of data, two sexes, and three age groups for each birthplace. The 22 data points for SSA-born case were derived from 11 years of data, two sexes, and one age group. In the stage two fitting scheme, the 22 data points for SSA-born cases were removed, as these cases were combined with OF-born cases for one foreign-born category. This left a total of 132 data points.

Fewer data points meant that the Chi square distribution to compare the GOF statistic against had fewer degrees of freedom. This would normally aid interpretation of the GOF statistic and comparison across studies, though in the case of results from this study, across both stage one and stage two, the GOF statistic compared to the Chi square distribution led to the same value, 1. Thus, the statistics did not help differentiate quality of fits across scenarios with different degrees of freedom. Instead, visual inspection of plots from stage one and stage two confirmed that stage two fits were better. Within a stage of fitting with the same number of data points, the GOF statistic was used to compare fits.

For illustration of fit quality, the model output versus observed number of tuberculosis notifications for the two best-fitting scenarios from this fitting scheme, scenarios 9 and

4, are shown in Figure 5-21 – 5-24. It is clear from these plots that foreign-born fits are generally good. UK-born fits vary with age category and sex and are imperfect for some of these categories.

For both scenarios 9 and 4, the number of cases of tuberculosis among UK-born individuals aged 0 – 14 years was predicted well by the model, despite those groups not being used for calculation of the GOF statistic in fitting. For those aged 15 – 44 years, the model predicts observed cases fairly well for females. For males, cases are substantially underestimated by the model. For those aged 45 – 64 years, the number of cases in males was predicted well by the model, but for females, this group was problematic. In both best-fitting scenarios, the model predicted an increasing number of cases in this group while the observed number of cases decreased from 1999 – 2009. For those aged 65 years and above, fits were also problematic; in males, the observed number of cases decreased. Model output showed some increases and decreases but overall, a stable number of cases from 1999 – 2009. Generally, the model overestimated the number of cases for this group, though numbers were reasonably similar to those observed. For females on the other hand, the model was much worse at predicting observed numbers of cases. In this group, the number of observed cases also decreased from 1999 – 2009, even more markedly than for males. The model predicted a trend of increasing case notifications over this time period, failing to fit the data at all, except when the upward and downward trend lines crossed in 2006.

For foreign-born cases, for all groups for males and females, model trends reproduced observed trends in notification data. For those aged 0 – 14 years, the model slightly overestimated the number of cases for both males and females, although this was a very small absolute number of cases. For those aged 15 – 44 years, the group in which the vast majority of cases were found, the upward trend in numbers of notification was reproduced well by the model for both males and females. The number of cases was also reproduced well by the model, apart from a slight overestimation near the end of the simulation when the observed number of cases levelled off for both males and females. The number of cases of tuberculosis in those aged 45 – 64 years was accurately predicted by the model for males and females, although slightly

overestimated by the model for males. For those aged 65 years and above, the model predicted the number of observed cases well for males and females.

Table 5-13: Results of model fitting to observed case notifications for stage two fits with a single foreign-born group. Results include estimated disease risks and the value of the goodness of fit (GOF) statistic measuring the quality of fit of the model to data. The GOF rank orders fits from the best (1) to the worst (10). Only the best-fitting replicate for each scenario is shown. Assumptions for the infection status of migrants upon entry to the UK and values for the contact rate are also given.

| | | | Disease risks for UK-born adult males by disease type | | | | | |
|---|---|---|---|---|---|---|---|---|
| Scenario | Infection status of migrants | Contact rate | Primary (%) | Reactivation (% per year) | Reinfection (%) | Foreign:UK-born | GOF | GOF rank |
| 3 | Scr1 | all = 8 | 10.0% | 0.012% | 3.4% | 2.33 | 3548 | 4 |
| 4 | Scr1 | all = 10 | 7.5% | 0.017% | 1.2% | 2.59 | 3474 | 2 |
| 8 | Scr2 | all = 8 | 9.0% | 0.012% | 3.8% | 2.39 | 3499 | 3 |
| 9 | Scr2 | all = 10 | 7.4% | 0.012% | 2.6% | 2.44 | 3449 | 1 |
| 13 | ARI low | all = 8 | 10.2% | 0.018% | 2.4% | 3.75 | 5390 | 9 |
| 14 | ARI low | all = 10 | 9.0% | 0.022% | 0.1% | 3.50 | 5418 | 10 |
| 16 | ARI med | all = 4 | 26.3% | 0.017% | 11.0% | 2.05 | 4556 | 7 |
| 17 | ARI med | all = 6 | 17.3% | 0.021% | 4.8% | 2.23 | 4617 | 8 |
| 22 | ARI high | all = 6 | 17.4% | 0.018% | 5.1% | 2.16 | 4217 | 5 |
| 23 | ARI high | all = 8 | 13.0% | 0.019% | 2.9% | 2.30 | 4248 | 6 |
| Mean | | | 12.7% | 0.017% | 3.7% | 2.6 | | |

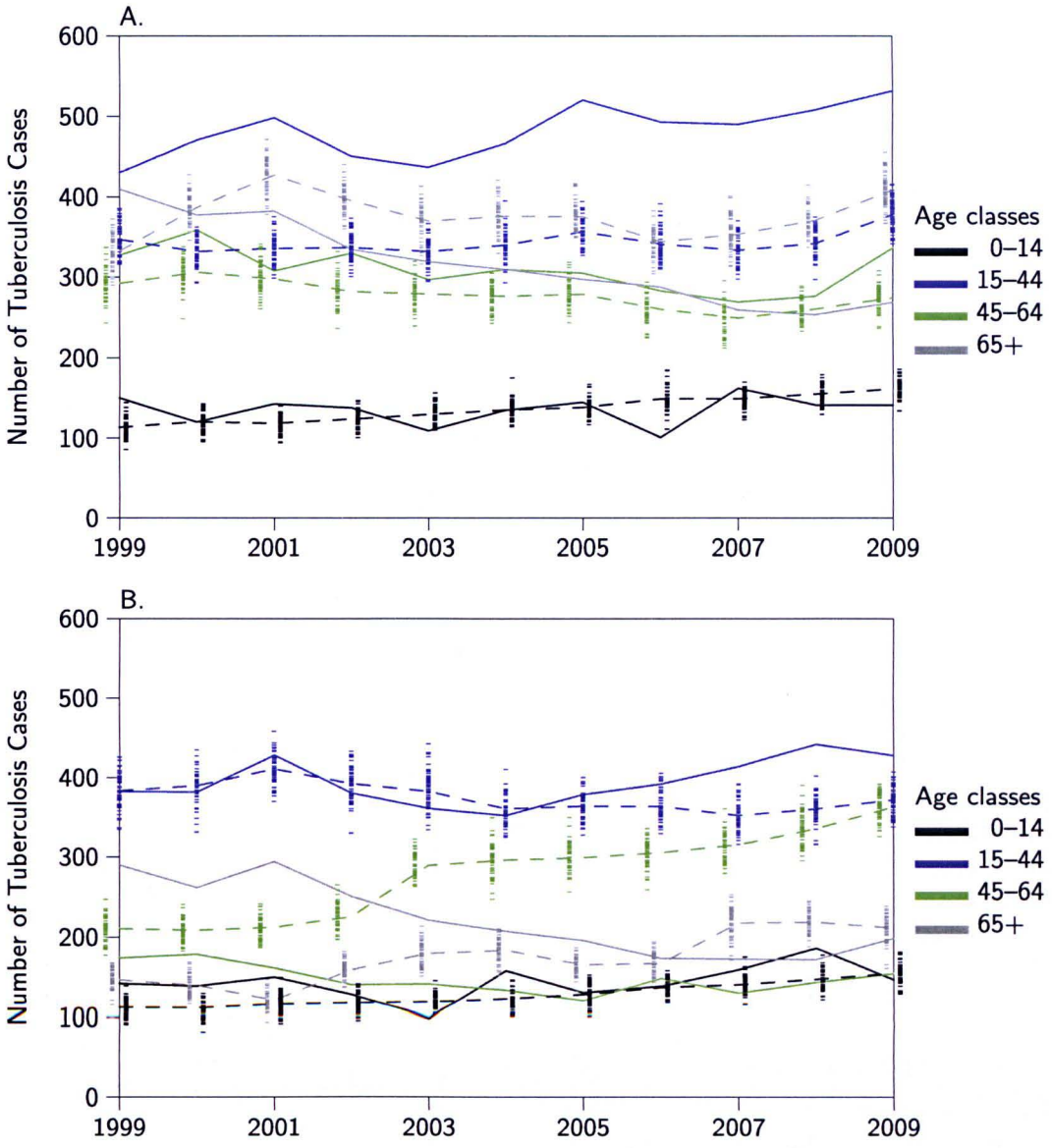


Figure 5-21: Model output and observed cases notified in England and Wales for UK-born males (A) and females (B) by age category for 1999 – 2009 for stage two fitting of a modified version of Scenario 9 for which a single foreign-born category was used. Sub-Saharan African-born and other foreign-born were combined during fitting. Model output follows the dashed line and individual runs of the model are denoted with a '-' (30 for each data point). Observed numbers of case notifications follow the solid line. Age categories are as follows: ages 0 – 14 years are in black; ages 15 – 44 years are in blue; ages 45 – 64 are in green; and ages 65 years and over are in grey.

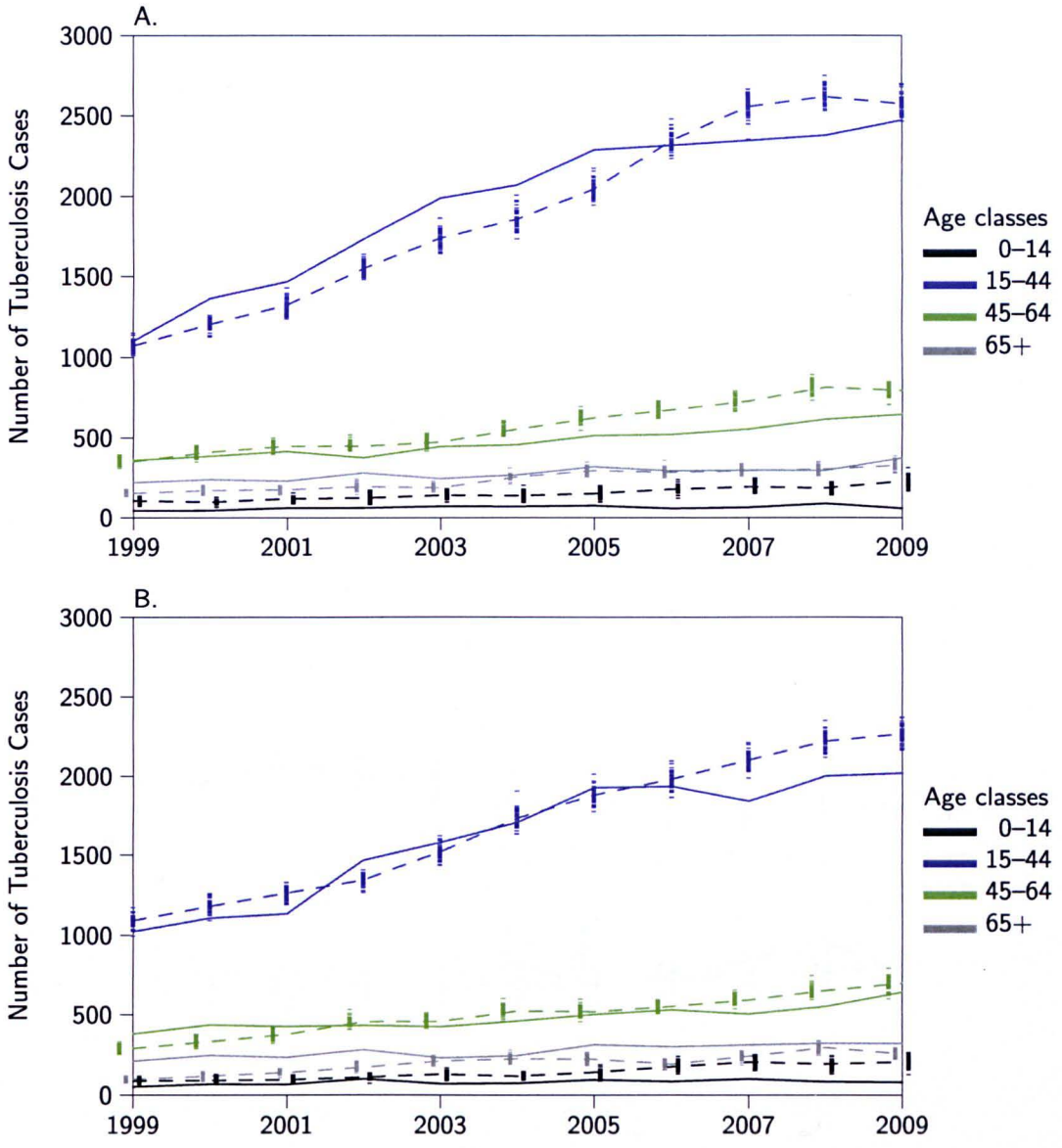


Figure 5-22: Model output and observed cases notified in England and Wales for foreign-born males (A) and females (B) by age category for 1999 – 2009, for stage two fitting of a modified version of Scenario 9 for which a single foreign-born category was used. Sub-Saharan African-born and other foreign-born were combined during fitting. Model output follows the dashed line and individual runs of the model are denoted with a '-' (30 for each data point). Observed numbers of case notifications follow the solid line. Age categories are as follows: ages 0 – 14 years are in black; ages 15 – 44 years are in blue; ages 45 – 64 are in green; and ages 65 years and over are in grey.

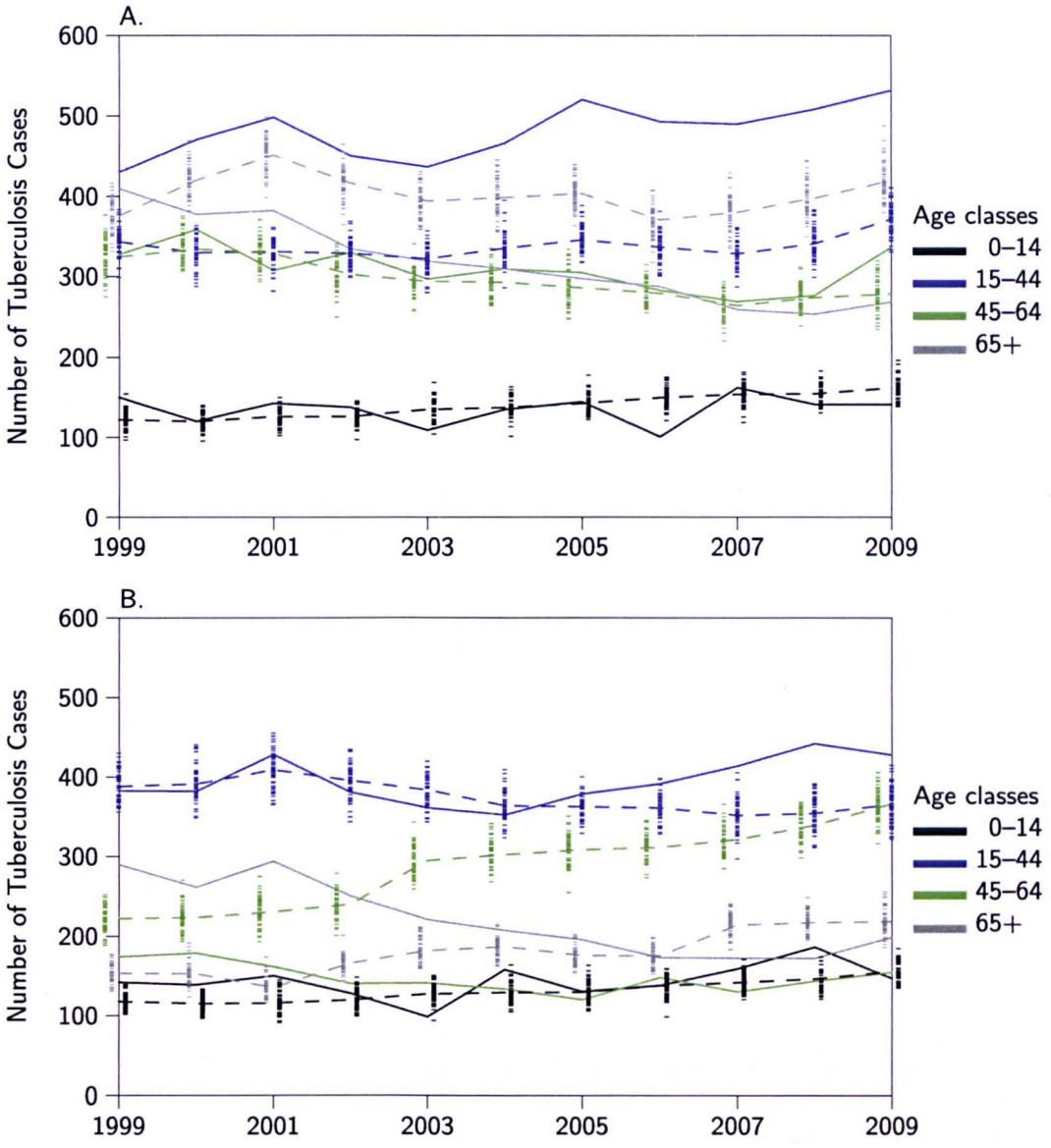


Figure 5-23: Model output and observed cases notified in England and Wales for UK-born males (A) and females (B) by age category for 1999 – 2009, for stage two fitting of a modified version of Scenario 4 for which a single foreign-born category was used. Sub-Saharan African-born and other foreign-born were combined during fitting. Model output follows the dashed line and individual runs of the model are denoted with a '-' (30 for each data point). Observed numbers of case notifications follow the solid line. Age categories are as follows: ages 0 – 14 years are in black; ages 15 – 44 years are in blue; ages 45 – 64 are in green; and ages 65 years and over are in grey.

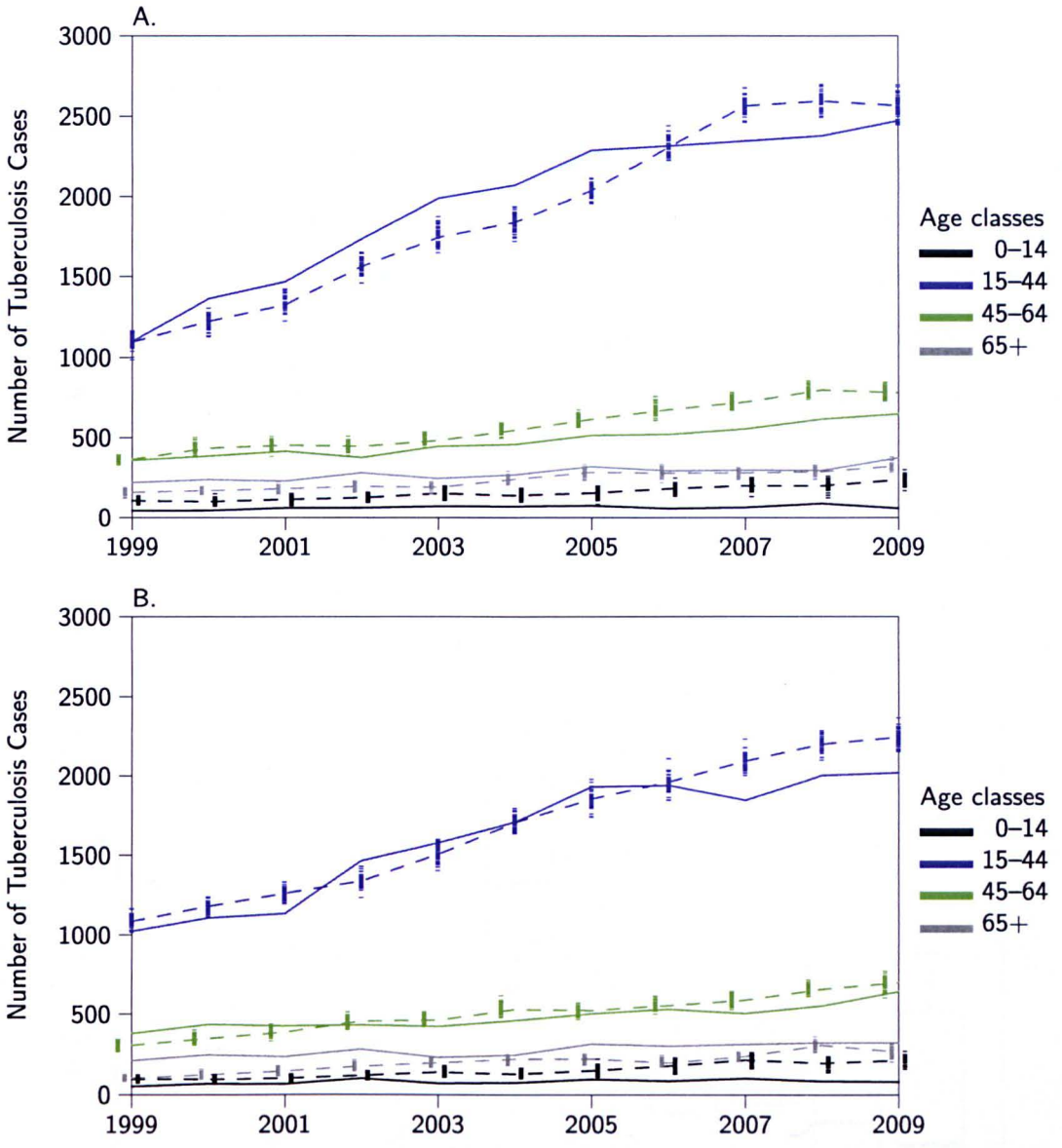


Figure 5-24: Model output and observed cases notified in England and Wales for foreign-born males (A) and females (B) by age category for 1999 – 2009, for stage two fitting of a modified version of Scenario 4, for which a single foreign-born category was used. Sub-Saharan African-born and other foreign-born were combined during fitting. Model output follows the dashed line and individual runs of the model are denoted with a '-' (30 for each data point). Observed numbers of case notifications follow the solid line. Age categories are as follows: ages 0 – 14 years are in black; ages 15 – 44 years are in blue; ages 45 – 64 are in green; and ages 65 years and over are in grey.

#### 5.2.2.1.2 Replicates

ANOVA performed on the 10 scenarios used for this variation of stage two fitting showed that there was significant variation in GOF statistics across the scenarios (p-value 0.01). Only when analysis was restricted to the four scenarios with lowest mean GOF statistic—scenarios 3, 4, 8, and 9—did the ANOVA show that GOF statistics do not significantly vary with fitting scenario (p-value 0.07). Note that scenarios 3, 4, 8, and 9 also have the absolute lowest GOF statistics when comparing only the lowest of the five replicates for each scenario.

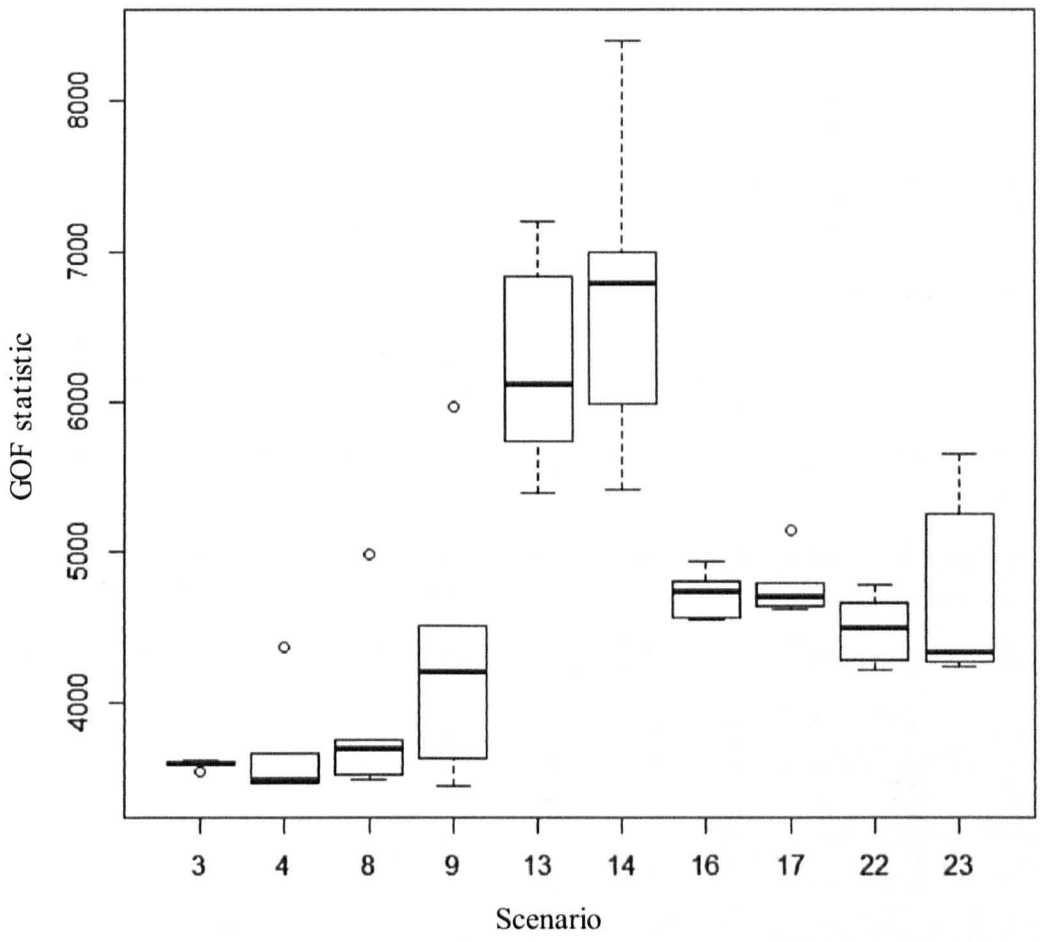


Figure 5-25: Box plot of the goodness of fit (GOF) statistics resulting from the 10 fitting scenarios with five replicates each for stage two fits. Analysis of variance (ANOVA) showed GOF statistics differed significantly among the 10 scenarios (p-value 0.0102). ANOVA performed on scenarios 3, 4, 8, and 9 showed that GOF statistics were not significantly different among those four scenarios (p-value 0.0662).

### *5.2.2.2 Disease risks*

Best-fitting disease risk estimates for the 10 scenarios are shown in Table 5-13. Since GOF statistics from scenarios 3, 4, 8, and 9 were the lowest and there were no significant differences between them, only these are presented. Results from these four scenarios gave a mean value for the risk of *Primary Disease* as 8.5% and 2.7% for *Reinfection Disease.* The annual rate of progression to *Reactivation Disease* from *Latent Infection* was estimated to be 0.013% on average. As shown in Table 5-15, the risk reduction for *Reinfection Disease* compared to *Primary Disease* averaged 75% among these four scenarios and 73% among all 10 scenarios, figures not shown.

Table 5-14: Estimated disease risks and mean values for the four best-fitting scenarios.

| | | | Disease risks for UK-born adult males by disease type | | | |
|---|---|---|---|---|---|---|
| Scenario | Infection status of migrants | Contact rate | Primary (%) | Reactivation (% per year) | Reinfection (%) | Foreign:UK-born |
| 9 | Scr2 | all = 10 | 7.4% | 0.012% | 2.6% | 2.44 |
| 4 | Scr1 | all = 10 | 7.5% | 0.017% | 1.2% | 2.59 |
| 8 | Scr2 | all = 8 | 9.0% | 0.012% | 3.8% | 2.39 |
| 3 | Scr1 | all = 8 | 10.0% | 0.012% | 3.4% | 2.33 |
| | | Mean | 8.5% | 0.013% | 2.7% | 2.44 |

Table 5-15: Comparison of disease risk estimates for *Primary Disease* and *Reinfection Disease* for the four best-fitting input parameter scenarios from stage two. Ratios between the risk of Reinfection and Primary Disease are reported, along with the corresponding reduction in risk for Reinfection Disease.

| Scenario | Primary (%) | Reinfection (%) | Risk ratio | Risk Reduction |
|---|---|---|---|---|
| 9 | 7.4% | 2.6% | 0.35 | 0.65 |
| 4 | 7.5% | 1.2% | 0.15 | 0.85 |
| 8 | 9.0% | 3.8% | 0.42 | 0.58 |
| 3 | 10.0% | 3.4% | 0.34 | 0.66 |
| Mean | | | 0.25 | 0.75 |

### *5.2.2.3 Proportion of cases due to recent transmission in the UK*

A summary of the proportion of cases due to recent transmission in the UK estimated by the model is shown in Table 5-16. As found in stage one fits, age-dependent trends differed between UK-born and foreign-born. UK-born cases aged 0 – 14 years had the lowest proportion of disease due to recent transmission in the UK, 95% on average. This proportion decreased with increasing age category, with an estimated 42% of cases due to recent transmission in the UK for those aged 65 years and above. For foreign-born individuals, the trend was reversed. Those aged from 0 – 44 years had a lower proportion of disease due to recent transmission, around 49%, than those aged 45 years and older, around 74%. Only considering the four best-fitting scenarios, and averaging over all age and sex categories for 1999 – 2009 resulted in the estimate that about 43% of foreign-born and 55% of UK-born cases, or 46% of cases overall, were due to recent transmission in the UK. There was little variation among the four scenarios, as estimates ranged from 52 – 57% for UK-born cases, 42 – 44% for foreign-born cases, and 45 – 47% overall.

Estimates of the proportion of disease due to recent transmission in the UK for input scenario 9, the best-fitting of the 10 scenarios in which a single foreign-born category was fit, are shown in Figure 5-26 and Figure 5-27. Plots show age- and sex-specific trends in the estimated proportion of cases due to recent transmission in the UK over 1999 – 2009. Age-specific trends in UK-born cases differed from trends in foreign-born cases. Estimates were higher for UK-born, and clearly declined with increasing age. For foreign-born cases, the age-specific trends were less clear, though generally, those under age 45 had lower estimated proportions of disease due to recent transmission, while those aged 45 years and older had a higher proportion of disease due to recent transmission in the UK.

Table 5-16: Proportion of cases due to recent transmission in the UK, averaged over all demographic categories and input scenarios for stage two fits using a single foreign-born category, 1999 – 2009.

| Age | Birthplace | Mean | 0.25 | 0.75 |
|---|---|---|---|---|
| 0 – 14 years | Foreign | 0.47 | 0.28 | 0.63 |
| 15 – 44 years | Foreign | 0.50 | 0.36 | 0.53 |
| 45 – 64 years | Foreign | 0.76 | 0.65 | 0.86 |
| 65 years+ | Foreign | 0.72 | 0.62 | 0.83 |
| 0 – 14 years | UK | 0.95 | 0.93 | 0.98 |
| 15 – 44 years | UK | 0.69 | 0.58 | 0.82 |
| 45 – 64 years | UK | 0.63 | 0.49 | 0.78 |
| 65 years+ | UK | 0.42 | 0.29 | 0.53 |

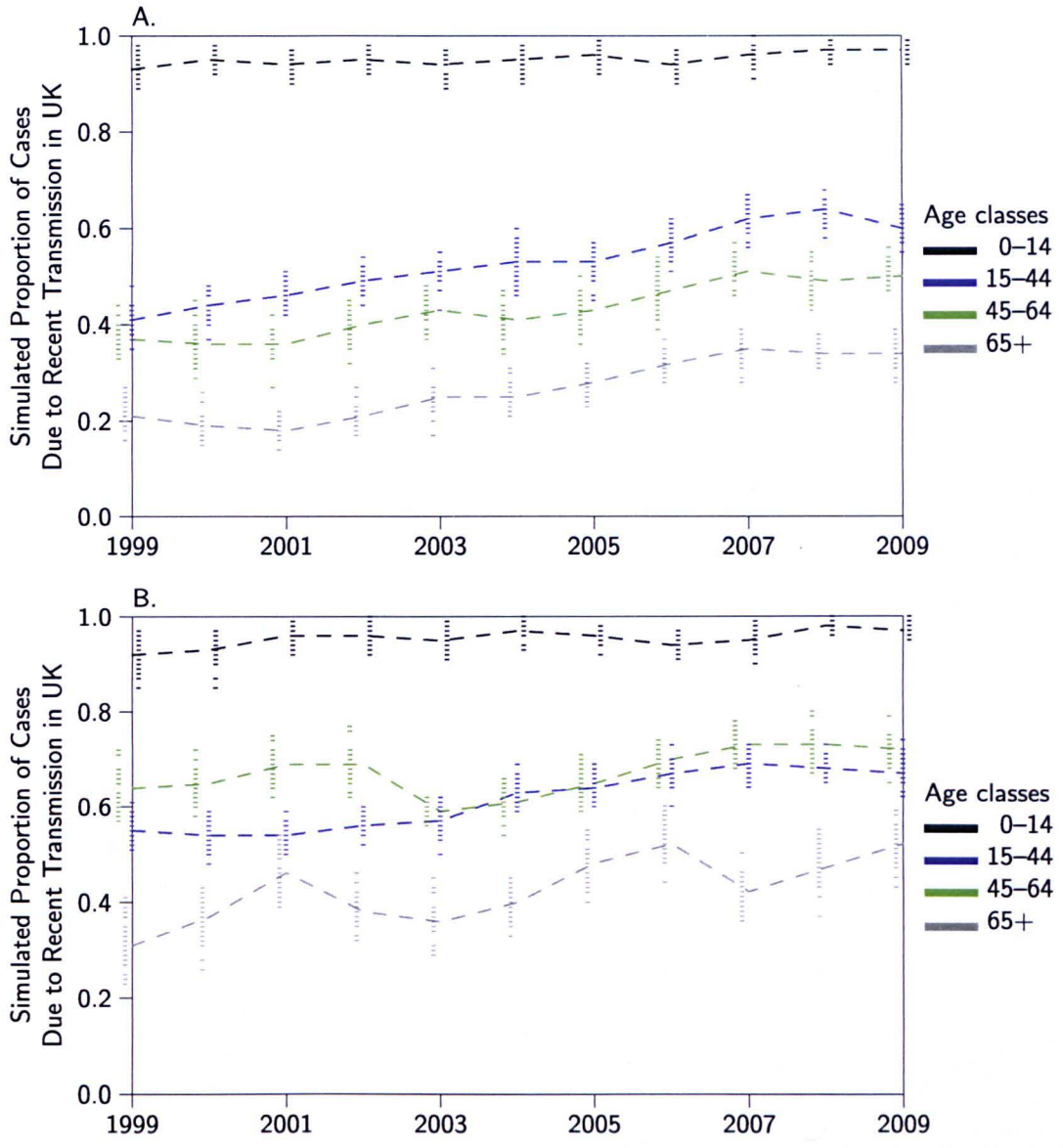


Figure 5-26: Estimates of the proportion of disease due to recent transmission in the UK for UK-born males (A) and females (B) by age category for scenario 9, the best-fitting of 10 stage two scenarios in which a single foreign-born category was used for fitting.

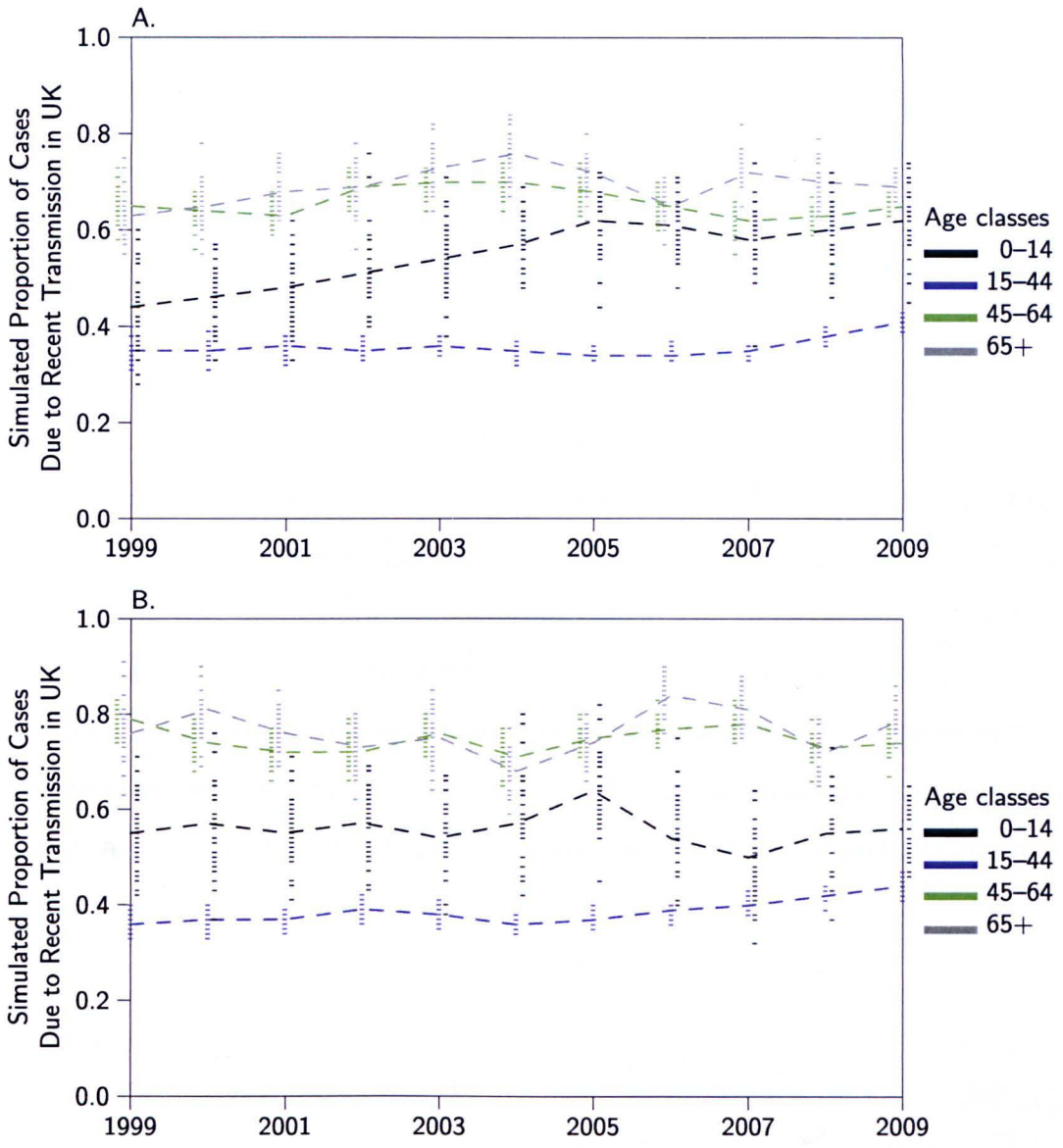


Figure 5-27: Estimates of the proportion of disease due to recent transmission in the UK for foreign-born males and females by age category by age category for scenario 9, the best-fitting of 10 stage two scenarios in which a single foreign-born category was used for fitting.

#### 5.2.2.3.1 Effect of contact rate and infection status of migrants

Again for this variation, results showed that several scenarios fit the data roughly equally well. Two-way ANOVA was performed for testing the effects of those parameters on GOF statistics and disease risks. As seen in the analyses on the stage one fits, only the infection status of migrants was significant (p-value $1.57 \times 10^{-11}$). Also, as seen in analyses on the stage one fits, both the contact rate and the infection status of migrants were significantly associated with primary disease risk estimates. For the contact rate, there was a negative coefficient for the linear regression, indicating that disease increased with decreasing contact rate. For *Reactivation Disease* risks, only the contact rate was significantly associated with estimates (p-value 0.002). The linear regression coefficient showed that there was a very small negative association between the variables, meaning increased contact rates resulted in decreased *Reactivation Disease* estimates. However, neither the contact rate nor assumptions about the infection status of migrants significantly affected *Reinfection Disease* risk estimates.

## 5.3 Discussion

Results showed the best fits to observed data were stage two fits for which a single foreign-born category was used in fitting. For this reason, these stage two results will be used to fulfil objectives and to inform subsequent modelling described in the thesis. Firstly, pertaining to objective two of this study, disease risk estimates were lower than earlier estimates for UK-born adult males, reflecting a decrease in disease risk following infection for UK-born individuals in recent years. Foreign-born disease risks were consistently higher than those of UK-born, on average 2.4 times higher than UK-born disease risks. The model fits suggested the contact rate was between eight and 10 effective contacts per year for infectious cases. Secondly, pertaining to objective three, estimates of the proportion of cases due to recent transmission in the UK were around 46% on average. Results are discussed further below, in separate sections discussing the quality of model fit to data, disease risk estimates, estimates of the proportion of cases due to recent transmission in the UK, and plausible assumptions about the contact rate and infection status of migrants. Also, limitations of the study are discussed. This section concludes with a discussion of the implications of results.

### 5.3.1 Summary and Interpretation of Findings

#### *5.3.1.1 Quality of model fit to data*

In stage one fits, model output followed major trends in the observed notification data, and reproduced observed data for some groups well, but fell short of accurately reproducing observed data for other demographic groups. Poor fits for some population groups, most notably for SSA-born individuals, suggested problems in one or more elements of the model, input data, or fitting process. Discrepancies between the model and observed data have many possible causes, as discussed below in Section 5.3.5.1.

In stage two fits under the variation using a single foreign-born category, the model output reproduced observed data much more satisfactorily. There are several potential reasons why fitting to the combined foreign-born group resulted in better fits to observed data. Combining SSA-born and OF-born cases may have improved fits since the model does not completely differentiate between these two groups in the

first place. It appears that the assumption in the model that OF-born individuals and HIV-negative SSA-born individuals have the same disease risks caused some constraint in model fitting. The numbers of SSA-born case notifications were generally underestimated in the model while the number of OF-born notifications fit to observed data more accurately. The OF-born group had many more cases than SSA-born and had more influence on disease risk parameters. Combining groups also meant fitting to fewer data points and larger population sizes for each data point. The increased population sizes may have improved fits but also resulted in more consistent results for best-fitting disease risks and for scenarios of the contact rate and the infection status of migrants, which fit observed data best. For these reasons, the parameter estimates and scenarios for subsequent modelling will use only the results from the stage two fits using a single foreign-born category.

### 5.3.2 Estimates of Disease Risk

Estimates for the risk of *Primary Disease* and *Reinfection Disease* for UK-born adult males were around 8.5% and 2.7%, respectively, among the four best-fitting of the 10 scenarios using a single foreign-born group. These were lower than estimates from England and Wales over the 20$^{th}$ century [19], which were 14% and 8% for *Primary Disease* and *Reinfection Disease* respectively. The estimate for *Primary Disease* was also lower than a recent estimate of 15%, obtained using models of tuberculosis dynamics in the US based on similar principles [223]. Estimates obtained here are closer to earlier estimates for adult males from the Netherlands, which were 5% and 2% for cumulative risks of *Primary Disease* and *Reinfection Disease*. The risk reduction for *Reinfection Disease* compared with *Primary Disease* estimated in this study was around 75%, similar to the 80% estimated in a recent literature review of evidence from cohort studies [306], but higher than estimates 62% [18], 40% and less than 35% [307] provided by previous modelling studies. Estimates for the annual rate of *Reactivation Disease* averaged approximately 0.01% among the four best-fitting of the 10 scenarios, which was lower than previous estimates for UK-born males of around 0.03% [19] and almost identical to estimates from recent models of tuberculosis dynamics in the US of around 0.01% [223].

However, for foreign-born individuals, taking into account the risk ratio estimated by the model for foreign-born versus UK-born risk meant disease risks for foreign-born individuals were consistently higher than risks for UK-born individuals. The risk ratio between foreign-born and UK-born disease risk was around 2.6 on average for the 10 scenarios, or 2.4 in the best four of these 10 scenarios. This ratio makes foreign-born disease risk estimates very similar to the risks estimated for UK-born males in England and Wales over the last century [19].

### 5.3.3 Proportion of Cases Due to Recent Transmission in the UK

The overall estimated proportion of cases due to recent transmission in the UK was high, around 46% on average for the four best-fitting scenarios. There was little variation among scenarios as estimates ranged from 45 – 47%. Estimates were higher than the recent estimates of the proportion of disease due to recent transmission in the Netherlands, estimated to be 40% for urban cases and 27% for rural cases using an analysis stratifying cases using characteristics including results from strain typing of isolates. The estimates obtained in this study were also higher, particularly for older age groups, compared with a previous modelling study using data from the Netherlands [9]. Estimates of the proportion of disease due to recent transmission have also been obtained by molecular epidemiological studies, which attempted to differentiate between disease due to recent infection and due to older or imported infections. The estimates obtained here were higher than estimates from the two UK studies estimating the proportion of cases due to recent transmission using RFLP data [166, 308]. However, these studies were limited by interpretation of the genotyping data and the still unclear relationship between genotype clustering and recent transmission, as discussed in Chapter 2, Section 2.4.3 and further in Chapter 7.

Note that male-female differences in the proportion of disease due to recent transmission can be largely explained by relative risks between adult males and females. In the model it was assumed that females are less likely to develop *Reactivation Disease* than males, very much less likely to develop *Reinfection Disease*, and equally likely to develop *Primary Disease*. Therefore, best-fitting disease risks will alter differences in the proportion of cases due to recent transmission in the UK. For

example, increased primary disease risk will mean a larger proportion of disease due to recent transmission for females.

Results among birthplace groups showed that UK-born individuals had a higher proportion of disease due to recent transmission than foreign-born individuals, as expected. Age-dependent trends in the proportion of disease due to recent transmission were also different between UK-born and foreign-born individuals. UK-born children have the highest proportion of disease due to recent transmission—almost all cases—whereas older UK-born individuals have a decreased proportion of disease due to recent transmission because they are more likely to have been infected long ago. Still, these proportions were high, even in older individuals.

The patterns for foreign-born individuals generally showed that those aged 15 – 44 years had the lowest estimates of the proportion of disease due to recent transmission. This finding seems reasonable because they are the group most likely to be the newest immigrants and are more likely to have an infection acquired abroad. Disease caused by recent infections acquired abroad did not contribute to the proportion of disease due to recent transmission in the UK. Foreign-born children also had a low proportion of cases due to recent transmission in the UK for similar reasons. Based on their ages, they would have been recent migrants, and therefore probably acquired their infection abroad. On the other hand, the elderly foreign-born have likely been in the UK for a long time and are likely to have an older infection, which carries relatively low risks for progression to disease. Since recent infections have a much higher risk of progression to disease, many of these cases are due to recent transmission, even though most *infections* in this group are probably older or imported. The estimates here may have been high due to a higher contact rate than is appropriate. Model estimates of the proportion of cases due to recent transmission in the UK were sensitive to the input scenarios, in particular to the contact rate and to disease risks. As would be expected, higher contact rates and higher risks of *Primary Disease* resulted in higher estimates for the proportion of disease due to recent transmission in the UK. The contact rate was explored further in Chapter 7, where genotyping data were used to help understand the relationship between genotype clustering and recent transmission.

### 5.3.4 Effect of Contact Rate and Infection Status of Migrants

The most appropriate contact rates and assumptions for the infection status of migrants are assumed to be those from the best fits of the model to observed data. Among the 10 scenarios tested for the stage two variation fitting a single foreign-born category, four were distinguished as better than the rest, when considering both the mean GOF statistics across five replicates—they were the four lowest—and also when considering only the best-fitting among the five replicates for each scenario—also the four lowest using this criterion. Therefore, these scenarios will form a starting point for contact rate values and distributions for infection status of migrants used in subsequent modelling in this thesis.

### 5.3.5 Limitations

#### *5.3.5.1 Discrepancies between model outputs and observed data for stage one fits*

One limitation of the work is that there were discrepancies between model output and observed data for stage one fitting runs designed *a priori*. There are several potential causes of these discrepancies, each of which should be considered for better understanding of the limitations of the model. For one, discrepancies may result because the fitting routine, including choice of variable parameters, may itself be limited. The fitting scheme was developed under constraints of lack of data and limited computer resources. This consideration, and the limited scope of the project generally, led to some simplifications that may have contributed to problems obtaining good fits of the model to observed data. It is possible that the model parameters left to vary during fitting were too few to achieve a good fit of model output to observed data. The ratio of disease risk between UK-born and foreign-born, *df*, and the ratios of disease risk between males and females, *sd*, were particularly restrictive and likely hindered better fitting of the model output to observed data across demographic groups. It may have been better to fit more variable disease risk parameters, even though some estimates would be based on little data.

Fit discrepancies could also be due to input parameters and input scenarios. Although these represent a range of plausible assumptions about the contact rate parameter and assumptions about the infection status of migrants, it is possible that the contact rate parameters do not cover the true best values even among 25 scenarios. With regard to other input data, it is possible that some other parameters were problematic. Due to insufficient or inaccurate observed data, some of the fixed parameters in the model could have been inaccurate and made it impossible to fit model output to observed data accurately. Major data limitations are discussed below in Section 5.3.5.2. Furthermore, the target data that the model is fit to could have been inaccurate. Limitations to these data are also discussed below in Section 5.3.5.2.

Other causes of the discrepancies between model output and observed data could be due to errors in the coding of the model itself, though steps were taken to avoid these types of errors, as discussed in Chapter 3, Section 3.4.

#### *5.3.5.2 Data and model assumptions*

Another potential cause of inadequate fits of the model in stage one of fitting were input parameter values and other assumptions in the model. Parameters were taken from literature sources or data whenever possible. However, some parameters had to be based on little or surmised data, including the infection status of the population in 1981, the infection status of migrants, and the contact rate. Other parameter values were uncertain for some demographic groups or uncertain for recent years. For example, the ratio of female to male disease risk was taken from an England and Wales study, but it is possible those values did not extend to the populations modelled here, since they were based on the white ethnic group in an earlier time period. Also, the age-specific proportion of cases that are smear-positive was taken from previous work, based on trial results which are now more than 50 years old. In addition, the vaccine efficacy assumptions for UK-born individuals were based on older trials. Other parameters were based on uncertain data including LFS estimates for population size estimates for small demographic categories, which were uncertain and affect case numbers and also notification rates. SSA-born migration and population sizes were the least certain, as was HIV prevalence among SSA-born migrants.

Finally, the assumption of lifelong immunity for those vaccinated may have been unrealistic, as well as the assumption that no migrants are vaccine-protected. This may have allowed an artificially high contact rate to fit the data, since such contact would commonly be wasted on vaccine-protected individuals. On the other hand, although the mechanism is wrong, it is likely that much of the population in the UK is effectively immune from M. tuberculosis infection due to isolation of the disease and its transmission to mainly high-risk groups of the population.

One possible shortcoming of the model is that it may have been stratified into more categories than supported in available data for parameters. Despite input data limitations, the model output was stratified into age, sex, and birthplace categories because these were each thought to be important factors in the natural history and epidemiology of tuberculosis. However, the model as it was parameterized may not have allowed necessary differences among categories to fit model output to observed data. For example, in many cases it was assumed that males and females had equal values for parameters, when in reality the case may be different. Specifically, estimates of infection and disease prevalence for the initial population in 1981 and infection and disease prevalence in immigrants (where applicable), the ARI was assumed equal for males and females. In addition, the contact rate was assumed equal for males and females, and both were equally susceptible to being the target of a transmission. In reality, there may actually have been differences between male and female behaviour, which could also interact with age, but the model and available input data did not specify so. This limitation is also true for birthplace.

#### *5.3.5.3 Other limitations*

Due to limited computer time, only five replicates of each fitting scenarios were run. The five replicates exhibited variance, though often the two or three replicates with the best GOF statistics had similar GOF statistics and parameter estimates. The patterns generally supported choosing the best fitting among the five replicates, though more replicates would have increased the likelihood that the fits and the resulting parameter estimates were the best possible for that scenario. Also due to limited computer hours, confidence bounds for the estimated parameters were not

calculated. These could have been estimated by running the model with randomly chosen parameter values many times (say thousands), accepting parameter estimates as part of 95% confidence interval if model output using those parameter estimates fell within some specified range, determined by the optimal deviance. Confidence limits were not generated mainly because the method requires thousands of runs of the model and computer hours were limited. Instead of using confidence limits, several best-fitting scenarios covering a range of estimates were tested for subsequent modelling in the thesis.

### 5.3.6 Implications of Findings

Estimates of the proportion of disease due to recent transmission in the UK over this time period may help assess *M. tuberculosis* infection control strategies. The estimated proportions of disease due to recent transmission show there is still motivation to work towards stopping the transmission of *M. tuberculosis* in the UK. Estimates for foreign-born proportions of disease due to recent transmission are lower than those for UK-born proportions, but many foreign-born cases are still likely to be due to transmission within the UK.

Disease risk estimates show there is a clear difference between risks in UK-born and foreign-born, with foreign-born risks 2.4 times higher on average. Higher disease risk for foreign-born individuals means that measures such as prophylactic treatment may be more cost-effective since more cases may be avoided for any number of treatments given. These newly estimated risks may help inform cost-effectiveness studies of interventions such as prophylactic treatment of latent infection.

Implications for subsequent modelling are multi-fold. Firstly, disease risk estimates span a wide range of values and it is problematic to identify exact values to be used. Several values should be used in subsequent modelling. Secondly, several combinations of the contact rate and assumptions about the infection status of migrants led to equally good fits of model to data. A range of these input scenarios should also be used in subsequent fitting. Thirdly, my results suggest that the model and current data do not have the ability to distinguish between separate groups of foreign-born individuals and that this population should be modelled as one group in

any further modelling if more data do not become available. As better screening data, immigration data, and HIV prevalence data become available, the model may provide better differentiation among immigrant groups.

# 6 Molecular Epidemiology of the West Midlands

This chapter describes the analysis of genotyping data from the West Midlands to fulfil objective four of the thesis. Here the molecular epidemiology of the West Midlands is studied from 2007 to 2011, using both 15-locus and 24-locus VNTR typing systems, to provide an analysis of the risk factors for clustering of isolates from tuberculosis cases and a population-based comparison between the two VNTR typing systems. A crude estimate of the proportion of cases due to recent transmission is also calculated from the VNTR typing data. Results provide population-level estimates of clustering and risk factors associated with clustering using 24-locus VNTR data and a background for the use of these data in the molecular epidemiological modelling described in Chapter 7.

## 6.1 Methods

### 6.1.1 Study Population, Data Collected, and Laboratory Methods

The study population included all tuberculosis cases from the West Midlands notified from January 2007 to December 2011. The region has a population size of approximately 5.6 million [309] and includes the cities of Birmingham, Coventry, and Wolverhampton (see Chapter 2, Section 2.3.2.2), all noted for relatively high rates of tuberculosis incidence in the UK [112]. In addition, the region as a whole has tuberculosis notification rates above the national average, at approximately 18.5 cases per 100,000 population in 2011 [105]. Mirroring national trends, notification rates in the region have increased since the late 1980s, with concurrent increases in the proportion of patients born abroad.

During this period, data on notified cases were maintained in the national ETS database (see Chapter 2, Section 2.3.2.3.1). For each case, ETS holds: demographic data including age, sex, world region of birth, ethnic group, and time from entry to the UK and tuberculosis diagnosis for foreign-born individuals; clinical details, including site of disease and year of notification; behavioural risk factors, including history of or current problem drug or alcohol use and history of or current homelessness or time spent in prison; and laboratory data, including culture positivity and drug sensitivity for cases in the study. The characteristics used as variables in the study are summarized in Table 6-1. In the ETS database, duplicate notifications and specimens from the same patient were collated if they occurred within 12 months of the initial notification or specimen date. Episodes of tuberculosis more than 12 months apart were captured as separate notifications.

Also during the study period, clinical specimens from suspected cases of tuberculosis in the West Midlands were routinely sent to the HPA Regional Centres for Mycobacteriology, Birmingham, for culturing, identification, strain typing, and drug susceptibility testing according to standard methods as described previously [310]. Briefly, specimens were first decontaminated, examined by microscopy, and then incubated in liquid culture. Positive cultures were identified with a DNA test which

detects *M. tuberculosis* complex DNA. Culture-positive isolates were tested for drug susceptibility to isoniazid, rifampicin, pyrazinamide, and ethambutol.

Table 6-1: Description of demographic and other characteristics of tuberculosis cases used for variables in analyses in the study.

| Variable name | Description | Categories |
|---|---|---|
| *sex* | Sex | Male, female |
| *age group* | Age group | 0 – 14, 15 – 44, 45 – 64, 65+ (years) |
| *birthplace* | World region of birth | UK, Europe, East Mediterranean, Africa, Americas, South Asia, East/Southeast Asia |
| *ethnicity* | Ethnic group | White, Black-Caribbean, Black-African, Black-Other, South Asian, Chinese, Mixed/Other |
| *time since entry* | Time from entry to UK to TB diagnosis | 0 – 1, 2 – 4, 5 – 9, 10+ (years) |
| *disease site* | Site of body affected by disease | Pulmonary (including pulmonary and extra-pulmonary), extra-pulmonary only |
| *drug sensitivity* | Sensitivity to the four first line antibiotics, isoniazid, rifampicin, pyrazinamide, and ethambutol | Sensitive, resistant to one or more drugs |
| *previous diagnosis* | History of TB diagnosis | Yes, no |
| *drug use* | History of or current problem drug use | Yes, no |
| *alcohol use* | History of or current problem alcohol use | Yes, no |
| *homelessness* | History of or current homelessness | Yes, no |
| *prison time* | History of or currently imprisoned | Yes, no |

From 2007 – 2009, isolates from culture-positive samples were routinely typed with a set of 15 VNTR loci, including the 12 original MIRU loci and ETR loci 'A', 'B', and 'C' [150]. From 2010, isolates from culture-positive samples were routinely typed with the standard set of 24 VNTR loci, which include the 15 loci used previously, plus an additional nine loci [152] (See also Chapter 2, Section 2.4.1.3). VNTR typing was

performed according to standard methods, as described previously [116, 140, 141, 310, 311].

To extend the dataset of 24-locus profiles, further typing with the additional nine loci was undertaken for isolates from 2007 – 2009 with 15-locus profiles. To conserve limited laboratory resources and time, only those 2007 – 2009 isolates that matched one or more other isolates in the study population, or were 'clustered' according to a cluster analysis on the 15 loci, were typed with an additional nine loci. These isolates were identified by performing a preliminary cluster analysis on the 15 loci for all isolates from 2007 – 2011 (details of this preliminary cluster analysis are described below, in Section 6.1.2.3).

In the laboratory, steps were taken to exclude isolates suspected of laboratory contamination (J. Evans, personal communication), however, after I obtained laboratory data, no cases were excluded from analysis due to suspected laboratory cross-contamination. This is because the dates specimens were processed in the laboratory were unknown.

Strain types and other laboratory data were linked to cases notified to ETS through a matching process based on patient names, birthdates and addresses, undertaken by the HPA Tuberculosis section, as described elsewhere [113]. The matching process allowed laboratory strain types from culture-confirmed isolates to be associated with demographic and other case data in ETS. Cases without a match to laboratory data included culture-negative cases, cases with no specimens sent to the laboratory, and cases for which the matching process failed.

### 6.1.2 Data Analysis

#### *6.1.2.1 Descriptive analysis*

The proportion of notified cases with a positive culture was calculated, as well as the proportion of notified cases that were eligible for VNTR cluster analysis. For the 15-locus analysis, eligible cases include those with 15-locus strain type profiles and no more than one of the 15 loci missing. For the 24-locus analysis, these include cases with 24-locus profiles missing no more than two loci, and cases with 15 locus-profiles,

which were unique in a dataset-wide cluster analysis on the 15 loci. For eligible cases and for all cases, the numbers and proportion of cases by demographic and other characteristics outlined in Table 6-1 were tabulated. Only cases that fulfilled criteria for inclusion in both the 15- and 24-locus analyses were used for subsequent analyses.

#### *6.1.2.2 Clustering definitions*

In this study, two definitions of clustering were used along with one index for estimating the proportion of cases due to recent transmission. First, according to the 'n method' [163], cases with strain type profiles that exactly matched one or more other profiles in the study population were considered 'clustered'. Cases with strain type profiles that did not have an exact match in the study population were considered 'unique'. A group of two or more cases with identical strain types in the study population formed a 'cluster'.

Second, the 'retrospective method' attempts to distinguish between source cases and recipients of infection. Under this definition, cases with strain type profiles that exactly match one or more other profiles from case(s) in the study population, which were notified previously, within some defined time period, were clustered. A case with a strain type profile that did not have a match in the study population from a case notified previously, within the defined time period, was considered unique. In the present study, the time period used was two years. Other studies have looked four years and one year prior to define retrospective clustering [164-166]. Here, the proportion of cases clustered by study duration justified the choice of a two-year definition, as shown in Appendix 10.17, because the proportion of cases clustered did not increase much after two years of genotyping data. The retrospective method eliminates some bias in cluster analysis that occurs for other definitions of clustering because cases notified at different times have different follow-up periods for assessing clustering. Using the retrospective method, the same time period for each isolate is used to look retrospectively for an isolate's match. However, this method results in a loss of data, with increasing losses for longer retrospective time periods considered. As defined here, the first two years of notified cases were excluded, as there were not enough preceding data to evaluate those cases as clustered or unique.

Lastly, clustering according to the 'n-1 method' [162] was calculated to provide a simple estimate of the proportion of cases due to recent transmission. The n-1 method considers any isolate for which there is another isolate with an identical strain type profile from a case notified previously as clustered. This means that within a cluster of cases with identical strain type profiles, all but the first of them to be notified are considered clustered, with the implicit assumption that the first case notified is the source of transmission.

#### *6.1.2.3 Molecular epidemiology*

Cluster analyses on strain type profiles were performed for both typing systems, 15-locus 24-locus VNTR, including a preliminary cluster analysis on the 15 loci. Methods were identical for each cluster analysis. VNTR profiles were imported into BioNumerics version 6.2 (Applied Maths, Kortrijik, Belgium) as categorical data 'experiments', separately for 15-and 24-locus VNTR. For each typing system, a cluster analysis was performed using the Unweighted Pair Group Method with Arithmetic Mean, Pearson's correlation, and 100% profile similarity cut-off, though alternate options in BioNumerics produced identical results since VNTR data are categorical and 100% profile similarity was used to define clusters. Missing loci were ignored, which meant profiles with missing loci were allowed to cluster with complete profiles.

For each cluster analysis, an identification number was assigned to each distinct strain type profile, such that identical profiles share the same identification number using the 'fill field with cluster number' script available with the BioNumerics software. The resulting dataset was analysed using Stata version 12.1 (Stata Corporation, College Station, TX). The proportion of isolates clustered, using both 15- and 24-locus VNTR for both the n method and the two-year retrospective method, was calculated by the demographic and other characteristics of cases found in Table 6-1. In addition, cluster size distributions were compared for 15- and 24-locus typing systems and an estimate of the proportion of cases due to recent transmission was calculated using the n-1 method for both typing systems.

A univariate analysis of factors associated with clustering was performed for both typing systems and definitions of clustering for all characteristics found in Table 6-1.

Maximum likelihood estimates of odds ratios (OR) with Wald tests with 95% confidence limits are reported. Significance was evaluated using p-values derived from the likelihood ratio chi-square test (LRT), with $p<0.05$ considered significant.

Multivariate logistic regression models were also constructed for each typing system and definition of clustering, with a subset of factors explored in the univariate tests. It was decided *a priori* to include the *age group* in the multivariate models since there was an expectation this age would impact cluster membership [166, 169, 312]. Multivariate models included other variables significantly associated with clustering in the univariate analysis (p-values < 0.05), less some exceptions. It was decided *a priori* that only one of the *region of birth* and *ethnicity* variables would be included in the model, with *region of birth* preferred if both significant. *Time since entry*, only applicable to foreign-born cases, was not eligible for inclusion in these multivariate models since that would reduce the models to foreign-born only (a separate model for foreign-born is described below). In addition, since the behavioural risk factors, *drug use, alcohol use, homelessness*, and *prison time,* were only collected for a subset of cases, these were left out of these multivariate models. The excluded factors were explored in additional multivariate models, described below. For factors included in the multivariate models, adjusted ORs and their 95% confidence limits were reported, with significance evaluated using p-values from the LRT.

For foreign-born cases, separate multivariate models were constructed to allow incorporation of the variable *time since entry*. These models were constructed using variables significantly associated with clustering in a univariate analysis of risk factors for clustering in foreign-born cases. Models were constructed for each definition of clustering, although for simplicity, only the 24-locus typing system was used. As in the initial models, behavioural risk factors were excluded from the foreign-born models because of the large amount of missing data for these variables.

Additional supplementary multivariate logistic regression models were constructed to explore behavioural risk factors significantly associated with clustering on the univariate tests. Again, models were constructed for each definition of clustering, though only using the 24-locus typing system. For each model, demographic and other

factors significantly associated with clustering in the initial multivariate model described above were included at the start. Behavioural factors were then added one at a time, using forward selection. Inclusions were accepted if the model was significantly improved according to the LRT test at the p=0.05 significance level.

## 6.2 Results

### 6.2.1 Study Population and Descriptive Analysis

From 2007 – 2011, there were 4,845 tuberculosis cases notified in the West Midlands. Of those, 2,749 (56.7%) had a positive culture, and isolates from 2,543 of them (92.5%) were typed with at least 15 loci. Of those typed, 2,423 (95.3%) had at most one missing locus and were eligible for the preliminary cluster analysis on the 15 loci.

The preliminary cluster analysis was performed on all the 2,423 eligible 15-locus profiles to determine which 15-locus profiles were clustered, and therefore should have isolates typed with the additional nine loci. The additional typing was carried out on as many of these as possible, including more than 1,000 isolates from 2007 – 2009 that formerly had only 15-locus profiles. Nevertheless, 108 isolates were clustered in the 15-locus analysis but not typed with the additional nine loci and were thus excluded from further analyses. The preliminary cluster analysis also revealed 319 isolates that were unique on the 15-locus cluster analysis and eligible for inclusion in the 24-locus analysis.

After the additional typing, there were 2,080 cases with 24-locus profiles, of which 88 cases were excluded due to having profiles missing more than two loci, leaving 1,992 eligible cases. With the additional 319 cases with unique 15-locus profiles, there were a total of 2,311 cases eligible for analysis using 24-locus profiles.

The intersection of the 2,311 cases eligible for 24-locus analysis and the 2,423 cases eligible for 15-locus analysis left 2,283 cases eligible for both analyses. These represent 83.1% of culture-positive isolates, or 47.1% of all notified cases in the study period, as depicted in the diagram in Figure 6-1.

The demographic and other characteristics of the 2,283 cases are summarized in Table 6-2. Demographic variables were missing for only a small minority of cases, though for some other variables, such as *previous diagnosis* and the four behavioural risk factors, there were a substantial portion of cases with missing data; see Appendix 6-B for details of the proportion missing by variable. The demographic and other characteristics for all 4,845 notified cases are found in Appendix 6-C.

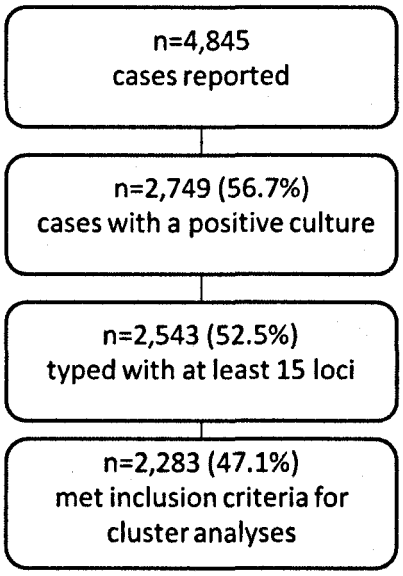


Figure 6-1: Diagram showing the proportion of notified cases included in cluster analyses. From 2007 – 2011, there were 4,845 cases reported in the West Midlands. Of those, 2,749 had a positive culture while the remainder had a negative culture result or were missing specimens or laboratory data. Of those with a positive culture, 2,543 were typed with at least 15 loci and 2,283 were eligible for inclusion in both 15- and 24-locus cluster analyses.

### 6.2.2 Molecular Epidemiology

Over all of the 2,283 cases analysed, 46.5% were clustered in the 24-locus analysis and 68.9% in the 15-locus analysis, using the n method. There were a total of 252 clusters in the 24-locus analysis, 135 of which included only two cases. There were a further 85 clusters with 3 – 5 cases, 18 clusters with 6 – 9 cases and 13 clusters with 11 – 49 cases. Only one 24-locus cluster had 50 or more cases. Under the 15-locus analysis, there were a total of 255 clusters, 97 of which included only two cases. There were 96 clusters with 3 – 5 cases, 33 clusters with 6 – 9 cases and 27 clusters with 10 – 49 cases using 15-locus profiles. Only two 15-locus clusters had 50 or more cases. The cluster

size distributions for the 15- and 24-locus typing systems are compared visually in Figure 6-2.

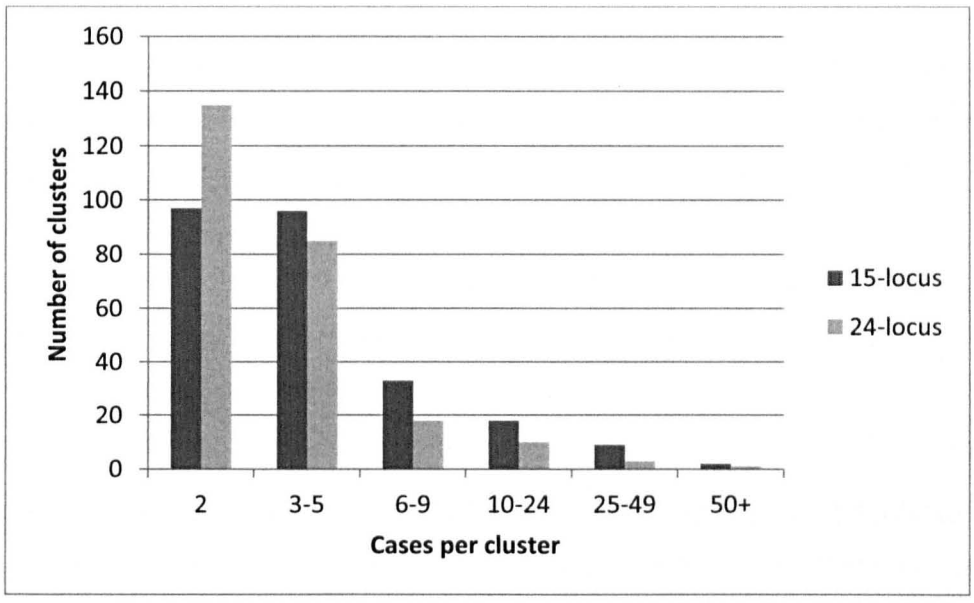


Figure 6-2: Cluster size distributions for 15- and 24-locus VNTR typing systems. Clusters were defined according to the n method and included all cases notified in the West Midlands from 2007 – 2011 that were typed with VNTR and met inclusion criteria for the study.

The proportion of cases clustered by demographic and other characteristics is shown in Table 6-2 for the analysis based on 24-locus profiles, using both the n method and the two-year retrospective method to define clustering. Table 6-1 also shows the ORs obtained from univariate analysis of variables associated with clustering for each definition of clustering with p-values obtained using the LRT.

Factors significantly associated with clustering using the n method included a*ge group, sex, birthplace, ethnicity, time since entry* (only applicable to foreign-born cases), *disease site, drug sensitivity, drug use, alcohol use,* and *prison time*. The risk of clustering increased with: a younger age group; male sex; UK birthplace; white ethnicity; pulmonary disease; infection with strains sensitive to antibiotic drugs; alcohol use; drug use; and present or past history of imprisonment.

Using the retrospective method, trends were consistent with those observed using the n method, although the proportions clustered were lower. The risk of clustering

increased with: younger age group; UK birthplace; white ethnicity; pulmonary disease; strains sensitive to antibiotic drugs; alcohol use; drug use; and prison time, as in the analysis using the n method. Sex was not significantly associated with clustering under the retrospective method analysis.

The same trends were observed using the 15-locus typing system for both the n method and retrospective method. The only differences were that neither *age group* nor *sex* was significantly associated with clustering for either definition of clustering (see results for 15-locus analyses in Appendix 6-D).

Multivariate logistic regression models for each typing system and the definition of clustering were used to test whether univariate associations with clustering held after adjustment for other variables. *Age group* was included in the models *a priori*. For the 24-locus typing system and both definitions of clustering, *birthplace*, *disease site*, and *drug sensitivity* were included in the model because they were significantly associated with clustering in the univariate analyses. *Sex* was also included in the n method model because it was significant in the univariate analysis. For both models, all variables remained significantly associated with clustering, with adjusted odds ratios reported in Table 6-2.

Results were similar under the 15-locus multivariate models. For both the n method and retrospective method, variables *birthplace*, *disease site*, and *drug sensitivity* remained significantly associated with clustering, with adjusted odds ratios reported in Appendix 6-D. *Age group*, which was included in both models but not significant in the univariate analysis, was not significantly associated with clustering (see Appendix 6-D).

Note that while *ethnicity* was also significantly associated with clustering for both typing systems and definitions of clustering, it was excluded from multivariate models *a priori* due for correlation with *region of birth*. In addition, although the variable *time since entry* (applicable to foreign-born cases) and several behavioural characteristics of cases were significantly associated with clustering for both typing systems and definitions of clustering, they were excluded and explored in two supplementary multivariate models.

The univariate analysis of factors associated with clustering for foreign-born using the 24-locus typing system and n method showed that longer *time since entry*, pulmonary *disease site*, drug sensitive strains, *drug use*, and *prison time* were significantly associated with clustering. OR and associated p-values are found in Table 6-3. A multivariate model was constructed with these factors, less the behavioural risk factors *drug use* and *prison time*, which were omitted as before due to missing data. Only the variables *sex*, *time since entry*, and *disease site* remained significant in the multivariate model. Adjusted ORs and associated p-values are also shown in Table 6-3. Using the 24-locus system and the retrospective method of clustering, only *drug use* and *alcohol use* were significantly associated with clustering in the univariate analysis for foreign-born cases (results not shown); a multivariate model was therefore not constructed using the retrospective method.

For the multivariate logistic regression model incorporating behavioural risk factors, only the n method and 24-locus typing system were used. The model included base variables *age group*, *birthplace*, *disease site*, and *drug sensitivity*, all significantly associated with clustering in the main multivariate model. Next, the behavioural risk factor *drug use,* which was most significantly associated with clustering in the univariate analysis (p-value 0.00), was added to the regression model. It significantly improved model fit, according to the LRT (p-value 0.01). None of the three remaining behavioural factors significantly improved model fit after *drug use* was added. Final adjusted ORs for the model including *drug use* are shown in Table 6-4.

Results using the retrospective method were very similar to results for the n method; again, only *drug use* improved the model fit significantly (results not shown).

Finally, the estimated proportion of cases due to recent transmission, or clustering according to the n-1 method, was 57.7% for 15-locus analysis and 35.5% for the 24-locus analysis.

Table 6-2: Demographic features and risk factors for clustering using 24-locus typing for cases notified in the West Midlands, by the n method and retrospective method of clustering. The n method results apply to all cases in the study population, 2007 – 2011. The retrospective method results only apply to cases from 2009 – 2011, as cases from 2007 – 2008 are only used to define clusters for later cases.

| | All cases, 07 – 11 | | Clustered cases, under the 'n method' | | | | | | All cases, 09 – 11 | | Clustered cases, under the 2-year 'retrospective method' | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | Col % | N | Row % | OR (95% CI) | p | aOR (95% CI) | p | N | Col % | N | % | OR (95% CI) | p | aOR (95% CI) | p |
| **Sex** | | | | | | | | | | | | | | | | |
| Male | 1,271 | 55.7 | 618 | 48.6 | 1.0 | 0.03 | 1.0 | 0.04 | 784 | 55.9 | 293 | 37.4 | 1.0 | 0.09 | 0.0 | |
| Female | 1,010 | 44.3 | 444 | 44.0 | 0.8 (0.7,1.0) | | 0.8 (0.7,1.0) | | 618 | 44.1 | 204 | 33.0 | 0.8 (0.7,1.0) | | 0.0 (0.0,0.0) | |
| Total | 2,281 | 100.0 | 1,062 | 46.6 | | | | | 1,402 | 100.0 | 497 | 35.5 | | | | |
| **Age group (years)** | | | | | | | | | | | | | | | | |
| 0-14 | 55 | 2.4 | 39 | 70.9 | 1.0 | 0.00 | 1.0 | 0.00 | 31 | 2.2 | 18 | 58.1 | 1.0 | 0.00 | 1.0 | 0.00 |
| 15-44 | 1,440 | 63.1 | 692 | 48.1 | 0.4 (0.2,0.7) | | 0.5 (0.3,0.9) | | 873 | 62.3 | 338 | 38.7 | 0.5 (0.2,0.9) | | 0.6 (0.3,1.3) | |
| 45-64 | 425 | 18.6 | 192 | 45.2 | 0.3 (0.2,0.6) | | 0.4 (0.2,0.8) | | 273 | 19.5 | 81 | 29.7 | 0.3 (0.1,0.7) | | 0.4 (0.2,0.8) | |
| 65 and over | 363 | 15.9 | 139 | 38.3 | 0.3 (0.1,0.5) | | 0.3 (0.2,0.6) | | 225 | 16.1 | 60 | 26.7 | 0.3 (0.1,0.6) | | 0.3 (0.1,0.7) | |
| Total | 2,283 | 100.0 | 1,062 | 46.5 | | | | | 1,402 | 100.0 | 497 | 35.5 | | | | |
| **Birthplace** | | | | | | | | | | | | | | | | |
| UK | 693 | 32.4 | 451 | 65.1 | 1.0 | 0.00 | 1.0 | 0.00 | 412 | 30.7 | 222 | 53.9 | 1.0 | 0.00 | 1.0 | 0.00 |
| Europe | 53 | 2.5 | 17 | 32.1 | 0.3 (0.1,0.5) | | 0.2 (0.1,0.4) | | 37 | 2.8 | 10 | 27.0 | 0.3 (0.1,0.7) | | 0.3 (0.1,0.6) | |
| East Mediterranean | 24 | 1.1 | 10 | 41.7 | 0.4 (0.2,0.9) | | 0.4 (0.2,1.0) | | 17 | 1.3 | 5 | 29.4 | 0.4 (0.1,1.0) | | 0.3 (0.1,1.0) | |
| Africa | 364 | 17.0 | 138 | 37.9 | 0.3 (0.3,0.4) | | 0.4 (0.3,0.5) | | 228 | 17.0 | 60 | 26.3 | 0.3 (0.2,0.4) | | 0.3 (0.2,0.4) | |
| Americas | 25 | 1.2 | 14 | 56.0 | 0.7 (0.3,1.5) | | 0.9 (0.4,2.0) | | 14 | 1.0 | 5 | 35.7 | 0.5 (0.2,1.4) | | 0.7 (0.2,2.2) | |

| | All cases, 07 – 11 | | Clustered cases, under the 'n method' | | | | | | All cases, 09 – 11 | | Clustered cases, under the 2-year 'retrospective method' | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | Col % | N | Row % | OR (95% CI) | p | aOR (95% CI) | p | N | Col % | N | % | OR (95% CI) | p | aOR (95% CI) | p |
| South Asia | 927 | 43.3 | 356 | 38.4 | 0.3 (0.3,0.4) | | 0.4 (0.3,0.5) | | 599 | 44.6 | 169 | 28.2 | 0.3 (0.3,0.4) | | 0.4 (0.3,0.5) | |
| East/Southeast Asia | 56 | 2.6 | 14 | 25.0 | 0.2 (0.1,0.3) | | 0.2 (0.1,0.3) | | 37 | 2.8 | 10 | 27.0 | 0.3 (0.1,0.7) | | 0.3 (0.1,0.6) | |
| Total | 2,142 | 100.0 | 1,000 | 46.7 | | | | | 1,344 | 100.0 | 481 | 35.8 | | | | |
| **Ethnicity** | | | | | | | | | | | | | | | | |
| White | 383 | 17.4 | 213 | 55.6 | 1.0 | 0.00 | | | 222 | 16.5 | 89 | 40.1 | 1.0 | 0.01 | | |
| Black-Caribbean | 84 | 3.8 | 61 | 72.6 | 2.1 (1.3,3.6) | | | | 37 | 2.7 | 20 | 54.1 | 1.8 (0.9,3.5) | | | |
| Black-African | 350 | 15.9 | 140 | 40.0 | 0.5 (0.4,0.7) | | | | 219 | 16.3 | 64 | 29.2 | 0.6 (0.4,0.9) | | | |
| Black-Other | 11 | 0.5 | 6 | 54.6 | 1.0 (0.3,3.2) | | | | 9 | 0.7 | 5 | 55.6 | 1.9 (0.5,7.1) | | | |
| South Asian | 1228 | 55.7 | 559 | 45.5 | 0.7 (0.5,0.8) | | | | 757 | 56.2 | 271 | 35.8 | 0.8 (0.6,1.1) | | | |
| Chinese | 18 | 0.8 | 4 | 22.2 | 0.2 (0.1,0.7) | | | | 10 | 0.7 | 1 | 10.0 | 0.2 (0.0,1.3) | | | |
| Mixed/Other | 129 | 5.9 | 51 | 39.5 | 0.5 (0.3,0.8) | | | | 94 | 7.0 | 36 | 38.3 | 0.9 (0.6,1.5) | | | |
| Total | 2203 | 100.0 | 1,034 | 46.9 | | | | | 1,348 | 100.0 | 486 | 36.1 | | | | |
| **Time since entry to UK to tuberculosis diagnosis (years)*** | | | | | | | | | | | | | | | | |
| 0 – 1 | 242 | 18.2 | 72 | 29.8 | 1.0 | 0.02 | | | 151 | 17.4 | 34 | 22.5 | 1.0 | 0.29 | | |
| 2 – 4 | 307 | 23.1 | 111 | 36.2 | 1.3 (0.9,1.9) | | | | 192 | 22.1 | 49 | 25.5 | 1.2 (0.7,1.9) | | | |
| 5 – 9 | 310 | 23.3 | 122 | 39.4 | 1.5 (1.1,2.2) | | | | 218 | 25.1 | 61 | 28.0 | 1.3 (0.8,2.2) | | | |
| 10 and over | 470 | 35.4 | 193 | 41.1 | 1.6 (1.2,2.3) | | | | 308 | 35.4 | 94 | 30.5 | 1.5 (1.0,2.4) | | | |
| Total | 1,329 | 100.0 | 498 | 37.5 | | | | | 869 | 100.0 | 238 | 27.4 | | | | |
| **Disease site** | | | | | | | | | | | | | | | | |
| Pulmonary | 1,505 | 66.3 | 783 | 52.0 | 1.0 | 0.00 | 1.0 | 0.00 | 931 | 66.6 | 363 | 39.0 | 1.0 | 0.00 | 1.0 | 0.00 |
| Extra-pulmonary | 766 | 33.7 | 277 | 36.2 | 0.5 (0.4,0.6) | | 0.6 (0.5,0.7) | | 467 | 33.4 | 134 | 28.7 | 0.6 (0.5,0.8) | | 0.7 (0.5,0.8) | |
| Total | 2,271 | 100.0 | 1,060 | 46.7 | | | | | 1,398 | 100.0 | 497 | 35.6 | | | | |
| **Drug sensitivity** | | | | | | | | | | | | | | | | |

| | All cases, 07 – 11 | | Clustered cases, under the 'n method' | | | | | | All cases, 09 – 11 | | Clustered cases, under the 2-year 'retrospective method' | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | Col % | N | Row % | OR (95% CI) | p | aOR (95% CI) | p | N | Col % | N | % | OR (95% CI) | p | aOR (95% CI) | p |
| Resistant to at least one drug | 110 | 4.9 | 36 | 32.7 | 1.0 | 0.00 | 1.0 | 0.01 | 78 | 5.6 | 13 | 16.7 | 1.0 | 0.00 | 1.0 | 0.01 |
| Sensitive | 2,159 | 95.2 | 1,022 | 47.3 | 1.8 (1.2,2.8) | | 1.8 (1.2,2.9) | | 1,310 | 94.4 | 481 | 36.7 | 2.9 (1.6,5.3) | | 2.3 (1.2,4.4) | |
| Total | 2,269 | 100.0 | 1,058 | 46.6 | | | | | 1,388 | 100.0 | 494 | 35.6 | | | | |
| **Previous diagnosis** | | | | | | | | | | | | | | | | |
| No | 1,445 | 86.7 | 680 | 47.1 | 1.0 | 0.53 | | | 1,059 | 85.1 | 376 | 35.5 | 1.0 (0.0,0.0) | 0.19 | | |
| Yes | 221 | 13.3 | 109 | 49.3 | 1.1 (0.8,1.5) | | | | 185 | 14.9 | 75 | 40.5 | 1.2 (0.9,1.7) | | | |
| Total | 1,666 | 100.0 | 789 | 47.4 | | | | | 1,244 | 100.0 | 451 | 36.3 | | | | |
| **History of or current problem drug use**** | | | | | | | | | | | | | | | | |
| No | 1,047 | 95.9 | 472 | 45.1 | 1.0 | 0.00 | | | 1,018 | 95.9 | 352 | 34.6 | 1.0 (0.0,0.0) | 0.00 | | |
| Yes | 45 | 4.1 | 36 | 80.0 | 4.9 (2.3,10.2) | | | | 44 | 4.1 | 34 | 77.3 | 6.4 (3.1,13.2) | | | |
| Total | 1,092 | 100.0 | 508 | 46.5 | | | | | 1,062 | 100.0 | 386 | 36.4 | | | | |
| **History of or current problem alcohol use**** | | | | | | | | | | | | | | | | |
| No | 1,025 | 96.7 | 466 | 45.5 | 1.0 | 0.00 | | | 997 | 96.6 | 352 | 35.3 | 1.0 (0.0,0.0) | 0.00 | | |
| Yes | 35 | 3.3 | 25 | 71.4 | 3.0 (1.4,6.3) | | | | 35 | 3.4 | 23 | 65.7 | 3.5 (1.7,7.1) | | | |
| Total | 1,060 | 100.0 | 491 | 46.3 | | | | | 1,032 | 100.0 | 375 | 36.3 | | | | |
| **History of or current homelessness**** | | | | | | | | | | | | | | | | |
| No | 1,066 | 96.9 | 494 | 46.3 | 1.0 | 0.45 | | | 1,034 | 96.8 | 373 | 36.1 | 1.0 (0.0,0.0) | 0.20 | | |
| Yes | 34 | 3.1 | 18 | 52.9 | 1.3 (0.7,2.6) | | | | 34 | 3.2 | 16 | 47.1 | 1.6 (0.8,3.1) | | | |
| Total | 1,100 | 100.0 | 512 | 46.6 | | | | | 1,068 | 100.0 | 389 | 36.4 | | | | |
| **History of or currently in prison**** | | | | | | | | | | | | | | | | |
| No | 1,007 | 95.9 | 465 | 46.2 | 1.0 | 0.00 | | | 980 | 95.9 | 346 | 35.3 | 1.0 (0.0,0.0) | 0.00 | | |
| Yes | 43 | 4.1 | 32 | 74.4 | 3.4 (1.7,6.8) | | | | 42 | 4.1 | 29 | 69.1 | 4.1 (2.1,8.0) | | | |
| Total | 1,050 | 100.0 | 497 | 47.3 | | | | | 1,022 | 100.0 | 375 | 36.7 | | | | |

| | All cases, 07 – 11 | | Clustered cases, under the 'n method' | | | | | All cases, 09 – 11 | | Clustered cases, under the 2-year 'retrospective method' | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | Col % | N | Row % | OR (95% CI) | p | aOR (95% CI) | p | N | Col % | N | % | OR (95% CI) | p | aOR (95% CI) | p |

*Foreign-born only, **Missing for 2007, 2008 and half of 2009 cases

Table 6-3: Demographic features and risk factors for clustering under the 24-locus typing system for foreign-born cases notified in the West Midlands 2007-2011, using the n method for clustering.

| | Total cases | | Clustering according to 24 loci, n method | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | N | Col % | N | % | OR | 95% CI | p | aOR | 95% CI | p |
| **Sex** | | | | | | | | | | |
| Male | 819 | 55.6 | 332 | 40.54 | 1.00 | | 0.02 | 1.00 | | 0.01 |
| Female | 654 | 44.4 | 227 | 34.71 | 0.78 | 0.63 – 0.96 | | 0.74 | 0.59 – 0.93 | |
| Total | 1,473 | 100 | 559 | 37.95 | | | | | | |
| **Age group** | | | | | | | | | | |
| 0-14 | 19 | 1.29 | 12 | 63.16 | 1.00 | | 0.09 | | | |
| 15-44 | 922 | 62.55 | 336 | 36.44 | 0.33 | 0.13 – 0.86 | | | | |
| 45-64 | 291 | 19.74 | 113 | 38.83 | 0.37 | 0.14 – 0.97 | | | | |
| 65 and over | 242 | 16.42 | 98 | 40.5 | 0.40 | 0.15 – 1.04 | | | | |
| Total | 1,474 | 100 | 559 | 37.92 | | | | | | |
| **Region of birth** | | | | | | | | | | |
| UK | n/a | n/a | n/a | n/a | n/a | n/a | n/a | | | |
| Europe | 3.66 | 3.66 | 17 | 32.08 | 1.00 | | 0.13 | | | |
| East Mediterranean | 1.66 | 5.31 | 10 | 41.67 | 1.51 | 0.56 – 4.09 | | | | |
| Africa | 25.12 | 30.43 | 138 | 37.91 | 1.29 | 0.70 – 2.39 | | | | |
| Americas | 1.73 | 32.16 | 14 | 56 | 2.70 | 1.01 – 7.17 | | | | |
| South Asia | 63.98 | 96.14 | 356 | 38.4 | 1.32 | 0.73 – 2.39 | | | | |
| East/Southeast Asia | 3.86 | 100 | 14 | 25 | 0.71 | 0.31 – 1.63 | | | | |
| Total | 1,449 | 100 | 549 | 37.89 | | | | | | |
| **Ethnicity** | | | | | | | | | | |
| White | 2.91 | 2.91 | 11 | 26.19 | 1.00 | | 0.09 | | | |
| Black Caribbean | 2.15 | 5.05 | 16 | 51.61 | 3.01 | 1.12 – 8.05 | | | | |
| Black-African | 22.84 | 27.89 | 130 | 39.39 | 1.83 | 0.89 – 3.77 | | | | |
| Black-Other | 0.48 | 28.37 | 3 | 42.86 | 2.11 | 0.41 – 10.98 | | | | |
| South Asian | 63.53 | 91.9 | 361 | 39.32 | 1.83 | 0.91 – 3.68 | | | | |
| Chinese | 1.11 | 93.01 | 4 | 25 | 0.94 | 0.25 – 3.53 | | | | |
| Mixed/Other | 6.99 | 100 | 29 | 28.71 | 1.14 | 0.50 – 2.56 | | | | |
| Total | 1,445 | 100 | 554 | 38.34 | | | | | | |
| **Years since entry to** | | | | | | | | | | |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **tuberculosis diagnosis*** | | | | | | | | | | |
| 0 – 1 | 242 | 18.21 | 72 | 29.75 | 1.00 | | 0.02 | 1.00 | | 0.03 |
| 2 – 4 | 307 | 23.1 | 111 | 36.16 | 1.34 | 0.93 – 1.92 | | 1.35 | 0.94 – 1.95 | |
| 5 – 9 | 310 | 23.33 | 122 | 39.35 | 1.53 | 1.07 – 2.19 | | 1.56 | 1.09 – 2 .24 | |
| 10 and over | 470 | 35.36 | 193 | 41.06 | 1.65 | 1.18 – 2.29 | | 1.60 | 1.14 – 2.24 | |
| Total | 1,329 | 100 | 498 | 37.47 | | | | | | |
| **Disease site** | | | | | | | | | | |
| Pulmonary | 875 | 59.69 | 365 | 41.71 | 1.00 | | 0.00 | 1.00 | | 0.00 |
| Extra pulmonary | 591 | 40.31 | 193 | 32.66 | 0.68 | 0.54 – 0.84 | | 0.72 | 0.57 – 0.91 | |
| Total | 1,466 | 100 | 558 | 38.06 | | | | | | |
| **Drug sensitivity** | | | | | | | | | | |
| Resistant ≥1 drug | 86 | 5.87 | 23 | 26.74 | 1.00 | | 0.02 | 1.00 | | 0.08 |
| Sensitive | 1,378 | 94.13 | 533 | 38.68 | 1.73 | 1.06 – 2.82 | | 1.55 | 0.93 – 2.59 | |
| Total | 1,464 | 100 | 556 | 37.98 | | | | | | |
| **Previous diagnosis** | | | | | | | | | | |
| No | 935 | 86.73 | 345 | 36.9 | 1.00 | | 0.25 | | | |
| Yes | 143 | 13.27 | 60 | 41.96 | 1.24 | 0.86 – 1.77 | | | | |
| Total | 1,078 | 100 | 405 | 37.57 | | | | | | |
| **History of or current problem drug use**** | | | | | | | | | | |
| No | 718 | 99.17 | 266 | 37.05 | 1.00 | | 0.02 | | | |
| Yes | 6 | 0.83 | 5 | 83.33 | 8.50 | 0.99 – 73.11 | | | | |
| Total | 724 | 100 | 271 | 37.43 | | | | | | |
| **History of or current problem alcohol use**** | | | | | | | | | | |
| No | 696 | 98.44 | 256 | 36.78 | 1.00 | | 0.07 | | | |
| Yes | 11 | 1.56 | 7 | 63.64 | 3.01 | 0.50 – 0.68 | | | | |
| Total | 707 | 100 | 263 | 37.2 | | | | | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **History of or current homelessness**** | | | | | | | |
| No | 712 | 97.53 | 266 | 37.36 | 1.00 | | 0.73 |
| Yes | 18 | 2.47 | 6 | 33.33 | 0.84 | 0.31 – 2.26 | |
| Total | 730 | 100 | 272 | 37.26 | | | |
| **History of or currently in prison**** | | | | | | | |
| No | 686 | 98.28 | 262 | 38.19 | 1.00 | | 0.05 |
| Yes | 12 | 1.72 | 8 | 66.67 | 3.24 | 0.97 – 10.85 | |
| Total | 698 | 100 | 270 | 38.68 | | | |

**Missing for 2007, 2008 and half of 2009 cases

Table 6-4: Adjusted odd ration (aOR) for a multivariate model constructed with factors significantly associated with clustering, plus additional behavioural factors added using a forward selection. Only drug use significantly improved the model; aORs for the model including drug use are reported.

| **Sex** | aOR | CI | CI |
|---|---|---|---|
| Male | 1.00 | | |
| Female | 0.71 | 0.54 | 0.94 |
| **Age group** | | | |
| 0 – 14 | 1.00 | | |
| 15 – 44 | 0.49 | 0.19 | 1.25 |
| 45 – 64 | 0.28 | 0.11 | 0.76 |
| 65 and over | 0.23 | 0.08 | 0.61 |
| **Region of birth** | | | |
| UK | 1.00 | | |
| Europe | 0.23 | 0.09 | 0.58 |
| East Mediterranean | 0.34 | 0.11 | 1.06 |
| Africa | 0.35 | 0.23 | 0.54 |
| Americas | 1.17 | 0.30 | 4.58 |
| South Asia | 0.44 | 0.32 | 0.60 |
| East/Southeast Asia | 0.26 | 0.11 | 0.62 |
| **Disease site** | | | |
| Pulmonary | 1.00 | | |
| Extra pulmonary | 0.69 | 0.51 | 0.92 |
| **Drug sensitivity** | | | |
| Resistant to one or more drugs | 1.00 | | |
| Sensitive | 1.65 | 0.86 | 3.17 |
| **History of or current problem drug use** | | | |
| No | 1.00 | | |
| Yes | 3.24 | 1.29 | 8.12 |

## 6.3 Discussion

For 15- and 24-locus VNTR typing systems and both definitions of clustering, cases with a UK birthplace, pulmonary disease, and drug sensitive strains were associated with higher risks of clustering. These associations remained significant even after adjustment for other risk factors in multivariate models. Younger age was also significantly associated with clustering for the 24-locus analysis, both in the univariate analysis and after adjustment for other factors. Drug use was significantly associated with clustering after adjustment for other factors in a model exploring behavioural risk factors collected for a subset of cases. For foreign-born individuals, time since arrival was significantly associated with clustering. Results are consistent with other studies, several of which have found younger age, native birthplace, pulmonary disease, drug use, and time since arrival for foreign-born individuals to be risk factors for clustering [168, 169, 313].

The association between drug sensitive strains and clustering has been found previously [170, 314, 315] and is supported by theoretical and empirical evidence [316, 317]. However, the reasons are still unclear and several studies have shown the opposite effect, whereby drug resistance is associated with clustering [318]. Fitness of the bacterial strain may depend on the type of drug resistance mutation and could therefore explain some of the variance in the effect of drug resistance on clustering across studies [319]. Several studies have shown that isoniazid-resistant strains with the prevalent *katG* mutation were at least as likely to be clustered as sensitive strains, however isoniazid-resistant strains without the mutation were less likely to be clustered [315, 319, 320].

It is possible that the association between clustering and drug sensitivity is not due to the drug sensitivity itself, but to other associated factors. The multivariate model used to conclude that drug sensitive strains were more likely to be clustered was adjusted for other risk factors, including birthplace. Still, it should also be noted that drug resistance is associated with foreign-birthplace. Since foreign-born individuals are less likely to cluster than UK-born individuals, this could help explain why drug sensitive strains are more likely to be clustered than drug resistant strains. Supporting this idea, the multivariate model for foreign-born individuals shows that, after adjustment for

other factors, drug resistance was *not* significantly associated with clustering in that model. However, this finding could simply reflect the reduced power to detect associations within a smaller sample size of only foreign-born cases.

Another factor significantly associated with clustering in univariate analyses was ethnicity, though this variable was not included in the multivariate model because of concerns over correlation with the birthplace variable. Across typing of both systems and definitions of clustering, all non-White ethnicities showed reduced risk of clustering, apart from Black-Caribbean individuals. This reduced risk is likely due to the immigration history of groups, with migration from the Caribbean peaking many years ago and reducing in recent years; most Black-Caribbean tuberculosis patients were born in the UK, and of those born outside the UK arrived many years ago.

The overall proportion of isolates clustered in this study was 46.5%, using 24-locus VNTR and the n method. This finding compares well to several RFLP-based clustering proportions found in studies from other developed countries. For example, 45% were clustered in the Netherlands over a six-year period using RFLP. A study of tuberculosis in New York City from 1990 – 1999 showed 48% clustered with RFLP [171], although in this study, most cases were US-born, not foreign-born, which may contribute to increased clustering proportions. In Baltimore, a 2.5 year study showed 46% of isolates clustered with RFLP [321]. These results are consistent with studies comparing the discriminatory ability of 24-locus VNTR with RFLP showing agreement between the two typing methods [149, 154, 322-324]. However, even among studies from developed countries, there is a large variation in the proportion clustered across RFLP studies [325], with proportions ranging from below 20% to more than 80% (e.g. [170, 326-328]), so the results of the present study contrast with some estimates.

It follows that the rough estimate of the proportion of disease due to recent transmission, which was 35.5% using the n-1 method and 24-locus VNTR, is similar to RFLP-based estimates from other developed countries [160, 171, 312, 321, 329, 330]. This estimate is also similar to an estimate of 38% from the Netherlands based on a more sophisticated interpretation of RFLP data combined with epidemiological data [331]. However, this percentage is higher than a recent estimate of about 25% of cases

due to recent transmission in the United States based on spoligotyping and 12-locus VNTR data combined with geospatial scanning [332].

In contrast with the results of this study, a national study of tuberculosis in the UK for cases notified in 1998 with isolates typed using RFLP found only 21% of cases clustered [166]. Although proportions clustered are discrepant, the 1998 study used only one year of data and was further limited by a low proportion of isolates with typing results—61% of the culture-confirmed cases, compared to 92.5% in the present study. As authors pointed out, both of these limitations reduce the proportion clustered and adjustments can be made for them. Following their work, correction for the study duration can be made assuming 72.5% of total clustering is identified within one year and correction for the proportion of culture-positive isolates typed results in an adjusted proportion clustered of 20.6%/0.725/0.61 = 46.6%, the corrected proportion clustered according to the n method. Correction for the proportion of culture-positive isolates typed in the present study results in an adjusted proportion clustered of 46.5/92.5 = 50.3%. Despite the disparate geographic areas considered and different typing systems used, the adjusted proportions clustered are similar between this study and the 1998 study.

The proportion of cases clustered in the present study is also significantly higher than the only other population-based molecular epidemiological study in the UK, a study of isolates from London spanning 2.5 years from 1995-1997. In that analysis, 22.7% of isolates were clustered using RFLP. The restricted geographical range and low proportion typed were limitations of the study however. Only 77% of culture-confirmed cases in London were typed, which means an even lower proportion of total cases were typed since the proportion culture-confirmed can be low (here it was 53%). A further 18% of isolates were excluded from the cluster analysis due to low copy number profiles. For these reasons, it is not surprisingly that the London study found a much lower proportion of isolates clustered.

As expected, the proportion of cases clustered was higher using 15-locus VNTR compared with 24-locus VNTR for any definition of clustering. Interestingly however, conclusions about risk factors for clustering are generally very similar between the two typing systems. This similarity suggests data from 15-locus typing may have some

utility in population-based studies of risk factors for clustering, even if 15-locus VNTR is not discriminatory enough to make conclusions about individual transmission events. Results were also very similar between the n method and retrospective method of clustering definitions, though the retrospective method did not identify as many significant risk factors for clustering. This result could partly be due to reduced sample sizes resulting from exclusion of the first two years of data, or a true effect resulting from the slightly different, and less biased, definition of clustering.

One major limitation to the study is the combined effect of a restricted geographical area of the study population and incomplete case ascertainment, both of which potentially decrease clustering proportions and odds ratios for clustering risk factors [8, 166]. The restricted geographical area means that some patients transmitting to or from patients in the study may have resided outside the study area and were not captured by the study. The study of molecular epidemiology in England in 1998 showed that restricting analyses to London only reduced clustering by nearly 40% [166]. However, as discussed above, that particular study was limited by a fairly low coverage of culture-positive patients and short study duration of only one year. If the study had been longer and the proportion of culture-positive cases had been higher, it is unlikely that restricting analyses to the London area would have resulted in such a large reduction in the clustering proportion. In summary, the effect of a restricted geographical area is difficult to disentangle from study size, including case ascertainment, though both are problematic.

The impact of incomplete case ascertainment is difficult to quantify. Incomplete ascertainment results from cases not being notified at all, cases with no laboratory data, and cases for which typing results were incomplete or missing. These missing cases also likely reduce clustering proportions. In this study, the low proportion of cases with a positive culture, only 56.7%, is of particular concern. This low proportion could be partly due to an increasing proportion of cases that are culture negative, or, more likely, due to cases missing from laboratory data.

A final limitation was that no cases were excluded due to laboratory contamination. None were excluded because dates of specimen processing were not available. This is unlikely to much impact results because of the low probability of laboratory error, as a

2002 study in the UK showed only 0.54 – 0.93% of cases with false positive results due to laboratory cross-contamination [333]. Laboratory contamination is discussed further in Appendix 10.18.

In conclusion, this descriptive study is one of the first large-scale, population-level molecular epidemiological studies using 24-locus VNTR. Results show that known risk factors for clustering of tuberculosis cases using RFLP were present in the study period in the West Midlands using 24-locus VNTR. Interestingly, the same risk factors for clustering were identified using both 24-locus and 15-locus VNTR, despite the reduced discriminatory ability of the 15-locus system. Using the n method and retrospective methods also resulted in similar conclusions. Crude estimates for the proportion of cases due to recent transmission in the UK established a baseline population estimate that can be compared to future estimates for this region using genotyping data and are also useful for comparison to model-derived estimates in Chapter 7.

# 7 West Midlands Molecular Epidemiological Modelling

This chapter describes the simulation of molecular epidemiological data from the West Midlands to explore the relationship between genotype clustering and recent transmission of *M. tuberculosis*. Genotyping of *M. tuberculosis* isolates from tuberculosis cases is potentially helpful for understanding transmission patterns in populations. However, the relationship between genotype clustering and recent transmission is unclear. The model described in Chapter 3 was used to simulate 24-locus VNTR genotyping data from the West Midlands by assigning each infected individual in the simulation a time, place and strain type of infection. A genotype cluster analysis was performed on simulated genotyping data and the proportion of cases clustered in the simulation was calculated. These model outputs were compared to observed molecular epidemiological data from the West Midlands from 2007 – 2011, described in Chapter 6, for a range of input parameter scenarios. For scenarios which corresponded best with observed data, implications for interpretation of clustering data, including positive and negative predictive values, as well as trends in the proportion of cases due to recent transmission are described.

## 7.1 Methods

### 7.1.1 Observed Data

Observed data used for this chapter included notification rates from the West Midlands, 2007 – 2011, presented in Chapter 2, Section 2.7.2 and reviewed below. Other observed data include 24-locus VNTR typing data associated with a subset of cases. The genotyping data are described in Chapter 6, though for modelling purposes, only the proportions clustered by age, sex, and birthplace were used.

#### *7.1.1.1 Notification rates by age, sex and birthplace*

Tuberculosis notification rates in the West Midlands from 2007 – 2011 were calculated for comparison with model output. The overall notification rate in the West Midlands for this time period was about 17 per 100,000 population, though as with the rates in England and Wales, UK-born rates are much lower than foreign-born rates. The UK-born notification rate over this period was 6.5 per 100,000 population while the foreign-born rate was 110 per 100,000 population. The UK-born notification rate in the West Midlands over this time period was higher than that seen in England and Wales from 1999 – 2009, which was about 4.5 per 100,00 overall. For the foreign-born individuals in the West Midlands from 2007 – 2011, overall notification rates were also higher than those in England and Wales from 1999 – 2009, where the rate was about 102 per 100,000 population.

The notification rates broken down by age are shown in Figure 7-1 for UK-born males and Figure 7-2 for UK-born females; rates for foreign-born are in Figure 7-3 for males and Figure 7-4 for females. Across age and sex categories, UK-born notification rates varied between about 1 per 100,000 to about 12 per 100,000, while foreign-born rates varied between about 20 per 100,000 to about 160 per 100,000. Particularly for UK-born cases, notification rates for the West Midlands varied more from year to year than England and Wales rates, likely due to the smaller numbers of cases in the West Midlands. This variation could have also been partly due to more uncertainty in population size estimates and fluctuation in those estimates.

Some trends in the notification rates for UK-born cases in the West Midlands differ from those seen in the whole of England and Wales. Notably, the notification rates for UK-born males and females aged 15 – 44 years were higher than in England and Wales over a similar time period. For males in the West Midlands, this ranged from about 8 – 10 per 100,000 from 2007 – 2011. For females, this ranged from about 7 – 11 per 100,000 from 2007 – 2011. In England and Wales, these ranged from about 4 – 6 per 100,000 for males and 4 – 5 per 100,000 for females from 1999 – 2009. Rates for those aged 15 – 44 years in the West Midlands were also higher than the rate for those aged 65 years and above, whereas in England and Wales, the notification rate in those aged 65 years and above were generally higher than that for those aged 15 – 44 years. This trend was always true for males and usually true for females, though the differences in notification rates between the age categories have decreased over time.

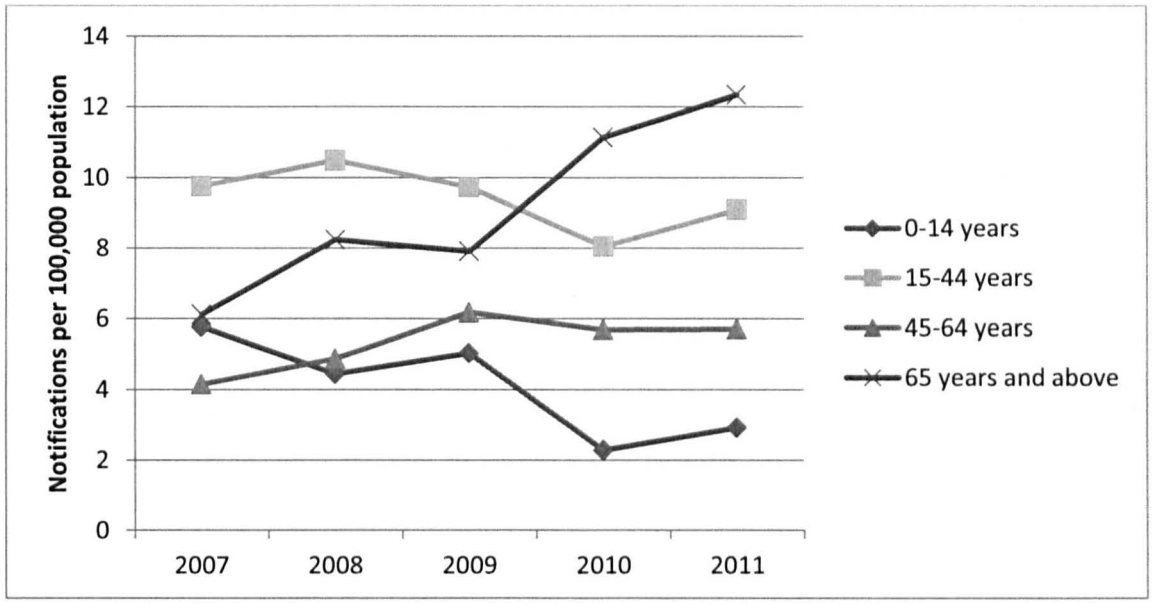


Figure 7-1: Tuberculosis notifications per 100,000 population for UK-born male cases reported in the West Midlands, 2007 – 2011.

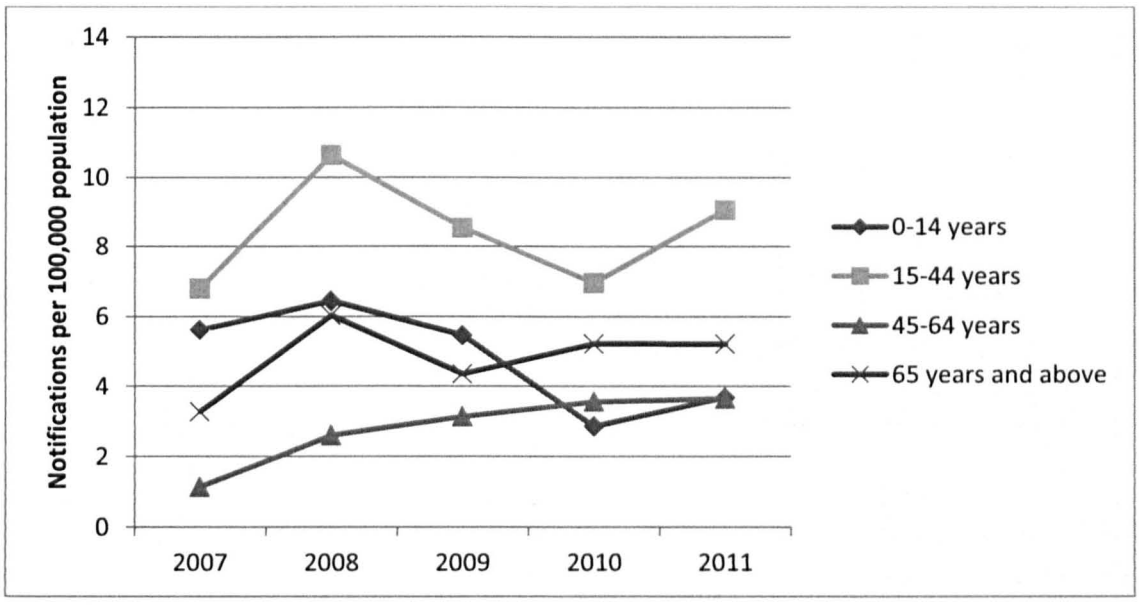


Figure 7-2: Tuberculosis notifications per 100,000 population for UK-born female cases reported in the West Midlands, 2007 – 2011.

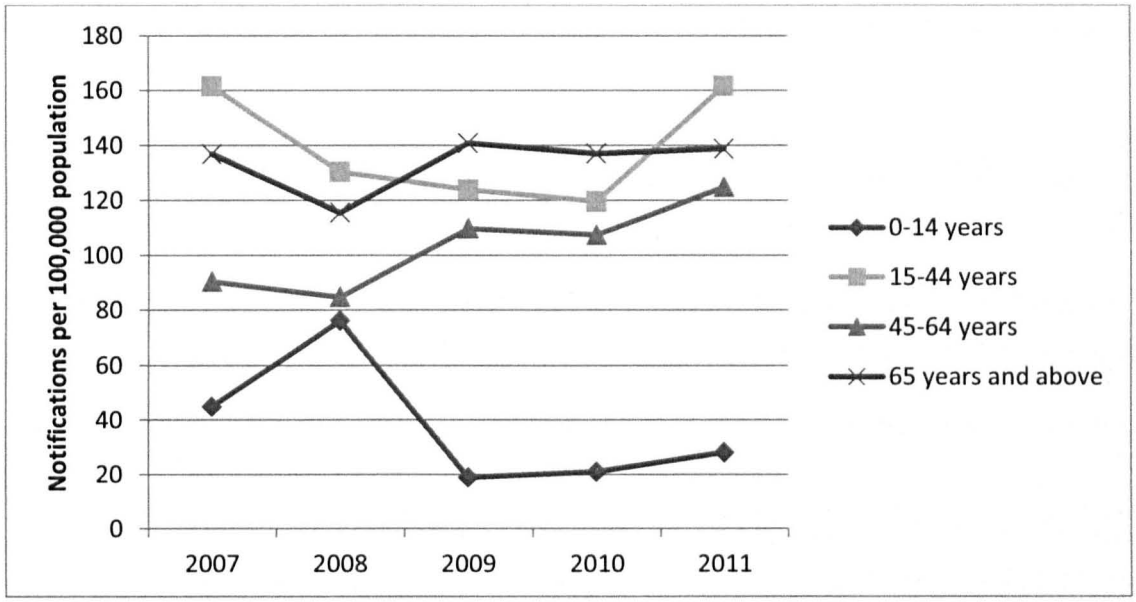


Figure 7-3: Tuberculosis notifications per 100,000 population for foreign-born male cases reported in the West Midlands, 2007 – 2011.

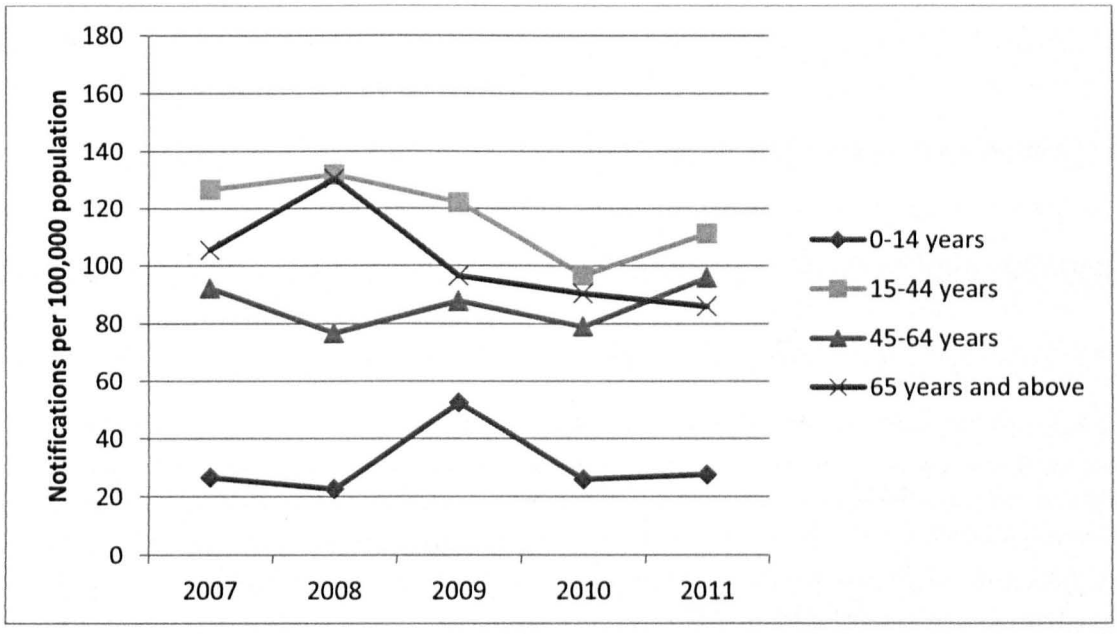


Figure 7-4: Tuberculosis notifications per 100,000 population for foreign-born female cases reported in the West Midlands, 2007 – 2011.

### *7.1.1.2 Proportion clustered by age, sex and birthplace*

As reported in Chapter 6, the overall proportion of cases clustered in the West Midlands from 2007 – 2011 was 46.5% using 24-locus VNTR genotyping data. Clustered cases were defined as those with an isolate with a 24-locus strain type profile that matched the profile of any other isolate in the study period, across cases from all demographic categories, according to the 'n method', as described in Chapter 6. For comparison with model output, the proportion of cases clustered is reported for each age category, sex, and birthplace for the West Midlands from 2007 – 2011. These proportions are found in Figure 7-5.

The proportion of cases clustered was generally higher for UK-born individuals than for those who were foreign-born, with about 65% of cases and about 38% of cases clustered in the two groups, respectively. Age-specific trends also differed. For UK-born individuals, the proportion clustered clearly decreased with age, with proportions clustered approximately halving from the youngest age group, 0 – 14 years, to the oldest age group, 65 years and above. For foreign-born individuals, those aged 0 – 14 years had a much higher proportion clustered than the three older age classes, but there was no clear age-dependent trend in the proportion clustered from age groups for those aged 15 years and above.

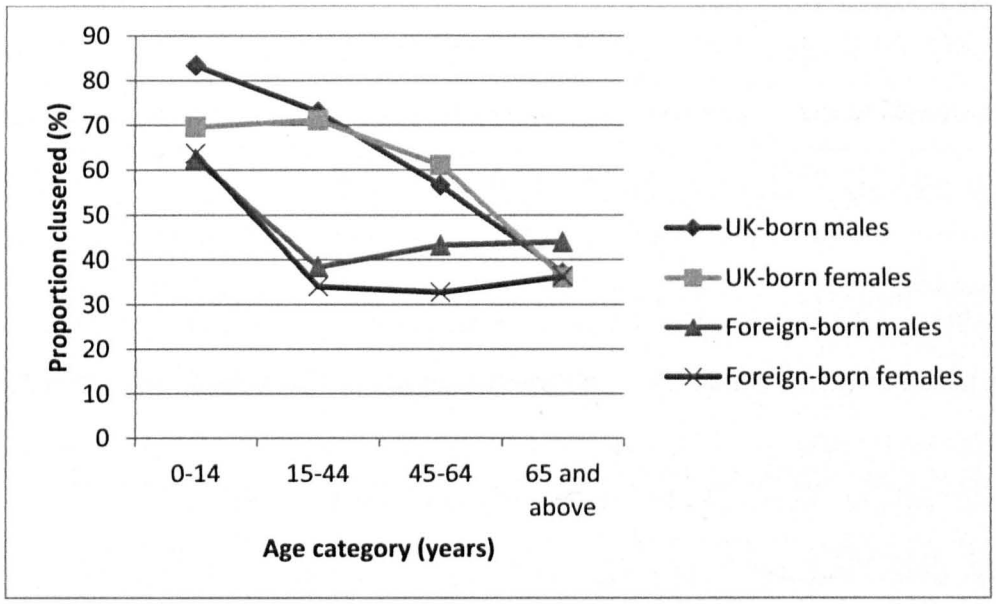


Figure 7-5: Proportion clustered (%) by age, sex and birthplace (UK-born and foreign-born) for the West Midlands, 2007 – 2011.

### 7.1.2 The Model

The model is described in Chapter 3, with model parameters and assumptions described in Chapter 4. Most parameter values and assumptions used for this application of the model were identical to those used for the England and Wales application of the model, though some parameter values and distributions were specific to the West Midlands, as indicated in Chapter 4. These include migration data and the proportion of cases that are pulmonary, plus genotyping-specific parameters not included in the England and Wales application of the model. For this application of the model, the full version of the model including genotype-related processes was used for simulation of molecular epidemiological data. Also, because of its importance in this chapter, the strain type assignment and mutation processes are briefly summarized below.

#### *7.1.2.1 Genotype-related processes*

Strain types were assigned to every infected and diseased individual in the simulation. During infection transmission, an infectious individual passed their strain type to the individual infected. This was true whether the recipient of infection had *Latent Infection* or was *Uninfected*. For assigning strains at model initialization and assigning strains to migrants entering the study population, strains were randomly selected from a distribution which represents the entire population of strains. The entire population included both observed strains—those strains from typed cases—and hypothesized unobserved strains—those strains assumed to exist in unreported or untyped disease cases or in all other infections that had not progressed to disease. Assumptions about strain type distributions were derived from observed strain typing data, as discussed below in Section 7.1.3.1.

Strain types in the model could mutate at any time during infection or disease. As discussed in Chapter 4, the mutation rate was defined here as the rate of change of the dominant 24-locus VNTR strain type in the body. The mutation rate took into account the rate at which mutants were produced and the probability that mutants were successful enough to become the dominant genotype in the body. For simplicity, the dominant strain was taken to be the strain that was transmitted and that was detected in culturing clinical specimens from disease sites.

For 24-locus VNTR strain types modelled here, mutants arose by a change in the number of repeats at any of the 24 loci in the strain type profile, although for simplicity and lack of appropriately detailed data, loci were not modelled individually. That would require a mutation rate specific to each locus and estimates of the genetic diversity at each locus, both of which are largely unknown for any given population of strains, including the population of the West Midlands. Estimates of the average mutation rate per locus per annum were used to infer a mutation rate per annum for entire 24-locus VNTR profiles, though in reality, loci have different mutation rates [334]. Because loci were not modelled individually, it was assumed that every mutation led to a new, unique 24-locus VNTR profile. In reality, it is possible that mutations to the VNTR profile could have resulted in a profile that matches as an existing strain in the population, a so-called 'homoplasic' mutation [335]. However, considering the relatively small population of strains, relatively large amount of diversity provided by 24-locus profiles, and relatively short time scales, the simplification seems reasonable. This conclusion is supported in a recent paper by Reyes et al. who studied the extent of homoplasic mutations in VNTR profiles by simulation of the mutation process nested in a transmission model [335]. Authors concluded that while there may be a significant amount of homoplasic mutations over long time periods, these are unlikely to substantially affect clustering statistics with sufficiently variable genetic markers such as 24-locus VNTR.

### 7.1.3 Key Input Parameters

Five model parameters that were particularly uncertain were varied, and model output under different combinations of variable parameters was compared with observed clustering proportions. Firstly, the mutation rate of 24-locus VNTR profiles was varied using five estimates from studies in the literature (see Section 1.1.1.1). Secondly, the strain type distribution assumptions were varied using three different assumptions based on observed strain type distributions (section Section 7.1.3.1). Lastly, three other parameters and distributions were varied together: disease risks, contact rates, and assumptions about the infection status of migrants upon entry to the UK. These parameters were varied together because the disease risks were estimated from England and Wales modelling, and different estimates were obtained under different values for the contact parameter and the infection status of migrants upon entry to

the UK. The four best-fitting scenarios from the England and Wales model, including all three parameters, were used here. Although, it should be noted that these were modified slightly by using increased contact numbers, as discussed in Section 7.1.4.2.

#### *7.1.3.1 Distributions of strain types at model initialization and for migrants*

For choosing strain types for all individuals at model initialization and for choosing strain types for migrants when they enter the UK, a distribution of strain type frequencies was necessary. Although there are data on *observed* strain type diversities for various populations, including those in the UK, these only include culture-positive, genotyped cases. The *unobserved* strain type diversity appears to be largely unexamined, but presumably, many strains are unobserved, including genotypes found in untyped disease cases and all infections that have not resulted in disease. Of course, empirical data on the distribution of unobserved strain types would be nearly impossible to obtain since isolation of *M. tuberculosis* from infection is not feasible, at least with current technology. However, to my knowledge, no studies have attempted to characterize this population even indirectly using models or techniques from population genetics or ecological tools for handling species diversity.

For lack of suitable alternative data, the observed strain type distributions from the West Midlands, 2007 – 2011, were used to generate plausible distributions for the total population of strain types, including observed and unobserved strain types. These distributions were created separately for each birthplace category, UK-born and foreign-born. UK-born and foreign-born strain type distributions were differentiated because the strain type distributions were expected to differ between these populations. The UK-born strain pool is expected to be a relatively distinct and stable population, though its distinctness and stability are influenced by contact between UK-born and other populations. The foreign-born strain pool is expected to contain a mixture of strains from countries around the world, giving the strain type pool vastly different properties, such as an increased total number of strains and higher variance in strains. In the simulation, the UK-born strain type distribution was used to assign

strains for UK-born individuals at initialization only. The foreign-born strain type distribution was used to assign strains for all migrants entering the UK at any point, including foreign-born individuals at model initialization and both UK-born and foreign-born individuals entering through migration at any point in the simulation. This method of assignment meant the foreign-born strain type pool covered more than 100 years of time and the entire world in geography.

One tool for characterizing observed strain type distributions and making inferences about the unknown population of strain types is a 'species abundance distribution' [336]. These plots are commonly used in ecology to study species abundance and diversity. The plots can take several related forms, including representing the abundance of each distinct species in the population on the y-axis, ordered left to right from most abundant to least abundant on the x-axis. In this application, a strain type is considered equivalent to a species and the abundance is the prevalence of a strain type in the population of known strains. For brevity, the terms 'UK-born strain' and 'foreign-born strain' are used below to mean 'strains in UK-born individuals' and 'strains in foreign-born individuals', respectively.

The observed species abundance plots for UK-born and foreign-born strains in the West Midlands are shown in Figure 7-6 and Figure 7-7. For UK-born cases, there were 379 distinct strains in the UK-born strain pool of 700 total cases. The most prevalent strain type accounted for 12.4% of all UK-born strains. The least prevalent were the 295 unique cases, each of which each accounted for 0.1% of all UK-born strains. For foreign-born cases, there were 1,128 distinct strains observed from 1,492 total foreign-born cases. The most prevalent of these accounted for about 0.2% of all foreign-born strains. The least prevalent were the 969 unique cases, each accounting for 0.07% of all strains. If the plots are examined with more resolution, it can be seen that the right tail of the distribution approaches a straight line. The line is formed by the lowest prevalence values, which include unique strains, clusters of two isolates, and clusters of three isolates.

Because empirical distributions miss a portion of strains in the entire population of infected and diseased individuals, the tail of the species abundance curve must be extended to estimate the number of unobserved strains. Only the tail of the

distribution was extended, both for simplicity and because rare strains are the ones most likely to be unobserved. Each tail was extended by fitting a straight line to the points in the right tail of the distribution, where it was approximately linear, and projecting to the x-intercept. This linear part of the line included at least the rightmost two prevalence values. The x-intercept then became a proxy for the total number of strains in the population, including observed and unobserved strains. The unobserved strains were then added to the empirical distribution, creating one occurrence for each new strain. The species abundance curve was then redrawn and used for model assumptions. For illustration, lines fit to the last two prevalence values for each distribution are shown in Figure 7-8 and Figure 7-10. Solving the equation for the straight line fit to all the points with the last two values of the UK-born curve results in the estimate that there are 700 distinct strains in the population of UK-born cases in the West Midlands, including the 379 observed and 321 unobserved strains. The new species abundance curve used in the model is shown in Figure 7-9. For foreign-born individuals, solving the equation for the line fit to all the points with the smallest two prevalence values resulted in an estimated 3,000 distinct strains in this population, including the 1,128 observed and 1,872 unobserved strains. The new curve accounting for all 3,000 strains is found in Figure 7-11. The same process was applied when fitting a line to the last three prevalence values of each distribution and resulted in 625 distinct strains for UK-born cases and 2,190 distinct strains for foreign-born cases (figures not shown).

Because this method resulted in conservative estimates of the total number of unobserved strains in the population, an extension of this method—choosing a larger number of total distinct strains beyond the intercept obtained by fitting the tail of the distribution as described above—was also used. This variation allowed for a high-diversity scenario for strain type distributions. The intercept estimates of 2,190 and 3,000 possible distinct strains for all migrants from all over the world for the long time span of infections included seemed particularly low, so those were increased to 5,000 for a less conservative measure. The 625 and 700 possible distinct strains for UK-born in the West Midlands seemed more reasonable, so those were increased only to 1,000 for the high diversity scenario. Of course, since most of these strains were rare, most were never observed in the model.

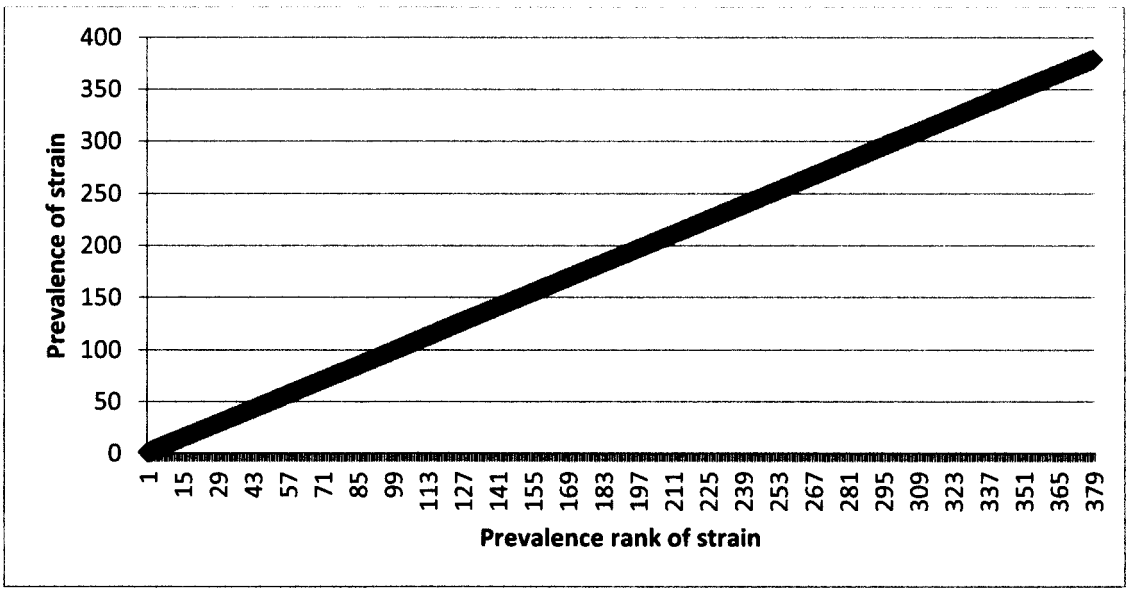


Figure 7-6: Strain type distribution observed for strains from UK-born cases in the West Midlands, 2007 – 2011. The prevalence of each strain in the population of all strains from UK-born cases is plotted. There were 379 distinct strains observed.

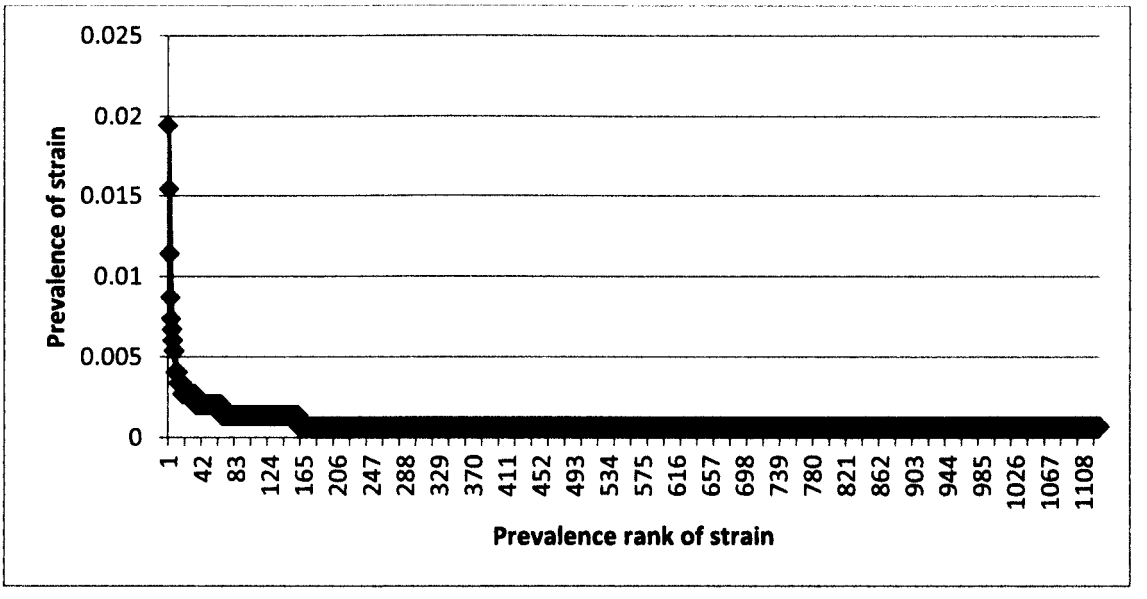


Figure 7-7: Strain type distribution observed for strains from foreign-born cases in the West Midlands, 2007 – 2011. The prevalence of each strain in the population of all strains from foreign-born cases is plotted. There were 1,128 distinct strains observed.

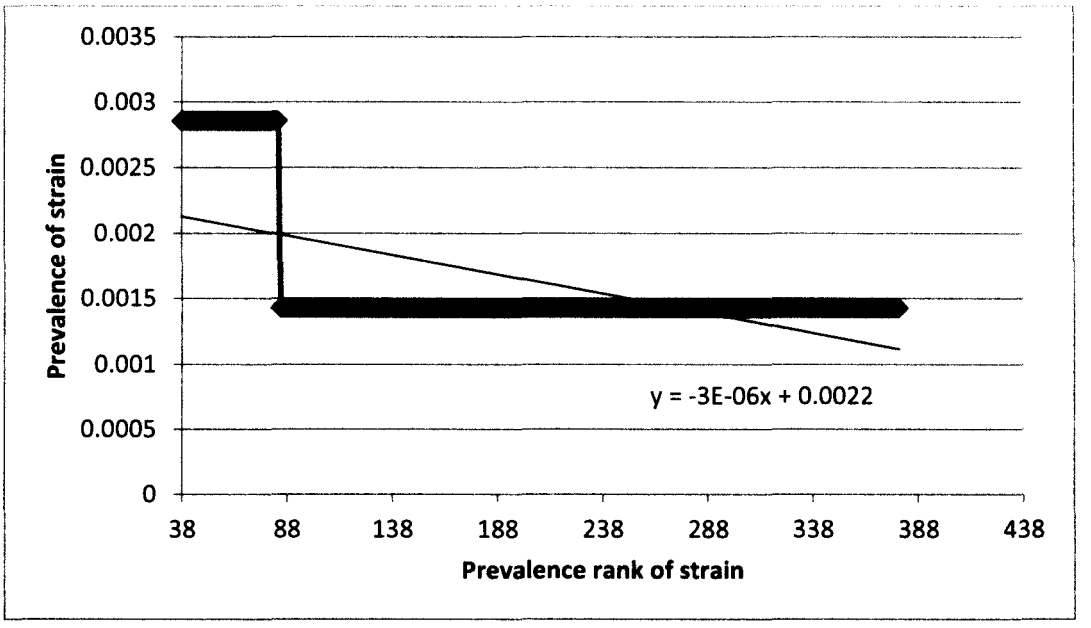


Figure 7-8: Excerpt from plot of strain type distribution observed for strains from UK-born cases in the West Midlands, 2007 – 2011, with straight line fitted to the last two prevalence values observed. The fitted line results in an intercept of 733, or an estimated 733 total strains in the population, including the 379 observed strains and 354 unobserved strains.

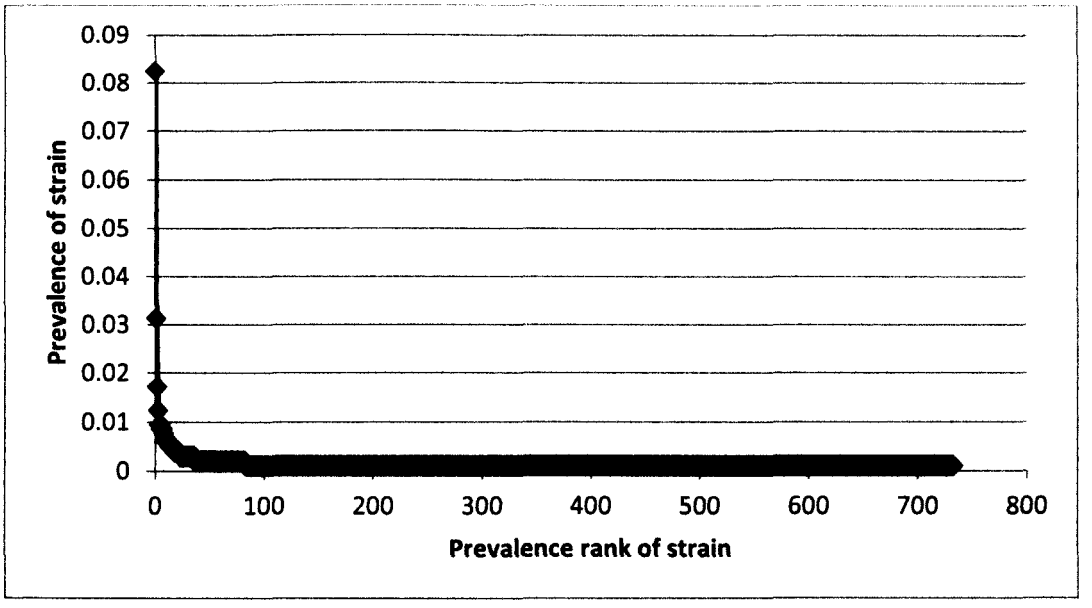


Figure 7-9: New strain type distribution for UK-born cases in the West Midlands, derived from extension of observed data by fitting a straight line to the last two points of the empirical distribution, resulting in 733 total distinct strains.

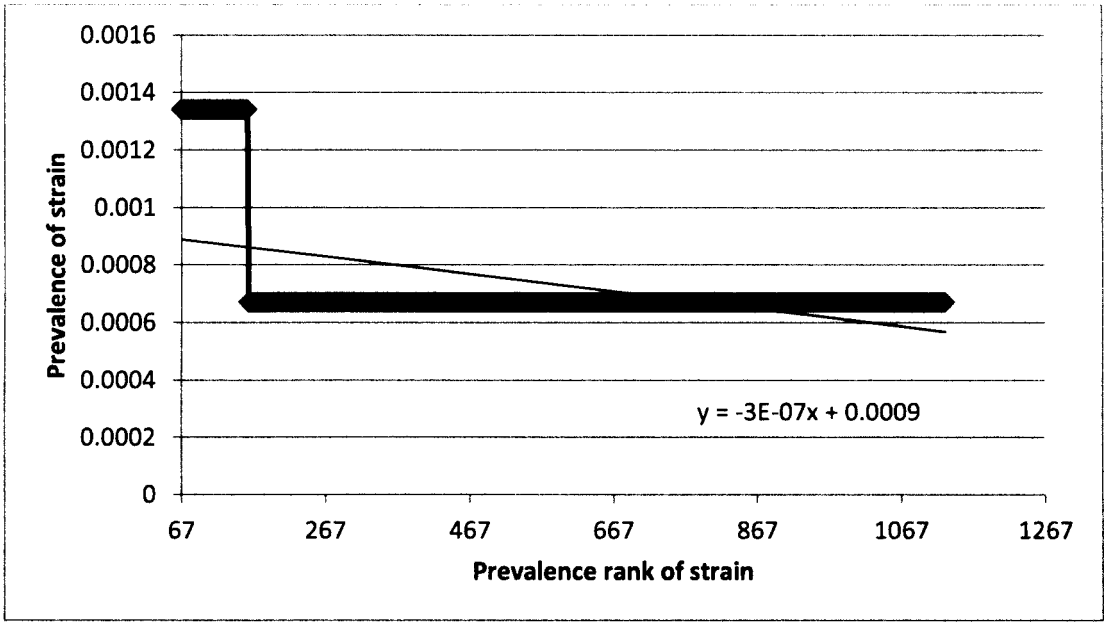


Figure 7-10: Excerpt from plot of strain type distribution observed for strains from foreign-born cases in the West Midlands, 2007 – 2011, with straight line fitted to the last two prevalence values. The fitted line results in an intercept of 3,000, or an estimated 3,000 total strains in the population, including the 1,128 observed strains and 1,872 unobserved strains.

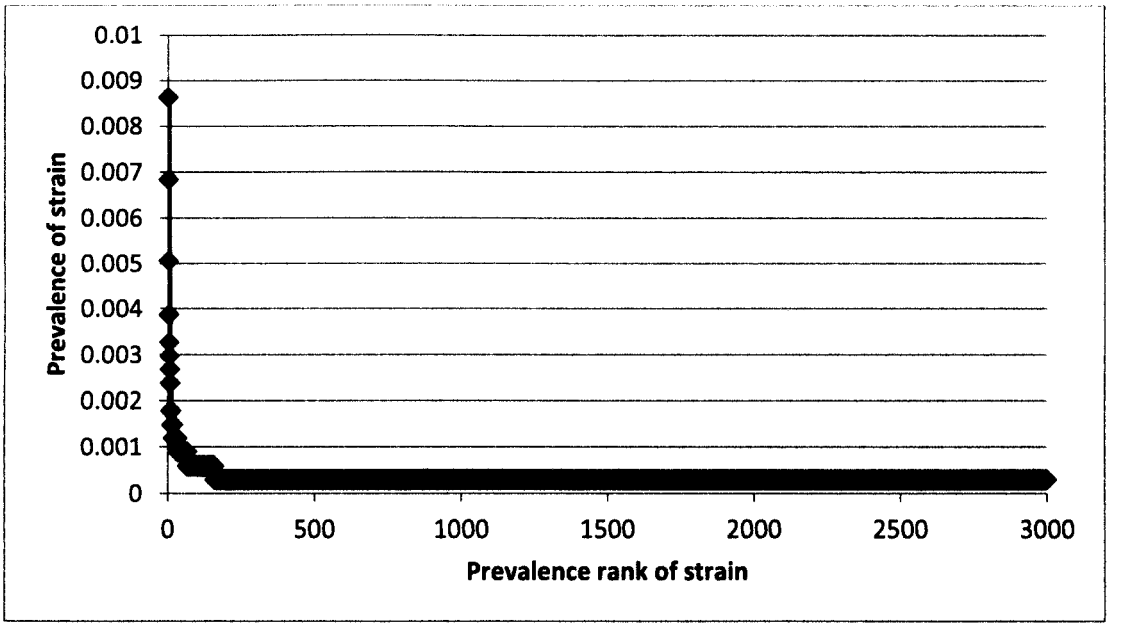


Figure 7-11: New strain type distribution for foreign-born cases in the West Midlands, derived from extension of observed data by fitting a straight line to the last two points of the empirical distribution, resulting in 3,000 total strains.

### *7.1.3.2 Mutation rate of 24-locus VNTR profiles*

As discussed in Chapter 2, Section 2.4.1.3.4, VNTR profile mutations are thought to occur due to replication errors. Slower replication of bacteria then leads to fewer mutations. In active disease, bacteria replicate at the fastest rate. It is thought that during *Latent Infection*, bacteria replicate much more slowly or even stop replicating [159]. In newer infections replication may be faster, including a period of rapid replication in primary infection, but this is likely to be on the order of weeks [337, 338]. Since *Recent Infection* and *Reinfection* states in the model include infections up to five years old, these are assumed to also have a reduced mutation rate, for simplicity equal to that in *Latent Infection*. Thus, it was assumed that two different mutation rates applied in the model, one for those with active disease and one for those with infection. Mutation rates for VNTR profiles in disease cases have been estimated in some recent studies, which are presented below. The mutation rate of VNTR loci involved in infection has not been studied, to my knowledge. Therefore, a study of the mutation rate of RFLP profiles in latent infection was compared to published estimates of the RFLP mutation rate in disease to estimate the relative rate of VNTR profiles mutation in infection, as described below.

#### 7.1.3.2.1 VNTR mutation rates in disease cases

At least four recent studies provide estimates for mutation rates of VNTR loci in disease cases. Generally, estimates were provided for the mutation rate per locus per year. These were then converted to the mutation rate as defined here, the rate of change per year for an entire 24-locus profile. The mutation rate for the entire 24-locus profile is given by the formula $1 - (1 - \mu_{locus})^{24}$, where $\mu_{locus}$ is the per locus per annum mutation rate.

The most recent estimates of VNTR mutation rates were reported by Aandahl et al., who analysed results of a tuberculosis transmission model combined with a strain mutation model in a Bayesian framework [157]. They provided estimates for two different versions of the strain mutation model. The 'linear' model assumed the mutation rate scales linearly with the number of repeats at a locus. The 'constant'

model assumed the mutation rate was independent of the number of repeats at a locus. Although authors concluded the linear model fits the data better, results from the constant model were used here since loci were not modelled individually. This estimate is $8.7 \times 10^{-3}$ mutations per locus per year, or for a 24-locus profile, a mutation rate of approximately 19% per year.

In 2010, Reyes and Tanaka estimated that the per-locus mutation rate is between $7 \times 10^{-4}$ and $1.5 \times 10^{-2}$ per year using a population genetic model, with estimates derived from relative genetic diversity observed between VNTR and other markers, plus assumptions about the mutation rates in those markers [155]. These estimates translated into per profile per year mutation rate estimates of approximately 1.7% and 30.4%. The midpoint between these estimates is a rate of change of about 17.2% per year for 24-locus profiles. The VNTR typing data used to derive estimates included 12 or more loci. It is likely that the average mutation rate per locus for 24-locus profiles was at the higher end of this spectrum since the additional loci used for 24-locus profiles are known to have increased diversity in copy numbers. More diversity in copy numbers generally means the locus is subject to higher mutation rates.

Wirth et al. looked at VNTR profiles from serial isolates of epidemiologically linked cases to estimate the per locus, per annum mutation rate of $1.4 \times 10^{-3}$ for five of the most variable loci [158]. They combined these data with a lower prior value for the mutation rate of $10^{-4}$, attempting to reflect the lower mutation rate for less variable loci in Bayesian inference to obtain a posterior estimate of the mutation rate of about $1.2 \times 10^{-4}$.

Lastly, Grant et al. used a mathematical modelling approach to estimate the mutation rate of VNTR loci per generation as $2.3 \times 10^{-8}$ [156]. This approach resulted in a per locus, per annum estimate of about $10^{-5}$ per locus per year or 0.0002 per profile per year. This estimate is far lower than those estimated in other work and does not explain the relatively common occurrence of mutation in VNTR profiles.

The four studies and results are summarized in Table 7-1, and two averages of the mutation rates are shown. The first is an average value for every rate reported in the table, while the second is the average which only considers one mutation rate from

each study. Those in bold were used as parameter values in the model, chosen to reflect the variety of rates estimated across studies.

Table 7-1: Mutation rates estimated for VNTR profiles from four recent studies. Studies generally provided per-locus, per-year rates of change, which were then converted to rates of change per profile, per-year for 24-locus VNTR.

| Study | Estimated mutation rate (per 24-locus profile per year) | Notes |
|---|---|---|
| Aandahl et al. 2012 [157] | **18.9%** | Value from constant model, since tuberculosis model does not model individual loci |
| Reyes and Tanaka 2010 [155] | 1.7% | Low estimate |
| Reyes and Tanaka 2010 [155] | 30.4% | High estimate |
| Reyes and Tanaka 2010 [155] | **17.2%** | Mid-value of low and high estimates |
| Wirth et al. 2008 [158] | **3.3%** | Estimate derived from data |
| Wirth et al. 2008 [158] | **0.3%** | Posterior estimate derived from data and a prior distribution to account for lower mutation rate in less variable loci |
| Grant et al. 2008 [156] | 0.02% | |
| Average, using all values | 10.3% | |
| Average, using one value per study | **9.9%** | |

#### 7.1.3.2.2 Relative mutation rates between infection and disease

Relative mutation rates between infection and disease for RFLP profiles were used in the model for relative rates between infection and disease for VNTR profiles. Lillebaek et al. compared 203 strains isolated from patients in the 1960s to over 4,000 strain isolates from patients in the 1990s to estimate that the mutation rate of strains in latent infection over long time periods was approximately 1.94% per year [130]. This mutation rate was compared with three studies estimating the mutation rate of RFLP profiles in disease cases [127-129], given an average ratio of the mutation rate in infection to the mutation rate in disease of 0.106, or 10.6%, as shown in Table 7-2. This value was fixed for all runs in this application of the simulation model.

Table 7-2: RFLP mutation rates and ratio of mutation rates in infection versus disease. The mutation rate in RFLP latent infection was estimated as 1.94% per profile per year by Lillebaek et al. [130]. Estimates of the mutation rates for disease cases are presented for three studies, with the average value in bold used in the model.

| Study | Mutation rate (per profile per year) | Infection mutation rate: disease mutation rate |
|---|---|---|
| Warren et al. 2002 [127] | 7.9% | 0.245 |
| de Boer et al. 1999 [128] | 21.7% | 0.090 |
| Rosenberg et al. 2003 [129] | 25.3% | 0.077 |
| **Average** | **18.3%** | **0.106** |

#### *7.1.3.3 Parameters obtained from England and Wales model fitting*

In the first instance, the four best scenarios obtained from fitting the England and Wales model to observed data were tested here, across variable mutation rates and assumptions for strain type distributions. The four scenarios include four disease risk parameters, contact rates, and assumptions about the infection status of migrants upon entry to the UK. The four scenarios specifying each of these parameters are indicated in Table 7-3.

Table 7-3: Best-fitting model input scenarios from England and Wales application of the model, which were used for baseline parameter values for the West Midlands application of the model. This table is an excerpt from (Table 5-22). These fits were obtained using a single foreign-born category during fitting. Scr1 and Scr2 are two different assumptions about the infection status of migrants to the UK, which were based on screening studies. See Chapter 5 for more details.

| | | | Disease risks for UK-born adult males by disease type | | | |
|---|---|---|---|---|---|---|
| Scenario | Infection status for migrants | Contact rate | Primary (%) | Reactivation (% per year) | Reinfection (%) | Foreign-born: UK-born |
| 3 | *Scr1* | 8 | 10.00% | 0.01% | 3.40% | 2.33 |
| 4 | *Scr1* | 10 | 7.50% | 0.02% | 1.20% | 2.59 |
| 8 | *Scr2* | 8 | 9.00% | 0.01% | 3.80% | 2.39 |
| 9 | *Scr2* | 10 | 7.40% | 0.01% | 2.60% | 2.44 |

### 7.1.4 Input Parameter Scenarios

#### *7.1.4.1 Stage one*

In stage one, 60 input scenarios were tested for comparison to observed data. These included the four input scenarios from England and Wales modelling shown in Table 7-3, each run with the five different mutation rates shown in bold in Table 7-1 and the three strain type distributions described in 7.1.3.1.

#### *7.1.4.2 Stage two*

Because runs of the model in stage one failed to achieve adequate fits to observed notification rates for several demographic categories, and because contact rate is one of the uncertain parameters that likely varies in different populations, a 50% increased contact rate was applied in stage two to improve correspondence with observed notification rates and clustering proportions. This increased contact rate was intended to reflect larger household sizes and different community contact structure than may have been found in England and Wales as a whole (see Section 7.3.5 for further discussion). Again, 60 input scenarios were tested, consisting of the four input scenarios from England and Wales modelling—with increased contact numbers—for each of the five different mutation rates and three strain type distributions.

In addition, the scenario with the best correspondence to observed data in stage two was tested with the contact rate reduced to the original value from the England and Wales model for each of the five mutation rates. Although problematic regarding fits to notification data, results were used to illustrate the effects of the increased contact rate on estimates of the proportion of cases due to recent transmission in the UK and the relationship between genotype clustering and recent transmission. This illustration also allowed for the comparison of results across the five mutation rates using a lower contact rate.

### 7.1.5 Model Output

Several types of model output were obtained for each of the input parameter scenarios tested. Notification rates per 100,000 were presented by age category, sex, and birthplace and compared to observed notification rates. The proportion clustered was presented for each age category, sex, and birthplace and also compared with

observed proportions clustered. In addition, the GOF of the model output to observed proportions clustered was calculated using the sum of least squares GOF statistic, reported to assist in assessing the fit of model output to observed clustering proportions. The 10 best-fitting scenarios defined by the lowest GOF statistics were used to summarize results.

Additional measures reported by the model included the proportion of cases infected recently in the UK for each age category, sex, and birthplace. As defined previously, recent infections were defined as those causing *Primary Disease* or *Reinfection Disease*, or infections that happened less than five years before disease onset. The positive and negative predictive values of clustering for identifying disease due to recent transmission in the UK were defined based on statistics used by Vynnycky et al. [161]. The statistics were defined slightly differently here because of interest in identifying cases due to both recent transmission and transmission occurring in the UK, rather than simply cases due to recent transmission. As such, the positive predictive value of clustering for identifying recent transmission was the proportion of cases that were in a cluster in a given period and had been infected or reinfected in the UK less than five years before disease onset. The negative predictive value of clustering was calculated as the proportion of cases that were not in a cluster in a given period that had been infected or last reinfected more than five years before disease onset, or were infected or last reinfected outside the UK. Both the positive and negative predictive values of clustering were reported for each age category, sex, and birthplace.

For each set of model parameters tested, 100 simulation runs were averaged and plotted for each type of model output. These were run using the computer hardware described in Chapter 3, Section 3.5.2.6.

## 7.2 Results

All runs of the model were first evaluated for correspondence with overall notification rates and clustering proportions, as well as the notification rates and clustering proportions by birthplace—UK-born and foreign-born—as a first step in assessment of model fits. Stage one runs of the model did not fit these overall proportions well. Notification rates for UK-born cases were lower than observed at around 4 per 100,000 population instead of the 6.5 per 100,000 observed. For UK-born males and females aged 15 – 44 years, notification rates were especially low compared to the observed rates, averaging around 4 per 100,000 population while the observed rate was around 9 per 100,0000 population. Foreign-born notification rates were also low, around 80 – 90 per 100,000 population instead of the 110 per 100,000 observed. Overall clustering proportions for initial scenarios were close to those observed for some of the scenarios for which the mutation rate was low, 3.3% or 0.3%. However, the clustering proportions were too low for most of the 60 scenarios tested.

Stage two runs of the model with the 50% increase in contact rate described earlier for all individuals fit observed notification and clustering data sufficiently well, though model output compared with observed notification rates by age, sex and birthplace categories shows some discrepancies. Rates for males aged 15 – 44 years were still lower than those observed, rates for females aged 45 – 64 years were high, and rates for foreign-born under 15 years of age were also high. Still, overall rates closely matched those observed and several model input scenarios fit the observed notification rates and clustering proportions roughly equally well.

Table 7-4 shows 10 of the best-fitting scenarios and resulting notification rates for UK-born and foreign-born cases. In those scenarios, the UK-born notification rates ranged from about 5 – 6 per 100,000 population, compared to 6.5 per 100,000 observed. The foreign-born rates ranged from 105 – 114 per 100,000 population, compared to 110 per 100,000 observed. Since this correspondence was deemed sufficient, the remainder of this section focuses on clustering and recent transmission output from the model.

Table 7-4: Results for simulated notification rates from the 10 best-fitting input parameter scenarios as measured by fit of simulated proportions clustered to observed proportions clustered. Input parameters and distributions include those taken from England and Wales modelling (infection status of migrants, disease risks as detailed in Table 7-3), a contact rate which was based on results from England and Wales simulations but increased by 50%, mutation rates based on literature values, and strain types distribution assumptions. Column headings which merit further description include: 1) Foreign:UK-born. This represents the ratio of disease risk in foreign-born individual to UK-born individuals and 2) The strain type distributions are named according to the total number of distinct strains in the total strain pool for UK-born at model initialization and for migrants entering the simulation at any time, including for foreign-born at model initialisation.

| | | Disease risk estimates from previous work | | | | | | | Notification rates per 100,000 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | Infection status of migrants | Primary | Reactivation (per year) | Reinfection | Foreign:UK-born | Contact rate | Mutation rate (per year) | Strain type distributions | Foreign-born | UK-born | All |
| 4 | Scr1 | 7.5% | 0.017% | 1.2% | 2.59 | 15 | 18.9% | UK-1000, Migrants-5000 | 114.6 | 6.0 | 15.9 |
| 6 | Scr1 | 7.5% | 0.017% | 1.2% | 2.59 | 15 | 17.2% | UK-1000, Migrants-5000 | 114.4 | 5.9 | 15.9 |
| 13 | Scr2 | 7.4% | 0.012% | 2.6% | 2.44 | 15 | 18.9% | UK-1000, Migrants-5000 | 105.1 | 4.6 | 13.8 |
| 33 | Scr2 | 9.0% | 0.012% | 3.8% | 2.39 | 12 | 18.9% | UK-1000, Migrants-5000 | 107.7 | 5.1 | 14.4 |
| 34 | Scr2 | 9.0% | 0.012% | 3.8% | 2.39 | 12 | 17.2% | UK-1000, Migrants-5000 | 114.1 | 6.0 | 15.9 |
| 36 | Scr1 | 10.0% | 0.012% | 3.4% | 2.33 | 12 | 18.9% | UK-1000, Migrants-5000 | 114.5 | 6.0 | 15.9 |

| | | Disease risk estimates from previous work | | | | | | | Notification rates per 100,000 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | Infection status of migrants | Primary | Reactivation (per year) | Reinfection | Foreign:UK-born | Contact rate | Mutation rate (per year) | Strain type distributions | Foreign-born | UK-born | All |
| 43 | Scr1 | 7.5% | 0.017% | 1.2% | 2.59 | 15 | 18.9% | UK-1000, Migrants-5000 | 105.0 | 4.6 | 13.7 |
| 44 | Scr2 | 9.0% | 0.012% | 3.8% | 2.39 | 12 | 18.9% | UK-1000, Migrants-5000 | 107.0 | 4.9 | 14.3 |
| 45 | Scr2 | 7.4% | 0.012% | 2.6% | 2.44 | 15 | 17.2% | UK-1000, Migrants-5000 | 105.2 | 4.6 | 13.8 |
| 46 | Scr1 | 7.5% | 0.017% | 1.2% | 2.59 | 15 | 17.2% | UK-1000, Migrants-5000 | 107.1 | 4.9 | 14.3 |
| Average | | | | | | | | | 109.5 | 5.2 | 14.8 |

### 7.2.1 Proportion Clustered

For each of the 10 scenarios with best correspondence to the observed age-specific proportions clustered, the simulated proportions clustered, and GOF statistics measuring the fit of these proportions to observed data are shown in Table 7-5. Trends in the age-specific proportions clustered, estimated proportion of cases due to recent transmission, and estimated predictive values of clustering were similar across the 10 best-fitting scenarios. For these reasons, age and sex-stratified results are discussed only for the best-fitting scenario, four, which is shown in Figure 7-12 – Figure 7-15. Bars 'a' show observed clustering proportions and bars 'b' show simulated clustering proportions, with dots above and below bar 'b' marking each of the 100 replicate runs of the model for the scenario. As shown in these figures, the simulation output did not differentiate clustering proportions between birthplaces and across age categories well. For UK-born cases, the model underestimated clustering in the young, though comes close to observed values in those aged 45 years and above. This trend was true for both males and females, though the underestimation of clustering proportions was worse for young males than young females. The overall percent of UK-born cases clustered in the simulation, 48%, was lower than the observed proportion of 65%.

For foreign-born cases, the simulation reproduced clustering proportions fairly well for males and less well for females. For all age categories apart from those aged 0 – 14 years, clustering proportions produced by the model were higher than those observed on average. For males, observed values generally fell within the range of 100 replicate runs. For females, observed values were generally lower than most of the 100 replicate runs for those age categories, apart from those aged 65 years and above. For those aged 0 – 14 years, the proportion clustered was underestimated by the model for both males and females, although replicate runs show there was uncertainty such that replicate runs ranged from 0 to 100%.

Table 7-5: Ten best-fitting scenarios for comparing clustering output with observed clustering proportions. Input parameters for each scenario can be found in Table 7-4. Simulated proportions clustered for each birthplace, UK-born and foreign-born are reported. The GOF statistic is also reported; this statistic is a measure of the fit of simulated clustering proportions to those observed using the sum of least squares, where the lower GOF indicates a better fit. Scenarios are ranked from best fit (1) to worst fit (10) of the 10 scenarios. Observed clustering was 46% overall, and 38% for foreign-born and 65% for UK-born.

| | Proportion clustered by birthplace | | | | |
|---|---|---|---|---|---|
| Scenario | Foreign | UK | All | GOF | GOF rank |
| 4 | 43% | 48% | 45% | 1.29 | 1 |
| 6 | 44% | 50% | 47% | 1.37 | 2 |
| 13 | 43% | 50% | 46% | 1.41 | 3 |
| 33 | 41% | 49% | 44% | 1.43 | 4 |
| 34 | 43% | 50% | 46% | 1.48 | 5 |
| 36 | 43% | 51% | 46% | 1.51 | 6 |
| 43 | 46% | 51% | 48% | 1.51 | 7 |
| 44 | 46% | 51% | 48% | 1.54 | 8 |
| 45 | 44% | 52% | 47% | 1.57 | 9 |
| 46 | 48% | 53% | 50% | 1.59 | 10 |
| Average | 44% | 51% | 47% | | |

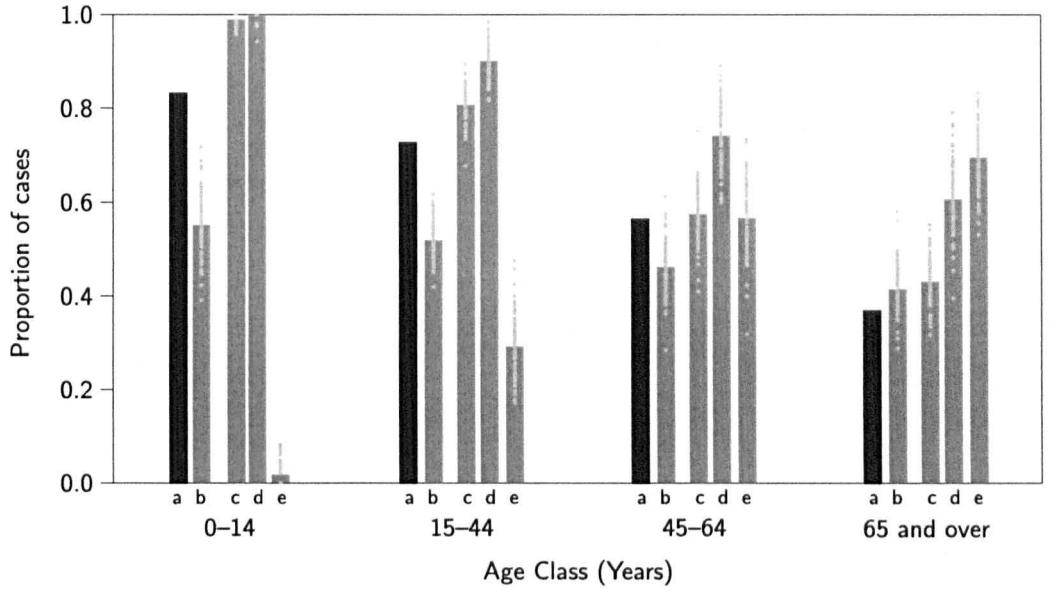


Figure 7-12: Observed and simulated clustering statistics by age category for UK-born males under model input scenario 4. Bars are labelled as follows: (a) observed proportion clustered, (b) simulated proportion clustered, where the bar depicts the mean and dots show individual runs of the simulation, (c) proportion of cases due to recent infection in the UK, (d) proportion of clustered isolates due to recent transmission in the UK, and (e) proportion of unique cases due to older infection or infection acquired abroad.

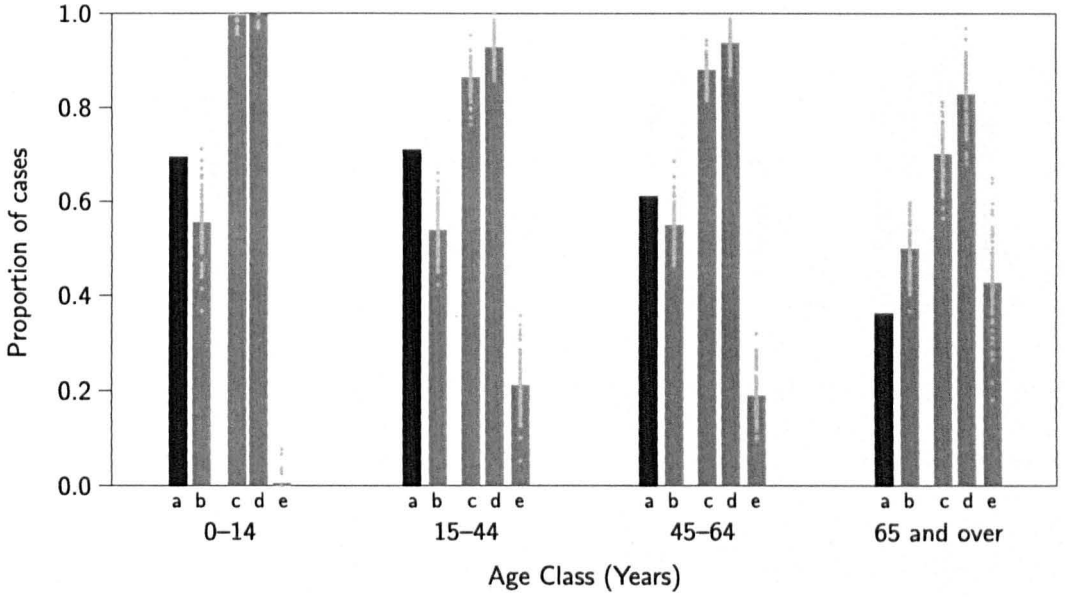


Figure 7-13: Observed and simulated clustering statistics by age category for UK-born females under model input scenario 4. Bars are as in Figure 7-12.

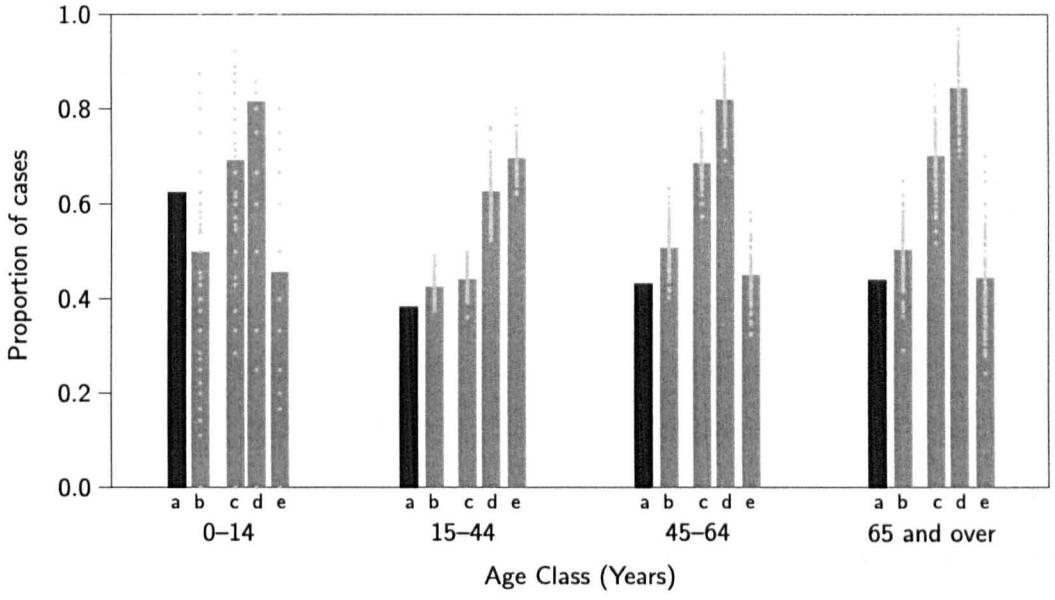


Figure 7-14: Observed and simulated clustering statistics by age category for foreign-born males under model input scenario 4. Bars are as in Figure 7-12.

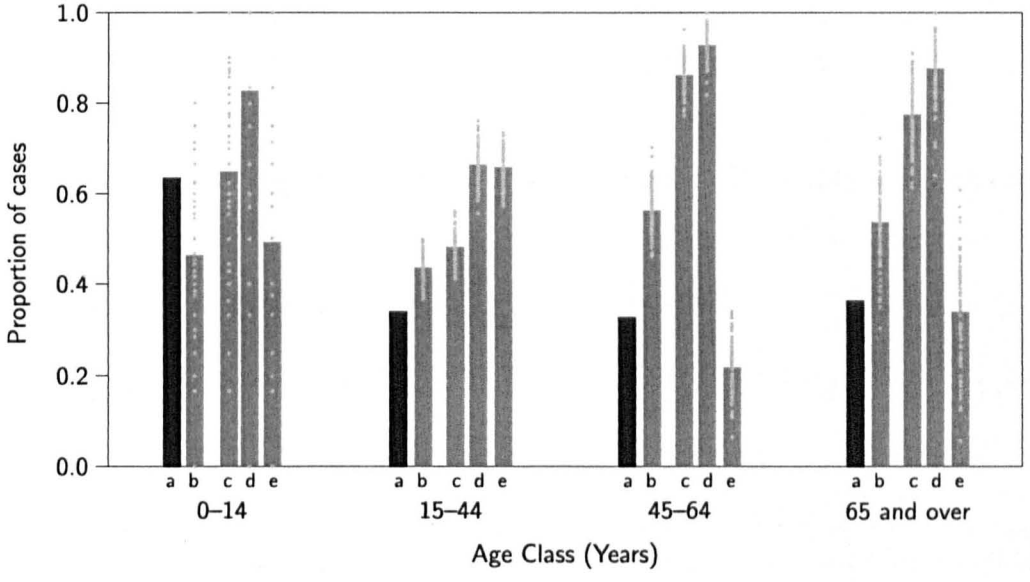


Figure 7-15: Observed and simulated clustering statistics by age category for foreign-born females under model input scenario 4. Bars are as in Figure 7-12.

### 7.2.2 Proportion of Cases Due to Recent Transmission

The estimated proportion of cases due to recent transmission, overall and for UK-born and foreign-born cases, are found in Table 7-6 for the 10 best-fitting scenarios. For scenario 4, the estimated proportion of cases due to recent transmission in the UK by age, sex and birthplace are shown in bars 'c' in Figure 7-12 – Figure 7-15. Overall, an estimated 63% of cases were due to recent transmission. Estimates differed between UK-born and foreign-born. For UK-born cases, around 76% on average were estimated to be due to recent transmission in the UK; this estimate was 56% on average for foreign-born cases. Age-dependent trends are clear in UK-born individuals, where the proportion of cases due to recent transmission was high in the youngest age class and decreased with increasing age. Almost 100% of cases in those aged 0 – 14 years were estimated to be due to recent transmission in the UK. On the other hand, 44% of cases in the oldest age class, those aged 65 years and above, were estimated to be due to recent transmission. For foreign-born cases, age-dependent trends in the proportion due to recent transmission in the UK were less clear. For males, age groups apart from those 15 – 44 years had similar estimated proportions, around 70%, while those aged 15 – 44 years had lower estimated proportions, just above 40%. For females, the two older age categories had a higher proportion of cases due to recent transmission, around 80%, than the younger age categories, at around 60%.

Table 7-6: Estimated proportion of cases due to recent transmission in the UK. Estimates obtained by fitting the model to 24-locus VNTR data from the West Midlands, 2007 – 2011. More information on parameter values for each scenario are found in Table 7-4.

| | Proportion of cases due to recent transmission | | |
|---|---|---|---|
| Scenario | Foreign-born | UK-born | All |
| 4 | 0.56 | 0.76 | 0.63 |
| 6 | 0.57 | 0.77 | 0.64 |
| 13 | 0.56 | 0.79 | 0.63 |
| 33 | 0.57 | 0.82 | 0.65 |
| 34 | 0.56 | 0.76 | 0.63 |
| 36 | 0.57 | 0.77 | 0.63 |
| 43 | 0.56 | 0.78 | 0.63 |
| 44 | 0.57 | 0.8 | 0.64 |
| 45 | 0.56 | 0.78 | 0.63 |
| 46 | 0.57 | 0.8 | 0.64 |
| Average | 0.57 | 0.78 | 0.64 |

Results from testing the best-fitting scenario, number four, with the reduced contact rate of 10 per year across the five values for the mutation rate of 24-locus profiles are found in Table 7-7. Although the scenario tested with a lowered contact rate was one of the initial fits that did not reproduce notification rates well, runs were tested to illustrate how conclusions may have changed if the contact rate had been lower. The results showed a decrease in the proportion of cases due to recent transmission with reduced contact rate, 45% overall. For foreign-born cases, the proportion due to recent transmission in the UK was estimated to be 39% and this proportion was 60% for UK-born. These values represent 30% and 20% reductions in the proportion of cases due to recent transmission for foreign-born and UK-born cases, respectively, with the reduced contact rate of 10 per year compared with 15 per year.

Table 7-7: Stage two fitting results of simulation runs using a modified version of scenario 4, using reduced contact rate of 10 per year across five values for the mutation rate. Contact rate was the same rate used in stage one fits.

| Input modification to scenario 4 | | Proportion clustered | | | | Proportion recent/UK | | |
|---|---|---|---|---|---|---|---|---|
| Mutation rate | Contact rate | Foreign-born | UK-born | All | GOF | Foreign-born | UK-born | All |
| 18.9% | 10 | 0.34 | 0.38 | 0.36 | 1.57 | 0.39 | 0.60 | 0.45 |
| 17.2% | 10 | 0.36 | 0.41 | 0.38 | 1.42 | 0.39 | 0.61 | 0.45 |
| 9.9% | 10 | 0.41 | 0.49 | 0.44 | 1.27 | 0.39 | 0.60 | 0.45 |
| 3.3% | 10 | 0.47 | 0.59 | 0.52 | 2.11 | 0.39 | 0.60 | 0.45 |
| 0.3% | 10 | 0.52 | 0.65 | 0.57 | 3.17 | 0.39 | 0.60 | 0.45 |

### 7.2.3 Predictive Value of Clustering

A comparison of simulated proportions clustered to the simulated proportion of cases due to recent transmission is found by comparing bar 'b' to bar 'c' for each age category, sex, and birthplace in Figure 7-12 – Figure 7-15. The simulated proportions clustered, shown in bars 'b', consistently underestimate the simulated proportion of cases due to recent transmission, shown in bars 'c', for all age groups, both sexes, and both birthplaces. These simulated proportions of cases clustered most notably underestimated the proportion of cases that were due to recent transmission in UK-born males and females aged less than 44 years, who represent the majority of cases in individuals born in the UK. The underestimation of the proportion of cases due to recent transmission was less severe for other categories. For foreign-born cases, the proportion of cases clustered was more similar to the proportion due to recent transmission for the 15 – 44 years age group, for both males and females, though still underestimated.

The positive predictive value of clustering for identifying recent transmission was calculated as the proportion of cases in a cluster that had been infected or reinfected in the UK within five years of disease onset in a given period. The positive predictive values of clustering varied with age and birthplace, as shown in bars 'd' for each birthplace, sex, and age category in Figure 7-12 – Figure 7-15. For UK-born cases, the positive predictive value of clustering decreased with age. This positive predictive value was very high for those aged 0 – 14 years and decreased to less than 60% for UK-

born males and to about 80% for UK-born females in the oldest age class, 65 years and above. For foreign-born cases, there was less of an age-dependent trend for the positive predictive value. The positive predictive value ranged from about 60 – 80% for males and females across age categories. There was a considerable amount of uncertainty in these estimates, indicated by the range covered by the 100 simulation replicates runs, each shown with a grey dot.

The negative predictive value of clustering was calculated as the proportion of cases not in a cluster in a given period that had been infected or last reinfected more than five years before disease onset or were infected or last reinfected outside the UK. The negative predictive value was lower for UK-born cases than for foreign-born cases, as shown in bars 'e' for each birthplace, sex, and age category in Figure 7-12 – Figure 7-15. For UK-born cases, this value increased dramatically with age for males, from nearly 0% in those aged 0 – 14 years to about 70% in those aged 65 years and above. The negative predictive value of clustering increased moderately with age for females, rising to just over 40% in the oldest age category. For foreign-born cases, age-dependent trends were less clear. For males, the negative predictive value of clustering was just above 40% for those aged 0 – 14 years, 45 – 64 years, and 65 years and above. For those aged 15 – 44 years, the negative predictive value was much higher at close to 70%. For females, a similar trend was observed, though there was more variation among age categories. Again, these estimates showed much uncertainty, indicated by the range covered by the 100 replicate runs of the simulation, each shown with a grey dot aligned with bars 'e'.

Results differed somewhat for the modified versions of scenario 4 with a reduced contact rate. As shown in Figure 7-16 – Figure 7-19, clustering did not always underestimate the proportion of cases due to recent transmission. The proportion clustered underestimated the proportion of cases due to recent transmission in the UK for all UK-born females and UK-born males under age 45 years. For foreign-born cases, clustering underestimated the proportion of cases due to recent transmission for all age categories apart from those aged 15 – 44 years.

Also shown in Figure 7-16 – Figure 7-19, the positive and negative predictive values of clustering also differed for the modified scenario. Trends in the positive predictive value of clustering were similar between the scenarios, though this values fell more

dramatically with age for UK-born cases in the modified scenario with a reduced contact rate, which resulted in a lower average value for the modified scenario. The age-dependent trends were less clear for foreign-born cases and generally similar those in the higher contact rate scenarios, though overall values were lower. For UK-born cases, trends in the negative predictive value of clustering were similar to those seen in the higher contact scenario, with a clear increase with increasing age groups for males and females. However, values were generally higher in the lower contact rate scenario. For foreign-born cases, again, age-dependent trends were not very clear but similar to patterns seen in the higher contact rate scenarios. Also, values were generally higher in the lower contact rate scenario.

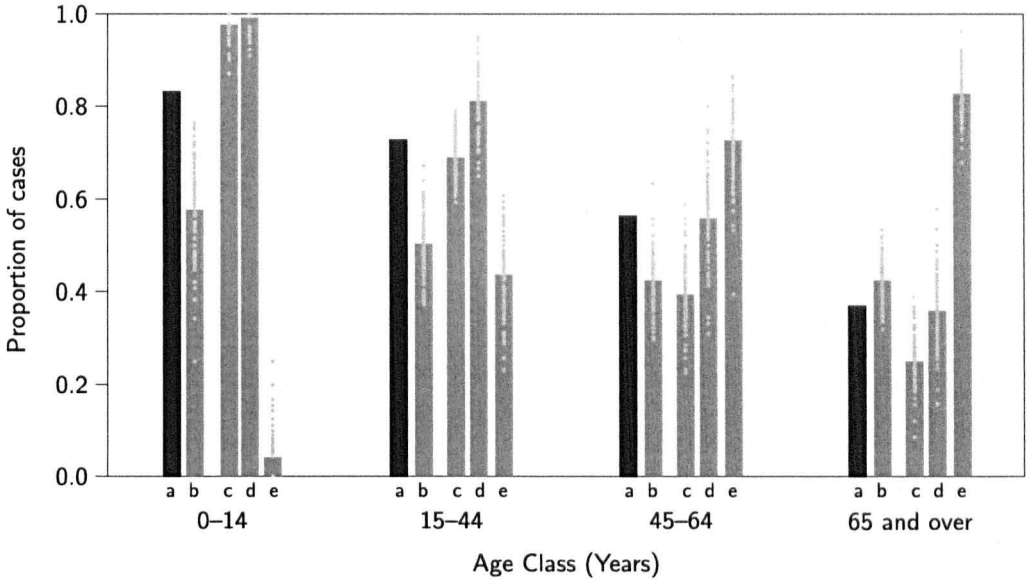


Figure 7-16: Observed and simulated clustering statistics by age category for UK-born males under a modified version of model input scenario 4, with a contact rate of 10 per year and a mutation rate of 9.9% per year. Bars are labelled as follows: (a) observed proportion clustered, (b) simulated proportion clustered, where the bar depicts the mean and dots show individual runs of the simulation, (c) proportion of cases due to recent infection in the UK, (d) proportion of clustered isolates due to recent transmission in the UK, and (e) proportion of unique cases due to older infection or infection acquired abroad.

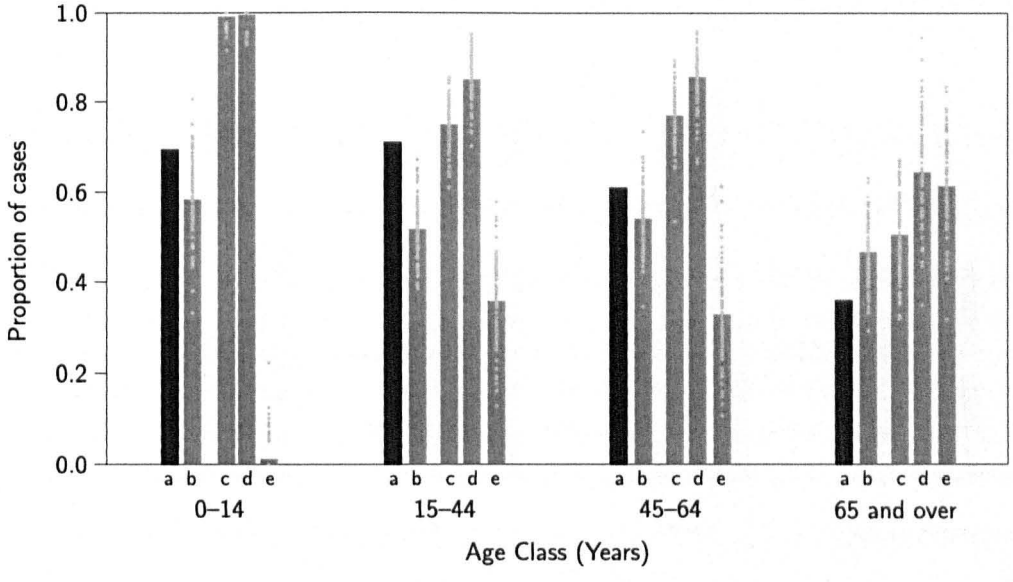


Figure 7-17: Observed and simulated clustering statistics by age category for UK-born females under a modified version of model input scenario 4, with a contact rate of 10 per year and a mutation rate of 9.9% per year. Bars are as in Figure 7-16.


Figure 7-18: Observed and simulated clustering statistics by age category for foreign-born males under a modified version of model input scenario 4, with a contact rate of 10 per year and a mutation rate of 9.9% per year. Bars are as in Figure 7-16.

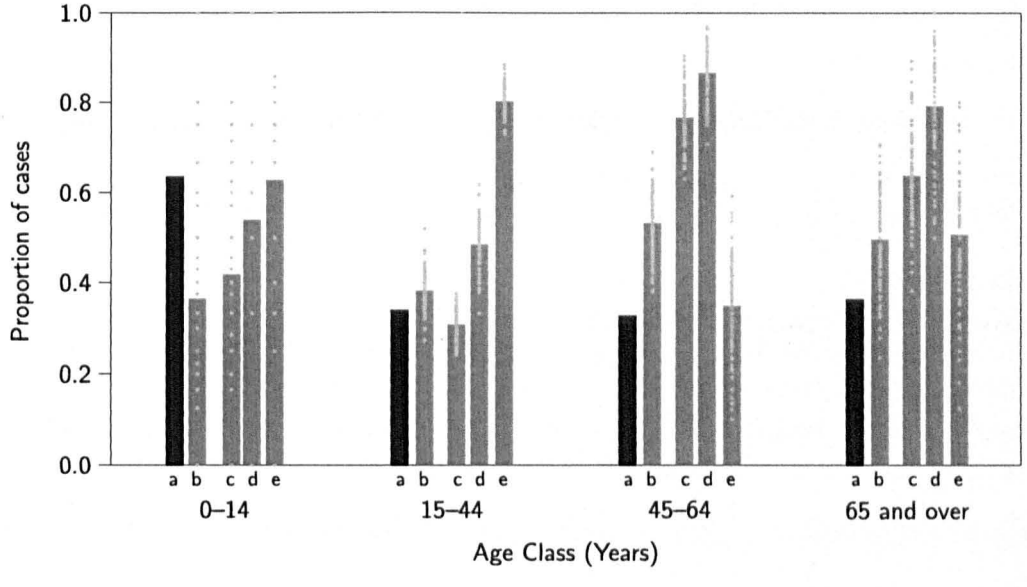


Figure 7-19: Observed and simulated clustering statistics by age category for foreign-born females under a modified version of model input scenario 4, with a contact rate of 10 per year and a mutation rate of 9.9% per year. Bars are as in Figure 7-16.

### 7.2.4 Mutation Rate, Strain Type Distributions, and Other Inputs

Mutation rates in the 10 best scenarios for correspondence to observed data included only the two highest of the five values tested, 17.2% and 18.9% per year. The lower mutation rates, especially the two lowest values tested, 3.3% and 0.3% per year, seemed to be incompatible with the observed clustering proportions across a range of input parameter values. However, results were sensitive to the contact rate assumed in the model. This was illustrated by results from runs using a reduced contact rate for best-fitting scenario 4 for each of the five mutation rates. With the reduced contact rate, the best-fitting mutation rate for scenario 4 was 9.9%, lower than the 18% in the higher contact rate version. Results are shown in Table 7-7.

The strain type distribution assumptions across each of the 10 the best-fitting scenarios were the two higher-diversity assumptions of the three distributions tested. However, the mutation rate was inversely correlated with the amount of diversity in the strain type distribution used. Increased mutation rates improved the fits for the lower diversity strain type assumptions.

Among best-fitting scenarios, the input scenarios from England and Wales modelling did not change simulation results much. These were comprised of disease risk estimates, the contact rate—which was increased for this stage of fitting, and the infection status of migrants.

## 7.3 Discussion

In stage two fitting, the model reproduced observed notification rates well due to the increased contact number. The overall proportions clustered in the simulation also corresponded with observed data well overall, though there were some discrepancies when data were stratified into age and sex classes. The proportion of cases due to recent transmission in the UK was estimated to be 63% across 10 of the best-fitting scenarios of stage two, though could be closer to 45% if the contact rate were lowered. The best-fitting scenarios resulted from high mutation rates for 24-locus VNTR profiles, 17-19% per year.

### 7.3.1 Fits to Clustering Proportions

Across several scenarios, the overall proportions clustered predicted by the model corresponded well to the overall proportions clustered in observed data. General trends, including the reduction in proportion clustered with increasing age for UK-born cases was present. However, age-dependent proportions clustered were not reproduced perfectly by the model. For foreign- born cases, observed clustering proportions were high for those aged 0 – 14 years, but roughly equal for other age groups. The trends in model estimates were similar to these, though lower than observed for those aged 0 – 14 years. It is likely that more fully age-specific disease risks and contact structure would have helped fit to age classes more closely. Risk of disease may be elevated in older individuals [28, 29, 223], whereas here, it was assumed that all individuals aged 20 years and over had the same risks of disease. In addition, contact rates would likely be lower for older individuals, though here were assumed equal across age groups [339, 340]. Also unrealistic was the assumption of homogeneous mixing among age groups, which, in reality, would likely be strongly associative by age [273, 340]. If, for example, higher reactivation risks and lower contact rates for the elderly were used, this age group would likely have had a lower clustering proportion and lower proportion due to recent transmission. Likewise, if the younger individuals had higher contact rates and lower reactivation disease risks, they would have had increased proportions clustered and proportion of cases due to recent transmission. These would have been more consistent with observed data.

Results also showed less differentiation between UK-born and foreign-born proportions clustered than was seen in observed data. This finding might have been due to inaccurate assumptions regarding the strain type distributions. For example, with increased diversity in the assumed foreign-born strain pool and decreased diversity in the UK-born strain pool, there would be more differentiation in clustering proportion between the groups. This outcome could also have been impacted by migration data and assumptions about the infection status of migrants upon entry to the UK. Increased migration of infected and diseased foreign-born individuals would have decreased clustering in this group, assuming there was sufficient strain diversity in the foreign-born strain pool.

### 7.3.2 Proportion of Cases Due to Recent Transmission in the UK

The estimated proportion of cases due to recent transmission in the UK, 63% overall, was higher than previous estimates found in genotyping studies in the UK [166, 308]. However, these estimates based on genotyping data would be expected to underestimate the proportion of cases due to recent transmission due to sampling bias [8, 163]. Also, although estimates obtained here were high, they still fell within the range of estimates of the proportion of cases due to recent transmission estimated from RFLP genotyping studies in developed countries [325], as discussed further in Section 6.3.

In any case, estimates obtained in these analyses could be thought of as an upper bound on the proportion of cases due to recent transmission in the UK. This consideration is because, in order to achieve adequate model fits to observed notification rates in the West Midlands, the contact rate alone was increased. In reality there are other factors that could have been altered to achieve the observed notification rates in simulation results. Other solutions which would not have increased the proportion of cases due to recent transmission as much, for example, modelling of individuals travelling to high burden countries, assuming increased reactivation risk in the elderly, or assuming higher infection and disease prevalence for migrants upon entry to the UK. In reality, a mixture of these and other factors was likely important. But since only the contact number was raised to achieve the observed notification rates in the simulation, it is likely that the estimates of recent transmission

from the model were at the upper limit of the true proportions because of this high contact rate.

The results from runs with a reduced contact number, as expected, showed a reduced proportion of cases due to recent transmission, 45% on average. These results may be interpreted as a lower bound of the estimates for the proportion of cases due to recent transmission in the UK. The estimates may accurately reflect the proportion of cases due to recent transmission in the UK if model parameters other than the contact rate were altered to achieve fits to observed notification data, as discussed above.

### 7.3.3 Predictive Value of Clustering

The relationship between strain type clustering and the proportion of cases due to recent transmission in the UK in the simulation results showed that the proportion clustered consistently underestimated the proportion of cases due to recent transmission. This finding raises concerns about the utility of clustering data if mutation rates are as high as the estimates of 17.2% and 18.9% per year, which led to best fits of the model to observed data.

These conclusions are sensitive to the contact rate, however. When the contact rate was decreased to 10 per year, the proportions clustered were better estimates of the true proportion of cases due to recent transmission because the mutation rate which allowed best correspondence to observed data had decreased. There was not a consistent underestimate of the proportion of cases due to recent transmission, and the age-dependent effects were more similar to those found in previous work [9]. However, for foreign-born aged 15-44 years, this value was again low, which is problematic since these represent the largest group of cases.

The positive predictive values of clustering showed clear age-dependent trends for UK-born cases, with the positive predictive value decreasing with age and the negative predictive value increasing with age. These findings are similar to trends found in previous work looking at the positive predictive value among different population trends in the ARI [161]. It appeared, however, that the magnitude of increase and decrease in these values by age category matched better to settings with high, constant ARIs rather than ARIs that have declined, as in the Netherlands [161]. This may have been due to the higher contact rate assumed. Direct comparisons were

difficult because of the different age categories used. Trends for the predictive values of clustering for foreign-born cases did not show clear age-dependency, nor correspondence to any of the ARI scenarios presented by Vynnycky et al. [161]. This is not completely unexpected, as a mixture of native-born and foreign-born were modelled here, whereas that study did not consider immigrants.

### 7.3.4 Mutation Rate, Strain Type Distributions, and Other Inputs

Results showed that only the highest mutation rates for 24-locus profiles, about 17 – 19% per year, and highest diversity assumptions for strain type distributions fit observed data best.

When the lower contact version of scenario 4 was run, the average mutation rate across literature values, 9.8%, fit the data fairly well, indicating that as contact rate assumptions change, lower mutation rates are possible. Still, these results showed that even under reduced contact rate scenarios , the lowest mutation rate estimates from the literature, 0.03 – 0.3% per year, resulted in clustering proportions that were too high. This finding suggests that these estimates may be implausibly low. It should be noted that correlation between the strain type diversity assumptions and the mutation rate means that, without a more precise and accurate estimate of one of them, conclusions about each independently should be interpreted with caution.

### 7.3.5 Limitations

The work described in this chapter has several limitations. Firstly, the initial comparison of model output to notification data showed poor correspondence. It is recognized that there are differences in the resident populations and tuberculosis epidemiology between the West Midlands and other areas of the UK, including England and Wales, which likely explain discrepancies between initial outputs of the model and observed data. As discussed in Chapter 2, the West Midlands has a large population of South Asian individuals, including UK-born South Asians. Although at a lower risk of infection than those born in South Asia, UK-born South Asians may have more exposure to *M. tuberculosis* than other UK-born individuals through travel to South Asia, from household contacts born in South Asia, and from more contacts with South Asian-born individuals in their communities. A 2007 report showed that those of

South Asian descent in Birmingham live in households of larger size than other demographic groups [341].

Although the increased contact number in stage two fitting led to correspondence between simulated and observed notification rates in the West Midlands, this was only one solution to obtain correspondence, of which many possible others could be explored. One major factor that may relate to the discrepant fits includes the model's sensitivity to migration data, which were particularly uncertain for the West Midlands. Until better data are available, it may be better to model larger regions when more genotyping data are available, and where migration estimates would be based on larger and more robust sample sizes. It is also possible that infection status of migrants upon entry to the UK could be different than for England and Wales, for example, there might be higher infection and disease prevalence in migrants to the region. Another factor may be that case reporting completeness in the West Midlands is better than in England and Wales because of the higher notification rates in the West Midlands. Clinicians in this region are likely more aware of tuberculosis than clinicians in other regions, due to the high notification rates and large South Asian population. A higher proportion of cases notified in the model would increase the notification rates estimated at least somewhat. Other characteristics of the model structure which could be amended for this region include incorporating UK-born travel to South Asia or other high burden areas. It is possible that exposure abroad contributes to increased tuberculosis incidence in some ethnic groups, such as South Asians.

If any one or more of these factors could have been explored, it is possible that the contact rate would not need to be increased as much to account for differences in epidemiology between the West Midlands and England and Wales; a lower contact rate would have likely reduced the proportion of cases due to recent transmission in the UK. These modifications could also allow the model to fit to clustering proportions with lower mutation rates and strain type distributions with reduced strain diversity.

Further limitations include several model assumptions that could be made more realistic. There is no age-dependent mixing implemented or other more realistic contact structures taken into account. This work suggests a high contact rate may be necessary for the model to reproduce the West Midlands epidemiology generally. However, it is possible that the contact rate is high for some groups and moderate or

low for others, including the elderly. A heterogeneous contact rate may also allow a better fit to observed age-dependent clustering patterns, though could also change conclusions. In addition, other contact heterogeneities taken into account in tuberculosis models include a household structure for contact patterns. There could also be a more sophisticated handling of the variance in contact rate from individual to individual.

In addition, several parameter values are highly uncertain or otherwise problematic. Migration data are particularly uncertain for the West Midlands and the model is particularly sensitive to these inputs. Unless more precise migration data are made available, results must be interpreted with caution. The population size of the region means that population categories divided by age, sex and birthplace have a relatively small number of data points per category, especially for foreign-born individuals, which causes several problems. The notification rates per 100,000 may be subject to error due to inaccurate population sizes for some demographic categories, particularly for the foreign-born population. Also, the other data, such as proportion clustered in these groups, will be uncertain.

Modelling methods and model structure issues to be considered for future research include the modelling of individual loci, which would be more realistic and is a natural extension of the work done here. Travel could be added to the model–especially in the West Midlands where there are many UK-born people of South Asian descent, it is possible some of these are travelling back to high-incidence countries and being exposed to *M. tub*erculosis there. Finally, formal fitting methods could have been used to estimate the mutation rate and possibly other parameters. Also, an uncertainty analysis could be formalized and extended.

### 7.3.6 Implications

This work is the first application of an IBM of tuberculosis dynamics to help better understand observed genotyping data, and, to my knowledge, the first application of a model to simulate observed genotyping data in a mixed population of native-born and foreign-born individuals. This analysis suggests a high proportion of cases that are due to recent transmission in the UK and that clustering proportions consistently underestimate the proportion of disease due to recent transmission in the UK.

However, conclusions are sensitive to the contact rate and this value, along with other factors that influence the epidemiology of tuberculosis in the West Midlands, should be better understood before making definitive conclusions. These analyses are intended to be a first exploration into the IBM fit to genotyping data. Several lines of future work are opened by this study.

# 8 Conclusions

This chapter concludes the thesis with a unifying discussion of the study. The aim and objectives are reviewed and the main findings reported. The specific new contributions to the field are highlighted. Broader implications of the work, study limitations, and suggested future work are also discussed.

## 8.1 Introduction

The primary aim of the thesis is to better understand tuberculosis dynamics and epidemiology in the UK. Modelling approaches were used to achieve the following six study objectives:

1) Construct an IBM of tuberculosis dynamics in the UK that is freely available for others to use and documented according to a standardized protocol
2) Estimate risks of disease for those with a recent infection, recent reinfection, and latent infection in both UK-born and foreign-born individuals using the model applied to notification data from England and Wales, 1999 – 2009. Also identify plausible assumptions for the effective contact rate and the infection status of migrants entering the UK
3) Estimate the proportion of cases due to recent transmission in the UK by application of the model to data from England and Wales.
4) Describe the molecular epidemiology of tuberculosis the West Midlands from 2007 – 2011
5) Estimate the proportion of tuberculosis cases due to recent transmission in the UK by application of the model to genotyping data from the West Midlands and explore the relationship between recent transmission in the UK and genotype clustering
6) Identify plausible assumptions about the mutation rate for 24-locus VNTR, as well as the strain type diversity for 24-locus VNTR profiles found in UK-born and foreign-born individuals by application of the model to genotyping data from the West Midlands

## 8.2 Summary of Findings

As covered in Chapter 3, objective one was accomplished by developing the model, describing it according to the ODD protocol and publishing the code in the appendix of this thesis and online (http://www.cbs.umn.edu/modeling).

Objectives two and three were addressed in Chapter 5 by fitting the model to notification data from England and Wales, 1999 – 2009. For objective two, disease risk estimates were lower than earlier estimates for UK-born adult males, perhaps reflecting a decrease in disease risk following infection for UK-born individuals in recent years. Foreign-born disease risks were consistently higher than those of UK-born. Best fits of the model show foreign-born disease risks are on average 2.4 times higher than risks for UK-born. The model fits suggest the contact rate was between eight and 10 effective contacts per year for infectious cases. Plausible scenarios for the infection status of migrants upon entry to the UK were identified. For objective three, estimates of the proportion of cases due to recent transmission in the UK were around 42 – 47% on average for best-fitting scenarios, which is higher than previous estimates from genotyping-based studies in the UK.

In Chapter 6, objective four was addressed. The molecular epidemiological analysis of 24-locus VNTR data from the West Midlands identified risk factors for clustering. The identified risk factors were largely consistent with findings from earlier RFLP studies. The results suggest 24-locus VNTR is useful for population-level analyses of risk factors for clustering. Lastly, the estimate for the proportion of cases due to recent transmission using genotyping data alone was 36% overall.

The final two objectives of the study were addressed in Chapter 7 by application of the model to genotyping data from the West Midlands, 2007 – 2011. The proportion of tuberculosis cases due to recent transmission in the UK were estimated to be in the range of 45 – 63%. Again, these are higher than previous estimates for the UK. Simulations also suggested genotyping data were likely to consistently underestimate the proportion of disease due to recent transmission, though this conclusion is sensitive to model parameters. The mutation rates that best fit these data were 17 – 19% per 24-locus VNTR profile per year, though the rate could be as low as 10% per

profile per year, depending on model assumptions. Plausible strain type distributions for UK-born and foreign-born cases were identified.

## 8.3 Original Contributions

Several new contributions to the field resulted from this study. The first contribution is the presentation of a new, freely available, rigorously tested model of tuberculosis dynamics, which is adaptable to suit other applications. Other contributions include the furthering of the understanding of tuberculosis epidemiology in the UK. These include: estimates of disease progression risks for foreign-born and UK-born individuals in recent years; estimates of the proportion of disease due to recent transmission in the UK for recent years using three different approaches; an exploratory appraisal of genetic typing methods for estimating the extent of recent transmission in a population comprised of foreign-born and UK-born individuals; and a population-scale molecular epidemiological analysis using 24-locus VNTR from the West Midlands. Lastly, methodological contributions that have stemmed from the work in this thesis include several recently published algorithms useful for simulation models generally [234-236, 252, 342].

## 8.4 Broader Implications of the Work

There are three main areas where this work has broader implications, 1) foundations for future modelling studies, 2) identifying priorities for data collection, and 3) disease prevention and control.

Because simulation modelling requires enormous effort, creating model software which is rigorously tested, well documented, and freely available is beneficial and important for epidemiology for several reasons, 1) this helps ensure the model and conclusions drawn from its use are correct 2) this helps ensure others can reproduce the work and 3) this allows others to build on the work, rather than starting anew. Especially for the latter reason, the impact of releasing model software will perhaps be longer lasting than conclusions regarding tuberculosis epidemiology. Also, the new published algorithms add to the computing literature and may be useful other simulations in epidemiology and other disciplines.

In addition to model development, parameterization of the model was a considerable effort because incorporating detailed data, especially when distributions of times to event are necessary, is difficult. For future models, some of the parameters estimated or derived in this study may be useful. These include disease risks, the effective contact rate, the mutation rate of 24-locus VNTR profiles, strain type distributions, and plausible assumptions for the infection status of migrants to the UK. Another part of model parameterization is the organization of the thousands of data points used as input parameters to generate these time-to-event distributions. The use of specialized software for helping to ensure input data are read and documented correctly is important. This work motivated the creation of a system designed for reading and documenting input files, which is presented in a published paper I contributed to [252].

Also related to model parameterization, this study identified some areas where data could be strengthened. Population sizes based on LFS data are highly uncertain for small demographic categories, for example, many estimates for age groups in the West Midlands and for Sub-Saharan Africans by age and sex categories throughout the UK. Migration data are also unreliable for many demographic categories, especially estimates for the West Midlands generally and some categories for England and Wales. These are based on the IPS, which only samples a fraction of travellers. Both types of data are important for the epidemiology and modelling of several diseases and for other concerns of the country. Such data are expensive to collect, but the benefits may be worth the cost of expanding the surveys. Also, although plausible assumptions for the infection status of migrants were established in the study, updated and improved screening data to better inform these parameters would be beneficial as these are critical for the understanding of tuberculosis epidemiology. Similarly, although results from this work showed that two of the highest and most recent estimates for the mutation rate of VNTR profiles were most plausible [155, 157], these conclusions are sensitive to other model parameters. Further study and data on mutation rates is recommended because of the importance to both model conclusions and the utility of genotyping data.

Lastly, for tuberculosis prevention and control, results have several potential implications. Estimates of the proportion of cases due to recent transmission were generally higher than estimated elsewhere, between 45 – 63% on average, depending on model parameters. However, most previous estimates were based on crude estimates from genotyping data, which are likely to underestimate the proportion of disease due to recent transmission. Furthermore, especially within high-risk subgroups of the population, it is possible that individuals are exposed to a very high risk of infection and this would explain the large estimated proportion of cases due to recent transmission in the UK. Even the lower estimates of this range indicate a substantial number of cases due to recent transmission in the UK, and that efforts to stop further transmission are worthwhile.

The interruption of transmission can be achieved by reducing the period of infectiousness of active cases by treating as many cases as possible, as early and effectively as possible. This involves first identifying infectious cases, which can be achieved through active case finding. That can involve mobile screening units or screening programs in prisons, homeless shelters, or other high-risk populations. Confirmed cases would then be given treatment effectively and followed up to ensure treatment is administered properly and effectively. These and other such interventions could be compared with cost-effectiveness analyses taking into account results from this study. In addition, estimates of the proportion of disease due to recent transmission established here can be used for monitoring the change in these estimates over time and may help evaluate the success of interventions. Estimates may also help predict future trends in tuberculosis incidence.

Results showed that disease risks were consistently higher in foreign-born individuals, which may have implications for treatment and control. For example, prophylactic treatment of latent infection in these groups may be more beneficial than in the general population, because a higher risk of disease means the average number of cases prevented would be higher for any number of treatments administered. They would also be more likely to benefit an individual patient. Additionally, the higher disease risk may change the cost-effectiveness of the intervention and so should be considered in cost-effectiveness studies.

Conclusions about the utility of 24-VNTR genotyping for identifying recent transmission showed genotyping data were likely to underestimate the proportion of cases due to recent transmission, confirming results of previous work. However, the descriptive molecular epidemiological analysis suggests both 15 and 24-locus VNTR are useful for identifying risk factors for clustering, and, to some extent, recent transmission on a population level. This analysis highlights the utility of genotyping data and the benefits of the national strain typing service in the UK. Another benefit of the strain typing service demonstrated in this study is the usefulness of the genotyping data in modelling studies to help better understand tuberculosis epidemiology.

## 8.5 Limitations

Limitations are discussed in each chapter, but some of the most important and generally applicable limitations of the study are reviewed here.

First and foremost, this work is limited by the quality of the data used for model inputs. There is a need for more and better-quality data, including migration data and population size estimates for some subgroups of the population. Migration data and population sizes stratified by age, sex, and birthplace are particularly important for understanding patterns in tuberculosis dynamics and are uncertain for many demographic categories. Assumptions about the infection status of migrants were used to determine the proportion of immigrants in each of the disease compartments when they arrive in the UK. Thus, estimates will impact model output, as more infections and disease entering the UK means more cases from these sources and generally lower infection and disease risks *within* the country. Other model assumptions come from parameter values which are uncertain. For example, vaccination was assumed to impart lifelong immunity for those vaccinated effectively, which is likely overly simplistic. The relative risk ratio for HIV infected persons was fixed at 7.0 for all disease types and HIV-positive persons, but this assumption is also overly simplistic. Even if HIV infection is not explicitly modelled, the relative risk ratio could be varied to reflect the changing risk ratio for different stages of HIV infection, different *M. tuberculosis* infection status, effects of HIV treatment, and other risk factors for HIV-associated tuberculosis. Similarly, disease risks in females were fixed by

relative risk ratios based on previous studies, since some of these values are quite low and higher values could be explored.

The 24-locus VNTR data used for the molecular epidemiological analysis and for comparison to model output for the West Midlands model were problematic due to a low case ascertainment. Less than 50% of notified cases were typed. This low percentage was mainly due to a low proportion of cases with a positive culture, due to both a low proportion of cases that were pulmonary, but also missing laboratory data on some pulmonary cases. In addition, the geographical area of the study population was also restricted. Both of these limitations lower estimates of the proportion of cases due to recent transmission in the UK using genotyping data and could help explain why model-based estimates were higher. Low case ascertainment could also mean that some risk factors for clustering were missed.

Another limitation of the study is the equal contact rate for all individuals and the nearly-homogeneous contact structure assumed in the model. These are overly simplistic. In reality, the contact rate will likely depend on age and other factors and true contact patterns reveal associative mixing, both age-dependent mixing and a contact structure which depends on households, schools, and workplaces. None of these factors were taken into account in the model, although they likely have an impact on age-dependent disease risk estimates, estimates for the proportion of cases due to recent transmission, and the interpretation of genotyping data. Also related to contacts and risk of infection is that the risk of travel to high-burden countries by UK-born and foreign-born individuals was ignored. The effects of this travel may be important sources of infection risk that could be modelled.

Another limit is that the effect of HIV is included in a simplistic manner. Because an increasing number of immigrants are co-infected with HIV, future models could take co-infection into account more fully. Unfortunately, HIV testing of tuberculosis patients is not routine, and therefore, HIV prevalence in tuberculosis cases is hard to estimate. Further research should explore the impact of HIV on tuberculosis incidence in the UK. Antibiotic treatment of tuberculosis and drug resistance of strains were also not explicitly included in this study, though are not central to the conclusions herein.

Some of the challenges in obtaining adequate fits of the model to observed data indicate there are features of the tuberculosis epidemiology, especially in some subgroups of the population, that were not fully understood. Furthermore, fitting to many age, sex, and birthplace categories was problematic given that input data did not fully specify parameter values for each of these categories.

## 8.6 Areas for Further Research

Several lines of future work are inspired by the findings in this thesis.

With the appropriate data, the model could be further stratified to take into account more detailed birthplace or ethnic groupings to better understand the differences in these subgroups for the epidemiology of tuberculosis in the UK. This stratification would almost certainly require more refined migration data and population size estimates than are currently available. More broadly, the model is designed to be flexible enough to take into account almost any type of data, for example, individual-level host genetic factors or health status information, so that infection and disease risks are more specific to individuals.

Other structural improvements to the model for future studies include explicit spatial structure and contact networks, both within the scope of this kind of model. A more refined contact network would greatly improve the model and more accurately reflect the heterogeneous *M. tuberculosis* infection risks for different subgroups of the population in the UK. Another extension of the model structure could include incorporation of more sophisticated modelling of strain type mutation, including the modelling of each locus individually, or even incorporating new understanding of *M. tuberculosis* mutation processes based on whole genome sequencing, once those data are available. Lastly, explicit treatment of co-infections, such as HIV, which would be natural extensions of the model, are worth exploring.

Another future application of the model could test how well the model fits data from other developed countries to compare and contrast drivers of tuberculosis incidence among countries. In the first instance, this comparison could be done by changing population sizes and migration data but keeping the majority of parameters the same. Some countries in Western Europe and the United States and Canada have migration patterns similar to those found in England and Wales, with migrants entering from both countries with both low-burdens and high-burdens of tuberculosis.

Also, it should soon be possible to extend molecular epidemiological analyses and modelling to a larger geographical and temporal scale, as universal prospective 24-locus VNTR typing is now implemented across England and Wales. With more

genotyping data, the conclusions here could be tested and could contribute to the evaluation of the strain typing service. Eventually, it would be interesting to model molecular epidemiology of tuberculosis worldwide. The expansion from the 60-million individuals this model presently handles for the UK to six billion or more for the entire globe is only about two orders of magnitude larger than the present simulations. At the historical growth rate for computing capacity of doubling every 18 months or so, it will only be ten years until that two orders of magnitude can be readily accommodated on supercomputers. The algorithmic methods employed in this study are capable of handling such an increase with only linear increases in memory and computational time.

Finally, a large-scale model such as this, with some categories with millions of individuals behaving deterministically and others with only a few individuals behaving stochastically, could be used for developing and evaluating new computer algorithms beyond the several that have already emerged from the study --- for example adaptive algorithms for state-space searching to support faster and more reliable parameter fitting, and new parallel-coded algorithms for event-based simulation.

## 8.7 Concluding Remarks

Although there are many dedicated clinicians, prevention and control scientists, and researchers working to reduce the tuberculosis burden in the UK and worldwide, the disease remains a problem. This modelling study provides some insight into the difficult-to-measure and unobservable features of tuberculosis dynamics in the UK, and helps advance tools for studying the disease in hopes of informing prevention, control, and future research. Ultimately, this research is motivated by saving and improving lives.

# 9 References

1. HPA, *Tuberculosis in the UK: Annual report on tuberculosis surveillance in the UK 2008*, 2008, Health Protection Agency Centre for Infections: London.
2. MRC, *National survey of notifications of tuberculosis in England and Wales in 1988. Medical Research Council Cardiothoracic Epidemiology Group.* Thorax, 1992. **47**(10): p. 770-5.
3. NICE, *Tuberculosis: Clinical Diagnosis and Management of Tuberculosis, and Measures for Its Prevention and Control*, 2011: London.
4. Dasgupta, K., Menzies, D. , *Cost-effectiveness of tuberculosis control strategies among immigrants and refugees.* European Respiratory Journal, 2005. **25**(6): p. 1107-1116.
5. Horsburgh, C.R., *Priorities for the Treatment of Latent Tuberculosis Infection in the United States.* New England Journal of Medicine, 2004. **350**: p. 2060-2067.
6. Glynn, J.R., et al., *Interpreting DNA fingerprint clusters of Mycobacterium tuberculosis. European Concerted Action on Molecular Epidemiology and Control of Tuberculosis.* The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease, 1999. **3**(12): p. 1055-60.
7. van der Spuy, G.D., P.D. van Helden, and R.M. Warren, *Effect of study duration on the interpretation of tuberculosis molecular epidemiology investigations.* Tuberculosis, 2009. **89**(3): p. 238-242.
8. Murray, M., *Sampling bias in the molecular epidemiology of tuberculosis.* Emerging Infectious Diseases, 2002. **8**(4): p. 363-369.
9. Vynnycky, E., et al., *The effect of age and study duration on the relationship between 'clustering' of DNA fingerprint patterns and the proportion of tuberculosis disease attributable to recent transmission.* Epidemiology and infection, 2001. **126**(1): p. 43-62.
10. Sepkowitz, K.A., *How contagious is tuberculosis?* Clinical infectious diseases : an official publication of the Infectious Diseases Society of America, 1996. **23**(5): p. 954-62.
11. Schluger, N.W. and W.N. Rom, *The host immune response to tuberculosis.* Am J Respir Crit Care Med, 1998. **157**(3 Pt 1): p. 679-91.
12. Chiang, C.Y. and L.W. Riley, *Exogenous reinfection in tuberculosis.* Lancet Infectious Diseases, 2005. **5**(10): p. 629-636.
13. Styblo, K., *Epidemiology of tuberculosis*, 1991, Royal Netherlands Tuberculosis Association: The Hague.
14. Sepkowitz, K.A., *How contagious is tuberculosis?* Clinical Infectious Diseases, 1996. **23**(5): p. 954-962.
15. Styblo, K., *The relationship between the risk of tuberculous infection and the risk of developing infectious tuberculosis.* Bull Int Union Tuberc, 1985. **60**(3-4): p. 117-119.

16. Vynnycky, E. and P.E. Fine, *Lifetime risks, incubation period, and serial interval of tuberculosis.* American journal of epidemiology, 2000. **152**(3): p. 247-63.
17. Comstock, G.W., *Untreated inactive pulmonary tuberculosis. Risk of reactivation.* Public Health Rep, 1962. **77**: p. 461-70.
18. Sutherland, I., E. Svandova, and S. Radhakrishna, *The Development of Clinical Tuberculosis Following Infection with Tubercle-Bacilli .1. A Theoretical-Model for the Development of Clinical Tuberculosis Following Infection, Linking from Data on the Risk of Tuberculous Infection and the Incidence of Clinical Tuberculosis in the Netherlands.* Tubercle, 1982. **63**(4): p. 255-268.
19. Vynnycky, E. and P.E. Fine, *The natural history of tuberculosis: the implications of age-dependent risks of disease and the role of reinfection.* Epidemiology and infection, 1997. **119**(2): p. 183-201.
20. Erkens, C.G., et al., *Tuberculosis contact investigation in low prevalence countries: a European consensus.* The European respiratory journal : official journal of the European Society for Clinical Respiratory Physiology, 2010. **36**(4): p. 925-49.
21. Lin, P.L. and J.L. Flynn, *Understanding latent tuberculosis: a moving target.* J Immunol, 2010. **185**(1): p. 15-22.
22. Barry, C.E., 3rd, et al., *The spectrum of latent tuberculosis: rethinking the biology and intervention strategies.* Nature reviews Microbiology, 2009. **7**(12): p. 845-55.
23. Wang, P.D., *Epidemiological trends of childhood tuberculosis in Taiwan, 1998-2005.* Int J Tuberc Lung Dis, 2008. **12**(3): p. 250-4.
24. Horsburgh, C.R., et al., *Revisiting Rates of Reactivation Tuberculosis A Population-based Approach.* American Journal of Respiratory and Critical Care Medicine, 2010. **182**(3): p. 420-425.
25. Nagelkerke, N.J., et al., *Tuberculosis and sexually transmitted infections.* Emerg Infect Dis, 2004. **10**(11): p. 2055-6.
26. Glynn, J.R., *Resurgence of tuberculosis and the impact of HIV infection.* British medical bulletin, 1998. **54**(3): p. 579-93.
27. Perez-Velez, C.M. and B.J. Marais, *CURRENT CONCEPTS Tuberculosis in Children.* New England Journal of Medicine, 2012. **367**(4): p. 348-361.
28. Stead, W.W. and J.P. Lofgren, *Does the Risk of Tuberculosis Increase in Old-Age.* Journal of Infectious Diseases, 1983. **147**(5): p. 951-955.
29. Rajagopalan, S., *Tuberculosis and aging: A global health problem.* Clinical Infectious Diseases, 2001. **33**(7): p. 1034-1039.
30. Johnston, B. and J. Conly, *The changing face of Canadian immigration: Implications for infectious diseases.* Can J Infect Dis Med Microbiol, 2008. **19**(4): p. 270-2.
31. Lorant, V., H. Van Oyen, and I. Thomas, *Contextual factors and immigrants' health status: Double jeopardy.* Health & Place, 2008. **14**(4): p. 678-692.

32. Cantwell, M.F., et al., *Tuberculosis and race/ethnicity in the United States - Impact of socioeconomic status.* American Journal of Respiratory and Critical Care Medicine, 1998. **157**(4): p. 1016-1020.
33. Bellamy, R., *Susceptibility to mycobacterial infections: the importance of host genetics.* Genes Immun, 2003. **4**(1): p. 4-11.
34. Kruijshaar, M.E. and I. Abubakar, *Increase in extra pulmonary tuberculosis in England and Wales 1999 - 2006.* Thorax, 2009.
35. Musellim, B., et al., *Comparison of extra-pulmonary and pulmonary tuberculosis cases: factors influencing the site of reactivation.* Int J Tuberc Lung Dis, 2005. **9**(11): p. 1220-3.
36. Cowie, R.L. and J.W. Sharpe, *Extra-pulmonary tuberculosis: a high frequency in the absence of HIV infection.* Int J Tuberc Lung Dis, 1997. **1**(2): p. 159-62.
37. Peto, H.M., et al., *Epidemiology of extrapulmonary tuberculosis in the United States, 1993-2006.* Clinical infectious diseases : an official publication of the Infectious Diseases Society of America, 2009. **49**(9): p. 1350-7.
38. Cailhol, J., B. Decludt, and D. Che, *Sociodemographic factors that contribute to the development of extrapulmonary tuberculosis were identified.* J Clin Epidemiol, 2005. **58**(10): p. 1066-71.
39. Forssbohm, M., et al., *Demographic characteristics of patients with extrapulmonary tuberculosis in Germany.* The European respiratory journal : official journal of the European Society for Clinical Respiratory Physiology, 2008. **31**(1): p. 99-105.
40. Kempainen, R., et al., *Mycobacterium tuberculosis disease in Somali immigrants in Minnesota.* Chest, 2001. **119**(1): p. 176-180.
41. Yang, H., et al., *Tuberculosis in Calgary, Canada, 1995-2002: site of disease and drug susceptibility.* The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease, 2005. **9**(3): p. 288-93.
42. Zhang, X., et al., *Effect of sex, age, and race on the clinical presentation of tuberculosis: a 15-year population-based study.* Am J Trop Med Hyg, 2011. **85**(2): p. 285-90.
43. Harries, A.D., *Tuberculosis and human immunodeficiency virus infection in developing countries.* Lancet, 1990. **335**(8686): p. 387-90.
44. Shafer, R.W., et al., *Extrapulmonary tuberculosis in patients with human immunodeficiency virus infection.* Medicine (Baltimore), 1991. **70**(6): p. 384-97.
45. Teale, C., J.M. Goldman, and S.B. Pearson, *The Association of Age with the Presentation and Outcome of Tuberculosis - a 5-Year Survey.* Age and Ageing, 1993. **22**(4): p. 289-293.
46. Borgdorff, M.W., et al., *Mortality among tuberculosis patients in the Netherlands in the period 1993-1995.* European Respiratory Journal, 1998. **11**(4): p. 816-820.

47. Fielder, J.F., et al., *A high tuberculosis case-fatality rate in a setting of effective tuberculosis control: implications for acceptable treatment success rates.* International Journal of Tuberculosis and Lung Disease, 2002. **6**(12): p. 1114-1117.
48. Bakhshi, S.S., J. Hawker, and S. Ali, *Tuberculosis mortality in notified cases from 1989-1995 in Birmingham.* Public Health, 1998. **112**(3): p. 165-168.
49. Vasankari, T., et al., *Treatment outcome of extra-pulmonary tuberculosis in Finland: a cohort study.* Bmc Public Health, 2010. **10**: p. 399.
50. Walpola, H.C., et al., *Tuberculosis-related deaths in Queensland, Australia, 1989-1998: characteristics and risk factors.* The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease, 2003. **7**(8): p. 742-50.
51. Glynn, J.R., et al., *Tuberculosis and survival of HIV-infected individuals by time since seroconversion.* Aids, 2010. **24**(7): p. 1067-1069.
52. Rutledge, C. and J. Crouch, *The ultimate results in 1694 cases of tuberculosis treated at the Modern Woodmen of America Sanatorium.* American Review of Tuberculosis, 1919. **2**: p. 755–756.
53. Brewer, T.F. and S.J. Heymann, *Long time due: reducing tuberculosis mortality in the 21st century.* Archives of medical research, 2005. **36**(6): p. 617-21.
54. Stead, W.W., *The origin and erratic global spread of tuberculosis. How the past explains the present and is the key to the future.* Clin Chest Med, 1997. **18**(1): p. 65-77.
55. Frieden, T.R., B.H. Lerner, and B.R. Rutherford, *Lessons from the 1800s: tuberculosis control in the new millennium.* Lancet, 2000. **355**(9209): p. 1088-1092.
56. Vynnycky, E. and P.E. Fine, *Interpreting the decline in tuberculosis: the role of secular trends in effective contact.* International journal of epidemiology, 1999. **28**(2): p. 327-34.
57. Fine, P.E.M., *VARIATION IN PROTECTION BY BCG - IMPLICATIONS OF AND FOR HETEROLOGOUS IMMUNITY.* Lancet, 1995. **346**(8986): p. 1339-1345.
58. Comstock, G.W. and C.E. Palmer, *Long-term results of BCG vaccination in the southern United States.* Am Rev Respir Dis, 1966. **93**(2): p. 171-83.
59. Baily, D.V.J., *Trial of Bcg Vaccines in South-India for Tuberculosis Prevention.* Indian Journal of Medical Research, 1979. **70**(Sep): p. 349-363.
60. Trunz, B.B., P. Fine, and C. Dye, *Effect of BCG vaccination on childhood tuberculous meningitis and miliary tuberculosis worldwide: a meta-analysis and assessment of cost-effectiveness.* Lancet, 2006. **367**(9517): p. 1173-80.

61. Hart, P.D. and I. Sutherland, *BCG and vole bacillus vaccines in the prevention of tuberculosis in adolescence and early adult life.* Br Med J, 1977. **2**(6082): p. 293-5.
62. Winston, C.A. and K. Mitruka, *Treatment duration for patients with drug-resistant tuberculosis, United States.* Emerg Infect Dis, 2012. **18**(7): p. 1201-2.
63. Burgos, M., et al., *Treatment of multidrug-resistant tuberculosis in San Francisco: an outpatient-based approach.* Clinical infectious diseases : an official publication of the Infectious Diseases Society of America, 2005. **40**(7): p. 968-75.
64. Davies, P.D., *Drug-resistant tuberculosis.* J R Soc Med, 2001. **94**(6): p. 261-3.
65. O'Brien, R., *The treatment of tuberculosis*, in *Tuberculosis: A Comprehensive International Approach*, L.H. Reichman, E. , Editor 1993, Marecel Dekker: New York. p. 207-240.
66. WHO, *WHO report on the tuberculosis epidemic 1997: Use DOTS more widely*, 1997, World Health Organization: Geneva.
67. WHO, *The Stop TB Strategy: Building on and enhancing DOTS to meet the TB-related Millennium Development Goals*, 2006, World Health Organization, Stop TB Partnership: Geneva.
68. WHO, *Global tuberculosis control : epidemiology, strategy, financing*, in *WHO Report*2009, World Health Organization: Geneva, Switzerland.
69. in *Tuberculosis: Clinical diagnosis and management of tuberculosis, and measures for its prevention and control*2006: London.
70. Evans, C.A., *GeneXpert--a game-changer for tuberculosis control?* PLoS Med, 2011. **8**(7): p. e1001064.
71. Orlando, G., et al., *Interferon-gamma releasing assay versus tuberculin skin testing for latent tuberculosis infection in targeted screening programs for high risk immigrants.* Infection, 2010. **38**(3): p. 195-204.
72. Lalvani, A. and M. Pareek, *A 100 year update on diagnosis of tuberculosis infection.* British medical bulletin, 2010. **93**: p. 69-84.
73. Rieder, H.L., *Methodological issues in the estimation of the tuberculosis problem from tuberculin surveys.* Tuber Lung Dis, 1995. **76**(2): p. 114-21.
74. Menzies, D., *Interpretation of repeated tuberculin tests. Boosting, conversion, and reversion.* Am J Respir Crit Care Med, 1999. **159**(1): p. 15-21.
75. Harling, R., et al., *Tuberculosis screening of asylum seekers: 1 years' experience at the Dover Induction Centres.* Public Health, 2007. **121**(11): p. 822-7.
76. Mazurek, G.H., et al., *Guidelines for using the QuantiFERON-TB Gold test for detecting Mycobacterium tuberculosis infection, United States.* MMWR Recomm Rep, 2005. **54**(RR-15): p. 49-55.

77. Chin, D.P., et al., *Differences in contributing factors to tuberculosis incidence in U.S. -born and foreign-born persons.* American journal of respiratory and critical care medicine, 1998. **158**(6): p. 1797-803.
78. de Vries, G.v.H., R. A. H., Richardus, J.H., *Impact of Mobile Radiographic Screening on Tuberculosis among Drug Users and Homeless Persons.* American Journal of Respiratory and Critical Care Medicine, 2007. **176**(2): p. 201-207.
79. Schwartzman, K., et al., *Domestic returns from investment in the control of tuberculosis in other countries.* N Engl J Med, 2005. **353**(10): p. 1008-20.
80. *Global Tuberculosis Report 2012*, 2012, World Health Organization: Geneva, Switzerland.
81. Zumla, A., et al., *Impact of HIV infection on tuberculosis.* Postgrad Med J, 2000. **76**(895): p. 259-68.
82. Bauer, J., et al., *Results from 5 years of nationwide DNA fingerprinting of Mycobacterium tuberculosis complex isolates in a country with a low incidence of M-tuberculosis infection.* Journal of Clinical Microbiology, 1998. **36**(1): p. 305-308.
83. Rieder, H.L., et al., *Tuberculosis control in Europe and international migration.* Eur Respir J, 1994. **7**(8): p. 1545-53.
84. Schneider, E. and K.G. Castro, *Tuberculosis trends in the United States, 1992-2001.* Tuberculosis (Edinb), 2003. **83**(1-3): p. 21-9.
85. WHO, *Global Tuberculosis Control: Surveillance, Planning, Financing.* , in *WHO Report*2003, World Health Organization: Geneva, Switzerland

.
86. Raviglione, M.C., et al., *Secular trends of tuberculosis in western Europe.* Bull World Health Organ, 1993. **71**(3-4): p. 297-306.
87. Gilbert, R.L., et al., *The impact of immigration on tuberculosis rates in the United Kingdom compared with other European countries.* International Journal of Tuberculosis and Lung Disease, 2009. **13**(5): p. 645-651.
88. Diel, R., S. Rusch-Gerdes, and S. Niemann, *Molecular epidemiology of tuberculosis among immigrants in Hamburg, Germany.* Journal of Clinical Microbiology, 2004. **42**(7): p. 2952-2960.
89. McKenna, M.T., E. McCray, and I. Onorato, *The epidemiology of tuberculosis among foreign-born persons in the United States, 1986 to 1993.* N Engl J Med, 1995. **332**(16): p. 1071-6.
90. Cowie, R.L. and J.W. Sharpe, *Tuberculosis among immigrants: interval from arrival in Canada to diagnosis. A 5-year study in southern Alberta.* CMAJ, 1998. **158**(5): p. 599-602.
91. Borgdorff, M.W., et al., *Tuberculosis elimination in the Netherlands.* Emerg Infect Dis, 2005. **11**(4): p. 597-602.
92. Lillebaek, T., et al., *Persistent high incidence of tuberculosis in immigrants in a low-incidence country.* Emerg Infect Dis, 2002. **8**(7): p. 679-84.

93. Fanning, E.A., *Globalization of tuberculosis.* CMAJ, 1998. **158**(5): p. 611-2.
94. Brudey, K., et al., *Molecular epidemiology of Mycobacterium tuberculosis in western Sweden.* Journal of Clinical Microbiology, 2004. **42**(7): p. 3046-3051.
95. Haase, I., et al., *Use of geographic and genotyping tools to characterise tuberculosis transmission in Montreal.* International Journal of Tuberculosis and Lung Disease, 2007. **11**(6): p. 632-638.
96. Zhou, Y.C., et al., *Projection of tuberculosis incidence with increasing immigration trends.* Journal of Theoretical Biology, 2008. **254**(2): p. 215-228.
97. Prevention, C.f.D.C.a., *Reported Tuberculosis in the United States, 2011*, 2011, U.S. Department of Health and Human Services: Atlanta, GA.
98. Falzon, D. and F. Belghiti, *Tuberculosis: still a concern for all countries in Europe.* Euro Surveill, 2007. **12**(3): p. E070322 1.
99. Mulder, C., E. Klinkenberg, and D. Manissero, *Effectiveness of tuberculosis contact tracing among migrants and the foreign-born population.* Euro Surveill, 2009. **14**(11).
100. Pareek, M., et al., *Evaluation of immigrant tuberculosis screening in industrialized countries.* Emerg Infect Dis, 2012. **18**(9): p. 1422-9.
101. Liu, Y., et al., *Overseas screening for tuberculosis in U.S.-bound immigrants and refugees.* The New England journal of medicine, 2009. **360**(23): p. 2406-15.
102. Lowenthal, P., et al., *Reduced importation of tuberculosis after the implementation of an enhanced pre-immigration screening protocol.* The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease, 2011. **15**(6): p. 761-6.
103. Wilson, L.G., *The historical decline of tuberculosis in Europe and America: its causes and significance.* J Hist Med Allied Sci, 1990. **45**(3): p. 366-96.
104. McKeown, T. and R.G. Record, *Reasons for the decline of Mortality in England and Wales during the nineteenth century.* Population Studies, 1962. **16**: p. 94-122.
105. *Tuberculosis in the UK: Annual report on tuberculosis surveillance in the UK*, 2012, Health Protection Agency: London.
106. Roe, J.T., *Tuberculosis in Indian immigrants.* Tubercle, 1959. **40**: p. 387-8.
107. Springett, V.H., et al., *Tuberculosis in immigrants in Birmingham, 1956-1957.* Br J Prev Soc Med, 1958. **12**(3): p. 135-40.
108. BTA, *Tuberculosis among immigrants to England and Wales: A national survey in 1965.* Tubercle, 1966. **47**.
109. Rose, C., *Tuberculosis at the end of the 20th century in England and Wales: results of a national survey in 1998 (vol 56, pg 173, 2001).* Thorax, 2001. **56**(6): p. 504-504.

110. ONS. http://www.statistics.gov.uk. 2009 November 2009].
111. *Office for National Statistics.* 2010]; Available from: http://www.statistics.gov.uk.
112. Kruijshaar, M.E., et al., *Evidence for a national problem: continued rise in tuberculosis case numbers in urban areas outside London.* Thorax, 2012. **67**(3): p. 275-277.
113. Ditah, I.C., et al., *Monitoring tuberculosis treatment outcome: analysis of national surveillance data from a clinical perspective.* Thorax, 2008. **63**(5): p. 440-446.
114. Pareek, M., et al., *Tuberculosis screening of migrants to low-burden nations: insights from evaluation of UK practice.* The European respiratory journal : official journal of the European Society for Clinical Respiratory Physiology, 2011. **37**(5): p. 1175-82.
115. Hogan, H., et al., *Screening of new entrants for tuberculosis: responses to port notifications.* J Public Health (Oxf), 2005. **27**(2): p. 192-5.
116. Supply, P., et al., *Automated high-throughput genotyping for study of global epidemiology of Mycobacterium tuberculosis based on mycobacterial interspersed repetitive units.* Journal of Clinical Microbiology, 2001. **39**(10): p. 3563-3571.
117. Salamon, H., et al., *Genetic distances for the study of infectious disease epidemiology.* American Journal of Epidemiology, 2000. **151**(3): p. 324-334.
118. Barnes, P.F. and M.D. Cave, *Molecular epidemiology of tuberculosis.* The New England journal of medicine, 2003. **349**(12): p. 1149-56.
119. Murray, M. and E. Nardell, *Molecular epidemiology of tuberculosis: achievements and challenges to current knowledge.* Bull World Health Organ, 2002. **80**(6): p. 477-82.
120. Walker, T.M., et al., *Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study.* Lancet Infect Dis, 2012.
121. Schurch, A.C. and D. van Soolingen, *DNA fingerprinting of Mycobacterium tuberculosis: From phage typing to whole-genome sequencing.* Infection Genetics and Evolution, 2012. **12**(4): p. 602-609.
122. Hanekom, M., et al., *Discordance between mycobacterial interspersed repetitive-unit-variable-number tandem-repeat typing and IS6110 restriction fragment length polymorphism genotyping for analysis of Mycobacterium tuberculosis Beijing strains in a setting of high incidence of tuberculosis.* Journal of Clinical Microbiology, 2008. **46**(10): p. 3338-3345.
123. Vanembden, J.D.A., et al., *Strain Identification of Mycobacterium-Tuberculosis by DNA Fingerprinting - Recommendations for a Standardized Methodology.* Journal of Clinical Microbiology, 1993. **31**(2): p. 406-409.

124. Cave, M.D., et al., *Stability of DNA fingerprint pattern produced with IS6110 in strains of Mycobacterium tuberculosis.* J Clin Microbiol, 1994. **32**(1): p. 262-6.
125. Kline, S.E., L.L. Hedemark, and S.F. Davies, *Outbreak of tuberculosis among regular patrons of a neighborhood bar.* N Engl J Med, 1995. **333**(4): p. 222-7.
126. Niemann, S., et al., *Stability of IS6110 restriction fragment length polymorphism patterns of Mycobacterium tuberculosis strains in actual chains of transmission.* Journal of Clinical Microbiology, 2000. **38**(7): p. 2563-2567.
127. Warren, R.M., et al., *Calculation of the stability of the IS6110 banding pattern in patients with persistent Mycobacterium tuberculosis disease.* Journal of Clinical Microbiology, 2002. **40**(5): p. 1705-1708.
128. de Boer, A.S., et al., *Analysis of rate of change of IS6110 RFLP patterns of Mycobacterium tuberculosis based on serial patient isolates.* Journal of Infectious Diseases, 1999. **180**(4): p. 1238-1244.
129. Rosenberg, N.A., A.G. Tsolaki, and M.M. Tanaka, *Estimating change rates of genetic markers using serial samples: applications to the transposon IS6110 in Mycobacterium tuberculosis.* Theoretical Population Biology, 2003. **63**(4): p. 347-363.
130. Lillebaek, T., et al., *Stability of DNA patterns and evidence of Mycobacterium tuberculosis reactivation occurring decades after the initial infection.* Journal of Infectious Diseases, 2003. **188**(7): p. 1032-1039.
131. Savine, E., et al., *Stability of variable-number tandem repeats of mycobacterial interspersed repetitive units from 12 loci in serial isolates of Mycobacterium tuberculosis.* Journal of Clinical Microbiology, 2002. **40**(12): p. 4561-4566.
132. van der Spuy, G.D., et al., *Use of genetic distance as a measure of ongoing transmission of Mycobacterium tuberculosis.* Journal of Clinical Microbiology, 2003. **41**(12): p. 5640-5644.
133. Warren, R., et al., *Genotyping of Mycobacterium tuberculosis with additional markers enhances accuracy in epidemiological studies.* Journal of Clinical Microbiology, 1996. **34**(9): p. 2219-2224.
134. Dale, J.W., et al., *Conservation of IS6110 sequence in strains of Mycobacterium tuberculosis with single and multiple copies.* Tuber Lung Dis, 1997. **78**(5-6): p. 225-7.
135. Allix-Beguec, C., et al., *Evaluation and strategy for use of MIRU-VNTRplus, a multifunctional database for online analysis of genotyping data and phylogenetic identification of Mycobacterium tuberculosis complex isolates.* Journal of Clinical Microbiology, 2008. **46**(8): p. 2692-2699.
136. Bauer, J., et al., *Usefulness of spoligotyping To discriminate IS6110 low-copy-number Mycobacterium tuberculosis complex strains cultured in Denmark.* J Clin Microbiol, 1999. **37**(8): p. 2602-6.

137. Fomukong, N., et al., *Differences in the prevalence of IS6110 insertion sites in Mycobacterium tuberculosis strains: low and high copy number of IS6110.* Tuber Lung Dis, 1997. **78**(2): p. 109-16.
138. Hawkey, P.M., et al., *Mycobacterial interspersed repetitive unit typing of Mycobacterium tuberculosis compared to IS6110-based restriction fragment length polymorphism analysis for investigation of apparently clustered cases of tuberculosis.* Journal of Clinical Microbiology, 2003. **41**(8): p. 3514-3520.
139. Supply, P., et al., *Identification of novel intergenic repetitive units in a mycobacterial two-component system operon.* Molecular Microbiology, 1997. **26**(5): p. 991-1003.
140. Supply, P., et al., *Variable human minisatellite-like regions in the Mycobacterium tuberculosis genome.* Molecular Microbiology, 2000. **36**(3): p. 762-771.
141. Frothingham, R. and W.A. Meeker-O'Connell, *Genetic diversity in the Mycobacterium tuberculosis complex based on variable numbers of tandem DNA repeats.* Microbiology-Uk, 1998. **144**: p. 1189-1196.
142. Filliol, I., et al., *Molecular typing of Mycobacterium tuberculosis based on variable number of tandem DNA repeats used alone and in association with spoligotyping.* Journal of Clinical Microbiology, 2000. **38**(7): p. 2520-2524.
143. Mazars, E., et al., *High-resolution minisatellite-based typing as a portable approach to global analysis of Mycobacterium tuberculosis molecular epidemiology.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(4): p. 1901-1906.
144. Alonso-Rodriguez, N., et al., *Evaluation of the new advanced 15-loci MIRU-VNTR genotyping tool in Mycobacterium tuberculosis molecular epidemiology studies.* Bmc Microbiology, 2008. **8**: p. -.
145. Scott, A.N., et al., *Sensitivities and specificities of spoligotyping and mycobacterial interspersed repetitive unit-variable-number tandem repeat typing methods for studying molecular epidemiology of tuberculosis.* Journal of Clinical Microbiology, 2005. **43**(1): p. 89-94.
146. Blackwood, K.S., J.N. Wolfe, and A.M. Kabani, *Application of mycobacterial interspersed repetitive unit typing to Manitoba tuberculosis cases: Can restriction fragment length polymorphism be forgotten?* Journal of Clinical Microbiology, 2004. **42**(11): p. 5001-5006.
147. Cowan, L.S., et al., *Evaluation of a two-step approach for large-scale, prospective genotyping of Mycobacterium tuberculosis isolates in the United States.* J Clin Microbiol, 2005. **43**(2): p. 688-95.
148. CDC, *Guide to the Application of Genotyping to Tuberculosis Prevention and Control*, U.D.o.H.a.H. Services, Editor 2004, National TB Controllers Association / CDC Advisory Group on Tuberculosis Genotyping: Atlanta, GA.

149. Maes, M., et al., *24-Locus MIRU-VNTR genotyping is a useful tool to study the molecular epidemiology of tuberculosis among Warao Amerindians in Venezuela.* Tuberculosis, 2008. **88**(5): p. 490-494.
150. Gibson, A., et al., *Can 15-locus mycobacterial interspersed repetitive unit-variable-number tandem repeat analysis provide insight into the evolution of Mycobacterium tuberculosis?* Applied and Environmental Microbiology, 2005. **71**(12): p. 8207-8213.
151. Gopaul, K.K., et al., *Progression toward an improved DNA amplification-based typing technique in the study of Mycobacterium tuberculosis epidemiology.* Journal of Clinical Microbiology, 2006. **44**(7): p. 2492-2498.
152. Supply, P., et al., *Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of Mycobacterium tuberculosis.* Journal of Clinical Microbiology, 2006. **44**(12): p. 4498-4510.
153. Oelemann, M.C., et al., *Assessment of an optimized mycobacterial interspersed repetitive-unit-variable-number tandem-repeat typing system combined with spoligotyping for population-based molecular epidemiology studies of tuberculosis.* Journal of Clinical Microbiology, 2007. **45**(3): p. 691-697.
154. Allix-Beguec, C., M. Fauville-Dufaux, and P. Supply, *Three-year population-based evaluation of standardized mycobacterial interspersed repetitive-unit-variable-number tandem-repeat typing of Mycobacterium tuberculosis.* Journal of Clinical Microbiology, 2008. **46**(4): p. 1398-1406.
155. Reyes, J.F. and M.M. Tanaka, *Mutation rates of spoligotypes and variable numbers of tandem repeat loci in Mycobacterium tuberculosis.* Infect Genet Evol, 2010. **10**(7): p. 1046-51.
156. Grant, A., et al., *Mathematical modelling of Mycobacterium tuberculosis VNTR loci estimates a very slow mutation rate for the repeats.* Journal of Molecular Evolution, 2008. **66**(6): p. 565-574.
157. Aandahl, R.Z., et al., *A Model-Based Bayesian Estimation of the Rate of Evolution of VNTR Loci in Mycobacterium tuberculosis.* Plos Computational Biology, 2012. **8**(6).
158. Wirth, T., et al., *Origin, spread and demography of the Mycobacterium tuberculosis complex.* PLoS Pathog, 2008. **4**(9): p. e1000160.
159. Munoz-Elias, E.J., et al., *Replication dynamics of Mycobacterium tuberculosis in chronically infected mice.* Infect Immun, 2005. **73**(1): p. 546-51.
160. Alland, D., et al., *Transmission of tuberculosis in New York City. An analysis by DNA fingerprinting and conventional epidemiologic methods.* N Engl J Med, 1994. **330**(24): p. 1710-6.
161. Vynnycky, E., et al., *Annual Mycobacterium tuberculosis infection risk and interpretation of clustering statistics.* Emerging infectious diseases, 2003. **9**(2): p. 176-83.

162. Small, P.M., et al., *The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods.* The New England journal of medicine, 1994. **330**(24): p. 1703-9.
163. Glynn, J.R., E. Vynnycky, and P.E. Fine, *Influence of sampling on estimates of clustering and recent transmission of Mycobacterium tuberculosis derived from DNA fingerprinting techniques.* American journal of epidemiology, 1999. **149**(4): p. 366-71.
164. Jasmer, R.M., et al., *A molecular epidemiologic analysis of tuberculosis trends in San Francisco, 1991-1997.* Annals of internal medicine, 1999. **130**(12): p. 971-8.
165. Cattamanchi, A., et al., *A 13-year molecular epidemiological analysis of tuberculosis in San Francisco.* International Journal of Tuberculosis and Lung Disease, 2006. **10**(3): p. 297-304.
166. Love, J., et al., *Molecular epidemiology of tuberculosis in England, 1998.* The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease, 2009. **13**(2): p. 201-7.
167. Tanaka, M.M., R. Phong, and A.R. Francis, *An evaluation of indices for quantifying tuberculosis transmission using genotypes of pathogen isolates.* Bmc Infectious Diseases, 2006. **6**: p. -.
168. Fok, A., et al., *Risk factors for clustering of tuberculosis cases: a systematic review of population-based molecular epidemiology studies.* International Journal of Tuberculosis and Lung Disease, 2008. **12**(5): p. 480-492.
169. Nava-Aguilera, E., et al., *Risk factors associated with recent transmission of tuberculosis: systematic review and meta-analysis.* International Journal of Tuberculosis and Lung Disease, 2009. **13**(1): p. 17-26.
170. Heldal, E., et al., *Risk factors for recent transmission of Mycobacterium tuberculosis.* European Respiratory Journal, 2003. **22**(4): p. 637-642.
171. Geng, E., et al., *Changes in the transmission of tuberculosis in New York City from 1990 to 1999.* New England Journal of Medicine, 2002. **346**(19): p. 1453-1458.
172. Glynn, J.R., et al., *Interpreting DNA fingerprint clusters of Mycobacterium tuberculosis.* International Journal of Tuberculosis and Lung Disease, 1999. **3**(12): p. 1055-1060.
173. Luciani, F., A.R. Francis, and M.M. Tanaka, *Interpreting genotype cluster sizes of Mycobacterium tuberculosis isolates typed with IS6110 and spoligotyping.* Infection Genetics and Evolution, 2008. **8**(2): p. 182-190.
174. Murray, M. and E. Nardell, *Molecular epidemiology of tuberculosis: achievements and challenges to current knowledge.* Bulletin of the World Health Organization, 2002. **80**(6): p. 477-482.

175. Murray, M., *Determinants of cluster distribution in the molecular epidemiology of tuberculosis.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(3): p. 1538-1543.
176. Salpeter, E.E. and S.R. Salpeter, *Mathematical model for the epidemiology of tuberculosis, with estimates of the reproductive number and infection-delay function.* American Journal of Epidemiology, 1998. **147**(4): p. 398-406.
177. Aparicio, J.P., A.F. Capurro, and C. Castillo-Chavez, *Transmission and dynamics of tuberculosis on generalized households.* Journal of Theoretical Biology, 2000. **206**(3): p. 327-341.
178. Williams, B.G., et al., *Antiretroviral therapy for tuberculosis control in nine African countries.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(45): p. 19485-19489.
179. Murray, C.J.L. and J.A. Salomon, *Modeling the impact of global tuberculosis control strategies.* Proceedings of the National Academy of Sciences of the United States of America, 1998. **95**(23): p. 13881-13886.
180. Ziv, E., C.L. Daley, and S.M. Blower, *Early therapy for latent tuberculosis infection.* American Journal of Epidemiology, 2001. **153**(4): p. 381-385.
181. Dye, C., et al., *Prospects for worldwide tuberculosis control under the WHO DOTS strategy. Directly observed short-course therapy.* Lancet, 1998. **352**(9144): p. 1886-91.
182. Pitman, R., B. Jarman, and R. Coker, *Tuberculosis transmission and the impact of intervention on the incidence of infection.* The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease, 2002. **6**(6): p. 485-91.
183. Blower, S.M., et al., *The Intrinsic Transmission Dynamics of Tuberculosis Epidemics.* Nature Medicine, 1995. **1**(8): p. 815-821.
184. Currie, C.S.M., et al., *Tuberculosis epidemics driven by HIV: is prevention better than cure?* Aids, 2003. **17**(17): p. 2501-2508.
185. Dye, C. and B.G. Williams, *Eliminating human tuberculosis in the twenty-first century.* Journal of the Royal Society, Interface / the Royal Society, 2008. **5**(23): p. 653-62.
186. Schulzer, M., et al., *An Estimate of the Future Size of the Tuberculosis Problem in Sub-Saharan Africa Resulting from Hiv-Infection.* Tubercle and Lung Disease, 1992. **73**(1): p. 52-58.
187. Debanne, S.M., et al., *Multivariate Markovian modeling of tuberculosis: Forecast for the United States.* Emerging Infectious Diseases, 2000. **6**(2): p. 148-157.
188. Waaler, H., A. Geser, and S. Andersen, *The use of mathematical models in the study of the epidemiology of tuberculosis.* Am J Public Health Nations Health, 1962. **52**: p. 1002-13.

189. Colijn, C., T. Cohen, and M. Murray, *Mathematical Models of Tuberculosis: Accomplishments and Future Challenges.* Biomat 2006, 2007: p. 123-148

388.

190. Castillo-Chavez, C. and B.J. Song, *Dynamical models of tuberculosis and their applications.* Mathematical Biosciences and Engineering, 2004. **1**(2): p. 361-404.
191. Aparicio, J.P. and C. Castillo-Chavez, *Mathematical Modelling of Tuberculosis Epidemics.* Mathematical Biosciences and Engineering, 2009. **6**(2): p. 209-237.
192. Waaler, H.T. and M.A. Piot, *The use of an epidemiological model for estimating the effectiveness of tuberculosis control measures. Sensitivity of the effectiveness of tuberculosis control measures to the coverage of the population.* Bull World Health Organ, 1969. **41**(1): p. 75-93.
193. Waaler, H.T., *Cost-benefit analysis of BCG-vaccination under various epidemiological situations.* Bull Int Union Tuberc, 1968. **41**: p. 42-52.
194. Waaler, H.T., *The economics of tuberculosis control.* Tubercle, 1968. **49**: p. Suppl:2-4.
195. Waaler, H.T., *Model simulation and decision-making in tuberculosis programmes.* Bull Int Union Tuberc, 1970. **43**: p. 337-44.
196. Waaler, H.T. and M.A. Piot, *Use of an epidemiological model for estimating the effectiveness of tuberculosis control measures. Sensitivity of the effectiveness of tuberculosis control measures to the social time preference.* Bull World Health Organ, 1970. **43**(1): p. 1-16.
197. Brogger, S., *Systems analysis in tuberculosis control: a model.* Am Rev Respir Dis, 1967. **95**(3): p. 419-34.
198. ReVelle, C.S., W.R. Lynn, and F. Feldmann, *Mathematical models for the economic allocation of tuberculosis control activities in developing nations.* Am Rev Respir Dis, 1967. **96**(5): p. 893-909.
199. Ferebee, S.H., *An epidemiological model of tuberculosis in the United States.* National Tuberculosis Bulletin, 1967. **January**: p. 4-7.
200. Heymann, S.J., *Modelling the efficacy of prophylactic and curative therapies for preventing the spread of tuberculosis in Africa.* Trans R Soc Trop Med Hyg, 1993. **87**(4): p. 406-11.
201. Brewer, T.F., et al., *Evaluation of tuberculosis control policies using computer simulation.* JAMA, 1996. **276**(23): p. 1898-903.
202. Nicas, M. and E. Seto, *A simulation model for occupational tuberculosis transmission.* Risk Anal, 1997. **17**(5): p. 609-16.
203. Sawert, H., et al., *Costs and benefits of improving tuberculosis control: the case of Thailand.* Soc Sci Med, 1997. **44**(12): p. 1805-16.
204. Aparicio, J.P., A.F. Capurro, and C. Castillo-Chavez, *Long-term dynamics and re-emergence of tuberculosis.* Mathematical Approaches for Emerging and Reemerging Infectious Diseases: An Introduction, 2002. **125**: p. 351-360

368.
205. Song, B., C. Castillo-Chavez, and J.P. Aparicio, *Tuberculosis models with fast and slow dynamics: the role of close and casual contacts.* Math Biosci, 2002. **180**: p. 187-205.
206. Cohen, T. and M. Murray, *Modeling epidemics of multidrug-resistant M-tuberculosis of heterogeneous fitness.* Nature Medicine, 2004. **10**(10): p. 1117-1121.
207. Hughes, G.R., C.S.M. Currie, and E.L. Corbett, *Modeling tuberculosis in areas of high HIV prevalence.* Proceedings of the 2006 Winter Simulation Conference, Vols 1-5, 2006: p. 459-465

2307.
208. Bacaer, N., et al., *Modeling the joint epidemics of TB and HIV in a South African township.* Journal of Mathematical Biology, 2008. **57**(4): p. 557-593.
209. Magombedze, G., W. Garira, and E. Mwenje, *In-vivo mathematical study of co-infection dynamics of HIV-1 and Mycobacterium tuberculosis.* Journal of Biological Systems, 2008. **16**(3): p. 357-394.
210. Rodrigues, P., M.G.M. Gomes, and C. Rebelo, *Drug resistance in tuberculosis - a reinfection model.* Theoretical Population Biology, 2007. **71**(2): p. 196-212.
211. Long, E.F., N.K. Vaidya, and M.L. Brandeau, *Controlling Co-Epidemics: Analysis of HIV and Tuberculosis Infection Dynamics.* Operations Research, 2008. **56**(6): p. 1366-1381.
212. Legrand, J., et al., *Modeling the impact of tuberculosis control strategies in highly endemic overcrowded prisons.* PLoS One, 2008. **3**(5): p. e2100.
213. Lietman, T. and S.M. Blower, *Potential impact of tuberculosis vaccines as epidemic control agents.* Clin Infect Dis, 2000. **30 Suppl 3**: p. S316-22.
214. Vynnycky, E. and P.E. Fine, *The long-term dynamics of tuberculosis and other diseases with long serial intervals: implications of and for changing reproduction numbers.* Epidemiology and infection, 1998. **121**(2): p. 309-24.
215. Tanaka, M.M., et al., *Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data.* Genetics, 2006. **173**(3): p. 1511-1520.
216. Mellor, G.R., C.S.M. Currie, and E.L. Corbett, *Incorporating Household Structure into a Discrete-Event Simulation Model of Tuberculosis and HIV.* Acm Transactions on Modeling and Computer Simulation, 2011. **21**(4).
217. Mills, H.L., T. Cohen, and C. Colijn, *Modelling the performance of isoniazid preventive therapy for reducing tuberculosis in HIV endemic settings: the effects of network structure.* Journal of the Royal Society Interface, 2011. **8**(63): p. 1510-1520.

218. Bishai, J.D., W.R. Bishai, and D.M. Bishai, *Heightened Vulnerability to MDR-TB Epidemics after Controlling Drug-Susceptible TB.* PLoS One, 2010. **5**(9).
219. Warrender, C., S. Forrest, and F. Koster, *Modeling intercellular interactions in early Mycobacterium infection.* Bull Math Biol, 2006. **68**(8): p. 2233-61.
220. Segovia-Juarez, J.L., S. Ganguli, and D. Kirschner, *Identifying control mechanisms of granuloma formation during M-tuberculosis infection using an agent-based model.* Journal of Theoretical Biology, 2004. **231**(3): p. 357-376.
221. Cohen, T., et al., *Exogenous re-infection and the dynamics of tuberculosis epidemics: local effects in a network model of transmission.* Journal of the Royal Society Interface, 2007. **4**(14): p. 523-531.
222. Cohen, T., et al., *Are survey-based estimates of the burden of drug resistant TB too low? Insight from a simulation study.* PLoS One, 2008. **3**(6): p. e2363.
223. Guzzetta, G., et al., *Modeling socio-demography to capture tuberculosis transmission dynamics in a low burden setting.* J Theor Biol, 2011. **289**: p. 197-205.
224. Wolleswinkel-van den Bosch, J.H., et al., *The impact of immigration on the elimination of tuberculosis in The Netherlands: a model based approach.* International Journal of Tuberculosis and Lung Disease, 2002. **6**(2): p. 130-136.
225. MCluskey, C.C. and P. van den Driessche, *Global analysis of two tuberculosis models.* Journal of Dynamics and Differential Equations, 2004. **16**(1): p. 139-166.
226. Jia, Z.W., et al., *Modeling the impact of immigration on the epidemiology of tuberculosis.* Theoretical Population Biology, 2008. **73**(3): p. 437-448.
227. Brauer, F. and P. van den Driessche, *Models for transmission of disease with immigration of infectives.* Mathematical Biosciences, 2001. **171**(2): p. 143-154.
228. Brooks-Pollock, E., T. Cohen, and M. Murray, *The Impact of Realistic Age Structure in Simple Models of Tuberculosis Transmission.* PLoS One, 2010. **5**(1).
229. Dube, C., J. Sanchez, and A. Reeves, *Adapting existing models of highly contagious diseases to countries other than their country of origin.* Revue Scientifique Et Technique-Office International Des Epizooties, 2011. **30**(2): p. 581-589.
230. Edwards, A., *Open-source science to enable drug discovery.* Drug Discovery Today, 2008. **13**(17-18): p. 731-733.
231. Patlak, M., *Open-Source Science Makes Headway.* J Natl Cancer Inst, 2010. **102**(16): p. 1221-1223.

232. Grimm, V., et al., *A standard protocol for describing individual-based and agent-based models.* Ecological Modelling, 2006. **198**(1-2): p. 115-126.
233. Grimm, V., et al., *The ODD protocol A review and first update.* Ecological Modelling, 2010. **221**(23): p. 2760-2768.
234. Lehman, C. and A. Keen, *Efficient pseudo-random numbers generated from any probability distribution.* Proceedings of the 2012 International Conference on Modeling, Simulation, and Visualization Methods, 2012. **12**: p. 121-127.
235. Lehman, C., A. Keen, and R. Barnes, *Trading space for time: Constant-speed algorithms for managing time in scientific simulations.* Proceedings of the 2012 International Conference on Scientific Computing, 2012.
236. Keen, A. and C. Lehman, *Trading space for time: Constant-speed algorithms for grouping objects in scientific simulations.* Proceedings of the 2012 International Conference on Scientific Computing, 2012. **12**: p. 146-151.
237. Sargent, R.G., *Verification and Validation of Simulation Models.* Proceedings of the 2009 Winter Simulation Conference (Wsc 2009 ), Vol 1-4, 2009: p. 162-176.
238. Turing, A.M., *Checking a large routine*, in *Report of a Conference on High Speed Automatic Calculating Machines*1949, University Mathematical Laboratory, Cambridge. p. 67-69.
239. *Program Verification: Fundamental Issues in Computer Science*1993, Dordrecht, Holland: Kluwer Academic Publishers.
240. Floyd, R.W., *Assigning meanings to programs.* Proceedings of Symposium on Applied Mathematics, 1967. **19**: p. 19-32.
241. Hoare, C.A.R., *An Axiomatic Basis for Computer Programming.* Communications of the Acm, 1969. **12**(10): p. 576-580, 583.
242. Lehtinen, A. and J. Kuorikoski, *Computing the perfect model: Why do economists shun simulation?* Philosophy of Science, 2007. **74**(3): p. 304-329.
243. Aumann, C.A., *A methodology for developing simulation models of complex systems.* Ecological Modelling, 2007. **202**(3-4): p. 385-396.
244. Refsgaard, J.C. and H.J. Henriksen, *Modelling guidelines - terminology and guiding principles.* Advances in Water Resources, 2004. **27**(1): p. 71-82.
245. Huang, Y.P., et al., *Agent-based scientific simulation.* Computing in Science & Engineering, 2005. **7**(1): p. 22-29.
246. Wilensky, U. and W. Rand, *Making Models Match: Replicating an Agent-Based Model.* Jasss-the Journal of Artificial Societies and Social Simulation, 2007. **10**(4).
247. Ormerod, P. and B. Rosewell, *Validation and Verification of Agent-Based Models in the Social Sciences.* Epistemological Aspects of Computer Simulation in the Social Sciences, 2009. **5466**: p. 130-140.

248. Masys, A.J., *Understanding Climate Change through Modelling and Simulation: A case for Verification, Validation and Accreditation.* 2006 IEEE EIC Climate Change Conference, Vols 1 and 2, 2006: p. 495-503.
249. Jagdev, H.S., J. Browne, and P. Jordan, *Verification and Validation Issues in Manufacturing Models.* Computers in Industry, 1995. **25**(3): p. 331-353.
250. Oreskes, N., K. Shraderfrechette, and K. Belitz, *Verification, Validation, and Confirmation of Numerical-Models in the Earth-Sciences.* Science, 1994. **263**(5147): p. 641-646.
251. Skvortsov, A.T., et al., *Epidemic Modelling: Validation of Agent-based Simulation by Using Simple Mathematical Models.* Modsim 2007: International Congress on Modelling and Simulation, 2007: p. 657-+.
252. Lehman, C., S. Williams, and A. Keen, *The Centinel data format: Reliably communicating through time and place.* Proceedings of the 2012 International Conference on Information and Knowledge Engineering, 2012. **12**: p. 47-53.
253. Schoenharl, T.W. and G. Madey, *Evaluation of measurement techniques for the validation of agent-based simulations against streaming data.* Computational Science - Iccs 2008, Pt 3, 2008. **5103**: p. 6-15.
254. Anderson, A.E., et al., *Validation of finite element predictions of cartilage contact pressure in the human hip joint.* Journal of Biomechanical Engineering-Transactions of the Asme, 2008. **130**(5).
255. Press, W.H., *Numerical recipes : the art of scientific computing*2007, New York: Cambridge University Press. xxi, 1235 p.
256. Press, W.H., *Numerical recipes in C : the art of scientific computing*1992, Cambridge Cambridgeshire ; New York: Cambridge University Press. xxvi, 994 p.
257. Vynnycky, E., et al., *Limited impact of tuberculosis control in Hong Kong: attributable to high risks of reactivation disease.* Epidemiology and infection, 2008. **136**(7): p. 943-52.
258. Clark, M. and E. Vynnycky, *The use of maximum likelihood methods to estimate the risk of tuberculous infection and disease in a Canadian First Nations population.* International Journal of Epidemiology, 2004. **33**(3): p. 477-484.
259. Lorant, V., H. Van Oyen, and I. Thomas, *Contextual factors and immigrants' health status: Double jeopardy.* Health & Place, 2008. **14**(4): p. 678-692.
260. Uitenbroek, D.G. and A.P. Verhoeff, *Life expectancy and mortality differences between migrant groups living in Amsterdam, the Netherlands.* Social Science & Medicine, 2002. **54**(9): p. 1379-1388.
261. Weitoft, G.R., et al., *Mortality statistics in immigrant research: method for adjusting underestimation of mortality.* International journal of epidemiology, 1999. **28**(4): p. 756-763.
262. Markides, K.S. and K. Eschbach, *Aging, migration, and mortality: Current status of research on the Hispanic paradox.* Journals of

Gerontology Series B-Psychological Sciences and Social Sciences, 2005. **60**: p. 68-75.
263. *Economic and Social Data Service.* [cited 2010; Available from: http://www.esds.ac.uk/government/lfs/
264. *International migration: Migrants entering or leaving the United Kingdom*

*and England and Wales, 1999*, in *Series MN no 26*2001, Office for National Statistics: London: The Stationery Office.
265. *Long-Term International Migration Estimates Methodology Document: 1991 onwards*, 2010, Office for National Statistics.
266. ONS. *International migration, long-term migration homepage.* 2012 [cited 2012 02/09]; Available from: http://www.ons.gov.uk/ons/taxonomy/index.html?nscl=Long-term+Migrants.
267. Vynnycky, E., *An investigation of the transmission dynamics of `M. tuberculosis'* 1996: p. 288.
268. Sutherland, I. and V.H. Springett, *The Effects of the Scheme for Bcg Vaccination of Schoolchildren in England and Wales and the Consequences of Discontinuing the Scheme at Various Dates.* Journal of Epidemiology and Community Health, 1989. **43**(1): p. 15-24.
269. WHO, *WHO vaccine-preventable diseases: monitoring system 2010 global summary*, 2010.
270. Van Leth, F., M. Van der Werf, and M. Borgdorff, *Prevalence of tuberculous infection and incidence of tuberculosis; a re-assessment of the Styblo rule.* Bulletin of the World Health Organization, 2008. **86**(1): p. 20-26.
271. Vynnycky, E. and P.E. Fine, *The annual risk of infection with Mycobacterium tuberculosis in England and Wales since 1901.* The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease, 1997. **1**(5): p. 389-96.
272. Vynnycky, E. and P.E.M. Fine, *Interpreting the decline in tuberculosis: the role of secular trends in effective contact.* International journal of epidemiology, 1999. **28**(2): p. 327-334.
273. Borgdorff, M.W., et al., *Transmission of tuberculosis between people of different ages in The Netherlands: an analysis using DNA fingerprinting.* The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease, 1999. **3**(3): p. 202-6.
274. Pienaar, E., et al., *A model of tuberculosis transmission and intervention strategies in an urban residential area.* Computational Biology and Chemistry, 2010. **34**(2): p. 86-96.
275. Porco, T.C. and S.M. Blower, *Quantifying the intrinsic transmission dynamics of tuberculosis.* Theoretical Population Biology, 1998. **54**(2): p. 117-132.

276. Colijn, C., T. Cohen, and M. Murray, *Emergent heterogeneity in declining tuberculosis epidemics.* Journal of Theoretical Biology, 2007. **247**(4): p. 765-774.
277. Jia, Z.W., et al., *Transmission models of tuberculosis in heterogeneous population.* Chinese Medical Journal, 2007. **120**(15): p. 1360-1365.
278. Baussano, I., et al., *Impact of immigration and HIV infection on tuberculosis incidence in an area of low tuberculosis prevalence.* Epidemiology and infection, 2006. **134**(6): p. 1353-9.
279. Vynnycky, E., et al., *The effect of age and study duration on the relationship between 'clustering' of DNA fingerprint patterns and the proportion of tuberculosis disease attributable to recent transmission.* Epidemiol Infect, 2001. **126**(1): p. 43-62.
280. Paynter, S., et al., *Patient and health service delays in initiating treatment for patients with pulmonary tuberculosis: retrospective cohort study.* The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease, 2004. **8**(2): p. 180-5.
281. Rodger, A., et al., *Delay in the diagnosis of pulmonary tuberculosis, London, 1998-2000: analysis of surveillance data.* BMJ, 2003. **326**(7395): p. 909-10.
282. Lewis, K.E., et al., *Delay in starting treatment for tuberculosis in east London.* Commun Dis Public Health, 2003. **6**(2): p. 133-8.
283. Oxlade, O., et al., *Developing a Tuberculosis Transmission Model That Accounts for Changes in Population Health.* Med Decis Making, 2010.
284. Nisar, M. and P.D.O. Davies, *Current Trends in Tuberculosis Mortality in England and Wales.* Thorax, 1991. **46**(6): p. 438-440.
285. Martineau, A.R., et al., *Decreasing tuberculosis case fatality in England and Wales, 1988-2001.* International Journal of Tuberculosis and Lung Disease, 2004. **8**(6): p. 737-742.
286. Humphries, M.J., et al., *Deaths Occurring in Newly Notified Patients with Pulmonary Tuberculosis in England and Wales.* British Journal of Diseases of the Chest, 1984. **78**(2): p. 149-158.
287. Ahmed, A.B., et al., *The growing impact of HIV infection on the epidemiology of tuberculosis in England and Wales: 1999-2003.* Thorax, 2007. **62**(8): p. 672-676.
288. Presanis, A.M., et al., *Insights into the rise in HIV infections, 2001 to 2008: a Bayesian synthesis of prevalence evidence.* Aids, 2010. **24**(18): p. 2849-2858.
289. van Hest, N.A.H., et al., *Record-linkage and capture-recapture analysis to estimate the incidence and completeness of reporting of tuberculosis in England 1999-2002.* Epidemiology and Infection, 2008. **136**(12): p. 1606-1616.
290. Sheldon, C.D., et al., *Notification of Tuberculosis - How Many Cases Are Never Reported.* Thorax, 1992. **47**(12): p. 1015-1018.

291. Vynnycky, E. and P.E.M. Fine, *The annual risk of infection with Mycobacterium tuberculosis in England and Wales since 1901.* International Journal of Tuberculosis and Lung Disease, 1997. **1**(5): p. 389-396.
292. HPA. *Statutory notifications of TB (NOIDs).* [cited 2010; Available from: http://www.hpa.org.uk/web/HPAweb&Page&HPAwebAutoListName/Page/1294739541762.
293. Kumar, D., et al., *Tuberculosis in England and Wales in 1993: results of a national survey. Public Health Laboratory Service/British Thoracic Society/Department of Health Collaborative Group.* Thorax, 1997. **52**(12): p. 1060-7.
294. Vynnycky, E. and R. White, *An Introduction to Infectious Disease Modelling*2010, Oxford, UK: Oxford University Press.
295. Ormerod, L.P., *Tuberculosis screening and prevention in new immigrants 1983-88.* Respir Med, 1990. **84**(4): p. 269-71.
296. Brimnes, N., *BCG vaccination and WHO's global strategy for tuberculosis control 1948-1983.* Soc Sci Med, 2008. **67**(5): p. 863-73.
297. Cauthen, G.M., A. Pio, and H.G. ten Dam, *Annual risk of tuberculous infection. 1988.* Bull World Health Organ, 2002. **80**(6): p. 503-11; discussion 501-2.
298. *Tuberculosis control in the era of the HIV epidemic: risk of tuberculosis infection in Tanzania, 1983-1998.* Int J Tuberc Lung Dis, 2001. **5**(2): p. 103-12.
299. Shrestha, K.B., et al., *First national tuberculin survey in Nepal.* Int J Tuberc Lung Dis, 2008. **12**(8): p. 909-15.
300. Gopi, P.G., R. Subramani, and P.R. Narayanan, *Trend in the prevalence of TB infection and ARTI after implementation of a DOTS programme in south India.* Int J Tuberc Lung Dis, 2006. **10**(3): p. 346-8.
301. Norval, P.Y., C. Roustit, and K.K. San, *From tuberculin to prevalence survey in Cambodia.* Int J Tuberc Lung Dis, 2004. **8**(3): p. 299-305.
302. El Ibiary, S., et al., *Trend in the annual risk of tuberculous infection in Egypt, 1950-1996.* Int J Tuberc Lung Dis, 1999. **3**(4): p. 294-9.
303. Sutherland, I., et al., *The Risk of Tuberculous Infection in the Netherlands from 1967 to 1979.* Tubercle, 1983. **64**(4): p. 241-253.
304. Salpeter, E.E. and S.R. Salpeter, *The annual rate of tuberculosis infection and the probability of tuberculosis activation in young adults of different races.* American Journal of Epidemiology, 1999. **149**(11): p. S10-S10.
305. Ormerod, L.P., *Is new immigrant screening for tuberculosis still worthwhile?* Journal of Infection, 1998. **37**(1): p. 39-40.
306. Andrews, J.R., et al., *Risk of progression to active tuberculosis following reinfection with Mycobacterium tuberculosis.* Clinical infectious diseases : an official publication of the Infectious Diseases Society of America, 2012. **54**(6): p. 784-91.

307. Brooks-Pollock, E., et al., *Epidemiologic inference from the distribution of tuberculosis cases in households in Lima, Peru.* The Journal of infectious diseases, 2011. **203**(11): p. 1582-9.
308. Maguire, H., et al., *Molecular epidemiology of tuberculosis in London 1995-7 showing low rate of active transmission.* Thorax, 2002. **57**(7): p. 617-622.
309. *2011 Census - Population and Household Estimates for England and Wales, March 2011*, 2012, Office for National Statistics.
310. Evans, J.T., et al., *Cluster of human tuberculosis caused by Mycobacterium bovis: evidence for person-to-person transmission in the UK.* Lancet, 2007. **369**(9569): p. 1270-1276.
311. Evans, J.T., *Molecular epidemiology of tuberculosis in the midlands*, in *School of Immunity and Infection, College of Medical and Dental Sciences*2011, University of Birmingham: United Kingdom.
312. Kempf, M.C., et al., *Long-term molecular analysis of tuberculosis strains in Alabama, a state characterized by a largely indigenous, low-risk population.* Journal of Clinical Microbiology, 2005. **43**(2): p. 870-878.
313. Sharnprapai, S., et al., *Genotyping analyses of tuberculosis cases in US- and foreign-born Massachusetts residents.* Emerging Infectious Diseases, 2002. **8**(11): p. 1239-1245.
314. van Soolingen, D., et al., *Molecular epidemiology of tuberculosis in the Netherlands: A nationwide study from 1993 through 1997.* Journal of Infectious Diseases, 1999. **180**(3): p. 726-736.
315. van Doorn, H.R., et al., *Public health impact of isoniazid-resistant Mycobacterium tuberculosis strains with a mutation at amino-acid position 315 of katG: a decade of experience in The Netherlands.* Clinical Microbiology and Infection, 2006. **12**(8): p. 769-775.
316. Verhagen, L.M., et al., *Mycobacterial Factors Relevant for Transmission of Tuberculosis.* Journal of Infectious Diseases, 2011. **203**(9): p. 1249-1255.
317. Rey-Jurado, E., et al., *Impaired fitness of Mycobacterium tuberculosis resistant isolates in a cell culture model of murine macrophages.* Journal of Antimicrobial Chemotherapy, 2011. **66**(10): p. 2277-2280.
318. Cohen, T., B. Sommers, and M. Murray, *The effect of drug resistance on the fitness of Mycobacterium tuberculosis.* Lancet Infect Dis, 2003. **3**(1): p. 13-21.
319. Fenner, L., et al., *Effect of Mutation and Genetic Background on Drug Resistance in Mycobacterium tuberculosis.* Antimicrobial Agents and Chemotherapy, 2012. **56**(6): p. 3047-3053.
320. Gagneux, S., et al., *Impact of bacterial genetics on the transmission of isoniazid-resistant Mycobacterium tuberculosis.* Plos Pathogens, 2006. **2**(6): p. 603-610.
321. Bishai, W.R., et al., *Molecular and geographic patterns of tuberculosis transmission after 15 years of directly observed therapy.* Jama-Journal of the American Medical Association, 1998. **280**(19): p. 1679-1684.

322. Bidovec-Stojkovic, U., M. Zolnir-Dovc, and P. Supply, *One year nationwide evaluation of 24-locus MIRU-VNTR genotyping on Slovenian Mycobacterium tuberculosis isolates.* Respiratory Medicine, 2011. **105**: p. S67-S73.
323. Valcheva, V., et al., *Utility of new 24-locus variable-number tandem-repeat typing for discriminating Mycobacterium tuberculosis clinical isolates collected in Bulgaria.* J Clin Microbiol, 2008. **46**(9): p. 3005-11.
324. Roetzer, A., et al., *Evaluation of Mycobacterium tuberculosis typing methods in a 4-year study in Schleswig-Holstein, Northern Germany.* J Clin Microbiol, 2011. **49**(12): p. 4173-8.
325. Houben, R.M.G.J. and J.R. Glynn, *Systematic review and analysis of population-based molecular epidemiological studies.* International Journal of Tuberculosis and Lung Disease, 2009. **13**(2): p. 275-275.
326. Hernandez-Garduno, E., et al., *Predictors of clustering of tuberculosis in Greater Vancouver: a molecular epidemiologic study.* Canadian Medical Association Journal, 2002. **167**(4): p. 349-352.
327. Pena, M.J., et al., *Epidemiology of tuberculosis on Gran Canaria: a 4 year population study using traditional and molecular approaches.* Thorax, 2003. **58**(7): p. 618-22.
328. Thomsen, V.O., T. Lillebaek, and F. Stenz, *Tuberculosis in Greenland--current situation and future challenges.* Int J Circumpolar Health, 2004. **63 Suppl 2**: p. 225-9.
329. Inigo, J., et al., *Recent transmission of tuberculosis in Madrid: application of capture-recapture analysis to conventional and molecular epidemiology.* Int J Epidemiol, 2003. **32**(5): p. 763-9.
330. van Deutekom, H., et al., *A molecular epidemiological approach to studying the transmission of tuberculosis in Amsterdam.* Clinical infectious diseases : an official publication of the Infectious Diseases Society of America, 1997. **25**(5): p. 1071-7.
331. de Vries, G., et al., *Transmission Classification Model To Determine Place and Time of Infection of Tuberculosis Cases in an Urban Area.* Journal of Clinical Microbiology, 2008. **46**(12): p. 3924-3930.
332. Moonan, P.K., et al., *Using Genotyping and Geospatial Scanning to Estimate Recent Mycobacterium tuberculosis Transmission, United States.* Emerging Infectious Diseases, 2012. **18**(3): p. 458-465.
333. Ruddy, M., et al., *Estimation of the rate of unrecognized cross-contamination with Mycobacterium tuberculosis in London microbiology laboratories.* Journal of Clinical Microbiology, 2002. **40**(11): p. 4100-4104.
334. Tanaka, M.M. and J.F. Reyes, *VNTR mutation in Mycobacterium tuberculosis: Lower rates for less variable loci.* Infection Genetics and Evolution, 2011. **11**(6): p. 1192-1192.
335. Reyes, J.F., C.H.S. Chan, and M.M. Tanaka, *Impact of homoplasy on variable numbers of tandem repeats and spoligotypes in*

*Mycobacterium tuberculosis.* Infection Genetics and Evolution, 2012. **12**(4): p. 811-818.
336. McGill, B.J., et al., *Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework.* Ecol Lett, 2007. **10**(10): p. 995-1015.
337. Glickman, M.S. and W.R. Jacobs, *Microbial pathogenesis of Mycobacterium tuberculosis: Dawn of a discipline.* Cell, 2001. **104**(4): p. 477-485.
338. Ahmad, S., *Pathogenesis, Immunology, and Diagnosis of Latent Mycobacterium tuberculosis Infection.* Clinical & Developmental Immunology, 2011.
339. Borgdorff, M.W., et al., *Transmission of mycobacterium tuberculosis depending on the age and sex of source cases.* American Journal of Epidemiology, 2001. **154**(10): p. 934-943.
340. Mossong, J., et al., *Social contacts and mixing patterns relevant to the spread of infectious diseases.* PLoS Med, 2008. **5**(3): p. 381-391.
341. Simpson, L., *Household forecasts for Birmingham, with an ethnic group dimension*, 2007, Centre for Census and Survey Research, University of Manchester.
342. Lehman, C. and A. Keen, *Symmetry and simplicity in simulation: Reducing complexity in alternate parallel-serial processing.* Proceedings of the 2012 International Conference on Parallel and Distributed Processing Techniques and Applications, 2012. **12**(1): p. 90-94.
343. Borrell, S., et al., *Factors Associated with Differences between Conventional Contact Tracing and Molecular Epidemiology in Study of Tuberculosis Transmission and Analysis in the City of Barcelona, Spain.* Journal of Clinical Microbiology, 2009. **47**(1): p. 198-204.
344. Easterbrook, P.J., et al., *High rates of clustering of strains causing tuberculosis in Harare, Zimbabwe: a molecular epidemiological study (vol 42, pg 4536, 2004).* Journal of Clinical Microbiology, 2004. **42**(12): p. 5965-5965.
345. Cacho Calvo, J., et al., *Ten-year population-based molecular epidemiological study of tuberculosis transmission in the metropolitan area of Madrid, Spain.* Int J Tuberc Lung Dis, 2005. **9**(11): p. 1236-41.
346. Kik, S.V., et al., *Tuberculosis outbreaks predicted by characteristics of first patients in a DNA fingerprint cluster.* American Journal of Respiratory and Critical Care Medicine, 2008. **178**(1): p. 96-104.
347. Guernier, V., et al., *Use of cluster-graphs from spoligotyping data to study genotype similarities and a comparison of three indices to quantify recent tuberculosis transmission among culture positive cases in French Guiana during a eight year period.* Bmc Infectious Diseases, 2008. **8**: p. -.

348. Borgdorff, M.W., et al., *Progress towards tuberculosis elimination: secular trend, immigration and transmission.* European Respiratory Journal, 2010. **36**(2): p. 339-347.
349. Franzetti, F., et al., *Genotyping analyses of tuberculosis transmission among immigrant residents in Italy.* Clinical Microbiology and Infection, 2010. **16**(8): p. 1149-1154.
350. Lawson, L., et al., *A Molecular Epidemiological and Genetic Diversity Study of Tuberculosis in Ibadan, Nnewi and Abuja, Nigeria.* PLoS One, 2012. **7**(6).
351. Small, P.M., et al., *Molecular Strain Typing of Mycobacterium-Tuberculosis to Confirm Cross-Contamination in the Mycobacteriology Laboratory and Modification of Procedures to Minimize Occurrence of False-Positive Cultures.* Journal of Clinical Microbiology, 1993. **31**(7): p. 1677-1682.

# 10 Appendix

## 10.1 Equations for ODE model written in R language

The following ODE model was designed to match the simplified version of the tuberculosis IBM described in Chapter 3, Section 3.4.1.1. This model was used for testing in the initial phases of model development as described in Section 3.4.1.

```
# SOLVING ODE VERSION OF TUBERCULOSIS IBM

b = 50000;                      # Births
im = 6000;                      # Total immigrants per year
e1 = 0.006;                     # Emigration rate per year
c = 20.0;                       # Effective contacts per year
v1 = 0.77;                      # Efficacy of vaccine
v2 = 0.75;                      # Proportion vaccinated at designated age
v3 = 13;                        # Average age of vaccination
p = 0.6;                        # Proportion of disease cases that are pulmonary
d1 = 0.04*2;                    # Progression to disease for Recent Infection.
d2 = 0.0002*2;                  # Progression to disease for Latent Infection.
d3 = 0.04*2;                    # Progression to disease for Reinfection.
r1 = 0.2;                       # Annual rate of transfer to latent (I2) from I1
                                # and I3
r2 = 1.0;                       # Annual rate of transfer to Latent Infection
                                # from disease classes
m1 = 0.01;                      # Mortality of non-disease classes
m2 = 0.02;                      # Mortality of disease classes

im1 = .9*im;                    # Proportion of immigrants Uninfected
im2 = .1*im;                    # Proportion of immigrants with Recent Infection

U0 = 9855145;   V0 = 0;         # Initial conditions
I1 = 144855;   I2 = 0; I3 = 0;
D1 = 0;  D2 = 0; D3 = 0;
D4 = 0;  D5 = 0; D6 = 0;

tgap=1.0; tp=-2*tgap;           # Time between displays
#t=0; tmax=44; dt=1/365.25;     # Time range and time step
t=0; tmax=44; dt=1/1000;        # Time range and time step

plotU0 = numeric(); plotV0 = numeric();
plotI1 = numeric(); plotI2 = numeric(); plotI3 = numeric();
plotD1 = numeric(); plotD2 = numeric(); plotD3 = numeric();
plotD4 = numeric(); plotD5 = numeric(); plotD6 = numeric();

i = 1;

while(t<=tmax)                  # Trace the solution with Euler's method
{
  N0  = U0 +V0 +I1 +I2 +I3 +D1 +D2 +D3 +D4 +D5 +D6;
  c1 = c*(D1+D2+D3)*(U0/N0);
  c2 = c*(D1+D2+D3)*(I2/N0);

  dU0dt = -e1*U0 -m1*U0 - (v1*v2/v3)*U0  + b + im1  - c1;
  dV0dt = -e1*V0 -m1*V0 + (v1*v2/v3)*U0;
  dI1dt = -e1*I1 -m1*I1 - d1*I1 - r1*I1 + im2       + c1;
  dI2dt = -e1*I2 -m1*I2 - d2*I2 + r1*(I1+I3) + r2*(D1+D2+D3+D4+D5+D6) - c2;
  dI3dt = -e1*I3 -m1*I3  -d3*I3  - r1*I3                              + c2;
  dD1dt = -e1*D1 -m2*D1  -r2*D1  + p*d1*I1;
  dD2dt = -e1*D2 -m2*D2  -r2*D2  + p*d2*I2;
  dD3dt = -e1*D3 -m2*D3  -r2*D3  + p*d3*I3;
  dD4dt = -e1*D4 -m2*D4  -r2*D4  + (1-p)*d1*I1;
  dD5dt = -e1*D5 -m2*D5  -r2*D5  + (1-p)*d2*I2;
  dD6dt = -e1*D6 -m2*D6  -r2*D6  + (1-p)*d3*I3;
```

```
 dU0 = dU0dt*dt;
 dV0 = dV0dt*dt;
 dI1 = dI1dt*dt;
 dI2 = dI2dt*dt;
 dI3 = dI3dt*dt;
 dD1 = dD1dt*dt;
 dD2 = dD2dt*dt;
 dD3 = dD3dt*dt;
 dD4 = dD4dt*dt;
 dD5 = dD5dt*dt;
 dD6 = dD6dt*dt;

 U0 = U0 + dU0;
 V0 = V0 + dV0;
 I1 = I1 + dI1;
 I2 = I2 + dI2;
 I3 = I3 + dI3;
 D1 = D1 + dD1;
 D2 = D2 + dD2;
 D3 = D3 + dD3;
 D4 = D4 + dD4;
 D5 = D5 + dD5;
 D6 = D6 + dD6;

  plotU0[i] = U0; plotV0[i] = V0; plotI1[i] = I1; plotI2[i] = I2; plotI3[i] =
I3;
  plotD1[i] = D1; plotD2[i] = D2; plotD3[i] = D3; plotD4[i] = D4; plotD5[i] =
D5;
  plotD6[i] = D6;

  i = i+1;

if(t-tp>=tgap) { print(c(t,U0, V0, I1, I2, I3, D1, D2, D3, D4, D5, D6)); tp =
t; }
  t = t+dt;
}

FINAL =  c(t, U0, V0, I1, I2, I3, D1, D2, D3, D4, D5, D6);
FINAL;

plot.new(); plot.window(xlim=c(0,i),ylim=c(0,5000000))
axis(1); axis(2)
title(main="TB IBM 12")
title(xlab="t"); title(ylab="U0(t), V0(t), I1(t), I2(t), I3(t), D1(t), D2(t),
D3(t), D4(t), D5(t), D6(t)")
box()
lines(plotU0,col="green"); lines(plotV0,col="orange");
lines(plotI1,col="red"); lines(plotI2,col="black"); lines(plotI3,col="blue");
lines(plotD1,col="red"); lines(plotD2,col="grey"); lines(plotD3,col="yellow");
lines(plotD4,col="violet"); lines(plotD5,col="magenta");
lines(plotD6,col="pink");
```

## 10.2 Centinel Input Data File Format Example

This file format and software was used for reading input files into the model. The example file provides HIV prevalence values assumed for Sub-Saharan Africa-born individuals in the model.

```
Dataset:      HIV Prevalence in Sub-Saharan-African-born immigrants.
Description:  These data are estimates of the prevalence of HIV in SSA
              immigrants entering the UK from 1981. Estimates are based on
              the assumption that the prevalence was zero in 1981 and
              increased linearly to overall prevalence in SSA immigrants
              estimated by Presanis et al. for 2001. Estimates from 2001 to
              2008 were taken directly from Presanis et al. estimates. The
              estimates of prevalence in SSA immigrants by Presanis et al.
              2008 prevalence estimates are extended to 2009-2011 for lack of
other
              data.
Label y:      Year, relative to 1981.
Label s:      Prevalence by sex, 0=male, 1=female.

|y  |s0     |s1
| 0 |.0000 |.0000
| 1 |.0007 |.0012
| 2 |.0013 |.0024
| 3 |.0020 |.0036
| 4 |.0026 |.0048
| 5 |.0033 |.0060
| 6 |.0039 |.0072
| 7 |.0046 |.0084
| 8 |.0052 |.0096
| 9 |.0059 |.0108
|10 |.0065 |.0119
|11 |.0072 |.0131
|12 |.0078 |.0143
|13 |.0085 |.0155
|14 |.0091 |.0167
|15 |.0098 |.0179
|16 |.0104 |.0191
|17 |.0111 |.0203
|18 |.0117 |.0215
|19 |.0124 |.0227
|20 |.0130 |.0239
|21 |.0214 |.0283
|22 |.0221 |.0328
|23 |.0185 |.0337
|24 |.0163 |.0317
|25 |.0176 |.0340
|26 |.0160 |.0325
|27 |.0155 |.0339
|28 |.0155 |.0339
|29 |.0155 |.0339
|30 |.0155 |.0339
```

## 10.3 Classification of LFS Respondents into Three Birthplace Categories

Table 10-1: Classification of Labour Force Survey (LFS) respondents' country of birth into the three birthplace categories for the model. Birthplace categories are United Kingdom-born (UK-born), Sub-Saharan Africa-born (SSA-born), and other foreign-born (OF-born). The table shows LFS classifications for 1981 data and corresponding birthplace category used in the model. For some LFS countries of birth, classification is ambiguous and categories are distributed among the three birthplace categories for the model based on assumptions described in the text. These categories are highlighted in bold and make up less than 1% of the total respondents in 1981. The third column of the table, 'Intermediary grouping for redistribution of categories' provides information that aids redistribution of the ambiguous categories. 1999 – 2009 LFS data fall into similar categories, but since data are more resolved, there are fewer ambiguities. LFS countries of birth in bold are those that are redistributed into one of the three birthplaces in the model, according to proportions stated.

| Country of birth from LFS | Birthplace category for model | Intermediary grouping for redistribution of categories |
|---|---|---|
| United Kingdom | UK | |
| Channel Islands | UK | |
| Isle of Man | UK | |
| Irish Republic | OF | |
| Canada | OF | |
| Australia | OF | |
| New Zealand | OF | |
| Kenya | SSA | New Commonwealth, Africa |
| Uganda | SSA | New Commonwealth, Africa |
| Tanzania | SSA | New Commonwealth, Africa |
| Malawi | SSA | New Commonwealth, Africa |
| Zambia | SSA | New Commonwealth, Africa |
| Zimbabwe | SSA | New Commonwealth, Africa |
| Botswana, Lesotho and Swaziland | SSA | New Commonwealth, Africa |
| Gambia | SSA | New Commonwealth, Africa |
| Ghana | SSA | New Commonwealth, Africa |
| Nigeria | SSA | New Commonwealth, Africa |
| Sierra Leone | SSA | New Commonwealth, Africa |
| Barbados | OF | New Commonwealth, non-Africa |
| Jamaica | OF | New Commonwealth, non-Africa |

| | | |
|---|---|---|
| Trinidad and Tobago | OF | New Commonwealth, non-Africa |
| West Indies | OF | New Commonwealth, non-Africa |
| Other Caribbean | OF | New Commonwealth, non-Africa |
| Belize | OF | New Commonwealth, non-Africa |
| Guyana | OF | New Commonwealth, non-Africa |
| Bangladesh | OF | New Commonwealth, non-Africa |
| India | OF | New Commonwealth, non-Africa |
| Sri Lanka | OF | New Commonwealth, non-Africa |
| Hong Kong | OF | New Commonwealth, non-Africa |
| Malaysia | OF | New Commonwealth, non-Africa |
| Singapore | OF | New Commonwealth, non-Africa |
| Cyprus | OF | New Commonwealth, non-Africa |
| Gibraltar | | New Commonwealth, non-Africa |
| Malta | OF | New Commonwealth, non-Africa |
| Seychelles | | New Commonwealth, Africa |
| Mauritius | | New Commonwealth, Africa |
| **Other new Commonwealth** | **Redistributed into OF/SSA based on ratio of New Commonwealth, non-African: New Commonwealth, African.** | |
| Algeria | OF | Foreign, North Africa |
| Morocco | OF | Foreign, North Africa |
| Tunisia | OF | Foreign, North Africa |
| Libya | OF | Foreign, North Africa |
| Egypt | OF | Foreign, North Africa |
| South Africa | SSA | Foreign, SSA |
| **Other Africa, foreign** | **Redistributed into OF/SSA, based on ratio of Foreign, North Africa: Foreign, South Africa.** | |
| USA | OF | |
| Caribbean | OF | |
| Central America | OF | |
| South America | OF | |
| Pakistan | OF | |

| | |
|---|---|
| Burma | OF |
| China | OF |
| Japan | OF |
| Philippines | OF |
| Vietnam | OF |
| Iran | OF |
| Israel | OF |
| Other Middle Eastern | OF |
| Other Asia, foreign | OF |
| Belgium | OF |
| Denmark | OF |
| France | OF |
| Italy | OF |
| Luxembourg | OF |
| Netherlands | OF |
| Germany, Federal Republic of | OF |
| Germany (PNS) | OF |
| Greece | OF |
| Albania | OF |
| Bulgaria | OF |
| German, Democratic republic of | OF |
| Czechoslovakia | OF |
| Hungary | OF |
| Poland | OF |
| Romania | OF |
| Austria | OF |
| Switzerland | OF |
| Portugal | OF |
| Spain | OF |
| Finland | OF |
| Norway | OF |
| Sweden | OF |
| Yugoslavia | OF |

| | |
|---|---|
| Other Europe | OF |
| Turkey | OF |
| USSR | OF |
| **Rest of the world** | **Redistributed into UK/OF/SSA, based on proportion of each after other redistributions.** |
| **At sea/in the air** | **Redistributed into UK/OF/SSA, based on proportion of each after other redistributions.** |
| **Not stated/no reply** | **Redistributed into UK/OF/SSA, based on proportion of each after other redistributions.** |
| **Not known** | **Redistributed into UK/OF/SSA, based on proportion of each after other redistributions.** |

## 10.4 Population Size Estimates Derived from Analysis of LFS Data and Comparison to Other Sources

Table 10-2: Population size estimates for England and Wales in 1981, used for model initialization. Estimates were computed using Labour Force Survey (LFS) data for one-year age classes, amalgamated into larger age classes for presentation here. See text for LFS analysis methods.

| | UK-born | | OF-born | | SSA-born | |
|---|---|---|---|---|---|---|
| Age class (years) | Males | Females | Males | Females | Males | Females |
| 0-9 | 3,100,097 | 2,942,244 | 64,623 | 58,380 | 8,689 | 7,949 |
| 10-19 | 3,889,003 | 3,714,639 | 120,282 | 114,955 | 35,546 | 38,165 |
| 20-29 | 3,196,662 | 3,125,856 | 253,385 | 273,498 | 63,513 | 60,742 |
| 30-39 | 3,159,263 | 3,108,019 | 241,041 | 258,896 | 36,345 | 35,447 |
| 40-49 | 2,495,251 | 2,467,101 | 245,384 | 241,086 | 20,066 | 16,073 |
| 50-59 | 2,528,484 | 2,636,362 | 237,742 | 232,400 | 9,653 | 9,375 |
| 60+ | 3,899,090 | 5,493,322 | 205,322 | 229,967 | 4,968 | 9,123 |
| Total | 22,267,850 | 23,487,543 | 1,367,779 | 1,409,182 | 178,780 | 176,874 |

Table 10-3: Population size estimates (thousands) from Labour Force Survey for England and Wales by age, sex, and birthplace.

| | | UK-born | | OF-born | | SSA-born | |
|---|---|---|---|---|---|---|---|
| Year | Age class (years) | Male | Female | Male | Female | Male | Female |
| 1999 | 0 – 14 | 4,958 | 4,693 | 124 | 116 | 23 | 22 |
| | 15 – 44 | 9,598 | 9,631 | 792 | 883 | 183 | 205 |
| | 45 – 64 | 5,500 | 5,524 | 431 | 537 | 66 | 61 |
| | 65+ | 3,087 | 4,230 | 246 | 302 | 13 | 12 |
| | Total | 23,143 | 24,078 | 1,592 | 1,839 | 284 | 300 |
| 2000 | 0 – 14 | 4,929 | 4,695 | 117 | 110 | 26 | 26 |
| | 15 – 44 | 9,597 | 9,604 | 873 | 929 | 161 | 212 |
| | 45 – 64 | 5,520 | 5,593 | 458 | 529 | 78 | 73 |
| | 65+ | 3,106 | 4,224 | 251 | 297 | 16 | 19 |
| | Total | 23,152 | 24,115 | 1,699 | 1,865 | 281 | 330 |
| 2001 | 0 – 14 | 4,895 | 4,640 | 137 | 112 | 24 | 23 |
| | 15 – 44 | 9,574 | 9,613 | 885 | 970 | 219 | 232 |
| | 45 – 64 | 5,572 | 5,663 | 470 | 519 | 71 | 76 |
| | 65+ | 3,129 | 4,202 | 272 | 330 | 14 | 15 |
| | Total | 23,170 | 24,118 | 1,765 | 1,931 | 328 | 346 |
| 2002 | 0 – 14 | 4,819 | 4,612 | 133 | 125 | 37 | 28 |
| | 15 – 44 | 9,581 | 9,571 | 967 | 1,013 | 229 | 237 |
| | 45 – 64 | 5,614 | 5,697 | 470 | 536 | 82 | 97 |
| | 65+ | 3,162 | 4,198 | 279 | 339 | 16 | 19 |
| | Total | 23,176 | 24,079 | 1,849 | 2,013 | 364 | 381 |
| 2003 | 0 – 14 | 4,774 | 4,564 | 158 | 142 | 33 | 35 |
| | 15 – 44 | 9,536 | 9,495 | 1,041 | 1,060 | 251 | 295 |
| | 45 – 64 | 5,688 | 5,760 | 452 | 537 | 91 | 104 |
| | 65+ | 3,217 | 4,215 | 265 | 332 | 17 | 20 |
| | Total | 23,215 | 24,034 | 1,916 | 2,071 | 391 | 453 |
| 2004 | 0 – 14 | 4,761 | 4,562 | 140 | 128 | 33 | 34 |

| Year | Age class (years) | UK-born | | OF-born | | SSA-born | |
|---|---|---|---|---|---|---|---|
| | | Male | Female | Male | Female | Male | Female |
| | 15 – 44 | 9,607 | 9,516 | 1,026 | 1,083 | 264 | 285 |
| | 45 – 64 | 5,715 | 5,798 | 488 | 568 | 96 | 107 |
| | 65+ | 3,259 | 4,255 | 262 | 305 | 18 | 18 |
| | Total | 23,343 | 24,131 | 1,915 | 2,084 | 412 | 444 |
| | 0 – 14 | 4,706 | 4,492 | 159 | 158 | 38 | 34 |
| | 15 – 44 | 9,615 | 9,510 | 1,103 | 1,161 | 279 | 310 |
| 2005 | 45 – 64 | 5,766 | 5,861 | 503 | 577 | 113 | 118 |
| | 65+ | 3,289 | 4,224 | 267 | 342 | 26 | 26 |
| | Total | 23,376 | 24,088 | 2,032 | 2,237 | 455 | 487 |
| | 0 – 14 | 4,668 | 4,445 | 175 | 165 | 41 | 38 |
| | 15 – 44 | 9,482 | 9,456 | 1,275 | 1,237 | 291 | 340 |
| 2006 | 45 – 64 | 5,845 | 5,951 | 516 | 598 | 125 | 120 |
| | 65+ | 3,320 | 4,211 | 273 | 355 | 24 | 29 |
| | Total | 23,316 | 24,063 | 2,239 | 2,355 | 481 | 527 |
| | 0 – 14 | 4,651 | 4,399 | 176 | 197 | 42 | 35 |
| | 15 – 44 | 9,391 | 9,349 | 1,431 | 1,378 | 296 | 333 |
| 2007 | 45 – 64 | 5,910 | 6,030 | 532 | 621 | 130 | 130 |
| | 65+ | 3,360 | 4,224 | 279 | 360 | 28 | 31 |
| | Total | 23,312 | 24,002 | 2,418 | 2,555 | 496 | 529 |
| | 0 – 14 | 4,675 | 4,450 | 185 | 170 | 40 | 49 |
| | 15 – 44 | 9,318 | 9,227 | 1,504 | 1,472 | 303 | 321 |
| 2008 | 45 – 64 | 5,956 | 6,097 | 551 | 627 | 156 | 158 |
| | 65+ | 3,424 | 4,237 | 286 | 391 | 32 | 39 |
| | Total | 23,374 | 24,011 | 2,526 | 2,660 | 531 | 567 |
| | 0 – 14 | 4,640 | 4,457 | 216 | 192 | 42 | 43 |
| | 15 – 44 | 9,329 | 9,144 | 1,545 | 1,528 | 291 | 342 |
| 2009 | 45 – 64 | 6,024 | 6,151 | 567 | 632 | 154 | 195 |
| | 65+ | 3,485 | 4,300 | 298 | 398 | 42 | 31 |
| | Total | 23,477 | 24,053 | 2,626 | 2,750 | 530 | 610 |

Table 10-4: Population size estimates for the West Midlands in 1981, used for model initialization.Estimates were computed using Labour Force Survey (LFS) data for one-year age classes, amalgamated into larger age classes for presentation here. See text for LFS analysis methods.

| | UK-born | | OF-born | | SSA-born | |
|---|---|---|---|---|---|---|
| Age class (years) | Males | Females | Males | Females | Males | Females |
| 0 – 9 | 339,697 | 318,881 | 7,412 | 6,467 | 450 | 1,432 |
| 10 – 19 | 426,111 | 405,400 | 11,614 | 12,074 | 5,354 | 5,806 |
| 20 – 29 | 324,237 | 315,536 | 37,059 | 37,522 | 6,439 | 5,042 |
| 30 – 39 | 333,120 | 320,017 | 29,376 | 34,422 | 1,898 | 1,900 |
| 40 – 49 | 263,937 | 257,517 | 34,853 | 30,918 | 941 | 1,178 |
| 50 – 59 | 260,338 | 273,106 | 39,995 | 27,900 | 1,251 | 701 |
| 60+ | 379,256 | 525,820 | 25,177 | 19,376 | 491 | 0 |
| Total | 2,326,696 | 2,416,277 | 185,486 | 168,679 | 16,824 | 16,059 |

Table 10-5: West Midlands population size estimates for calculation of notification rates, 2007 – 2011. Estimates were obtained from analysis of the Labour Force Survey.

| | | UK-born | | OF-born | |
|---|---|---|---|---|---|
| Year | Age | Male | Female | Male | Female |
| 2007 | 0 – 14 | 484,625 | 462,860 | 17,811 | 18,768 |
| | 15 – 44 | 964,035 | 957,430 | 115,785 | 112,914 |
| | 45 – 64 | 601,825 | 620,005 | 60,851 | 54,239 |
| | 65+ | 343,609 | 427,947 | 32,130 | 42,687 |
| | Total | 2,394,094 | 2,468,242 | 226,577 | 228,608 |
| 2008 | 0 – 14 | 494,330 | 465,286 | 15,748 | 13,228 |
| | 15 – 44 | 925,366 | 930,817 | 147,173 | 141,822 |
| | 45 – 64 | 594,774 | 617,257 | 74,327 | 63,919 |
| | 65+ | 352,010 | 430,665 | 32,003 | 45,978 |
| | Total | 2,366,480 | 2,444,025 | 269,251 | 264,947 |
| 2009 | 0 – 14 | 476,383 | 457,347 | 26,426 | 22,844 |
| | 15 – 44 | 925,399 | 914,393 | 150,950 | 153,112 |
| | 45 – 64 | 599,452 | 608,526 | 72,826 | 77,509 |
| | 65+ | 354,039 | 435,925 | 38,340 | 46,643 |
| | Total | 2,355,273 | 2,416,191 | 288,542 | 300,108 |
| 2010 | 0 – 14 | 481,254 | 457,019 | 19,045 | 23,067 |
| | 15 – 44 | 920,537 | 920,876 | 155,518 | 142,779 |
| | 45 – 64 | 614,898 | 619,334 | 63,244 | 73,414 |
| | 65+ | 359,688 | 440,477 | 40,850 | 47,577 |
| | Total | 2,376,377 | 2,437,706 | 278,657 | 286,837 |
| 2011 | 0 – 14 | 479,968 | 462,142 | 24,900 | 21,673 |
| | 15 – 44 | 924,747 | 917,097 | 138,638 | 144,889 |
| | 45 – 64 | 612,903 | 631,159 | 70,479 | 69,896 |
| | 65+ | 372,955 | 440,762 | 38,887 | 56,971 |
| | Total | 2,390,573 | 2,451,160 | 272,904 | 293,429 |

Table 10-6: Comparison of 1981 population estimates from Labour Force Survey analysis with 1981 census for England and Wales (thousands of persons). Note, the 1981 census does not allow estimation of the Sub-Saharan African population directly; rather, these estimates come with some assumptions.

| | UK-Born | | | OF-Born | | | SSA-born | | |
|---|---|---|---|---|---|---|---|---|---|
| | Males | Females | Total | Males | Females | Total | Males | Females | Total |
| LFS | 22,268 | 23,488 | 45,755 | 1,368 | 1,409 | 2,777 | 179 | 177 | 356 |
| Census estimate | 22,029 | 23,274 | 45,303 | 1,427 | 1,465 | 2,892 | 169 | 158 | 327 |

Table 10-7: Comparison of 2001 population estimates from Labour Force Survey analysis with 2001 census for England and Wales (thousands of persons) for United Kingdom-born (UK-born), other foreign-born (OF-born), and Sub-Saharan Africa-born (SSA-born).

| | UK-born | | OF-born | | SSA-born | |
|---|---|---|---|---|---|---|
| | Males | Females | Males | Females | Males | Females |
| LFS | 23,170 | 24,118 | 1,765 | 1,931 | 328 | 346 |
| Census | 23,144 | 24,263 | 1,824 | 2,072 | 359 | 382 |

Table 10-8: Comparison of 2004 – 2009 population estimates from Labour Force Survey (LFS) analysis with Annual Population Survey (APS) estimates for England and Wales (thousands of persons). LFS estimates are based on independent analysis of LFS data. APS estimates are official estimates published by the Office for National Statistics.

| | UK-born | | Foreign-born | | Total | |
|---|---|---|---|---|---|---|
| Year | LFS | APS | LFS | APS | LFS | APS |
| 2004 | 47,474 | 47,382 | 4,855 | 4,958 | 52,330 | 52,340 |
| 2005 | 47,463 | 47,426 | 5,212 | 5,268 | 52,675 | 52,694 |
| 2006 | 47,379 | 47,310 | 5,601 | 5,691 | 52,980 | 53,001 |
| 2007 | 47,314 | 47,337 | 5,998 | 5,991 | 53,312 | 53,328 |
| 2008 | 47,385 | 47,419 | 6,284 | 6,294 | 53,669 | 53,713 |
| 2009 | 47,530 | 47,557 | 6,516 | 6,488 | 54,046 | 54,045 |

## 10.5 Case Fatality Rates

Table 10-9: Estimated case fatality rates by year, age category, and site of disease used in the model.

| | Non-pulmonary case fatality rate (%) by age category | | | | | Pulmonary case fatality rate (%) by age category | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Year | 0-14 years | 15-34 years | 35-54 years | 55-74 years | 75+ years | 0-14 years | 15-34 years | 35-54 years | 55-74 years | 75+ years |
| 1981 | 0.51% | 0.81% | 3.03% | 10.26% | 36.80% | 0.51% | 0.84% | 4.18% | 15.54% | 38.82% |
| 1982 | 0.32% | 0.63% | 3.73% | 10.64% | 35.63% | 0.32% | 0.66% | 5.15% | 16.11% | 37.58% |
| 1983 | 0.65% | 0.34% | 2.64% | 9.62% | 32.23% | 0.65% | 0.35% | 3.65% | 14.56% | 34.00% |
| 1984 | 0.67% | 0.78% | 2.97% | 8.90% | 34.23% | 0.67% | 0.82% | 4.10% | 13.47% | 36.11% |
| 1985 | 0.70% | 0.94% | 2.73% | 10.05% | 33.96% | 0.70% | 0.97% | 3.76% | 15.22% | 35.82% |
| 1986 | 0.62% | 0.72% | 2.39% | 8.76% | 33.95% | 0.62% | 0.75% | 3.29% | 13.26% | 35.82% |
| 1987 | 0.46% | 0.87% | 3.11% | 10.29% | 31.49% | 0.46% | 0.91% | 4.29% | 15.58% | 33.22% |
| 1988 | 0.00% | 0.51% | 3.10% | 11.51% | 35.32% | 0.00% | 0.53% | 4.27% | 17.43% | 37.26% |
| 1989 | 0.23% | 0.80% | 2.43% | 9.41% | 33.45% | 0.23% | 0.83% | 3.35% | 14.25% | 35.29% |
| 1990 | 0.50% | 0.72% | 2.15% | 9.23% | 28.70% | 0.50% | 0.75% | 2.97% | 13.97% | 30.27% |
| 1991 | 0.22% | 0.64% | 2.87% | 8.92% | 38.94% | 0.22% | 0.67% | 3.96% | 13.50% | 41.07% |
| 1992 | 1.01% | 0.56% | 2.47% | 8.88% | 29.84% | 1.01% | 0.58% | 3.41% | 13.44% | 31.48% |
| 1993 | 0.71% | 0.69% | 2.61% | 8.83% | 28.81% | 0.71% | 0.72% | 3.60% | 13.37% | 30.39% |
| 1994 | 0.00% | 0.94% | 2.70% | 9.71% | 27.67% | 0.00% | 0.97% | 3.72% | 14.70% | 29.19% |
| 1995 | 1.05% | 1.00% | 2.43% | 10.22% | 31.03% | 1.05% | 1.05% | 3.35% | 15.47% | 32.73% |

| Year | Non-pulmonary case fatality rate (%) by age category | | | | | Pulmonary case fatality rate (%) by age category | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0-14 years | 15-34 years | 35-54 years | 55-74 years | 75+ years | 0-14 years | 15-34 years | 35-54 years | 55-74 years | 75+ years |
| 1996 | 0.27% | 0.59% | 2.58% | 9.27% | 29.68% | 0.27% | 0.62% | 3.56% | 14.03% | 31.31% |
| 1997 | 0.22% | 0.63% | 2.51% | 8.00% | 27.86% | 0.22% | 0.66% | 3.46% | 12.12% | 29.39% |
| 1998 | 0.00% | 0.72% | 2.44% | 8.10% | 29.19% | 0.00% | 0.75% | 3.36% | 12.26% | 30.79% |
| 1999 | 0.00% | 0.70% | 2.13% | 7.69% | 29.51% | 0.00% | 0.73% | 2.93% | 11.64% | 31.13% |
| 2000 | 0.77% | 0.56% | 1.78% | 7.36% | 28.33% | 0.77% | 0.58% | 2.46% | 11.14% | 29.89% |
| 2001 | 0.00% | 0.89% | 2.00% | 8.23% | 25.79% | 0.00% | 0.93% | 2.75% | 12.46% | 27.21% |
| 2002 | 0.00% | 0.89% | 1.93% | 7.99% | 25.35% | 0.00% | 0.93% | 2.67% | 12.09% | 26.74% |
| 2003 | 0.00% | 0.89% | 1.90% | 7.89% | 24.98% | 0.00% | 0.93% | 2.63% | 11.95% | 26.35% |
| 2004 | 0.00% | 0.89% | 1.85% | 7.79% | 24.60% | 0.00% | 0.93% | 2.55% | 11.79% | 25.95% |
| 2005 | 0.00% | 0.89% | 1.82% | 7.61% | 24.21% | 0.00% | 0.93% | 2.51% | 11.52% | 25.54% |
| 2006 | 0.00% | 0.89% | 1.76% | 7.40% | 23.86% | 0.00% | 0.93% | 2.43% | 11.20% | 25.16% |
| 2007 | 0.00% | 0.89% | 1.73% | 7.26% | 23.48% | 0.00% | 0.93% | 2.39% | 11.00% | 24.77% |
| 2008 | 0.00% | 0.89% | 1.69% | 7.05% | 23.15% | 0.00% | 0.93% | 2.33% | 10.68% | 24.42% |
| 2009 | 0.00% | 0.89% | 1.64% | 6.87% | 22.78% | 0.00% | 0.93% | 2.27% | 10.40% | 24.03% |
| 2010 | 0.00% | 0.89% | 1.60% | 6.70% | 22.41% | 0.00% | 0.93% | 2.21% | 10.14% | 23.64% |
| 2011 | 0.00% | 0.89% | 1.56% | 6.53% | 22.06% | 0.00% | 0.93% | 2.15% | 9.88% | 23.27% |

## 10.6 HIV Prevalence in SSA-born

Table 10-10: Human Immunodeficiency Virus (HIV) prevalence assumed in Sub-Saharan Africa-born migrants entering the UK 1981-2008. Assumptions were based on 2001 to 2008 estimates from Presanis et al. 2010 [288].

| | Assumed HIV prevalence (%) | |
|---|---|---|
| Year | Males | Females |
| 1981 | 0.00% | 0.00% |
| 1982 | 0.07% | 0.12% |
| 1983 | 0.13% | 0.24% |
| 1984 | 0.20% | 0.36% |
| 1985 | 0.26% | 0.48% |
| 1986 | 0.33% | 0.60% |
| 1987 | 0.39% | 0.72% |
| 1988 | 0.46% | 0.84% |
| 1989 | 0.52% | 0.96% |
| 1990 | 0.59% | 1.08% |
| 1991 | 0.65% | 1.19% |
| 1992 | 0.72% | 1.31% |
| 1993 | 0.78% | 1.43% |
| 1994 | 0.85% | 1.55% |
| 1995 | 0.91% | 1.67% |
| 1996 | 0.98% | 1.79% |
| 1997 | 1.04% | 1.91% |
| 1998 | 1.11% | 2.03% |
| 1999 | 1.17% | 2.15% |
| 2000 | 1.24% | 2.27% |
| 2001 | 1.30% | 2.39% |
| 2002 | 2.14% | 2.83% |
| 2003 | 2.21% | 3.28% |
| 2004 | 1.85% | 3.37% |
| 2005 | 1.63% | 3.17% |
| 2006 | 1.76% | 3.40% |
| 2007 | 1.60% | 3.25% |
| 2008 | 1.55% | 3.39% |

## 10.7 Infection State Probabilities for Model Initialization in 1981

Table 10-11: Proportion by infection state for UK-born males in 1981.

| | UK-born males, proportion of individuals in each infection state in 1981 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Age class (years) | *Uninfected* | *Immune* | *Recent Infection* | *Latent Infection* | *Reinfection* | *Primary Disease* | *Reactivation Disease* | *Reinfection Disease* |
| 0 | 0.999867 | 0.000000 | 0.000063 | 0.000000 | 0.000000 | 0.000070 | 0.000000 | 0.000000 |
| 1 | 0.999571 | 0.000000 | 0.000353 | 0.000000 | 0.000000 | 0.000076 | 0.000000 | 0.000000 |
| 2 | 0.999231 | 0.000000 | 0.000693 | 0.000000 | 0.000000 | 0.000076 | 0.000000 | 0.000000 |
| 3 | 0.998840 | 0.000000 | 0.001084 | 0.000000 | 0.000000 | 0.000076 | 0.000000 | 0.000000 |
| 4 | 0.998391 | 0.000000 | 0.001533 | 0.000000 | 0.000000 | 0.000076 | 0.000000 | 0.000000 |
| 5 | 0.997875 | 0.000000 | 0.001781 | 0.000267 | 0.000000 | 0.000076 | 0.000000 | 0.000000 |
| 6 | 0.997282 | 0.000000 | 0.001780 | 0.000861 | 0.000000 | 0.000076 | 0.000000 | 0.000000 |
| 7 | 0.996602 | 0.000000 | 0.001779 | 0.001542 | 0.000001 | 0.000076 | 0.000001 | 0.000000 |
| 8 | 0.995820 | 0.000000 | 0.001778 | 0.002324 | 0.000002 | 0.000075 | 0.000001 | 0.000000 |
| 9 | 0.994922 | 0.000000 | 0.001777 | 0.003222 | 0.000003 | 0.000075 | 0.000001 | 0.000000 |
| 10 | 0.993890 | 0.000000 | 0.001776 | 0.004253 | 0.000005 | 0.000074 | 0.000002 | 0.000000 |
| 11 | 0.992706 | 0.000000 | 0.001774 | 0.005437 | 0.000007 | 0.000074 | 0.000002 | 0.000000 |
| 12 | 0.991347 | 0.000000 | 0.001772 | 0.006796 | 0.000009 | 0.000073 | 0.000003 | 0.000000 |
| 13 | 0.989786 | 0.000000 | 0.001770 | 0.008355 | 0.000012 | 0.000072 | 0.000004 | 0.000000 |
| 14 | 0.987995 | 0.000000 | 0.001768 | 0.010146 | 0.000015 | 0.000072 | 0.000004 | 0.000000 |
| 15 | 0.985940 | 0.000000 | 0.001765 | 0.012200 | 0.000019 | 0.000071 | 0.000005 | 0.000000 |
| 16 | 0.983583 | 0.000000 | 0.001761 | 0.014556 | 0.000024 | 0.000070 | 0.000006 | 0.000001 |
| 17 | 0.980881 | 0.000000 | 0.001757 | 0.017257 | 0.000028 | 0.000069 | 0.000007 | 0.000001 |
| 18 | 0.977783 | 0.000000 | 0.001753 | 0.020354 | 0.000034 | 0.000067 | 0.000008 | 0.000001 |
| 19 | 0.974233 | 0.000000 | 0.001748 | 0.023902 | 0.000041 | 0.000066 | 0.000009 | 0.000001 |
| 20 | 0.970168 | 0.000000 | 0.001742 | 0.027966 | 0.000048 | 0.000064 | 0.000011 | 0.000001 |
| 21 | 0.965515 | 0.000000 | 0.001735 | 0.032617 | 0.000057 | 0.000063 | 0.000012 | 0.000001 |

| UK-born males, proportion of individuals in each infection state in 1981 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Age class (years) | *Uninfected* | *Immune* | *Recent Infection* | *Latent Infection* | *Reinfection* | *Primary Disease* | *Reactivation Disease* | *Reinfection Disease* |
| 22 | 0.960193 | 0.000000 | 0.001727 | 0.037938 | 0.000066 | 0.000061 | 0.000014 | 0.000001 |
| 23 | 0.954109 | 0.000000 | 0.001717 | 0.044020 | 0.000077 | 0.000059 | 0.000016 | 0.000002 |
| 24 | 0.947160 | 0.000000 | 0.001706 | 0.050967 | 0.000090 | 0.000057 | 0.000018 | 0.000002 |
| 25 | 0.939231 | 0.000000 | 0.001694 | 0.058894 | 0.000105 | 0.000054 | 0.000020 | 0.000002 |
| 26 | 0.930194 | 0.000000 | 0.001680 | 0.067928 | 0.000121 | 0.000052 | 0.000022 | 0.000002 |
| 27 | 0.919907 | 0.000000 | 0.001663 | 0.078213 | 0.000140 | 0.000049 | 0.000024 | 0.000002 |
| 28 | 0.908214 | 0.000000 | 0.001644 | 0.089904 | 0.000162 | 0.000047 | 0.000027 | 0.000003 |
| 29 | 0.894945 | 0.000000 | 0.001622 | 0.103171 | 0.000186 | 0.000044 | 0.000029 | 0.000003 |
| 30 | 0.879916 | 0.000000 | 0.001597 | 0.118197 | 0.000214 | 0.000041 | 0.000032 | 0.000003 |
| 31 | 0.863161 | 0.000000 | 0.001569 | 0.134949 | 0.000245 | 0.000038 | 0.000035 | 0.000004 |
| 32 | 0.845859 | 0.000000 | 0.001539 | 0.152248 | 0.000277 | 0.000035 | 0.000037 | 0.000004 |
| 33 | 0.828217 | 0.000000 | 0.001509 | 0.169888 | 0.000310 | 0.000033 | 0.000039 | 0.000004 |
| 34 | 0.810241 | 0.000000 | 0.001478 | 0.187861 | 0.000343 | 0.000031 | 0.000041 | 0.000004 |
| 35 | 0.791943 | 0.000000 | 0.001446 | 0.206158 | 0.000377 | 0.000029 | 0.000043 | 0.000004 |
| 36 | 0.773331 | 0.000000 | 0.001413 | 0.224768 | 0.000412 | 0.000027 | 0.000045 | 0.000005 |
| 37 | 0.754420 | 0.000000 | 0.001380 | 0.243678 | 0.000447 | 0.000025 | 0.000047 | 0.000005 |
| 38 | 0.735222 | 0.000000 | 0.001345 | 0.262874 | 0.000483 | 0.000023 | 0.000048 | 0.000005 |
| 39 | 0.715754 | 0.000000 | 0.001311 | 0.282340 | 0.000519 | 0.000022 | 0.000050 | 0.000005 |
| 40 | 0.696032 | 0.000000 | 0.001276 | 0.302061 | 0.000555 | 0.000020 | 0.000051 | 0.000005 |
| 41 | 0.676075 | 0.000000 | 0.001240 | 0.322016 | 0.000593 | 0.000019 | 0.000052 | 0.000005 |
| 42 | 0.655904 | 0.000000 | 0.001203 | 0.342186 | 0.000630 | 0.000018 | 0.000053 | 0.000005 |
| 43 | 0.635541 | 0.000000 | 0.001167 | 0.362548 | 0.000668 | 0.000016 | 0.000054 | 0.000006 |
| 44 | 0.615009 | 0.000000 | 0.001130 | 0.383079 | 0.000706 | 0.000015 | 0.000055 | 0.000006 |
| 45 | 0.594335 | 0.000000 | 0.001092 | 0.403752 | 0.000745 | 0.000014 | 0.000056 | 0.000006 |
| 46 | 0.573545 | 0.000000 | 0.001054 | 0.424541 | 0.000783 | 0.000013 | 0.000057 | 0.000006 |

| UK-born males, proportion of individuals in each infection state in 1981 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Age class (years) | *Uninfected* | *Immune* | *Recent Infection* | *Latent Infection* | *Reinfection* | *Primary Disease* | *Reactivation Disease* | *Reinfection Disease* |
| 47 | 0.552668 | 0.000000 | 0.001016 | 0.445417 | 0.000822 | 0.000012 | 0.000058 | 0.000006 |
| 48 | 0.531734 | 0.000000 | 0.000978 | 0.466350 | 0.000861 | 0.000012 | 0.000059 | 0.000006 |
| 49 | 0.510777 | 0.000000 | 0.000940 | 0.487307 | 0.000900 | 0.000011 | 0.000059 | 0.000006 |
| 50 | 0.489828 | 0.000000 | 0.000902 | 0.508254 | 0.000939 | 0.000010 | 0.000060 | 0.000006 |
| 51 | 0.468924 | 0.000000 | 0.000864 | 0.529158 | 0.000978 | 0.000009 | 0.000061 | 0.000006 |
| 52 | 0.448100 | 0.000000 | 0.000826 | 0.549981 | 0.001017 | 0.000009 | 0.000061 | 0.000006 |
| 53 | 0.427393 | 0.000000 | 0.000788 | 0.570688 | 0.001056 | 0.000008 | 0.000062 | 0.000006 |
| 54 | 0.406842 | 0.000000 | 0.000750 | 0.591238 | 0.001094 | 0.000007 | 0.000062 | 0.000006 |
| 55 | 0.386485 | 0.000000 | 0.000713 | 0.611595 | 0.001132 | 0.000007 | 0.000063 | 0.000006 |
| 56 | 0.366362 | 0.000000 | 0.000676 | 0.631717 | 0.001169 | 0.000006 | 0.000063 | 0.000007 |
| 57 | 0.346513 | 0.000000 | 0.000639 | 0.651565 | 0.001206 | 0.000006 | 0.000064 | 0.000007 |
| 58 | 0.326978 | 0.000000 | 0.000603 | 0.671100 | 0.001243 | 0.000005 | 0.000064 | 0.000007 |
| 59 | 0.307796 | 0.000000 | 0.000568 | 0.690281 | 0.001279 | 0.000005 | 0.000065 | 0.000007 |
| 60 | 0.289007 | 0.000000 | 0.000533 | 0.709070 | 0.001314 | 0.000005 | 0.000065 | 0.000007 |
| 61 | 0.270649 | 0.000000 | 0.000500 | 0.727427 | 0.001348 | 0.000004 | 0.000065 | 0.000007 |
| 62 | 0.252760 | 0.000000 | 0.000467 | 0.745316 | 0.001381 | 0.000004 | 0.000066 | 0.000007 |
| 63 | 0.235375 | 0.000000 | 0.000435 | 0.762700 | 0.001414 | 0.000003 | 0.000066 | 0.000007 |
| 64 | 0.218529 | 0.000000 | 0.000404 | 0.779546 | 0.001445 | 0.000003 | 0.000066 | 0.000007 |
| 65 | 0.202255 | 0.000000 | 0.000374 | 0.795820 | 0.001475 | 0.000003 | 0.000066 | 0.000007 |
| 66 | 0.186582 | 0.000000 | 0.000345 | 0.811493 | 0.001505 | 0.000003 | 0.000067 | 0.000007 |
| 67 | 0.171536 | 0.000000 | 0.000317 | 0.826538 | 0.001533 | 0.000002 | 0.000067 | 0.000007 |
| 68 | 0.157143 | 0.000000 | 0.000290 | 0.840931 | 0.001559 | 0.000002 | 0.000067 | 0.000007 |
| 69 | 0.143422 | 0.000000 | 0.000265 | 0.854651 | 0.001585 | 0.000002 | 0.000067 | 0.000007 |
| 70 | 0.130392 | 0.000000 | 0.000241 | 0.867682 | 0.001609 | 0.000002 | 0.000068 | 0.000007 |
| 71 | 0.118065 | 0.000000 | 0.000218 | 0.880008 | 0.001632 | 0.000002 | 0.000068 | 0.000007 |

| UK-born males, proportion of individuals in each infection state in 1981 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Age class (years) | *Uninfected* | *Immune* | *Recent Infection* | *Latent Infection* | *Reinfection* | *Primary Disease* | *Reactivation Disease* | *Reinfection Disease* |
| 72 | 0.106450 | 0.000000 | 0.000197 | 0.891623 | 0.001654 | 0.000001 | 0.000068 | 0.000007 |
| 73 | 0.095554 | 0.000000 | 0.000177 | 0.902519 | 0.001674 | 0.000001 | 0.000068 | 0.000007 |
| 74 | 0.085376 | 0.000000 | 0.000158 | 0.912697 | 0.001693 | 0.000001 | 0.000068 | 0.000007 |
| 75 | 0.075913 | 0.000000 | 0.000140 | 0.922159 | 0.001711 | 0.000001 | 0.000068 | 0.000007 |
| 76 | 0.067158 | 0.000000 | 0.000124 | 0.930914 | 0.001727 | 0.000001 | 0.000068 | 0.000007 |
| 77 | 0.059098 | 0.000000 | 0.000109 | 0.938974 | 0.001742 | 0.000001 | 0.000068 | 0.000007 |
| 78 | 0.051718 | 0.000000 | 0.000096 | 0.946354 | 0.001756 | 0.000001 | 0.000069 | 0.000007 |
| 79 | 0.045014 | 0.000000 | 0.000083 | 0.953058 | 0.001769 | 0.000001 | 0.000069 | 0.000007 |
| 80 | 0.039045 | 0.000000 | 0.000072 | 0.959027 | 0.001780 | 0.000000 | 0.000069 | 0.000007 |
| 81 | 0.033761 | 0.000000 | 0.000062 | 0.964311 | 0.001790 | 0.000000 | 0.000069 | 0.000007 |
| 82 | 0.029099 | 0.000000 | 0.000054 | 0.968972 | 0.001798 | 0.000000 | 0.000069 | 0.000007 |
| 83 | 0.024999 | 0.000000 | 0.000046 | 0.973072 | 0.001806 | 0.000000 | 0.000069 | 0.000007 |
| 84 | 0.021405 | 0.000000 | 0.000040 | 0.976666 | 0.001813 | 0.000000 | 0.000069 | 0.000007 |
| 85 | 0.018265 | 0.000000 | 0.000034 | 0.979807 | 0.001818 | 0.000000 | 0.000069 | 0.000007 |
| 86 | 0.015531 | 0.000000 | 0.000029 | 0.982541 | 0.001824 | 0.000000 | 0.000069 | 0.000007 |
| 87 | 0.013159 | 0.000000 | 0.000024 | 0.984913 | 0.001828 | 0.000000 | 0.000069 | 0.000007 |
| 88 | 0.011108 | 0.000000 | 0.000021 | 0.986964 | 0.001832 | 0.000000 | 0.000069 | 0.000007 |
| 89 | 0.009341 | 0.000000 | 0.000017 | 0.988730 | 0.001835 | 0.000000 | 0.000069 | 0.000007 |
| 90 | 0.007825 | 0.000000 | 0.000014 | 0.990246 | 0.001838 | 0.000000 | 0.000069 | 0.000007 |
| 91 | 0.006529 | 0.000000 | 0.000012 | 0.991542 | 0.001840 | 0.000000 | 0.000069 | 0.000007 |
| 92 | 0.005426 | 0.000000 | 0.000010 | 0.992645 | 0.001842 | 0.000000 | 0.000069 | 0.000007 |
| 93 | 0.004490 | 0.000000 | 0.000008 | 0.993581 | 0.001844 | 0.000000 | 0.000069 | 0.000007 |
| 94 | 0.003700 | 0.000000 | 0.000007 | 0.994371 | 0.001846 | 0.000000 | 0.000069 | 0.000007 |
| 95 | 0.003036 | 0.000000 | 0.000006 | 0.995035 | 0.001847 | 0.000000 | 0.000069 | 0.000007 |
| 96 | 0.002480 | 0.000000 | 0.000005 | 0.995591 | 0.001848 | 0.000000 | 0.000069 | 0.000007 |

| UK-born males, proportion of individuals in each infection state in 1981 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Age class (years) | *Uninfected* | *Immune* | *Recent Infection* | *Latent Infection* | *Reinfection* | *Primary Disease* | *Reactivation Disease* | *Reinfection Disease* |
| 97 | 0.002016 | 0.000000 | 0.000004 | 0.996055 | 0.001849 | 0.000000 | 0.000069 | 0.000007 |
| 98 | 0.001632 | 0.000000 | 0.000003 | 0.996440 | 0.001849 | 0.000000 | 0.000069 | 0.000007 |
| 99 | 0.001314 | 0.000000 | 0.000002 | 0.996757 | 0.001850 | 0.000000 | 0.000069 | 0.000007 |
| 100 | 0.001054 | 0.000000 | 0.000002 | 0.997018 | 0.001851 | 0.000000 | 0.000069 | 0.000007 |
| 101 | 0.000844 | 0.000000 | 0.000002 | 0.997227 | 0.001851 | 0.000000 | 0.000069 | 0.000007 |
| 102 | 0.000677 | 0.000000 | 0.000001 | 0.997395 | 0.001851 | 0.000000 | 0.000069 | 0.000007 |
| 103 | 0.000542 | 0.000000 | 0.000001 | 0.997529 | 0.001851 | 0.000000 | 0.000069 | 0.000007 |
| 104 | 0.000434 | 0.000000 | 0.000001 | 0.997637 | 0.001852 | 0.000000 | 0.000069 | 0.000007 |
| 105 | 0.000348 | 0.000000 | 0.000001 | 0.997723 | 0.001852 | 0.000000 | 0.000069 | 0.000007 |
| 106 | 0.000279 | 0.000000 | 0.000001 | 0.997792 | 0.001852 | 0.000000 | 0.000069 | 0.000007 |
| 107 | 0.000224 | 0.000000 | 0.000000 | 0.997848 | 0.001852 | 0.000000 | 0.000069 | 0.000007 |
| 108 | 0.000179 | 0.000000 | 0.000000 | 0.997892 | 0.001852 | 0.000000 | 0.000069 | 0.000007 |
| 109 | 0.000144 | 0.000000 | 0.000000 | 0.997928 | 0.001852 | 0.000000 | 0.000069 | 0.000007 |
| 110 | 0.000115 | 0.000000 | 0.000000 | 0.997956 | 0.001852 | 0.000000 | 0.000069 | 0.000007 |
| 111 | 0.000092 | 0.000000 | 0.000000 | 0.997979 | 0.001852 | 0.000000 | 0.000069 | 0.000007 |
| 112 | 0.000074 | 0.000000 | 0.000000 | 0.997997 | 0.001852 | 0.000000 | 0.000069 | 0.000007 |
| 113 | 0.000059 | 0.000000 | 0.000000 | 0.998012 | 0.001852 | 0.000000 | 0.000069 | 0.000007 |
| 114 | 0.000047 | 0.000000 | 0.000000 | 0.998024 | 0.001852 | 0.000000 | 0.000069 | 0.000007 |
| 115 | 0.000038 | 0.000000 | 0.000000 | 0.998033 | 0.001852 | 0.000000 | 0.000069 | 0.000007 |
| 116 | 0.000030 | 0.000000 | 0.000000 | 0.998041 | 0.001852 | 0.000000 | 0.000069 | 0.000007 |
| 117 | 0.000024 | 0.000000 | 0.000000 | 0.998047 | 0.001852 | 0.000000 | 0.000069 | 0.000007 |
| 118 | 0.000020 | 0.000000 | 0.000000 | 0.998052 | 0.001852 | 0.000000 | 0.000069 | 0.000007 |
| 119 | 0.000016 | 0.000000 | 0.000000 | 0.998056 | 0.001852 | 0.000000 | 0.000069 | 0.000007 |
| 120 | 0.000013 | 0.000000 | 0.000000 | 0.998059 | 0.001852 | 0.000000 | 0.000069 | 0.000007 |

Table 10-12: Proportion by infection state for UK-born females in 1981.

| | UK-born females, proportion of individuals in each infection state in 1981 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Age class (years) | *Uninfected* | *Immune* | *Recent Infection* | *Latent Infection* | *Reinfection* | *Primary Disease* | *Reactivation Disease* | *Reinfection Disease* |
| 0 | 0.999867 | 0.000000 | 0.000063 | 0.000000 | 0.000000 | 0.000070 | 0.000000 | 0.000000 |
| 1 | 0.999571 | 0.000000 | 0.000353 | 0.000000 | 0.000000 | 0.000076 | 0.000000 | 0.000000 |
| 2 | 0.999231 | 0.000000 | 0.000693 | 0.000000 | 0.000000 | 0.000076 | 0.000000 | 0.000000 |
| 3 | 0.998840 | 0.000000 | 0.001084 | 0.000000 | 0.000000 | 0.000076 | 0.000000 | 0.000000 |
| 4 | 0.998391 | 0.000000 | 0.001533 | 0.000000 | 0.000000 | 0.000076 | 0.000000 | 0.000000 |
| 5 | 0.997875 | 0.000000 | 0.001781 | 0.000267 | 0.000000 | 0.000076 | 0.000000 | 0.000000 |
| 6 | 0.997282 | 0.000000 | 0.001780 | 0.000861 | 0.000000 | 0.000076 | 0.000000 | 0.000000 |
| 7 | 0.996602 | 0.000000 | 0.001779 | 0.001542 | 0.000001 | 0.000076 | 0.000000 | 0.000000 |
| 8 | 0.995820 | 0.000000 | 0.001778 | 0.002325 | 0.000001 | 0.000076 | 0.000000 | 0.000000 |
| 9 | 0.994922 | 0.000000 | 0.001776 | 0.003223 | 0.000003 | 0.000076 | 0.000000 | 0.000000 |
| 10 | 0.993890 | 0.000000 | 0.001774 | 0.004255 | 0.000005 | 0.000076 | 0.000000 | 0.000000 |
| 11 | 0.992706 | 0.000000 | 0.001772 | 0.005439 | 0.000007 | 0.000076 | 0.000000 | 0.000000 |
| 12 | 0.991347 | 0.000000 | 0.001770 | 0.006798 | 0.000010 | 0.000076 | 0.000000 | 0.000000 |
| 13 | 0.989786 | 0.000000 | 0.001767 | 0.008358 | 0.000012 | 0.000076 | 0.000001 | 0.000000 |
| 14 | 0.987995 | 0.000000 | 0.001764 | 0.010149 | 0.000016 | 0.000076 | 0.000001 | 0.000000 |
| 15 | 0.985940 | 0.000000 | 0.001760 | 0.012204 | 0.000020 | 0.000075 | 0.000001 | 0.000000 |
| 16 | 0.983583 | 0.000000 | 0.001756 | 0.014561 | 0.000024 | 0.000075 | 0.000001 | 0.000000 |
| 17 | 0.980881 | 0.000000 | 0.001751 | 0.017263 | 0.000029 | 0.000075 | 0.000001 | 0.000000 |
| 18 | 0.977783 | 0.000000 | 0.001745 | 0.020361 | 0.000035 | 0.000075 | 0.000001 | 0.000000 |
| 19 | 0.974233 | 0.000000 | 0.001739 | 0.023910 | 0.000042 | 0.000075 | 0.000002 | 0.000000 |
| 20 | 0.970168 | 0.000000 | 0.001732 | 0.027975 | 0.000049 | 0.000074 | 0.000002 | 0.000000 |
| 21 | 0.965515 | 0.000000 | 0.001723 | 0.032627 | 0.000058 | 0.000074 | 0.000002 | 0.000000 |
| 22 | 0.960193 | 0.000000 | 0.001714 | 0.037949 | 0.000068 | 0.000074 | 0.000003 | 0.000000 |
| 23 | 0.954109 | 0.000000 | 0.001703 | 0.044033 | 0.000079 | 0.000073 | 0.000003 | 0.000000 |

| Age class (years) | UK-born females, proportion of individuals in each infection state in 1981 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Uninfected* | *Immune* | *Recent Infection* | *Latent Infection* | *Reinfection* | *Primary Disease* | *Reactivation Disease* | *Reinfection Disease* |
| 24 | 0.947160 | 0.000000 | 0.001691 | 0.050981 | 0.000092 | 0.000073 | 0.000004 | 0.000000 |
| 25 | 0.939231 | 0.000000 | 0.001676 | 0.058909 | 0.000107 | 0.000072 | 0.000004 | 0.000000 |
| 26 | 0.930194 | 0.000000 | 0.001660 | 0.067945 | 0.000124 | 0.000071 | 0.000005 | 0.000000 |
| 27 | 0.919907 | 0.000000 | 0.001642 | 0.078232 | 0.000143 | 0.000071 | 0.000006 | 0.000000 |
| 28 | 0.908214 | 0.000000 | 0.001621 | 0.089924 | 0.000165 | 0.000070 | 0.000006 | 0.000000 |
| 29 | 0.894945 | 0.000000 | 0.001597 | 0.103193 | 0.000189 | 0.000069 | 0.000007 | 0.000000 |
| 30 | 0.879916 | 0.000000 | 0.001570 | 0.118220 | 0.000218 | 0.000068 | 0.000009 | 0.000000 |
| 31 | 0.863161 | 0.000000 | 0.001540 | 0.134974 | 0.000249 | 0.000067 | 0.000010 | 0.000000 |
| 32 | 0.845859 | 0.000000 | 0.001509 | 0.152274 | 0.000281 | 0.000065 | 0.000011 | 0.000000 |
| 33 | 0.828217 | 0.000000 | 0.001478 | 0.169915 | 0.000314 | 0.000064 | 0.000012 | 0.000000 |
| 34 | 0.810241 | 0.000000 | 0.001446 | 0.187889 | 0.000348 | 0.000063 | 0.000014 | 0.000000 |
| 35 | 0.791943 | 0.000000 | 0.001413 | 0.206187 | 0.000382 | 0.000061 | 0.000015 | 0.000000 |
| 36 | 0.773331 | 0.000000 | 0.001380 | 0.224797 | 0.000416 | 0.000060 | 0.000016 | 0.000000 |
| 37 | 0.754420 | 0.000000 | 0.001346 | 0.243707 | 0.000452 | 0.000059 | 0.000018 | 0.000000 |
| 38 | 0.735222 | 0.000000 | 0.001311 | 0.262903 | 0.000488 | 0.000057 | 0.000019 | 0.000000 |
| 39 | 0.715754 | 0.000000 | 0.001277 | 0.282370 | 0.000524 | 0.000056 | 0.000021 | 0.000000 |
| 40 | 0.696032 | 0.000000 | 0.001241 | 0.302090 | 0.000561 | 0.000054 | 0.000022 | 0.000000 |
| 41 | 0.676075 | 0.000000 | 0.001206 | 0.322045 | 0.000598 | 0.000053 | 0.000023 | 0.000000 |
| 42 | 0.655904 | 0.000000 | 0.001170 | 0.342214 | 0.000636 | 0.000051 | 0.000025 | 0.000000 |
| 43 | 0.635541 | 0.000000 | 0.001133 | 0.362576 | 0.000674 | 0.000050 | 0.000027 | 0.000000 |
| 44 | 0.615009 | 0.000000 | 0.001097 | 0.383106 | 0.000712 | 0.000048 | 0.000028 | 0.000000 |
| 45 | 0.594335 | 0.000000 | 0.001060 | 0.403779 | 0.000750 | 0.000047 | 0.000030 | 0.000000 |
| 46 | 0.573545 | 0.000000 | 0.001023 | 0.424567 | 0.000789 | 0.000045 | 0.000031 | 0.000000 |
| 47 | 0.552668 | 0.000000 | 0.000985 | 0.445442 | 0.000828 | 0.000043 | 0.000033 | 0.000000 |
| 48 | 0.531734 | 0.000000 | 0.000948 | 0.466374 | 0.000867 | 0.000042 | 0.000034 | 0.000000 |

| UK-born females, proportion of individuals in each infection state in 1981 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Age class (years) | *Uninfected* | *Immune* | *Recent Infection* | *Latent Infection* | *Reinfection* | *Primary Disease* | *Reactivation Disease* | *Reinfection Disease* |
| 49 | 0.510777 | 0.000000 | 0.000911 | 0.487330 | 0.000906 | 0.000040 | 0.000036 | 0.000000 |
| 50 | 0.489828 | 0.000000 | 0.000873 | 0.508277 | 0.000945 | 0.000039 | 0.000038 | 0.000000 |
| 51 | 0.468924 | 0.000000 | 0.000836 | 0.529180 | 0.000984 | 0.000037 | 0.000039 | 0.000000 |
| 52 | 0.448100 | 0.000000 | 0.000799 | 0.550002 | 0.001023 | 0.000035 | 0.000041 | 0.000000 |
| 53 | 0.427393 | 0.000000 | 0.000762 | 0.570707 | 0.001062 | 0.000034 | 0.000042 | 0.000000 |
| 54 | 0.406842 | 0.000000 | 0.000725 | 0.591257 | 0.001100 | 0.000032 | 0.000044 | 0.000000 |
| 55 | 0.386485 | 0.000000 | 0.000689 | 0.611612 | 0.001138 | 0.000031 | 0.000045 | 0.000000 |
| 56 | 0.366362 | 0.000000 | 0.000653 | 0.631733 | 0.001176 | 0.000029 | 0.000047 | 0.000000 |
| 57 | 0.346513 | 0.000000 | 0.000617 | 0.651581 | 0.001213 | 0.000028 | 0.000049 | 0.000000 |
| 58 | 0.326978 | 0.000000 | 0.000583 | 0.671114 | 0.001249 | 0.000026 | 0.000050 | 0.000000 |
| 59 | 0.307796 | 0.000000 | 0.000548 | 0.690294 | 0.001285 | 0.000025 | 0.000052 | 0.000000 |
| 60 | 0.289007 | 0.000000 | 0.000515 | 0.709082 | 0.001320 | 0.000023 | 0.000053 | 0.000000 |
| 61 | 0.270649 | 0.000000 | 0.000482 | 0.727438 | 0.001354 | 0.000022 | 0.000054 | 0.000000 |
| 62 | 0.252760 | 0.000000 | 0.000450 | 0.745326 | 0.001388 | 0.000020 | 0.000056 | 0.000000 |
| 63 | 0.235375 | 0.000000 | 0.000419 | 0.762709 | 0.001420 | 0.000019 | 0.000057 | 0.000000 |
| 64 | 0.218529 | 0.000000 | 0.000389 | 0.779553 | 0.001452 | 0.000018 | 0.000059 | 0.000000 |
| 65 | 0.202255 | 0.000000 | 0.000360 | 0.795826 | 0.001482 | 0.000016 | 0.000060 | 0.000000 |
| 66 | 0.186582 | 0.000000 | 0.000332 | 0.811499 | 0.001511 | 0.000015 | 0.000061 | 0.000000 |
| 67 | 0.171536 | 0.000000 | 0.000305 | 0.826543 | 0.001539 | 0.000014 | 0.000062 | 0.000000 |
| 68 | 0.157143 | 0.000000 | 0.000280 | 0.840935 | 0.001566 | 0.000013 | 0.000063 | 0.000000 |
| 69 | 0.143422 | 0.000000 | 0.000255 | 0.854654 | 0.001592 | 0.000012 | 0.000065 | 0.000000 |
| 70 | 0.130392 | 0.000000 | 0.000232 | 0.867683 | 0.001616 | 0.000011 | 0.000066 | 0.000000 |
| 71 | 0.118065 | 0.000000 | 0.000210 | 0.880010 | 0.001639 | 0.000010 | 0.000067 | 0.000000 |
| 72 | 0.106450 | 0.000000 | 0.000190 | 0.891623 | 0.001661 | 0.000009 | 0.000068 | 0.000000 |
| 73 | 0.095554 | 0.000000 | 0.000170 | 0.902519 | 0.001681 | 0.000008 | 0.000068 | 0.000000 |

| | UK-born females, proportion of individuals in each infection state in 1981 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Age class (years) | *Uninfected* | *Immune* | *Recent Infection* | *Latent Infection* | *Reinfection* | *Primary Disease* | *Reactivation Disease* | *Reinfection Disease* |
| 74 | 0.085376 | 0.000000 | 0.000152 | 0.912696 | 0.001700 | 0.000007 | 0.000069 | 0.000000 |
| 75 | 0.075913 | 0.000000 | 0.000135 | 0.922158 | 0.001718 | 0.000006 | 0.000070 | 0.000000 |
| 76 | 0.067158 | 0.000000 | 0.000120 | 0.930912 | 0.001734 | 0.000005 | 0.000071 | 0.000000 |
| 77 | 0.059098 | 0.000000 | 0.000105 | 0.938971 | 0.001749 | 0.000005 | 0.000071 | 0.000000 |
| 78 | 0.051718 | 0.000000 | 0.000092 | 0.946350 | 0.001763 | 0.000004 | 0.000072 | 0.000000 |
| 79 | 0.045014 | 0.000000 | 0.000080 | 0.953054 | 0.001776 | 0.000004 | 0.000073 | 0.000000 |
| 80 | 0.039045 | 0.000000 | 0.000070 | 0.959023 | 0.001787 | 0.000003 | 0.000073 | 0.000000 |
| 81 | 0.033761 | 0.000000 | 0.000060 | 0.964306 | 0.001797 | 0.000003 | 0.000073 | 0.000000 |
| 82 | 0.029099 | 0.000000 | 0.000052 | 0.968967 | 0.001805 | 0.000002 | 0.000074 | 0.000000 |
| 83 | 0.024999 | 0.000000 | 0.000044 | 0.973067 | 0.001813 | 0.000002 | 0.000074 | 0.000000 |
| 84 | 0.021405 | 0.000000 | 0.000038 | 0.976661 | 0.001820 | 0.000002 | 0.000074 | 0.000000 |
| 85 | 0.018265 | 0.000000 | 0.000033 | 0.979801 | 0.001825 | 0.000001 | 0.000075 | 0.000000 |
| 86 | 0.015531 | 0.000000 | 0.000028 | 0.982535 | 0.001831 | 0.000001 | 0.000075 | 0.000000 |
| 87 | 0.013159 | 0.000000 | 0.000023 | 0.984907 | 0.001835 | 0.000001 | 0.000075 | 0.000000 |
| 88 | 0.011108 | 0.000000 | 0.000020 | 0.986958 | 0.001839 | 0.000001 | 0.000075 | 0.000000 |
| 89 | 0.009341 | 0.000000 | 0.000017 | 0.988724 | 0.001842 | 0.000001 | 0.000075 | 0.000000 |
| 90 | 0.007825 | 0.000000 | 0.000014 | 0.990240 | 0.001845 | 0.000001 | 0.000076 | 0.000000 |
| 91 | 0.006529 | 0.000000 | 0.000012 | 0.991536 | 0.001847 | 0.000001 | 0.000076 | 0.000000 |
| 92 | 0.005426 | 0.000000 | 0.000010 | 0.992639 | 0.001849 | 0.000000 | 0.000076 | 0.000000 |
| 93 | 0.004490 | 0.000000 | 0.000008 | 0.993574 | 0.001851 | 0.000000 | 0.000076 | 0.000000 |
| 94 | 0.003700 | 0.000000 | 0.000007 | 0.994364 | 0.001853 | 0.000000 | 0.000076 | 0.000000 |
| 95 | 0.003036 | 0.000000 | 0.000005 | 0.995028 | 0.001854 | 0.000000 | 0.000076 | 0.000000 |
| 96 | 0.002480 | 0.000000 | 0.000004 | 0.995585 | 0.001855 | 0.000000 | 0.000076 | 0.000000 |
| 97 | 0.002016 | 0.000000 | 0.000004 | 0.996048 | 0.001856 | 0.000000 | 0.000076 | 0.000000 |
| 98 | 0.001632 | 0.000000 | 0.000003 | 0.996433 | 0.001856 | 0.000000 | 0.000076 | 0.000000 |

| UK-born females, proportion of individuals in each infection state in 1981 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Age class (years) | *Uninfected* | *Immune* | *Recent Infection* | *Latent Infection* | *Reinfection* | *Primary Disease* | *Reactivation Disease* | *Reinfection Disease* |
| 99 | 0.001314 | 0.000000 | 0.000002 | 0.996750 | 0.001857 | 0.000000 | 0.000076 | 0.000000 |
| 100 | 0.001054 | 0.000000 | 0.000002 | 0.997011 | 0.001858 | 0.000000 | 0.000076 | 0.000000 |
| 101 | 0.000844 | 0.000000 | 0.000002 | 0.997220 | 0.001858 | 0.000000 | 0.000076 | 0.000000 |
| 102 | 0.000677 | 0.000000 | 0.000001 | 0.997388 | 0.001858 | 0.000000 | 0.000076 | 0.000000 |
| 103 | 0.000542 | 0.000000 | 0.000001 | 0.997522 | 0.001859 | 0.000000 | 0.000076 | 0.000000 |
| 104 | 0.000434 | 0.000000 | 0.000001 | 0.997630 | 0.001859 | 0.000000 | 0.000076 | 0.000000 |
| 105 | 0.000348 | 0.000000 | 0.000001 | 0.997716 | 0.001859 | 0.000000 | 0.000076 | 0.000000 |
| 106 | 0.000279 | 0.000000 | 0.000000 | 0.997785 | 0.001859 | 0.000000 | 0.000076 | 0.000000 |
| 107 | 0.000224 | 0.000000 | 0.000000 | 0.997841 | 0.001859 | 0.000000 | 0.000076 | 0.000000 |
| 108 | 0.000179 | 0.000000 | 0.000000 | 0.997885 | 0.001859 | 0.000000 | 0.000076 | 0.000000 |
| 109 | 0.000144 | 0.000000 | 0.000000 | 0.997921 | 0.001859 | 0.000000 | 0.000076 | 0.000000 |
| 110 | 0.000115 | 0.000000 | 0.000000 | 0.997949 | 0.001859 | 0.000000 | 0.000076 | 0.000000 |
| 111 | 0.000092 | 0.000000 | 0.000000 | 0.997972 | 0.001859 | 0.000000 | 0.000076 | 0.000000 |
| 112 | 0.000074 | 0.000000 | 0.000000 | 0.997990 | 0.001859 | 0.000000 | 0.000076 | 0.000000 |
| 113 | 0.000059 | 0.000000 | 0.000000 | 0.998005 | 0.001859 | 0.000000 | 0.000076 | 0.000000 |
| 114 | 0.000047 | 0.000000 | 0.000000 | 0.998017 | 0.001859 | 0.000000 | 0.000076 | 0.000000 |
| 115 | 0.000038 | 0.000000 | 0.000000 | 0.998026 | 0.001859 | 0.000000 | 0.000076 | 0.000000 |
| 116 | 0.000030 | 0.000000 | 0.000000 | 0.998034 | 0.001859 | 0.000000 | 0.000076 | 0.000000 |
| 117 | 0.000024 | 0.000000 | 0.000000 | 0.998040 | 0.001859 | 0.000000 | 0.000076 | 0.000000 |
| 118 | 0.000020 | 0.000000 | 0.000000 | 0.998045 | 0.001859 | 0.000000 | 0.000076 | 0.000000 |
| 119 | 0.000016 | 0.000000 | 0.000000 | 0.998049 | 0.001859 | 0.000000 | 0.000076 | 0.000000 |
| 120 | 0.000013 | 0.000000 | 0.000000 | 0.998052 | 0.001860 | 0.000000 | 0.000076 | 0.000000 |

Table 10-13: Proportion by infection state for foreign-born males in 1981.

| Age class (years) | Foreign-born males, proportion of individuals by infection state in 1981 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Uninfected* | *Immune* | *Recent Infection* | *Latent Infection* | *Reinfection* | *Primary Disease* | *Reactivation Disease* | *Reinfection Disease* |
| 0 | 0.99582 | 0.00000 | 0.00351 | 0.00000 | 0.00000 | 0.00066 | 0.00000 | 0.00000 |
| 1 | 0.98751 | 0.00000 | 0.01183 | 0.00000 | 0.00000 | 0.00066 | 0.00000 | 0.00000 |
| 2 | 0.97926 | 0.00000 | 0.02008 | 0.00000 | 0.00000 | 0.00066 | 0.00000 | 0.00000 |
| 3 | 0.97108 | 0.00000 | 0.02826 | 0.00000 | 0.00000 | 0.00066 | 0.00000 | 0.00000 |
| 4 | 0.96297 | 0.00000 | 0.03637 | 0.00000 | 0.00000 | 0.00066 | 0.00000 | 0.00000 |
| 5 | 0.95493 | 0.00000 | 0.04023 | 0.00414 | 0.00003 | 0.00066 | 0.00000 | 0.00000 |
| 6 | 0.94695 | 0.00000 | 0.03989 | 0.01235 | 0.00014 | 0.00066 | 0.00000 | 0.00000 |
| 7 | 0.93904 | 0.00000 | 0.03956 | 0.02043 | 0.00030 | 0.00065 | 0.00000 | 0.00001 |
| 8 | 0.93120 | 0.00000 | 0.03923 | 0.02836 | 0.00054 | 0.00065 | 0.00000 | 0.00002 |
| 9 | 0.92342 | 0.00000 | 0.03891 | 0.03617 | 0.00084 | 0.00064 | 0.00000 | 0.00002 |
| 10 | 0.91571 | 0.00000 | 0.03859 | 0.04387 | 0.00117 | 0.00063 | 0.00000 | 0.00003 |
| 11 | 0.90806 | 0.00000 | 0.03826 | 0.05150 | 0.00151 | 0.00062 | 0.00000 | 0.00004 |
| 12 | 0.90048 | 0.00000 | 0.03794 | 0.05907 | 0.00184 | 0.00061 | 0.00000 | 0.00005 |
| 13 | 0.89296 | 0.00000 | 0.03762 | 0.06658 | 0.00217 | 0.00060 | 0.00000 | 0.00006 |
| 14 | 0.88550 | 0.00000 | 0.03731 | 0.07402 | 0.00250 | 0.00060 | 0.00000 | 0.00007 |
| 15 | 0.87811 | 0.00000 | 0.03699 | 0.08141 | 0.00283 | 0.00059 | 0.00000 | 0.00008 |
| 16 | 0.87077 | 0.00000 | 0.03668 | 0.08873 | 0.00315 | 0.00058 | 0.00000 | 0.00008 |
| 17 | 0.86350 | 0.00000 | 0.03637 | 0.09599 | 0.00348 | 0.00057 | 0.00000 | 0.00009 |
| 18 | 0.85629 | 0.00000 | 0.03607 | 0.10319 | 0.00379 | 0.00056 | 0.00000 | 0.00010 |
| 19 | 0.84914 | 0.00000 | 0.03576 | 0.11033 | 0.00411 | 0.00056 | 0.00000 | 0.00011 |
| 20 | 0.84204 | 0.00000 | 0.03546 | 0.11741 | 0.00442 | 0.00055 | 0.00000 | 0.00012 |
| 21 | 0.83501 | 0.00000 | 0.03517 | 0.12443 | 0.00473 | 0.00066 | 0.00000 | 0.00000 |
| 22 | 0.82804 | 0.00000 | 0.03487 | 0.13139 | 0.00504 | 0.00066 | 0.00000 | 0.00000 |
| 23 | 0.82112 | 0.00000 | 0.03458 | 0.13829 | 0.00534 | 0.00066 | 0.00000 | 0.00000 |

| Age class (years) | Foreign-born males, proportion of individuals by infection state in 1981 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Uninfected* | *Immune* | *Recent Infection* | *Latent Infection* | *Reinfection* | *Primary Disease* | *Reactivation Disease* | *Reinfection Disease* |
| 24 | 0.81426 | 0.00000 | 0.03429 | 0.14514 | 0.00564 | 0.00066 | 0.00000 | 0.00000 |
| 25 | 0.80746 | 0.00000 | 0.03400 | 0.15193 | 0.00594 | 0.00066 | 0.00001 | 0.00000 |
| 26 | 0.80022 | 0.00000 | 0.03370 | 0.15916 | 0.00626 | 0.00066 | 0.00001 | 0.00000 |
| 27 | 0.79197 | 0.00000 | 0.03335 | 0.16740 | 0.00662 | 0.00066 | 0.00001 | 0.00000 |
| 28 | 0.78258 | 0.00000 | 0.03295 | 0.17678 | 0.00703 | 0.00066 | 0.00001 | 0.00000 |
| 29 | 0.77191 | 0.00000 | 0.03250 | 0.18742 | 0.00750 | 0.00066 | 0.00001 | 0.00000 |
| 30 | 0.75983 | 0.00000 | 0.03199 | 0.19949 | 0.00803 | 0.00066 | 0.00001 | 0.00000 |
| 31 | 0.74615 | 0.00000 | 0.03142 | 0.21314 | 0.00863 | 0.00066 | 0.00001 | 0.00000 |
| 32 | 0.73149 | 0.00000 | 0.03080 | 0.22778 | 0.00927 | 0.00066 | 0.00001 | 0.00000 |
| 33 | 0.71654 | 0.00000 | 0.03017 | 0.24271 | 0.00992 | 0.00065 | 0.00001 | 0.00000 |
| 34 | 0.70129 | 0.00000 | 0.02953 | 0.25793 | 0.01059 | 0.00065 | 0.00001 | 0.00000 |
| 35 | 0.68577 | 0.00000 | 0.02887 | 0.27342 | 0.01127 | 0.00065 | 0.00001 | 0.00000 |
| 36 | 0.66997 | 0.00000 | 0.02820 | 0.28920 | 0.01197 | 0.00065 | 0.00001 | 0.00000 |
| 37 | 0.65391 | 0.00000 | 0.02753 | 0.30523 | 0.01267 | 0.00065 | 0.00001 | 0.00000 |
| 38 | 0.63760 | 0.00000 | 0.02684 | 0.32151 | 0.01338 | 0.00065 | 0.00001 | 0.00000 |
| 39 | 0.62105 | 0.00000 | 0.02614 | 0.33803 | 0.01411 | 0.00065 | 0.00001 | 0.00000 |
| 40 | 0.60428 | 0.00000 | 0.02544 | 0.35478 | 0.01484 | 0.00065 | 0.00002 | 0.00000 |
| 41 | 0.58730 | 0.00000 | 0.02472 | 0.37173 | 0.01559 | 0.00065 | 0.00002 | 0.00000 |
| 42 | 0.57012 | 0.00000 | 0.02400 | 0.38888 | 0.01634 | 0.00065 | 0.00002 | 0.00000 |
| 43 | 0.55277 | 0.00000 | 0.02326 | 0.40620 | 0.01710 | 0.00064 | 0.00002 | 0.00000 |
| 44 | 0.53527 | 0.00000 | 0.02253 | 0.42367 | 0.01787 | 0.00064 | 0.00002 | 0.00000 |
| 45 | 0.51763 | 0.00000 | 0.02178 | 0.44128 | 0.01864 | 0.00064 | 0.00002 | 0.00000 |
| 46 | 0.49989 | 0.00000 | 0.02104 | 0.45900 | 0.01942 | 0.00064 | 0.00002 | 0.00000 |
| 47 | 0.48205 | 0.00000 | 0.02028 | 0.47680 | 0.02020 | 0.00064 | 0.00003 | 0.00000 |
| 48 | 0.46416 | 0.00000 | 0.01953 | 0.49467 | 0.02098 | 0.00064 | 0.00003 | 0.00000 |

| Age class (years) | Foreign-born males, proportion of individuals by infection state in 1981 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Uninfected* | *Immune* | *Recent Infection* | *Latent Infection* | *Reinfection* | *Primary Disease* | *Reactivation Disease* | *Reinfection Disease* |
| 49 | 0.44623 | 0.00000 | 0.01877 | 0.51257 | 0.02177 | 0.00063 | 0.00003 | 0.00000 |
| 50 | 0.42829 | 0.00000 | 0.01802 | 0.53048 | 0.02255 | 0.00063 | 0.00003 | 0.00000 |
| 51 | 0.41037 | 0.00000 | 0.01726 | 0.54836 | 0.02334 | 0.00063 | 0.00003 | 0.00000 |
| 52 | 0.39251 | 0.00000 | 0.01651 | 0.56619 | 0.02412 | 0.00063 | 0.00004 | 0.00000 |
| 53 | 0.37474 | 0.00000 | 0.01576 | 0.58394 | 0.02490 | 0.00062 | 0.00004 | 0.00000 |
| 54 | 0.35708 | 0.00000 | 0.01502 | 0.60157 | 0.02567 | 0.00062 | 0.00004 | 0.00000 |
| 55 | 0.33957 | 0.00000 | 0.01428 | 0.61905 | 0.02644 | 0.00062 | 0.00005 | 0.00000 |
| 56 | 0.32224 | 0.00000 | 0.01355 | 0.63635 | 0.02719 | 0.00061 | 0.00005 | 0.00000 |
| 57 | 0.30513 | 0.00000 | 0.01283 | 0.65343 | 0.02794 | 0.00061 | 0.00005 | 0.00000 |
| 58 | 0.28827 | 0.00000 | 0.01212 | 0.67026 | 0.02868 | 0.00061 | 0.00006 | 0.00000 |
| 59 | 0.27170 | 0.00000 | 0.01143 | 0.68681 | 0.02941 | 0.00060 | 0.00006 | 0.00000 |
| 60 | 0.25544 | 0.00000 | 0.01074 | 0.70303 | 0.03012 | 0.00060 | 0.00007 | 0.00000 |
| 61 | 0.23954 | 0.00000 | 0.01007 | 0.71891 | 0.03081 | 0.00059 | 0.00007 | 0.00000 |
| 62 | 0.22402 | 0.00000 | 0.00942 | 0.73440 | 0.03149 | 0.00058 | 0.00008 | 0.00000 |
| 63 | 0.20893 | 0.00000 | 0.00878 | 0.74947 | 0.03215 | 0.00058 | 0.00008 | 0.00000 |
| 64 | 0.19427 | 0.00000 | 0.00817 | 0.76410 | 0.03279 | 0.00057 | 0.00009 | 0.00000 |
| 65 | 0.18009 | 0.00000 | 0.00757 | 0.77826 | 0.03341 | 0.00056 | 0.00010 | 0.00000 |
| 66 | 0.16642 | 0.00000 | 0.00700 | 0.79191 | 0.03401 | 0.00055 | 0.00011 | 0.00000 |
| 67 | 0.15327 | 0.00000 | 0.00644 | 0.80504 | 0.03459 | 0.00055 | 0.00012 | 0.00000 |
| 68 | 0.14067 | 0.00000 | 0.00591 | 0.81762 | 0.03514 | 0.00054 | 0.00013 | 0.00000 |
| 69 | 0.12863 | 0.00000 | 0.00541 | 0.82963 | 0.03567 | 0.00052 | 0.00014 | 0.00000 |
| 70 | 0.11718 | 0.00000 | 0.00493 | 0.84106 | 0.03617 | 0.00051 | 0.00015 | 0.00000 |
| 71 | 0.10633 | 0.00000 | 0.00447 | 0.85190 | 0.03664 | 0.00050 | 0.00016 | 0.00000 |
| 72 | 0.09608 | 0.00000 | 0.00404 | 0.86213 | 0.03709 | 0.00048 | 0.00018 | 0.00000 |
| 73 | 0.08644 | 0.00000 | 0.00363 | 0.87175 | 0.03751 | 0.00047 | 0.00019 | 0.00000 |

| Age class (years) | Foreign-born males, proportion of individuals by infection state in 1981 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Uninfected* | *Immune* | *Recent Infection* | *Latent Infection* | *Reinfection* | *Primary Disease* | *Reactivation Disease* | *Reinfection Disease* |
| 74 | 0.07742 | 0.00000 | 0.00325 | 0.88076 | 0.03791 | 0.00045 | 0.00021 | 0.00000 |
| 75 | 0.06901 | 0.00000 | 0.00290 | 0.88915 | 0.03827 | 0.00043 | 0.00023 | 0.00000 |
| 76 | 0.06121 | 0.00000 | 0.00257 | 0.89693 | 0.03861 | 0.00041 | 0.00025 | 0.00000 |
| 77 | 0.05402 | 0.00000 | 0.00227 | 0.90412 | 0.03893 | 0.00039 | 0.00027 | 0.00000 |
| 78 | 0.04741 | 0.00000 | 0.00199 | 0.91072 | 0.03922 | 0.00037 | 0.00029 | 0.00001 |
| 79 | 0.04137 | 0.00000 | 0.00174 | 0.91674 | 0.03948 | 0.00035 | 0.00031 | 0.00001 |
| 80 | 0.03594 | 0.00000 | 0.00151 | 0.92216 | 0.03972 | 0.00032 | 0.00034 | 0.00001 |
| 81 | 0.03113 | 0.00000 | 0.00131 | 0.92697 | 0.03993 | 0.00030 | 0.00036 | 0.00001 |
| 82 | 0.02688 | 0.00000 | 0.00113 | 0.93121 | 0.04012 | 0.00027 | 0.00038 | 0.00001 |
| 83 | 0.02313 | 0.00000 | 0.00097 | 0.93496 | 0.04028 | 0.00025 | 0.00041 | 0.00001 |
| 84 | 0.01984 | 0.00000 | 0.00083 | 0.93824 | 0.04042 | 0.00022 | 0.00043 | 0.00001 |
| 85 | 0.01696 | 0.00000 | 0.00071 | 0.94112 | 0.04055 | 0.00020 | 0.00045 | 0.00001 |
| 86 | 0.01445 | 0.00000 | 0.00061 | 0.94362 | 0.04066 | 0.00018 | 0.00048 | 0.00001 |
| 87 | 0.01226 | 0.00000 | 0.00052 | 0.94580 | 0.04075 | 0.00016 | 0.00050 | 0.00001 |
| 88 | 0.01037 | 0.00000 | 0.00044 | 0.94769 | 0.04084 | 0.00014 | 0.00052 | 0.00001 |
| 89 | 0.00874 | 0.00000 | 0.00037 | 0.94932 | 0.04091 | 0.00012 | 0.00053 | 0.00001 |
| 90 | 0.00734 | 0.00000 | 0.00031 | 0.95072 | 0.04097 | 0.00010 | 0.00055 | 0.00001 |
| 91 | 0.00613 | 0.00000 | 0.00026 | 0.95192 | 0.04102 | 0.00009 | 0.00056 | 0.00001 |
| 92 | 0.00511 | 0.00000 | 0.00021 | 0.95295 | 0.04107 | 0.00008 | 0.00058 | 0.00001 |
| 93 | 0.00424 | 0.00000 | 0.00018 | 0.95382 | 0.04111 | 0.00006 | 0.00059 | 0.00001 |
| 94 | 0.00350 | 0.00000 | 0.00015 | 0.95455 | 0.04114 | 0.00005 | 0.00060 | 0.00001 |
| 95 | 0.00288 | 0.00000 | 0.00012 | 0.95517 | 0.04116 | 0.00005 | 0.00061 | 0.00001 |
| 96 | 0.00236 | 0.00000 | 0.00010 | 0.95569 | 0.04119 | 0.00004 | 0.00062 | 0.00001 |
| 97 | 0.00192 | 0.00000 | 0.00008 | 0.95613 | 0.04121 | 0.00003 | 0.00062 | 0.00001 |
| 98 | 0.00156 | 0.00000 | 0.00007 | 0.95649 | 0.04122 | 0.00003 | 0.00063 | 0.00001 |

| Age class (years) | Foreign-born males, proportion of individuals by infection state in 1981 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Uninfected* | *Immune* | *Recent Infection* | *Latent Infection* | *Reinfection* | *Primary Disease* | *Reactivation Disease* | *Reinfection Disease* |
| 99 | 0.00126 | 0.00000 | 0.00005 | 0.95679 | 0.04124 | 0.00002 | 0.00063 | 0.00001 |
| 100 | 0.00101 | 0.00000 | 0.00004 | 0.95704 | 0.04125 | 0.00002 | 0.00064 | 0.00001 |
| 101 | 0.00081 | 0.00000 | 0.00003 | 0.95724 | 0.04126 | 0.00001 | 0.00064 | 0.00001 |
| 102 | 0.00065 | 0.00000 | 0.00003 | 0.95740 | 0.04126 | 0.00001 | 0.00064 | 0.00001 |
| 103 | 0.00052 | 0.00000 | 0.00002 | 0.95753 | 0.04127 | 0.00001 | 0.00064 | 0.00001 |
| 104 | 0.00042 | 0.00000 | 0.00002 | 0.95763 | 0.04127 | 0.00001 | 0.00064 | 0.00001 |
| 105 | 0.00033 | 0.00000 | 0.00001 | 0.95771 | 0.04128 | 0.00001 | 0.00065 | 0.00001 |
| 106 | 0.00027 | 0.00000 | 0.00001 | 0.95778 | 0.04128 | 0.00000 | 0.00065 | 0.00001 |
| 107 | 0.00021 | 0.00000 | 0.00001 | 0.95783 | 0.04128 | 0.00000 | 0.00065 | 0.00001 |
| 108 | 0.00017 | 0.00000 | 0.00001 | 0.95787 | 0.04128 | 0.00000 | 0.00065 | 0.00001 |
| 109 | 0.00014 | 0.00000 | 0.00001 | 0.95791 | 0.04128 | 0.00000 | 0.00065 | 0.00001 |
| 110 | 0.00011 | 0.00000 | 0.00000 | 0.95794 | 0.04129 | 0.00000 | 0.00065 | 0.00001 |
| 111 | 0.00009 | 0.00000 | 0.00000 | 0.95796 | 0.04129 | 0.00000 | 0.00065 | 0.00001 |
| 112 | 0.00007 | 0.00000 | 0.00000 | 0.95798 | 0.04129 | 0.00000 | 0.00065 | 0.00001 |
| 113 | 0.00006 | 0.00000 | 0.00000 | 0.95799 | 0.04129 | 0.00000 | 0.00065 | 0.00001 |
| 114 | 0.00005 | 0.00000 | 0.00000 | 0.95800 | 0.04129 | 0.00000 | 0.00065 | 0.00001 |
| 115 | 0.00004 | 0.00000 | 0.00000 | 0.95801 | 0.04129 | 0.00000 | 0.00065 | 0.00001 |
| 116 | 0.00003 | 0.00000 | 0.00000 | 0.95802 | 0.04129 | 0.00000 | 0.00065 | 0.00001 |
| 117 | 0.00002 | 0.00000 | 0.00000 | 0.95802 | 0.04129 | 0.00000 | 0.00065 | 0.00001 |
| 118 | 0.00002 | 0.00000 | 0.00000 | 0.95803 | 0.04129 | 0.00000 | 0.00065 | 0.00001 |
| 119 | 0.00002 | 0.00000 | 0.00000 | 0.95803 | 0.04129 | 0.00000 | 0.00065 | 0.00001 |
| 120 | 0.00001 | 0.00000 | 0.00000 | 0.95803 | 0.04129 | 0.00000 | 0.00065 | 0.00001 |

Table 10-14: Proportion by infection state for foreign-born females in 1981.

| Foreign-born females, proportion by infection state in 1981 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Age category (years) | *Uninfected* | *Immune* | *Recent Infection* | *Latent Infection* | *Reinfection* | *Primary Disease* | *Reactivation Disease* | *Reinfection Disease* |
| 0 | 0.99582 | 0.00000 | 0.00351 | 0.00000 | 0.00000 | 0.00066 | 0.00000 | 0.00000 |
| 1 | 0.98751 | 0.00000 | 0.01183 | 0.00000 | 0.00000 | 0.00066 | 0.00000 | 0.00000 |
| 2 | 0.97926 | 0.00000 | 0.02008 | 0.00000 | 0.00000 | 0.00066 | 0.00000 | 0.00000 |
| 3 | 0.97108 | 0.00000 | 0.02826 | 0.00000 | 0.00000 | 0.00066 | 0.00000 | 0.00000 |
| 4 | 0.96297 | 0.00000 | 0.03637 | 0.00000 | 0.00000 | 0.00066 | 0.00000 | 0.00000 |
| 5 | 0.95493 | 0.00000 | 0.04023 | 0.00414 | 0.00003 | 0.00066 | 0.00000 | 0.00000 |
| 6 | 0.94695 | 0.00000 | 0.03989 | 0.01235 | 0.00014 | 0.00066 | 0.00000 | 0.00000 |
| 7 | 0.93904 | 0.00000 | 0.03956 | 0.02043 | 0.00030 | 0.00065 | 0.00000 | 0.00001 |
| 8 | 0.93120 | 0.00000 | 0.03923 | 0.02836 | 0.00054 | 0.00065 | 0.00000 | 0.00002 |
| 9 | 0.92342 | 0.00000 | 0.03891 | 0.03617 | 0.00084 | 0.00064 | 0.00000 | 0.00002 |
| 10 | 0.91571 | 0.00000 | 0.03859 | 0.04387 | 0.00117 | 0.00063 | 0.00000 | 0.00003 |
| 11 | 0.90806 | 0.00000 | 0.03827 | 0.05150 | 0.00150 | 0.00062 | 0.00000 | 0.00004 |
| 12 | 0.90048 | 0.00000 | 0.03795 | 0.05908 | 0.00183 | 0.00061 | 0.00000 | 0.00005 |
| 13 | 0.89296 | 0.00000 | 0.03764 | 0.06659 | 0.00215 | 0.00060 | 0.00000 | 0.00006 |
| 14 | 0.88550 | 0.00000 | 0.03733 | 0.07403 | 0.00247 | 0.00060 | 0.00000 | 0.00007 |
| 15 | 0.87811 | 0.00000 | 0.03702 | 0.08142 | 0.00279 | 0.00059 | 0.00000 | 0.00008 |
| 16 | 0.87077 | 0.00000 | 0.03671 | 0.08874 | 0.00311 | 0.00058 | 0.00000 | 0.00008 |
| 17 | 0.86350 | 0.00000 | 0.03641 | 0.09600 | 0.00342 | 0.00057 | 0.00000 | 0.00009 |
| 18 | 0.85629 | 0.00000 | 0.03611 | 0.10320 | 0.00374 | 0.00056 | 0.00000 | 0.00010 |
| 19 | 0.84914 | 0.00000 | 0.03581 | 0.11035 | 0.00405 | 0.00056 | 0.00000 | 0.00011 |
| 20 | 0.84204 | 0.00000 | 0.03551 | 0.11743 | 0.00435 | 0.00055 | 0.00000 | 0.00012 |
| 21 | 0.83501 | 0.00000 | 0.03510 | 0.12445 | 0.00478 | 0.00066 | 0.00000 | 0.00000 |

| Foreign-born females, proportion by infection state in 1981 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Age category (years) | *Uninfected* | *Immune* | *Recent Infection* | *Latent Infection* | *Reinfection* | *Primary Disease* | *Reactivation Disease* | *Reinfection Disease* |
| 22 | 0.82804 | 0.00000 | 0.03480 | 0.13141 | 0.00509 | 0.00066 | 0.00000 | 0.00000 |
| 23 | 0.82112 | 0.00000 | 0.03451 | 0.13831 | 0.00539 | 0.00066 | 0.00000 | 0.00000 |
| 24 | 0.81426 | 0.00000 | 0.03421 | 0.14516 | 0.00570 | 0.00066 | 0.00000 | 0.00000 |
| 25 | 0.80746 | 0.00000 | 0.03392 | 0.15195 | 0.00600 | 0.00066 | 0.00001 | 0.00000 |
| 26 | 0.80022 | 0.00000 | 0.03361 | 0.15919 | 0.00632 | 0.00066 | 0.00001 | 0.00000 |
| 27 | 0.79197 | 0.00000 | 0.03326 | 0.16743 | 0.00669 | 0.00066 | 0.00001 | 0.00000 |
| 28 | 0.78258 | 0.00000 | 0.03286 | 0.17680 | 0.00710 | 0.00066 | 0.00001 | 0.00000 |
| 29 | 0.77191 | 0.00000 | 0.03240 | 0.18745 | 0.00757 | 0.00066 | 0.00001 | 0.00000 |
| 30 | 0.75983 | 0.00000 | 0.03188 | 0.19952 | 0.00811 | 0.00066 | 0.00001 | 0.00000 |
| 31 | 0.74615 | 0.00000 | 0.03130 | 0.21317 | 0.00871 | 0.00066 | 0.00001 | 0.00000 |
| 32 | 0.73149 | 0.00000 | 0.03067 | 0.22781 | 0.00936 | 0.00066 | 0.00001 | 0.00000 |
| 33 | 0.71654 | 0.00000 | 0.03003 | 0.24274 | 0.01003 | 0.00065 | 0.00001 | 0.00000 |
| 34 | 0.70129 | 0.00000 | 0.02938 | 0.25796 | 0.01070 | 0.00065 | 0.00001 | 0.00000 |
| 35 | 0.68577 | 0.00000 | 0.02872 | 0.27346 | 0.01139 | 0.00065 | 0.00001 | 0.00000 |
| 36 | 0.66997 | 0.00000 | 0.02804 | 0.28924 | 0.01209 | 0.00065 | 0.00001 | 0.00000 |
| 37 | 0.65391 | 0.00000 | 0.02735 | 0.30527 | 0.01280 | 0.00065 | 0.00001 | 0.00000 |
| 38 | 0.63760 | 0.00000 | 0.02666 | 0.32156 | 0.01352 | 0.00065 | 0.00001 | 0.00000 |
| 39 | 0.62105 | 0.00000 | 0.02595 | 0.33808 | 0.01425 | 0.00065 | 0.00001 | 0.00000 |
| 40 | 0.60428 | 0.00000 | 0.02523 | 0.35483 | 0.01500 | 0.00065 | 0.00002 | 0.00000 |
| 41 | 0.58730 | 0.00000 | 0.02450 | 0.37178 | 0.01575 | 0.00065 | 0.00002 | 0.00000 |
| 42 | 0.57012 | 0.00000 | 0.02377 | 0.38893 | 0.01651 | 0.00065 | 0.00002 | 0.00000 |
| 43 | 0.55277 | 0.00000 | 0.02303 | 0.40626 | 0.01728 | 0.00064 | 0.00002 | 0.00000 |
| 44 | 0.53527 | 0.00000 | 0.02228 | 0.42373 | 0.01805 | 0.00064 | 0.00002 | 0.00000 |
| 45 | 0.51763 | 0.00000 | 0.02153 | 0.44134 | 0.01883 | 0.00064 | 0.00002 | 0.00000 |
| 46 | 0.49989 | 0.00000 | 0.02077 | 0.45906 | 0.01962 | 0.00064 | 0.00002 | 0.00000 |
| 47 | 0.48205 | 0.00000 | 0.02001 | 0.47687 | 0.02041 | 0.00064 | 0.00003 | 0.00000 |

| Foreign-born females, proportion by infection state in 1981 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Age category (years) | *Uninfected* | *Immune* | *Recent Infection* | *Latent Infection* | *Reinfection* | *Primary Disease* | *Reactivation Disease* | *Reinfection Disease* |
| 48 | 0.46416 | 0.00000 | 0.01924 | 0.49474 | 0.02120 | 0.00064 | 0.00003 | 0.00000 |
| 49 | 0.44623 | 0.00000 | 0.01848 | 0.51264 | 0.02200 | 0.00063 | 0.00003 | 0.00000 |
| 50 | 0.42829 | 0.00000 | 0.01771 | 0.53055 | 0.02279 | 0.00063 | 0.00003 | 0.00000 |
| 51 | 0.41037 | 0.00000 | 0.01695 | 0.54843 | 0.02358 | 0.00063 | 0.00003 | 0.00000 |
| 52 | 0.39251 | 0.00000 | 0.01618 | 0.56626 | 0.02438 | 0.00063 | 0.00004 | 0.00000 |
| 53 | 0.37474 | 0.00000 | 0.01543 | 0.58401 | 0.02516 | 0.00062 | 0.00004 | 0.00000 |
| 54 | 0.35708 | 0.00000 | 0.01467 | 0.60164 | 0.02594 | 0.00062 | 0.00004 | 0.00000 |
| 55 | 0.33957 | 0.00000 | 0.01393 | 0.61913 | 0.02672 | 0.00062 | 0.00005 | 0.00000 |
| 56 | 0.32224 | 0.00000 | 0.01319 | 0.63642 | 0.02749 | 0.00061 | 0.00005 | 0.00000 |
| 57 | 0.30513 | 0.00000 | 0.01246 | 0.65351 | 0.02824 | 0.00061 | 0.00005 | 0.00000 |
| 58 | 0.28827 | 0.00000 | 0.01174 | 0.67034 | 0.02899 | 0.00061 | 0.00006 | 0.00000 |
| 59 | 0.27170 | 0.00000 | 0.01103 | 0.68688 | 0.02973 | 0.00060 | 0.00006 | 0.00000 |
| 60 | 0.25544 | 0.00000 | 0.01034 | 0.70311 | 0.03045 | 0.00060 | 0.00007 | 0.00000 |
| 61 | 0.23954 | 0.00000 | 0.00967 | 0.71898 | 0.03115 | 0.00059 | 0.00007 | 0.00000 |
| 62 | 0.22402 | 0.00000 | 0.00901 | 0.73447 | 0.03184 | 0.00058 | 0.00008 | 0.00000 |
| 63 | 0.20893 | 0.00000 | 0.00837 | 0.74954 | 0.03250 | 0.00058 | 0.00008 | 0.00000 |
| 64 | 0.19427 | 0.00000 | 0.00775 | 0.76416 | 0.03315 | 0.00057 | 0.00009 | 0.00000 |
| 65 | 0.18009 | 0.00000 | 0.00715 | 0.77831 | 0.03378 | 0.00056 | 0.00010 | 0.00000 |
| 66 | 0.16642 | 0.00000 | 0.00657 | 0.79196 | 0.03439 | 0.00055 | 0.00011 | 0.00000 |
| 67 | 0.15327 | 0.00000 | 0.00602 | 0.80508 | 0.03497 | 0.00055 | 0.00012 | 0.00000 |
| 68 | 0.14067 | 0.00000 | 0.00549 | 0.81765 | 0.03553 | 0.00054 | 0.00013 | 0.00000 |
| 69 | 0.12863 | 0.00000 | 0.00498 | 0.82966 | 0.03606 | 0.00052 | 0.00014 | 0.00000 |
| 70 | 0.11718 | 0.00000 | 0.00451 | 0.84108 | 0.03657 | 0.00051 | 0.00015 | 0.00000 |
| 71 | 0.10633 | 0.00000 | 0.00406 | 0.85191 | 0.03705 | 0.00050 | 0.00016 | 0.00000 |
| 72 | 0.09608 | 0.00000 | 0.00363 | 0.86213 | 0.03750 | 0.00048 | 0.00018 | 0.00000 |
| 73 | 0.08644 | 0.00000 | 0.00323 | 0.87173 | 0.03793 | 0.00047 | 0.00019 | 0.00000 |

| Foreign-born females, proportion by infection state in 1981 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Age category (years) | *Uninfected* | *Immune* | *Recent Infection* | *Latent Infection* | *Reinfection* | *Primary Disease* | *Reactivation Disease* | *Reinfection Disease* |
| 74 | 0.07742 | 0.00000 | 0.00286 | 0.88073 | 0.03833 | 0.00045 | 0.00021 | 0.00000 |
| 75 | 0.06901 | 0.00000 | 0.00252 | 0.88910 | 0.03870 | 0.00043 | 0.00023 | 0.00000 |
| 76 | 0.06121 | 0.00000 | 0.00221 | 0.89687 | 0.03904 | 0.00041 | 0.00025 | 0.00000 |
| 77 | 0.05402 | 0.00000 | 0.00192 | 0.90404 | 0.03936 | 0.00039 | 0.00027 | 0.00000 |
| 78 | 0.04741 | 0.00000 | 0.00166 | 0.91061 | 0.03965 | 0.00037 | 0.00029 | 0.00001 |
| 79 | 0.04137 | 0.00000 | 0.00143 | 0.91662 | 0.03992 | 0.00035 | 0.00031 | 0.00001 |
| 80 | 0.03594 | 0.00000 | 0.00122 | 0.92201 | 0.04016 | 0.00032 | 0.00034 | 0.00001 |
| 81 | 0.03113 | 0.00000 | 0.00104 | 0.92680 | 0.04037 | 0.00030 | 0.00036 | 0.00001 |
| 82 | 0.02688 | 0.00000 | 0.00088 | 0.93102 | 0.04056 | 0.00027 | 0.00038 | 0.00001 |
| 83 | 0.02313 | 0.00000 | 0.00074 | 0.93474 | 0.04073 | 0.00025 | 0.00041 | 0.00001 |
| 84 | 0.01984 | 0.00000 | 0.00063 | 0.93800 | 0.04087 | 0.00022 | 0.00043 | 0.00001 |
| 85 | 0.01696 | 0.00000 | 0.00052 | 0.94085 | 0.04100 | 0.00020 | 0.00045 | 0.00001 |
| 86 | 0.01445 | 0.00000 | 0.00044 | 0.94334 | 0.04111 | 0.00018 | 0.00048 | 0.00001 |
| 87 | 0.01226 | 0.00000 | 0.00037 | 0.94550 | 0.04121 | 0.00016 | 0.00050 | 0.00001 |
| 88 | 0.01037 | 0.00000 | 0.00031 | 0.94737 | 0.04129 | 0.00014 | 0.00052 | 0.00001 |
| 89 | 0.00874 | 0.00000 | 0.00025 | 0.94898 | 0.04136 | 0.00012 | 0.00053 | 0.00001 |
| 90 | 0.00734 | 0.00000 | 0.00021 | 0.95037 | 0.04142 | 0.00010 | 0.00055 | 0.00001 |
| 91 | 0.00613 | 0.00000 | 0.00017 | 0.95155 | 0.04148 | 0.00009 | 0.00056 | 0.00001 |
| 92 | 0.00511 | 0.00000 | 0.00014 | 0.95256 | 0.04152 | 0.00008 | 0.00058 | 0.00001 |
| 93 | 0.00424 | 0.00000 | 0.00012 | 0.95342 | 0.04156 | 0.00006 | 0.00059 | 0.00001 |
| 94 | 0.00350 | 0.00000 | 0.00010 | 0.95415 | 0.04159 | 0.00005 | 0.00060 | 0.00001 |
| 95 | 0.00288 | 0.00000 | 0.00008 | 0.95476 | 0.04162 | 0.00005 | 0.00061 | 0.00001 |
| 96 | 0.00236 | 0.00000 | 0.00006 | 0.95527 | 0.04164 | 0.00004 | 0.00062 | 0.00001 |
| 97 | 0.00192 | 0.00000 | 0.00005 | 0.95570 | 0.04166 | 0.00003 | 0.00062 | 0.00001 |
| 98 | 0.00156 | 0.00000 | 0.00004 | 0.95606 | 0.04168 | 0.00003 | 0.00063 | 0.00001 |
| 99 | 0.00126 | 0.00000 | 0.00003 | 0.95635 | 0.04169 | 0.00002 | 0.00063 | 0.00001 |

| Foreign-born females, proportion by infection state in 1981 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Age category (years) | *Uninfected* | *Immune* | *Recent Infection* | *Latent Infection* | *Reinfection* | *Primary Disease* | *Reactivation Disease* | *Reinfection Disease* |
| 100 | 0.00101 | 0.00000 | 0.00003 | 0.95660 | 0.04170 | 0.00002 | 0.00064 | 0.00001 |
| 101 | 0.00081 | 0.00000 | 0.00002 | 0.95679 | 0.04171 | 0.00001 | 0.00064 | 0.00001 |
| 102 | 0.00065 | 0.00000 | 0.00002 | 0.95695 | 0.04172 | 0.00001 | 0.00064 | 0.00001 |
| 103 | 0.00052 | 0.00000 | 0.00001 | 0.95708 | 0.04172 | 0.00001 | 0.00064 | 0.00001 |
| 104 | 0.00042 | 0.00000 | 0.00001 | 0.95718 | 0.04173 | 0.00001 | 0.00064 | 0.00001 |
| 105 | 0.00033 | 0.00000 | 0.00001 | 0.95726 | 0.04173 | 0.00001 | 0.00065 | 0.00001 |
| 106 | 0.00027 | 0.00000 | 0.00001 | 0.95733 | 0.04174 | 0.00000 | 0.00065 | 0.00001 |
| 107 | 0.00021 | 0.00000 | 0.00001 | 0.95738 | 0.04174 | 0.00000 | 0.00065 | 0.00001 |
| 108 | 0.00017 | 0.00000 | 0.00000 | 0.95742 | 0.04174 | 0.00000 | 0.00065 | 0.00001 |
| 109 | 0.00014 | 0.00000 | 0.00000 | 0.95745 | 0.04174 | 0.00000 | 0.00065 | 0.00001 |
| 110 | 0.00011 | 0.00000 | 0.00000 | 0.95748 | 0.04174 | 0.00000 | 0.00065 | 0.00001 |
| 111 | 0.00009 | 0.00000 | 0.00000 | 0.95750 | 0.04174 | 0.00000 | 0.00065 | 0.00001 |
| 112 | 0.00007 | 0.00000 | 0.00000 | 0.95752 | 0.04174 | 0.00000 | 0.00065 | 0.00001 |
| 113 | 0.00006 | 0.00000 | 0.00000 | 0.95753 | 0.04174 | 0.00000 | 0.00065 | 0.00001 |
| 114 | 0.00005 | 0.00000 | 0.00000 | 0.95754 | 0.04175 | 0.00000 | 0.00065 | 0.00001 |
| 115 | 0.00004 | 0.00000 | 0.00000 | 0.95755 | 0.04175 | 0.00000 | 0.00065 | 0.00001 |
| 116 | 0.00003 | 0.00000 | 0.00000 | 0.95756 | 0.04175 | 0.00000 | 0.00065 | 0.00001 |
| 117 | 0.00002 | 0.00000 | 0.00000 | 0.95757 | 0.04175 | 0.00000 | 0.00065 | 0.00001 |
| 118 | 0.00002 | 0.00000 | 0.00000 | 0.95757 | 0.04175 | 0.00000 | 0.00065 | 0.00001 |
| 119 | 0.00002 | 0.00000 | 0.00000 | 0.95757 | 0.04175 | 0.00000 | 0.00065 | 0.00001 |
| 120 | 0.00001 | 0.00000 | 0.00000 | 0.95758 | 0.04175 | 0.00000 | 0.00065 | 0.00001 |

# 10.8 Data for the Infection Status of Migrants

Table 10-15: Proportion of immigrants, by age class, with Tine test reaction grades 0 – 4 from Ormerod et al. [295]. Also shown is 'n', the total number tested in each age class. These data were used to produce infection state probabilities for migrants to England and Wales following the screening method, schemes Scr1 and Scr2, described in Section 4.2.8.1.

| Tine test grade | 0 – 4 yrs | 5 – 14 yrs | 15 – 29 yrs | 30 – 44 yrs | 45 – 64 yrs | 65+ yrs |
|---|---|---|---|---|---|---|
| 0 | 0.58 | 0.37 | 0.25 | 0.16 | 0.10 | 0.09 |
| 1 | 0.34 | 0.44 | 0.44 | 0.46 | 0.52 | 0.32 |
| 2 | 0.06 | 0.13 | 0.16 | 0.19 | 0.15 | 0.22 |
| 3 | 0.00 | 0.04 | 0.07 | 0.08 | 0.06 | 0.19 |
| 4 | 0.02 | 0.03 | 0.08 | 0.10 | 0.17 | 0.18 |
| (n) | 110 | 211 | 1057 | 178 | 89 | 45 |

## 10.9 Fits to Notification Rates per 100,000 Population for Stage One

Although the model was not fit to the number of notifications per 100,000 population per year, results from fitting scenarios 14 and 22 are used to illustrate how notification rates predicted by the model compared to observed notification rates. Based on visual inspection, fits to notification rates were generally worse than fits to the number of case notifications.

UK-born individuals had the lowest notification rates per 100,000 population per year among the three birthplace groups. Notification rates are below 15 per 100,000 cases per year for all sex and age groups from 1999-2009, as shown in Figure 10-1 – Figure 10-4 for Scenarios 14 and 22. For UK-born males under both scenarios 14 and 22, observed notification rates are predicted fairly well by the model, apart from overestimation of the notification rate in those aged 65 years and above. Under scenario 14, notification rates are overestimated by the model for most of the time period, from about 2002-2009. Under scenario 22, trends are very similar although the model consistently overestimates the notification rate in those aged 65 years and above. The model predicts notification rates from about 13-15 per 100,000 per year, while observed notification rate fall from less than 13 per 100,000 to about 8 per 100,000 from 1999-2009 in this age group. For UK-born males aged 15-44 years, the model underestimated the rate in slightly for both scenarios. Overall, qualitative trends fit observed trends reasonably well and the absolute differences between notification rates predicted by the model and those observed are small, on the order of less than a few cases per 100,000 per year.

For UK-born females, model predictions are less consistent with observed notification rates, although still within a few cases per 100,000 population per year. Model predictions are particularly problematic for older individuals in both scenarios 14 and 22. For those aged 65 years and above, rates are underestimated for most of the 1999-2009 time period in both scenarios 14 and 22, and, notably, the decreasing trend in observed notification rates in this group is not reproduced by the model for either scenario 14 or 22. For the 45-64 age group, the model actually predicts an increased notification rate from 1999 – 2009 under both scenarios, while the observed notification rate *decreases* over this time period. Notification rates in this age group

are over-estimated by the model from 1999-2009. In both age groups and under both scenarios, notification rates are within a few cases per 100,000 per year of observed notification rates.

OF-born have notification rates much higher than those in UK-born, generally ranging from about 25-125 per 100,000 per year across the sex and age categories as shown in Figure 10-2 to Figure 10-5. Notification rates are highest in those aged 15-44 years, and these rates and trends are reproduced fairly well by the model for both males and females, though fits differ between the two scenarios 14 and 22. In scenario 14, notification rates in this group are underestimated by the model for males and overestimated for females. In scenario 22, notification rates in the model reproduce observed data until 2006-2009, when observed notification rates drop and are overestimated by the model. As in scenarios 14, for females under scenario 22, the notification rates are consistently overestimated by the model from 1999-2009, with difference between model and observed rates ranging from about 20 – 50 cases per 100,000 per year.

In both males and females under scenario 14, notification rates in those aged 45-64 years are overestimated by the model. Under scenario 22, rates are slightly overestimated for males and very close to observed for females. Although the model was not fit to notifications in children, the model consistently and greatly overestimates the notification rates in those aged 0-14 years, for both males and females and under scenarios 14 and 22. For those aged 65 years and above, the model reproduces observed trends and values of the notification rates fairly well, except under scenario 22, where notification rates in females aged 65 and older are underestimated by the model, often around 25 cases per 100,000 per year below observed notification rates.

SSA-born notification rates are highest of the three birthplace groups, approaching 300 cases per 100,000 population per year in the 15-44 age group, the only age group used for fitting the model to observed notifications. For both scenarios 14 and 22, observed notification rates in this age group are greatly underestimated by the model, with the largest discrepancies predicting about 100 fewer cases per 100,000 per year.

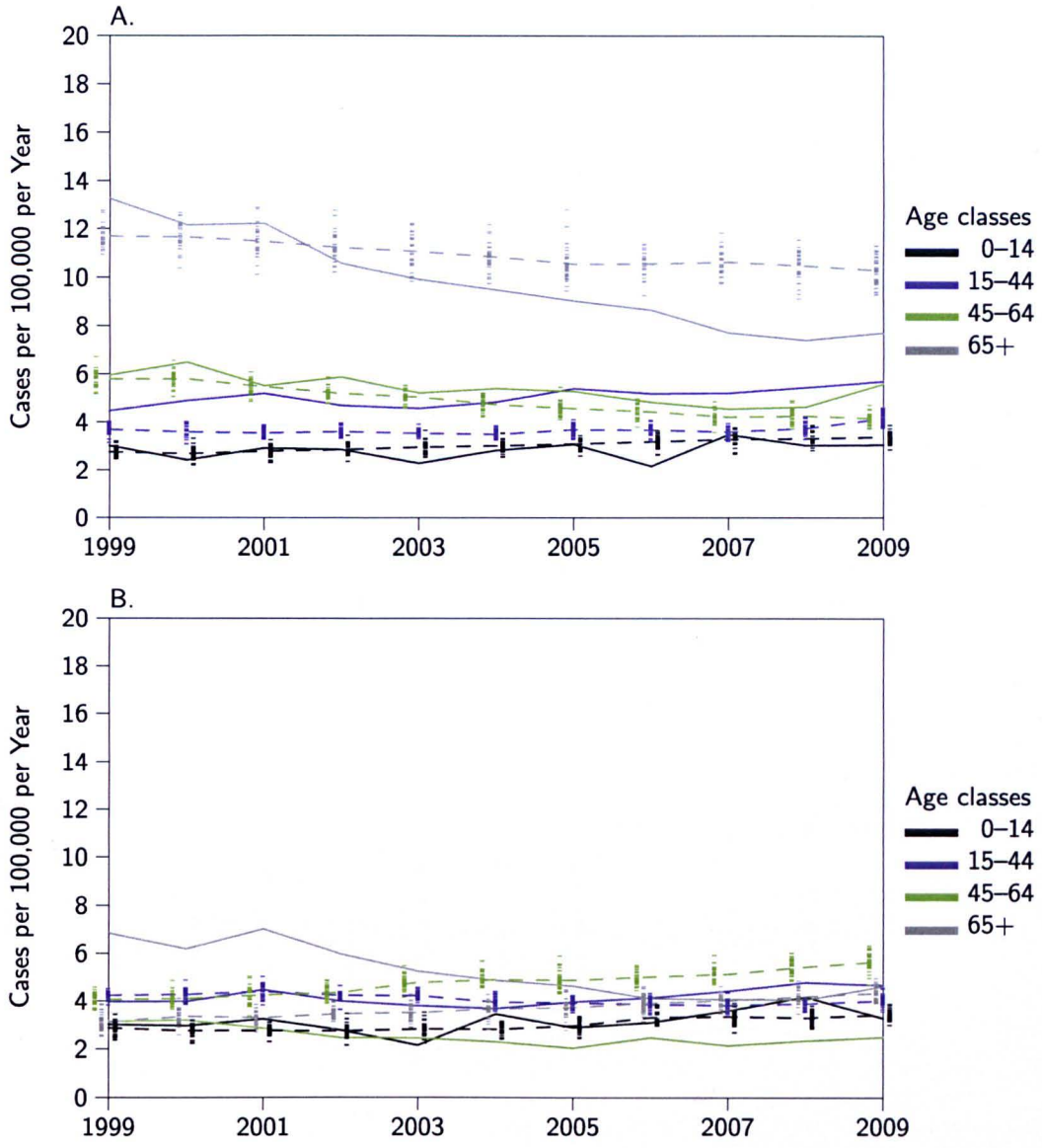


Figure 10-1: Simulated and observed cases per 100,000 per year in England and Wales for UK-born males (A) and females (B) by age category from 1999 – 2009, under fitting Scenario 14. Averaged model output follows the dashed line and individual runs of the model are denoted with a '-' (there are 30 for each data point). Observed numbers of case notifications follow the solid line. Age categories are as follows: ages 0 – 14 years are in black; ages 15 – 44 years are in blue; ages 45 – 64 are in green; and ages 65 years and over are in grey.

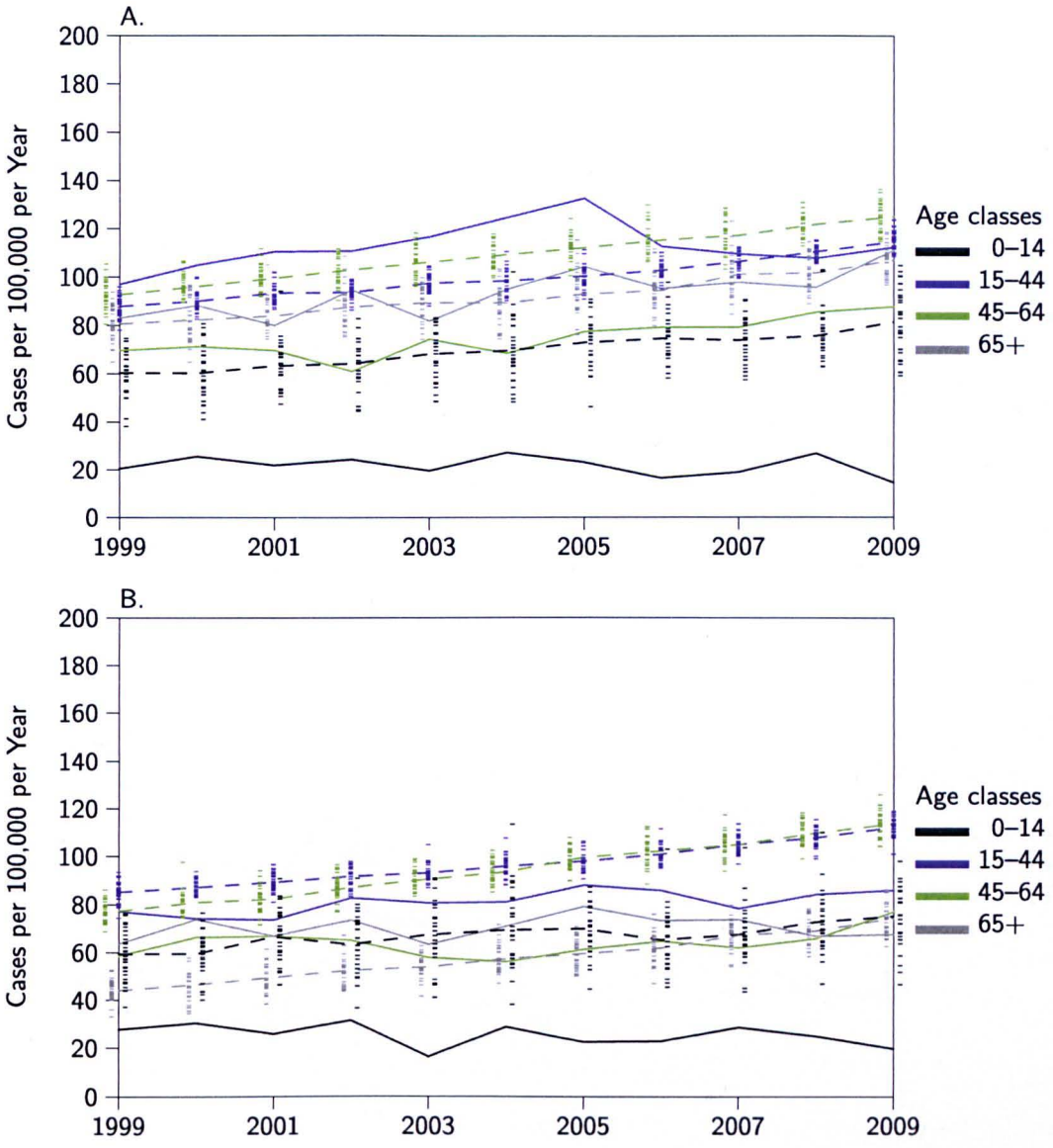


Figure 10-2: Simulated and observed cases per 100,000 per year in England and Wales for other foreign-born males (A) and females (B) by age category from 1999 – 2009, under fitting Scenario 14. Averaged model output follows the dashed line and individual runs of the model are denoted with a '-' (there are 30 for each data point). Observed numbers of case notifications follow the solid line. Age categories are as follows: ages 0 – 14 years are in black; ages 15 – 44 years are in blue; ages 45 – 64 are in green; and ages 65 years and over are in grey.

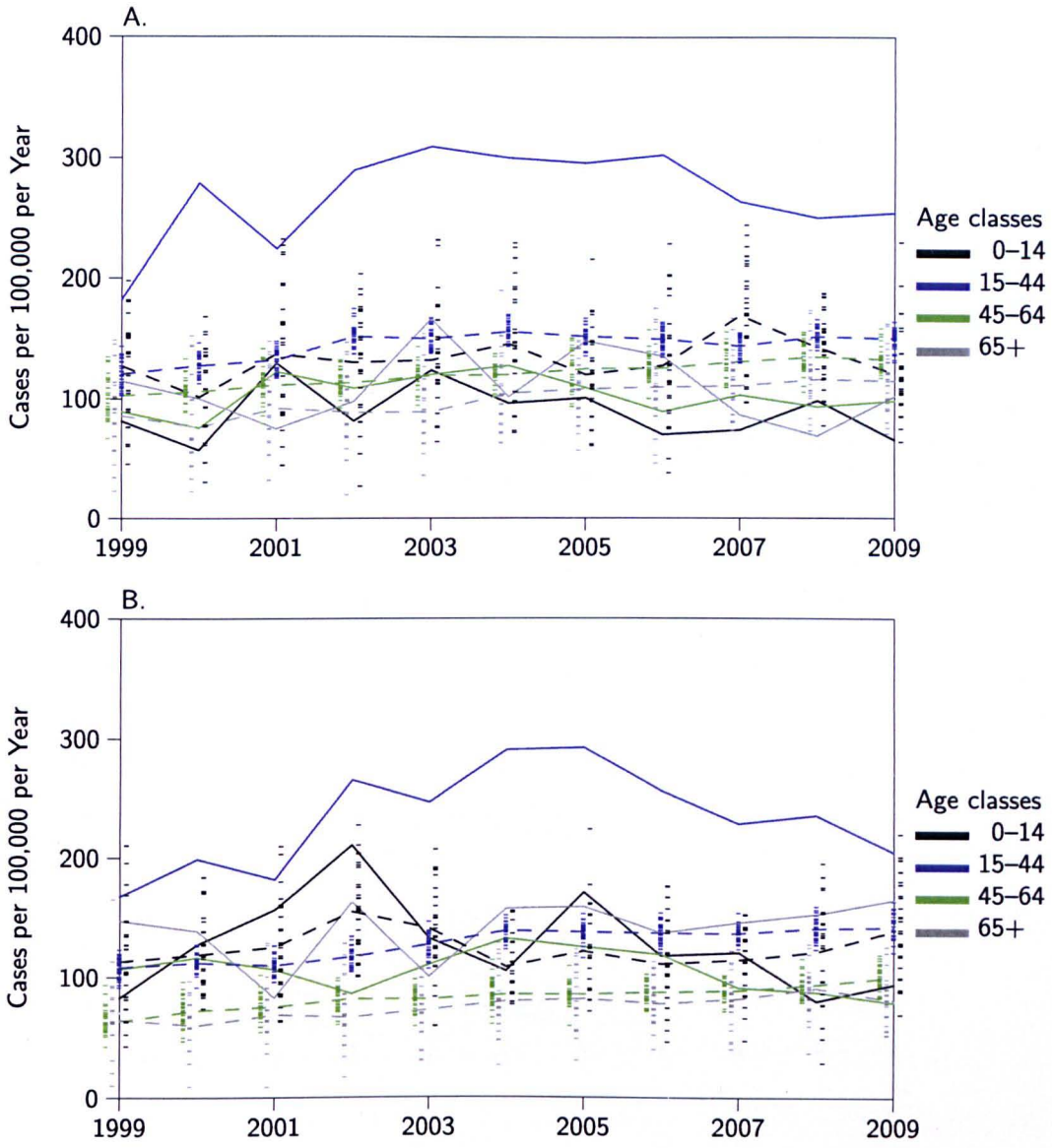


Figure 10-3: Simulated and observed cases per 100,000 per year in England and Wales for SSA-born males (A) and females (B) by age category from 1999 – 2009, under fitting Scenario 14. Averaged model output follows the dashed line and individual runs of the model are denoted with a '-' (there are 30 for each data point). Observed numbers of case notifications follow the solid line. Age categories are as follows: ages 0 – 14 years are in black; ages 15 – 44 years are in blue; ages 45 – 64 are in green; and ages 65 years and over are in grey.

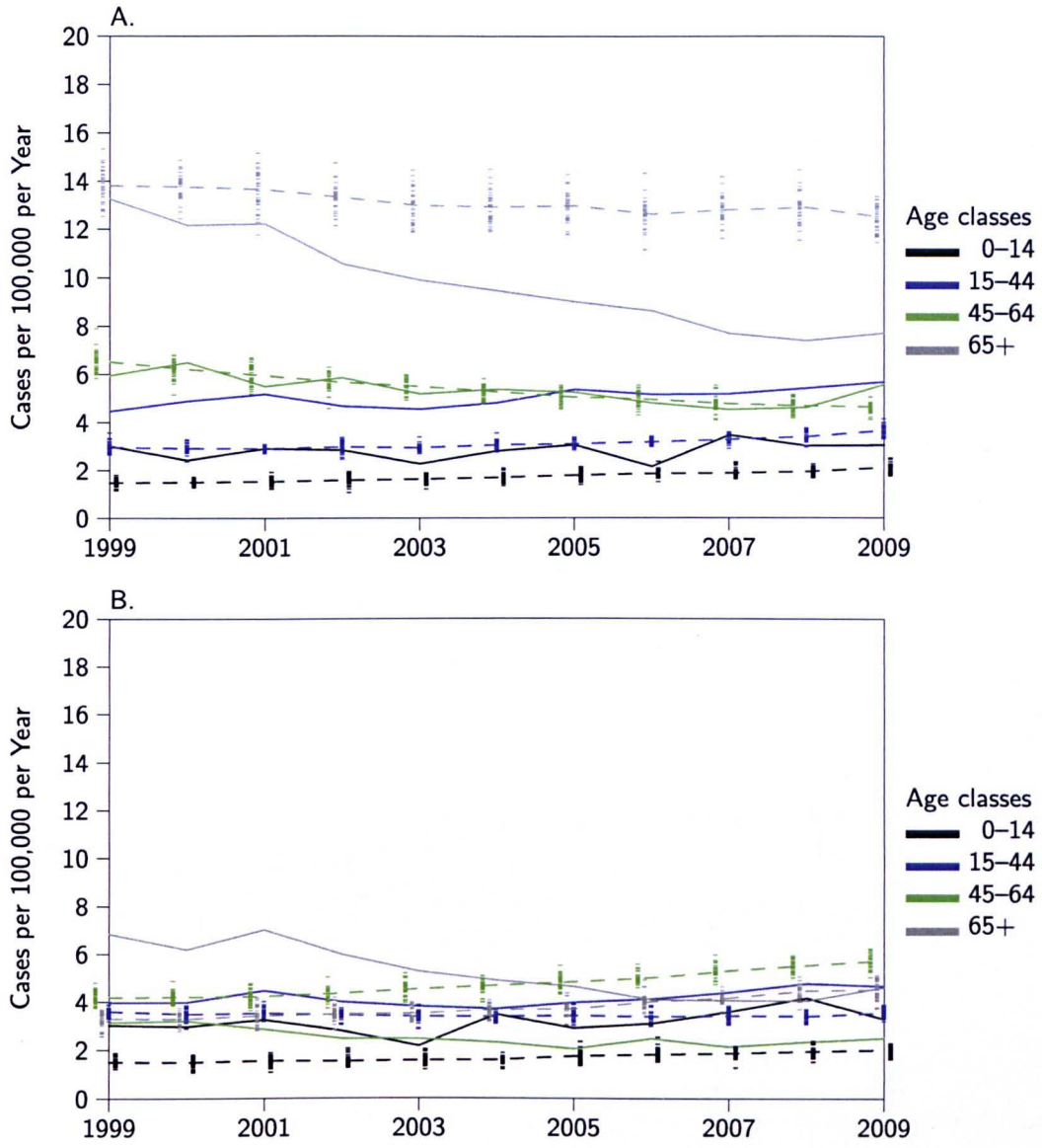


Figure 10-4: Simulated and observed cases per 100,000 per year in England and for UK-born males (A) and females (B) by age category for 1999 – 2009, under fitting Scenario 22. The averaged model output follows the dashed line and individual runs of the model are denoted with a '-' (30 for each data point). Observed numbers of case notifications follow the solid line. Age categories are as follows: ages 0 – 14 years are in black; ages 15 – 44 years are in blue; ages 45 – 64 are in green; and ages 65 years and over are in grey.

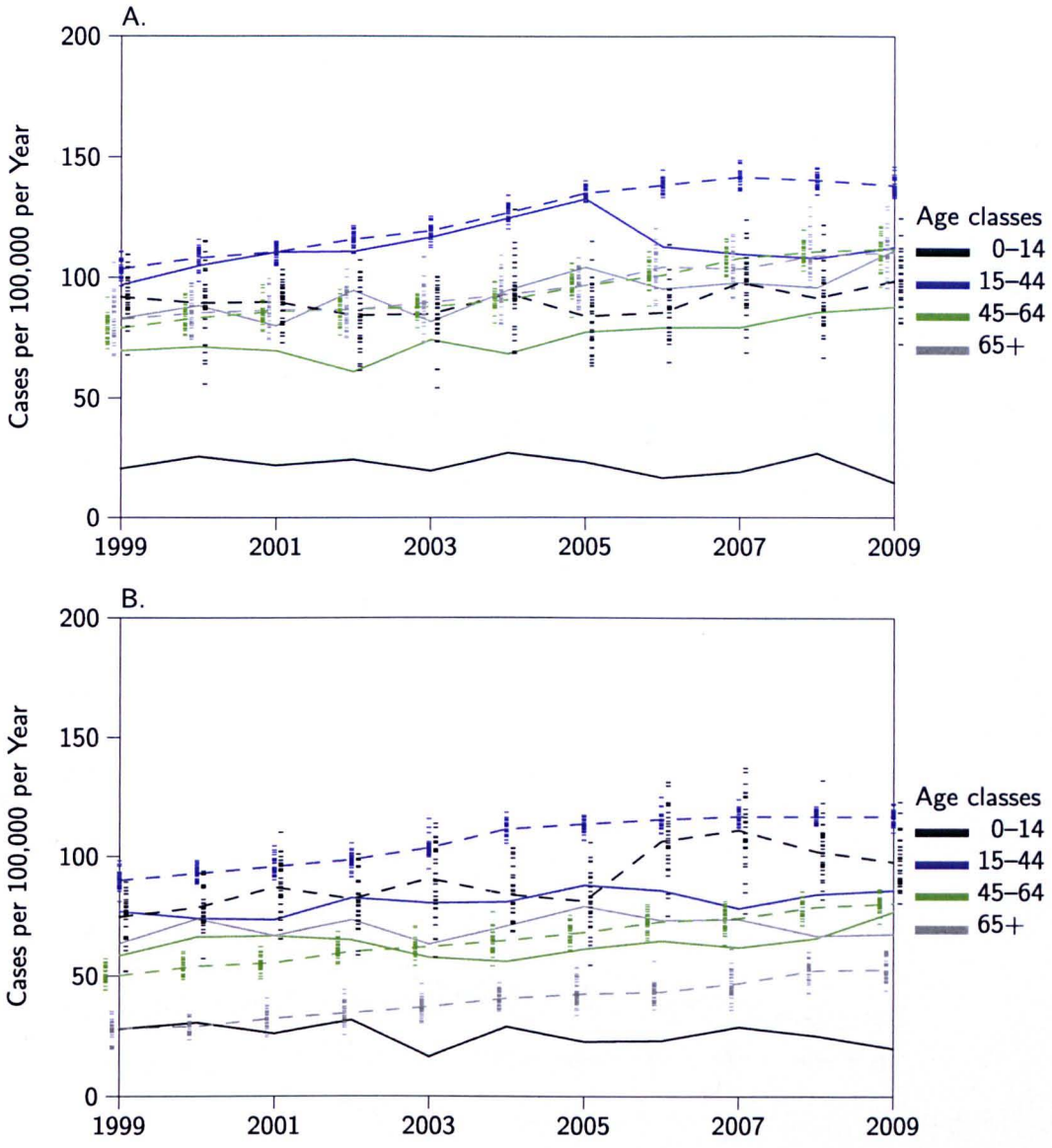


Figure 10-5: Simulated and observed cases per 100,000 per year in England and Wales for other foreign-born males (A) and females (B) by age category for 1999 – 2009, under fitting Scenario 22. The averaged model output follows the dashed line and individual runs of the model are denoted with a '-' (30 for each data point). Observed numbers of case notifications follow the solid line. Age categories are as follows: ages 0 – 14 years are in black; ages 15 – 44 years are in blue; ages 45 – 64 are in green; and ages 65 years and over are in grey.

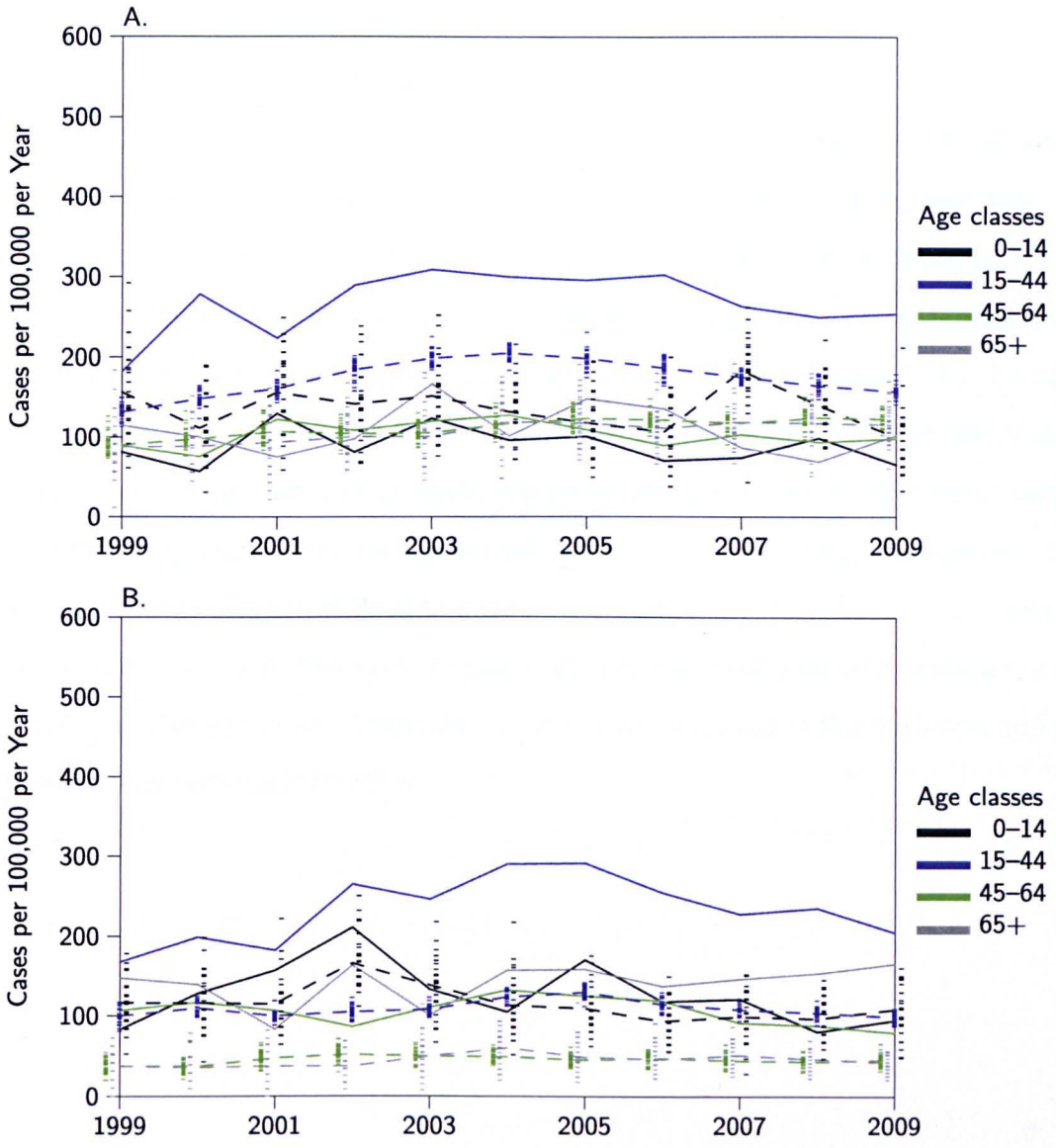


Figure 10-6: Simulated and observed cases per 100,000 per year in England and Wales for SSA-born males (A) and females (B) by age category from 1999 – 2009, under fitting Scenario 22. The averaged model output follows the dashed line and individual runs of the model are denoted with a '-' (30 for each data point). Observed numbers of case notifications follow the solid line. Age categories are as follows: ages 0 – 14 years are in black; ages 15 – 44 years are in blue; ages 45 – 64 are in green; and ages 65 years and over are in grey.

## 10.10 Stage Two Fitting Results Using Increased HIV Prevalence for SSA-Born Migrants

For this variation in stage two fitting, each of the 10 scenarios was run with an altered distribution for HIV prevalence in SSA-born immigrants to the UK each year. HIV prevalence was increased by 50% above original values, which were based on the estimated HIV prevalence of all SSA-born individuals living in the UK, as opposed to the HIV prevalence in new immigrants upon entry to the UK, as discussed in 4.2.4 and 5.1.5.1. As the results found in Table 10-16 show, this increased HIV prevalence did not improve upon model fits and or much impact resulting best-fitting parameter values. The best-fitting scenario had a higher GOF statistic than the lowest in stage one. This meant that this variation did not fit data as well as the best-fitting scenario of the 25 scenarios in stage one. This finding suggested there was no reason to replace the HIV prevalence distribution used originally with the version used in this variation and this variation was not studied further.

*Table 10-16: Results from fitting model to observed data in stage two fits with increased HIV prevalence in Sub-Sahara Africa-born immigrants. Risks were estimated for UK-born adult males for the three disease types and for the risk ratio between disease risk in foreign-born and UK-born adults.*

| | | | Disease risks by type of disease | | | | | |
|---|---|---|---|---|---|---|---|---|
| Scenario | Infection status of migrants | Contact rate | Primary (%) | Reactivation (% per year) | Reinfection (%) | Foreign:UK-born | GOF | GOF rank |
| 3 | Scr1(OR7) | all=8 | 10.0% | 0.012% | 8.0% | 2.13 | 11055 | 10 |
| 4 | Scr1(OR7) | all=10 | 7.2% | 0.016% | 3.0% | 2.63 | 10802 | 9 |
| 8 | Scr2(OR9) | all=8 | 8.3% | 0.016% | 4.1% | 2.48 | 10156 | 6 |
| 9 | Scr2(OR9) | all=10 | 9.3% | 0.010% | 7.2% | 2.21 | 10144 | 5 |
| 13 | ARI low | all=8 | 9.1% | 0.023% | 0.0% | 4.14 | 9276 | 1 |
| 14 | ARI low | all=10 | 10.7% | 0.015% | 0.5% | 3.01 | 9569 | 2 |
| 16 | ARI med | all=4 | 27.6% | 0.018% | 17.0% | 1.84 | 9755 | 3 |
| 17 | ARI med | all=6 | 18.8% | 0.018% | 9.8% | 1.94 | 10767 | 8 |
| 22 | ARI high | all=6 | 16.4% | 0.018% | 12.7% | 2.09 | 9831 | 4 |
| 23 | ARI high | all=8 | 12.7% | 0.013% | 16.0% | 1.99 | 10550 | 7 |

*Table 10-17: Disease risks for UK-born and foreign-born males and females under the stage two fitting scheme in which an increased HIV prevalence in Sub-Sahara Africa-born immigrants was used. Disease risks for UK-born were estimated by the model, as was a risk ratio between foreign-born and UK-born disease risk. Foreign-born disease risks were calculated by multiplying that risk ratio by UK-born risks. Female disease risks were calculated by multiplying those risks by risk ratios between males and females, see Table xx /text).*

| | Disease risks for UK-born males | | | Disease risks for foreign-born males | | | Disease risks for UK-born females | | | Disease risks for foreign-born females | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Scen-ario | Primary (%) | React. (% per year) | Reinf. (%) | Primary (%) | React. (% per year) | Reinf. (%) | Primary (%) | React. (% per year) | Reinf. (%) | Primary (%) | React. (% per year) | Reinf. (%) |
| 3 | 10.0 | 0.012 | 8.0 | 21.4 | 0.026 | 17.0 | 10.0 | 0.002 | 0.01 | 21.4 | 0.004 | 0.02 |
| 4 | 7.2 | 0.016 | 3.0 | 18.9 | 0.042 | 7.8 | 7.2 | 0.003 | 0.00 | 18.9 | 0.007 | 0.01 |
| 8 | 8.3 | 0.016 | 4.1 | 20.6 | 0.041 | 10.2 | 8.3 | 0.003 | 0.00 | 20.6 | 0.007 | 0.01 |
| 9 | 9.3 | 0.010 | 7.2 | 20.6 | 0.023 | 15.9 | 9.3 | 0.002 | 0.01 | 20.6 | 0.004 | 0.02 |
| 13 | 9.1 | 0.023 | 0.0 | 37.7 | 0.094 | 0.1 | 9.1 | 0.004 | 0.00 | 37.7 | 0.015 | 0.00 |
| 14 | 10.7 | 0.015 | 0.5 | 32.1 | 0.047 | 1.5 | 10.7 | 0.002 | 0.00 | 32.1 | 0.007 | 0.00 |
| 16 | 27.6 | 0.018 | 17.0 | 50.6 | 0.033 | 31.2 | 27.6 | 0.003 | 0.02 | 50.6 | 0.005 | 0.04 |
| 17 | 18.8 | 0.018 | 9.8 | 36.4 | 0.035 | 18.9 | 18.8 | 0.003 | 0.01 | 36.4 | 0.006 | 0.02 |
| 22 | 16.4 | 0.018 | 12.7 | 34.3 | 0.037 | 26.6 | 16.4 | 0.003 | 0.02 | 34.3 | 0.006 | 0.03 |
| 23 | 12.7 | 0.013 | 16.0 | 25.2 | 0.026 | 31.8 | 12.7 | 0.002 | 0.02 | 25.2 | 0.004 | 0.04 |

## 10.11 Stage Two Fitting Results For Altered infection Status For SSA-born Individuals In 1981

For this variation, each of the 10 scenarios was run with an altered distribution of infection state probabilities for the initial population in 1981. Specifically, this was altered to increase the infection and disease prevalence in SSA-born individuals, as described in Chapter 5, Section 5.1.5.2. As with the first variation tested in stage two fitting, visual inspection of fits to observed data showed this variation did not improve the quality of fits. Furthermore, the quality of fits as measured by the GOF statistic was not improved, as shown in the results in Table 10-18. The best-fitting scenario of the 10 did fit better than the best-fitting of the 25 stage one scenarios, however on average the fitting statistic values were similar, and qualitatively, best-fits were similar to those obtained in stage one fits. This variation was not studied further.

*Table 10-18: Results of stage two fits with increased prevalence of infection and disease (tuberculosis) at model initialization for Sub-Sahara Africa-born immigrants.*

| Scenario | Infection status of migrants | Contact rate | Disease risks for UK-born adult males by type: Primary (%) | Reactivation (% per year) | Reinfection (%) | Risk ratio, Foreign:UK-born | GOF | GOF rank |
|---|---|---|---|---|---|---|---|---|
| 3 | Scr1 | all=8 | 8.4 | 0.013 | 2.3 | 2.82 | 10730 | 10 |
| 4 | Scr1 | all=10 | 6.0 | 0.016 | 0.7 | 3.26 | 10555 | 9 |
| 8 | Scr2 | all=8 | 7.5 | 0.011 | 4.2 | 2.77 | 9944 | 6 |
| 9 | Scr2 | all=10 | 5.5 | 0.014 | 1.2 | 3.30 | 9796 | 4 |
| 13 | ARI low | all=8 | 10.3 | 0.015 | 1.5 | 3.72 | 8849 | 2 |
| 14 | ARI low | all=10 | 7.5 | 0.019 | 0.0 | 4.14 | 8537 | 1 |
| 16 | ARI med | all=4 | 20.7 | 0.020 | 5.3 | 2.55 | 9801 | 5 |
| 17 | ARI med | all=6 | 15.3 | 0.015 | 10.8 | 2.32 | 10251 | 7 |
| 22 | ARI high | all=6 | 15.2 | 0.013 | 14.0 | 2.22 | 9702 | 3 |
| 23 | ARI high | all=8 | 10.9 | 0.018 | 5.7 | 2.56 | 10336 | 8 |

## 10.12 Stage two fitting Results for Six Estimated Disease Risk Parameters

For this variation of stage two fitting, six disease risks were estimated, three for UK-born as in stage one fits, plus three analogous risks for foreign-born. The three risks estimated for foreign-born replaced the parameter, *df*, the parameter the ratio of disease risk between UK-born and foreign-born for all three disease types. Results for the 10 scenarios run with these six parameters show no improvement in fits according to the GOF statistics obtained. The average GOF was higher than in this stage two variation of fitting than across the 25 fitting scenarios in stage one, despite fewer degrees of freedom for the GOF statistic since there were more variable parameters. For this reason, results for this variation are described briefly but not illustrated with all plots. Table 10-19 shows GOF statistics and disease risk estimates for these fits.

Table 10-19: Results of model fitting to observed case notifications for the stage two fitting scheme with six disease risk parameters estimated. Only the best-fitting replicate for each scenario is shown.

| | | | Disease risks for UK-born adult males by disease type | | | Disease risks for foreign-born adult males by disease type | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | Infection status of migrants | Contact rate | Primary (%) | React. (% per year) | Reinf. (%) | Primary (%) | React. (% per year) | Reinf. (%) | GOF | GOF rank |
| 3 | Scr1 | 8 | 9.6 | 0.014 | 2.2 | 22.5 | 0.062 | 5.5 | 11007 | 6 |
| 4 | Scr1 | 10 | 7.9 | 0.014 | 3.6 | 19.0 | 0.042 | 6.3 | 11062 | 7 |
| 8 | Scr2 | 8 | 9.7 | 0.008 | 9.6 | 19.6 | 0.040 | 16.6 | 10237 | 4 |
| 9 | Scr2 | 10 | 7.7 | 0.011 | 4.5 | 18.1 | 0.035 | 6.9 | 10318 | 5 |
| 13 | ARI low | 8 | 10.9 | 0.020 | 2.1 | 38.5 | 0.082 | 0.3 | 9445 | 2 |
| 14 | ARI low | 10 | 8.7 | 0.020 | 1.9 | 32.1 | 0.069 | 0.9 | 9353 | 1 |
| 16 | ARI med | 4 | 30.1 | 0.015 | 5.7 | 51.1 | 0.085 | 1.8 | 10071 | 3 |
| 17 | ARI med | 6 | 14.3 | 0.016 | 8.6 | 29.7 | 0.050 | 4.6 | 11910 | 10 |
| 22 | ARI high | 6 | 13.7 | 0.020 | 0.0 | 28.8 | 0.075 | 0.1 | 11574 | 8 |
| 23 | ARI high | 8 | 14.0 | 0.021 | 0.0 | 28.7 | 0.070 | 0.1 | 11650 | 9 |

## 10.13 Quality of Fits to Notification Rates for Stage Two Fitting with Single Foreign-born Category

As shown in Figure 10-7 and Figure 10-8, comparing case notification rates generated by the model to those observed showed a different picture than when comparing the model and data using numbers of tuberculosis case notifications.

For UK-born males, the notification rates per 100,000 population predicted by the model are fairly close to observed notification rates, as shown in Figure 10-7 (A). The model overestimated the notification rate for those aged 65 years and above, the group with the highest notification rates for UK-born. Also, the model underestimated the notification rate for those aged 15-44, but is otherwise fairly good.

For UK-born females, the model predicted notification rates fit observed case notification rates less well, as shown in Figure 10-7 (B). The age categories 45-64 years and 65 years and above are especially problematic, as seen when comparing the number of case notifications predicted and observed. Again, for both of these groups, the model predicted an increasing trend, while the observed data showed decreasing trends over 1999-2009.

For foreign-born, simulated notification rates followed trends in observed notification rates fairly well. For males aged 15-44 years, this still meant that simulated notification rates were high or low by as much as 20 cases per 100,000 per year, depending on the year, but averaged rates were similar to those observed. For females, simulated rates were closer to observed rates. The fits to notification rates in children were the worst, as the model consistently over-estimated the number of cases in this group. For some runs, the rate was overestimated by as much as 50 cases per 100,000 for some runs of the simulation. Notification rates for older individuals were under-estimated by the model for several of the earlier years of fitting.

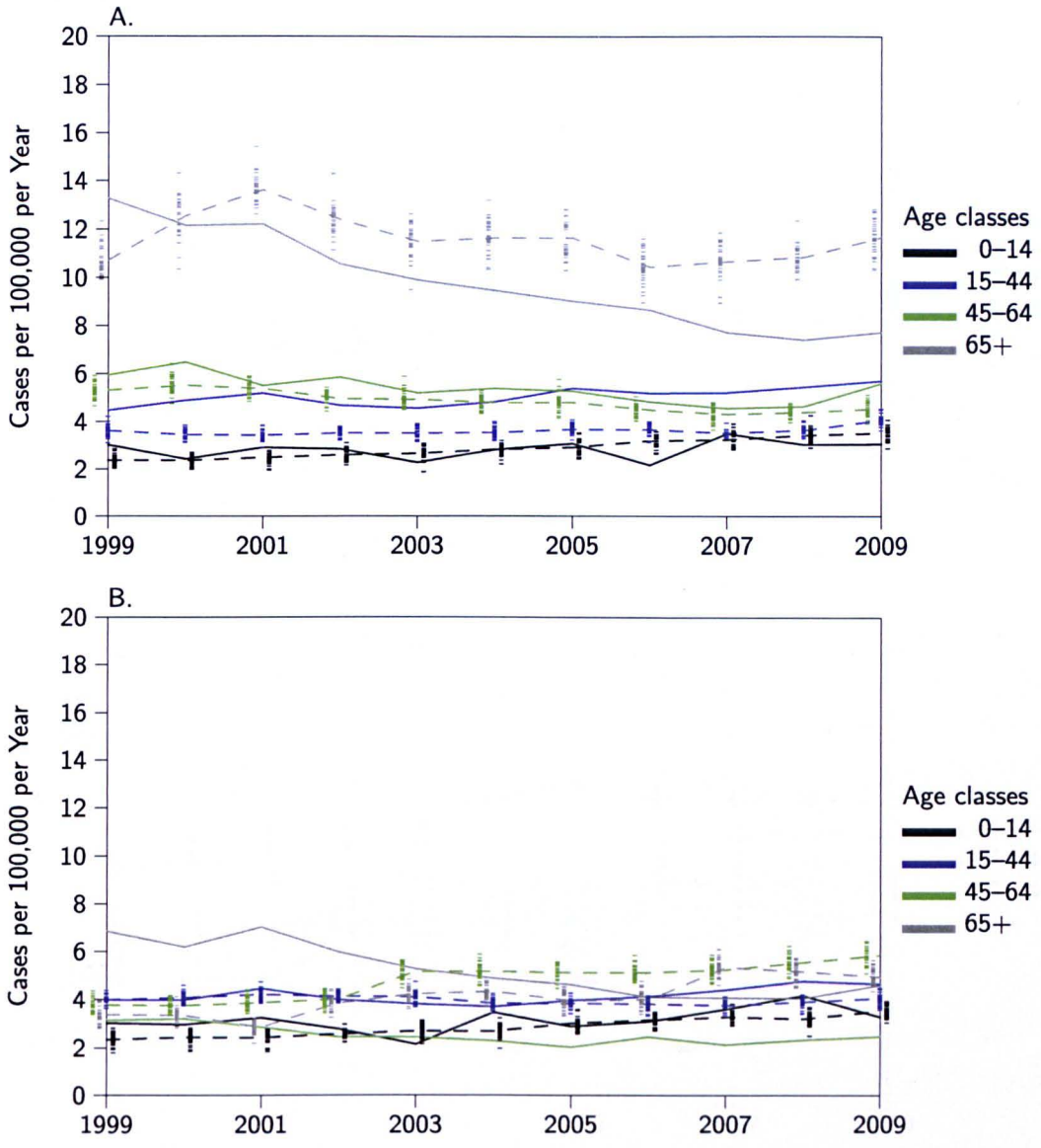


Figure 10-7: Observed versus simulated notification rates per 100,000 population per year for UK-born in England and Wales, 1999 – 2009, for stage two fitting of Scenario 9, for which a single foreign-born category was used. Sub-Saharan African-born and other foreign-born were combined during fitting. Model output follows the dashed line and individual runs of the model are denoted with a '-' (there are 30 for each data point). Observed notification rates follow the solid line. Age categories are as follows: ages 0 – 14 years are in black; ages 15 – 44 years are in blue; ages 45 – 64 are in green; and ages 65 years and over are in grey.

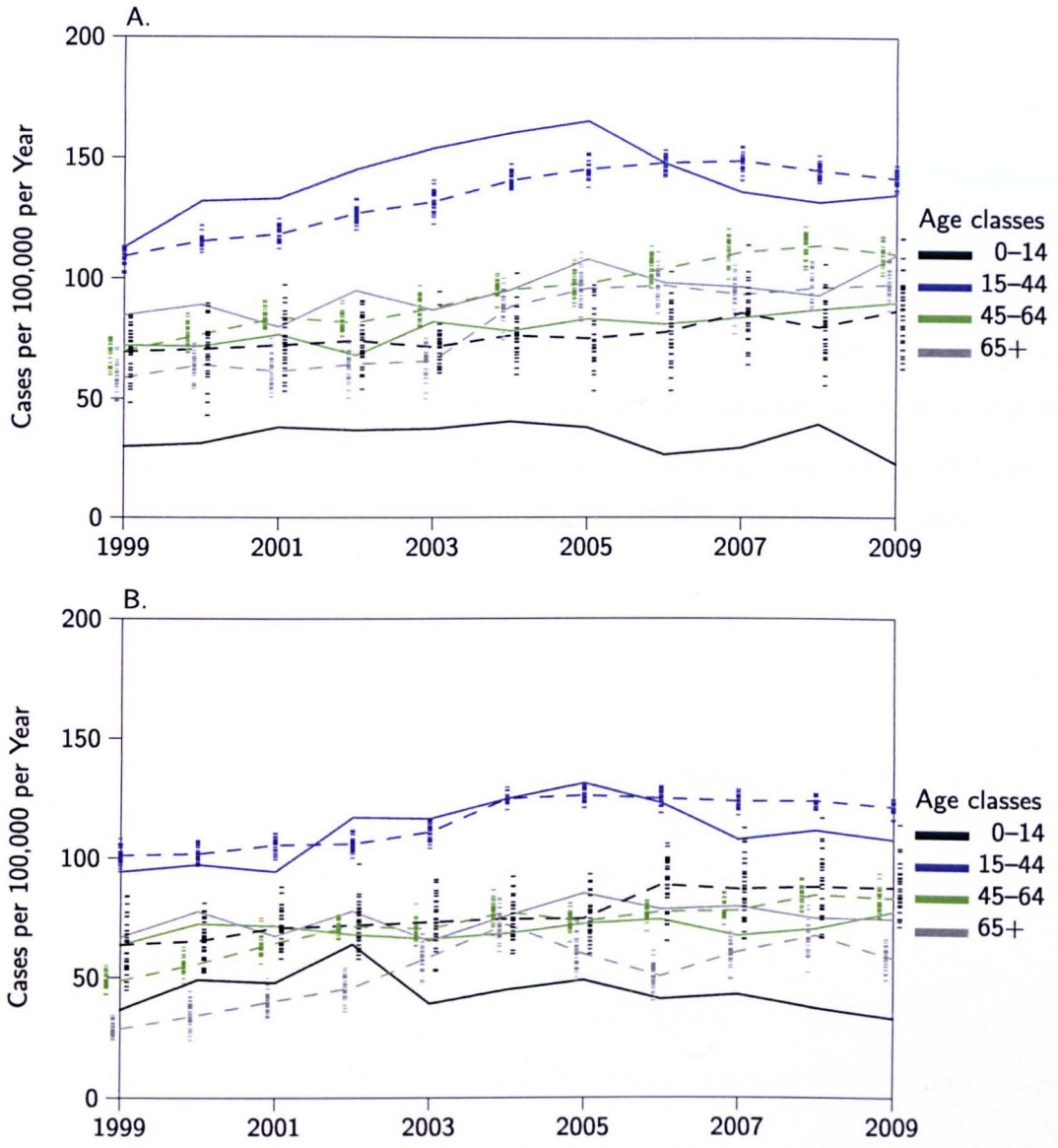


Figure 10-8: Observed versus simulated notification rates per 100,000 population per year for foreign-born in England and Wales, 1999 – 2009, for stage two fitting of Scenario 9, for which a single foreign-born category was used. Sub-Saharan African-born and other foreign-born were combined during fitting. Model output follows the dashed line and individual runs of the model are denoted with a '-' (there are 30 for each data point). Observed notification rates follow the solid line. Age categories are as follows: ages 0 – 14 years are in black; ages 15 – 44 years are in blue; ages 45 – 64 are in green; and ages 65 years and over are in grey.

## 10.14 Missing Data for West Midlands Cases, 2007-2011.

Most cases in the study had complete data on demographic characteristics. Generally, less than 1% of patients were missing data on these characteristics, apart from a few exceptions. There were 3.5% and 6.2% of cases missing data on *ethnicity* and *birthplace*, respectively. *Time since entry* was missing for nearly 10% of foreign-born individuals (and is not applicable to UK-born). Of non-demographic characteristics, there were more missing data; information on *previous diagnosis* was missing for 27% of cases, although collection of data on this feature of cases improved from 2007 – 2011, with less than 5% of 2011 cases missing that information. Also, more than 50% of cases were missing data on the behavioural risk factors, d*rug use, alcohol use, homelessness,* and *prison time,* as these were only collected from mid-2009.

# 10.15 Characteristics of Cases Notified in the West Midlands, 2007-2011.

Table 10-20: Characteristics of all 4,845 cases notified in the West Midlands, 2007-2011.

| Variable | N | % |
|---|---|---|
| **Year notified** | | |
| 2007 | 938 | 19.36 |
| 2008 | 1,015 | 20.95 |
| 2009 | 1,009 | 20.83 |
| 2010 | 872 | 18 |
| 2011 | 1,011 | 20.87 |
| Total | 4,845 | 100 |
| **Sex** | | |
| Male | 2,638 | 54.47 |
| Female | 2,205 | 45.53 |
| Total | 4,843 | 100 |
| **Age group** | | |
| 0 – 14 | 286 | 5.9 |
| 15 – 44 | 2,769 | 57.15 |
| 45 – 64 | 966 | 19.94 |
| 65 and over | 824 | 17.01 |
| Total | 4,845 | 100 |
| **Region of birth** | | |
| UK | 1,555 | 34.76 |
| Europe | 101 | 2.26 |
| East Mediterranean | 45 | 1.01 |
| Africa | 700 | 15.65 |
| Americas | 57 | 1.27 |
| South Asia | 1,912 | 42.75 |
| East/Southeast Asia | 103 | 2.3 |
| Total | 4,473 | 100 |
| **Ethnicity** | | |
| White | 880 | 18.81 |
| Black-Caribbean | 173 | 3.7 |
| Black-African | 711 | 15.2 |
| Black-Other | 18 | 0.38 |
| South Asian | 2,602 | 55.61 |
| Chinese | 47 | 1 |
| Mixed/Other | 248 | 5.3 |
| Total | 4,679 | 100 |
| **Years since entry to tuberculosis diagnosis*** | | |
| 0 – 1 | 435 | 16.35 |
| 2 – 4 | 573 | 21.54 |

| Variable | N | % |
|---|---|---|
| 5 – 9 | 606 | 22.78 |
| 10 and over | 1,046 | 39.32 |
| Total | 2,660 | 100 |
| **Disease site** | | |
| Pulmonary, with or without extra-pulmonary | 2,632 | 55.14 |
| Extra-pulmonary only | 2,141 | 44.86 |
| Total | 4,773 | 100 |
| **Drug sensitivity** | | |
| Resistant to one or more drugs | 147 | 5.42 |
| Sensitive | 2,567 | 94.58 |
| Total | 2,714 | 100 |
| **Previous diagnosis** | | |
| No | 2,920 | 85.56 |
| Yes | 493 | 14.44 |
| Total | 3,413 | 100 |
| **History of or current problem drug use**** | | |
| No | 2,181 | 96.8 |
| Yes | 72 | 3.2 |
| Total | 2,253 | 100 |
| **History of or current problem alcohol use**** | | |
| No | 2,138 | 97.31 |
| Yes | 59 | 2.69 |
| Total | 2,197 | 100 |
| **History of or current homelessness**** | | |
| No | 2,201 | 98.08 |
| Yes | 43 | 1.92 |
| Total | 2,244 | 100 |
| **History of or currently in prison**** | | |
| No | 2,084 | 97.11 |
| Yes | 62 | 2.89 |
| Total | 2,146 | 100 |

*Foreign-born only

**Missing for 2007, 2008 and half of 2009 cases

## *10.16 Demographic Characteristics and Risk Factors for Clustering Using 15-locus VNTR for Cases in the West Midlands, 2007-2011.*

Table 10-21: Demographic features and risk factors for clustering under the 15-locus typing system for cases notified in the West Midlands, using the n and retrospective methods of clustering. The n method results apply to all cases in the study, 2007 – 2011. The retrospective method only applies to cases from 2009 – 2011.

| | **All cases, 07 – 11** | | **Clustered cases, 'n method'** | | | | | | **All cases, 09 – 11** | | **Clustered cases, two-year 'retrospective method'** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | Col % | N | Row % | OR (95% CI) | p | aOR (95% CI) | p | N | Col % | N | % | OR (95% CI) | p | aOR (95% CI) | p |
| **Sex** | | | | | | | | | | | | | | | | |
| Male | 1,271 | 55.7 | 894 | 70.3 | 1.0 | 0.10 | | | 784 | 55.9 | 484 | 61.7 | 1.0 | 0.09 | | |
| Female | 1,010 | 44.3 | 678 | 67.1 | 0.9 (0.7,1.0) | | | | 618 | 44.1 | 354 | 57.3 | 0.8 (0.7,1.0) | | | |
| Total | 2,281 | 100.0 | 1572 | 68.9 | | | | | 1,402 | 100.0 | 838 | 59.8 | | | | |
| **Age group (years)** | | | | | | | | | | | | | | | | |
| 0 – 14 | 55 | 2.4 | 44 | 80.0 | 1.0 | 0.11 | 1.0 | 0.33 | 31 | 2.2 | 21 | 67.7 | 1.0 | 0.70 | 1.0 | |
| 15 – 44 | 1,440 | 63.1 | 984 | 68.3 | 0.5 (0.3,1.1) | | 0.8 (0.4,1.5) | | 873 | 62.3 | 524 | 60.0 | 0.7 (0.3,1.5) | | 0.9 (0.4,2.0) | 0.31 |
| 45 – 64 | 425 | 18.6 | 304 | 71.5 | 0.6 (0.3,1.3) | | 0.8 (0.4,1.6) | | 273 | 19.5 | 164 | 60.1 | 0.7 (0.3,1.6) | | 0.8 (0.3,1.8) | |
| 65 and over | 363 | 15.9 | 241 | 66.4 | 0.5 (0.2,1.0) | | 0.6 (0.3,1.3) | | 225 | 16.1 | 129 | 57.3 | 0.6 (0.3,1.4) | | 0.7 (0.3,1.6) | |
| Total | 2,283 | 100.0 | 1573 | 68.9 | | | | | 1,402 | 100.0 | 838 | 59.8 | | | | |

| | *All cases, 07 – 11* | | Clustered cases, 'n method' | | | | | | *All cases, 09 – 11* | | Clustered cases, two-year 'retrospective method' | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | Col % | N | Row % | OR (95% CI) | p | aOR (95% CI) | p | N | Col % | N | % | OR (95% CI) | p | aOR (95% CI) | p |
| **Region of birth** | | | | | | | | | | | | | | | | |
| UK | 693 | 32.4 | 571 | 82.4 | 1.0 | 0.00 | 1.0 | 0.00 | 412 | 30.7 | 308 | 74.8 | 1.0 | 0.00 | 1.0 | 0.00 |
| Europe | 53 | 2.5 | 32 | 60.4 | 0.3 (0.2,0.6) | | 0.3 (0.2,0.5) | | 37 | 2.8 | 17 | 46.0 | 0.3 (0.1,0.6) | | 0.3 (0.1,0.5) | |
| East Mediterranean | 24 | 1.1 | 15 | 62.5 | 0.4 (0.2,0.8) | | 0.4 (0.2,0.9) | | 17 | 1.3 | 10 | 58.8 | 0.5 (0.2,1.3) | | 0.5 (0.2,1.3) | |
| Africa | 364 | 17.0 | 213 | 58.5 | 0.3 (0.2,0.4) | | 0.3 (0.2,0.4) | | 228 | 17.0 | 107 | 46.9 | 0.3 (0.2,0.4) | | 0.3 (0.2,0.4) | |
| Americas | 25 | 1.2 | 19 | 76.0 | 0.7 (0.3,1.7) | | 0.8 (0.3,2.0) | | 14 | 1.0 | 10 | 71.4 | 0.8 (0.3,2.7) | | 1.0 (0.3,3.3) | |
| South Asia | 927 | 43.3 | 321 | 65.4 | 0.4 (0.3,0.5) | | 0.5 (0.4,0.6) | | 599 | 44.6 | 339 | 56.6 | 0.4 (0.3,0.6) | | 0.5 (0.4,0.6) | |
| East/Southeast Asia | 56 | 2.6 | 31 | 44.6 | 0.2 (0.1,0.3) | | 0.2 (0.1,0.3) | | 37 | 2.8 | 15 | 40.5 | 0.2 (0.1,0.5) | | 0.2 (0.1,0.4) | |
| Total | 2,142 | 100.0 | 1481 | 69.1 | | | | | 1,344 | 100.0 | 806 | 60.0 | | | | |
| **Ethnicity** | | | | | | | | | | | | | | | | |
| White | 383 | 17.4 | 296 | 77.3 | 1.0 | 0.00 | | | 222 | 16.5 | 146 | 65.8 | 1.0 | 0.03 | | |
| Black-Caribbean | 84 | 3.8 | 72 | 85.7 | 1.8 (0.9,3.4) | | | | 37 | 2.7 | 27 | 73.0 | 1.4 (0.6,3.1) | | | |
| Black-African | 350 | 15.9 | 209 | 59.7 | 0.4 (0.3,0.6) | | | | 219 | 16.3 | 111 | 50.7 | 0.5 (0.4,0.8) | | | |
| Black-Other | 11 | 0.5 | 7 | 63.6 | 0.5 (0.1,1.8) | | | | 9 | 0.7 | 5 | 55.6 | 0.7 (0.2,2.5) | | | |

| | ***All cases, 07 – 11*** | | **Clustered cases, 'n method'** | | | | | | ***All cases, 09 – 11*** | | **Clustered cases, two-year 'retrospective method'** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | Col % | N | Row % | OR (95% CI) | p | aOR (95% CI) | p | N | Col % | N | % | OR (95% CI) | p | aOR (95% CI) | p |
| South Asian | 1228 | 55.7 | 847 | 69.0 | 0.7 (0.5,0.9) | | | | 757 | 56.2 | 462 | 61.0 | 0.8 (0.6,1.1) | | | |
| Chinese | 18 | 0.8 | 13 | 72.2 | 0.8 (0.3,2.2) | | | | 10 | 0.7 | 6 | 60.0 | 0.8 (0.2,2.9) | | | |
| Mixed/Other | 129 | 5.9 | 79 | 61.2 | 0.5 (0.3,0.7) | | | | 94 | 7.0 | 54 | 57.5 | 0.7 (0.4,1.2) | | | |
| Total | 2203 | 100.0 | 1523 | 69.1 | | | | | 1,348 | 100.0 | 811 | 60.2 | | | | |
| **Years since entry to tuberculosis diagnosis*** | | | | | | | | | | | | | | | | |
| 0 – 1 | 242 | 18.2 | 131 | 54.1 | 1.0 | 0.00 | | | 151 | 17.4 | 66 | 43.7 | 1.0 | 0.02 | | |
| 2 – 4 | 307 | 23.1 | 200 | 65.2 | 1.6 (1.1,2.2) | | | | 192 | 22.1 | 108 | 56.3 | 1.7 (1.1,2.5) | | | |
| 5 – 9 | 310 | 23.3 | 186 | 60.0 | 1.3 (0.9,1.8) | | | | 218 | 25.1 | 109 | 50.0 | 1.3 (0.8,2.0) | | | |
| 10 and over | 470 | 35.4 | 317 | 67.5 | 1.8 (1.3,2.4) | | | | 308 | 35.4 | 179 | 58.1 | 1.8 (1.2,2.6) | | | |
| Total | 1,329 | 100.0 | 834 | 62.8 | | | | | 869 | 100.0 | 462 | 53.2 | | | | |
| **Disease site** | | | | | | | | | | | | | | | | |
| Pulmonary | 1,505 | 66.3 | 1093 | 72.6 | 1.0 | 0.00 | 1.0 | 0.00 | 931 | 66.6 | 585 | 62.8 | 1.0 | 0.00 | 1.0 | 0.01 |
| Extra-pulmonary | 766 | 33.7 | 475 | 62.0 | 0.6 (0.5,0.7) | | 0.7 (0.6,0.8) | | 467 | 33.4 | 252 | 54.0 | 0.7 (0.6,0.9) | | 0.7 (0.6,0.9) | |
| Total | 2,271 | 100.0 | 1568 | 69.0 | | | | | 1,398 | 100.0 | 837 | 59.9 | | | | |

| | ***All cases, 07 – 11*** | | **Clustered cases, 'n method'** | | | | | | ***All cases, 09 – 11*** | | **Clustered cases, two-year 'retrospective method'** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | Col % | N | Row % | OR (95% CI) | p | aOR (95% CI) | p | N | Col % | N | % | OR (95% CI) | p | aOR (95% CI) | p |
| **Drug sensitivity** | | | | | | | | | | | | | | | | |
| Resistant to at least one drug | 110 | 4.9 | 58 | 52.7 | 1.0 | 0.00 | 1.0 | 0.00 | 78 | 5.6 | 32 | 41.0 | 1.0 | 0.00 | 1.0 | 0.01 |
| Sensitive | 2,159 | 95.2 | 1507 | 69.8 | 2.1 (1.4,3.0) | | 1.9 (1.2,2.8) | | 1,310 | 94.4 | 798 | 60.9 | 2.2 (1.4,3.6) | | 1.9 (1.2,3.2) | |
| Total | 2,269 | 100.0 | 1565 | 69.0 | | | | | 1,388 | 100.0 | 830 | 59.8 | | | | |
| **Previous diagnosis** | | | | | | | | | | | | | | | | |
| No | 1,445 | 86.7 | 1004 | 69.5 | 1.0 | 0.31 | | | 1,059 | 85.1 | 633 | 59.8 | 1.0 (0.0,0.0) | 0.24 | | |
| Yes | 221 | 13.3 | 161 | 72.9 | 1.2 (0.9,1.6) | | | | 185 | 14.9 | 119 | 64.3 | 1.2 (0.9,1.7) | | | |
| Total | 1,666 | 100.0 | 1165 | 69.9 | | | | | 1,244 | 100.0 | 752 | 60.5 | | | | |
| **History of or current problem drug use**** | | | | | | | | | | | | | | | | |
| No | 1,047 | 95.9 | 713 | 68.1 | 1.0 | 0.00 | | | 1,018 | 95.9 | 603 | 59.2 | 1.0 (0.0,0.0) | 0.00 | | |
| Yes | 45 | 4.1 | 39 | 86.7 | 3.0 (1.3,7.3) | | | | 44 | 4.1 | 38 | 86.4 | 4.4 (1.8,10.4) | | | |
| Total | 1,092 | 100.0 | 752 | 68.9 | | | | | 1,062 | 100.0 | 641 | 60.4 | | | | |

| | All cases, 07 – 11 | | Clustered cases, 'n method' | | | | | All cases, 09 – 11 | | Clustered cases, two-year 'retrospective method' | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | Col % | N | Row % | OR (95% CI) | p | aOR (95% CI) | p | N | Col % | N | % | OR (95% CI) | p | aOR (95% CI) | p |
| **History of or current problem alcohol use**** | | | | | | | | | | | | | | | | |
| No | 1,025 | 96.7 | 698 | 68.1 | 1.0 | 0.00 | | | 997 | 96.6 | 592 | 59.4 | 1.0 (0.0,0.0) | 0.00 | | |
| Yes | 35 | 3.3 | 32 | 91.4 | 5.0 (1.5,16.4) | | | | 35 | 3.4 | 30 | 85.7 | 4.1 (1.6,10.7) | | | |
| Total | 1,060 | 100.0 | 730 | 68.9 | | | | | 1,032 | 100.0 | 622 | 60.3 | | | | |
| **History of or current homelessness**** | | | | | | | | | | | | | | | | |
| No | 1,066 | 96.9 | 732 | 68.7 | 1.0 | 0.54 | | | 1,034 | 96.8 | 616 | 59.6 | 1.0 (0.0,0.0) | 0.19 | | |
| Yes | 34 | 3.1 | 25 | 73.5 | 1.3 (0.6,2.7) | | | | 34 | 3.2 | 24 | 70.6 | 1.6 (0.8,3.4) | | | |
| Total | 1,100 | 100.0 | 757 | 68.8 | | | | | 1,068 | 100.0 | 640 | 59.9 | | | | |
| **History of or currently in prison**** | | | | | | | | | | | | | | | | |
| No | 1,007 | 95.9 | 698 | 69.3 | 1.0 | 0.00 | | | 980 | 95.9 | 589 | 60.1 | 1.0 (0.0,0.0) | 0.00 | | |
| Yes | 43 | 4.1 | 38 | 88.4 | 3.4 (1.3,8.6) | | | | 42 | 4.1 | 36 | 85.7 | 4.0 (1.7,9.5) | | | |
| Total | 1,050 | 100.0 | 736 | 70.1 | | | | | 1,022 | 100.0 | 625 | 61.2 | | | | |

*Foreign-born only, **Missing for 2007, 2008 and half of 2009 cases

## 10.17 Proportion of Isolates Clustered by Study Duration

The proportion clustered by study duration is shown in Figure 10-9, with clustering defined according to the n method and the 24-locus VNTR typing system. Clustering increases with increasing study duration. Note that since the retrospective method already requires exclusion of two years'-worth of data, an analogous plot using this method would only have three data points, so was not produced. This analysis could be used to support the two-year definition of retrospective method, as more than 80% of total clustering over the five years period is attained after two years, illustrated in Figure 10-10.

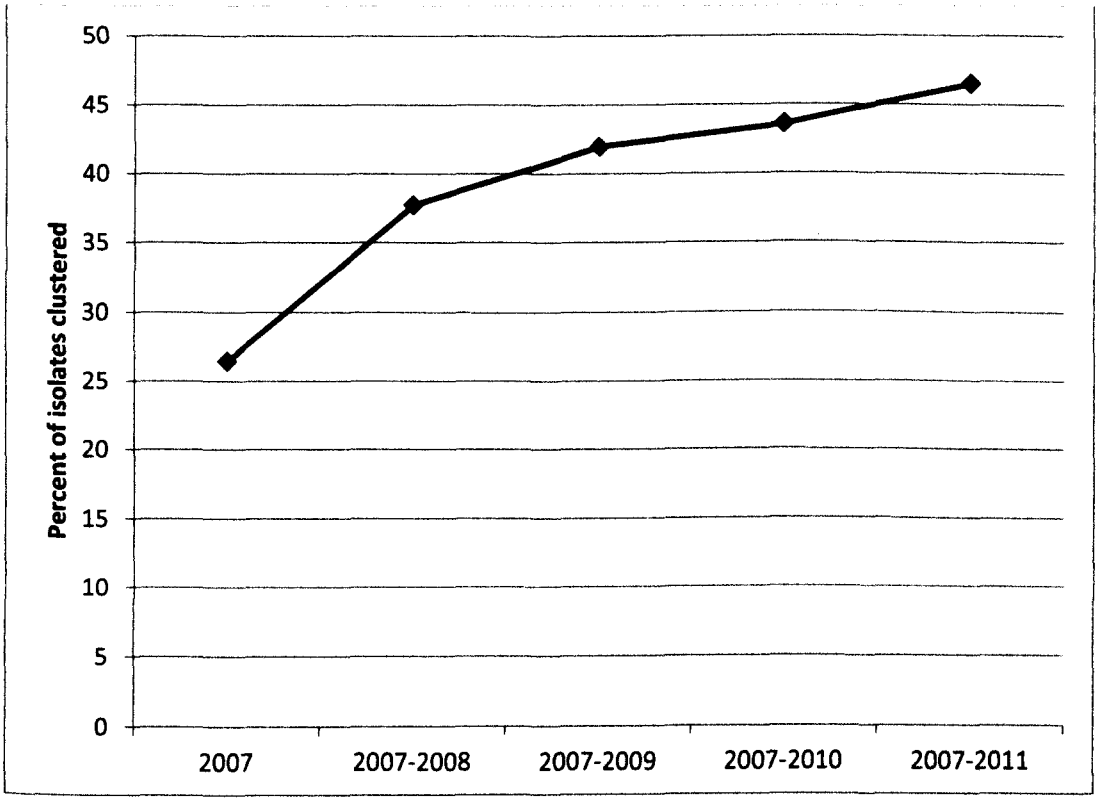


Figure 10-9: Percentage of isolates clustered by study duration, where clustering is defined according to the n method, using the 24-locus VNTR typing system.

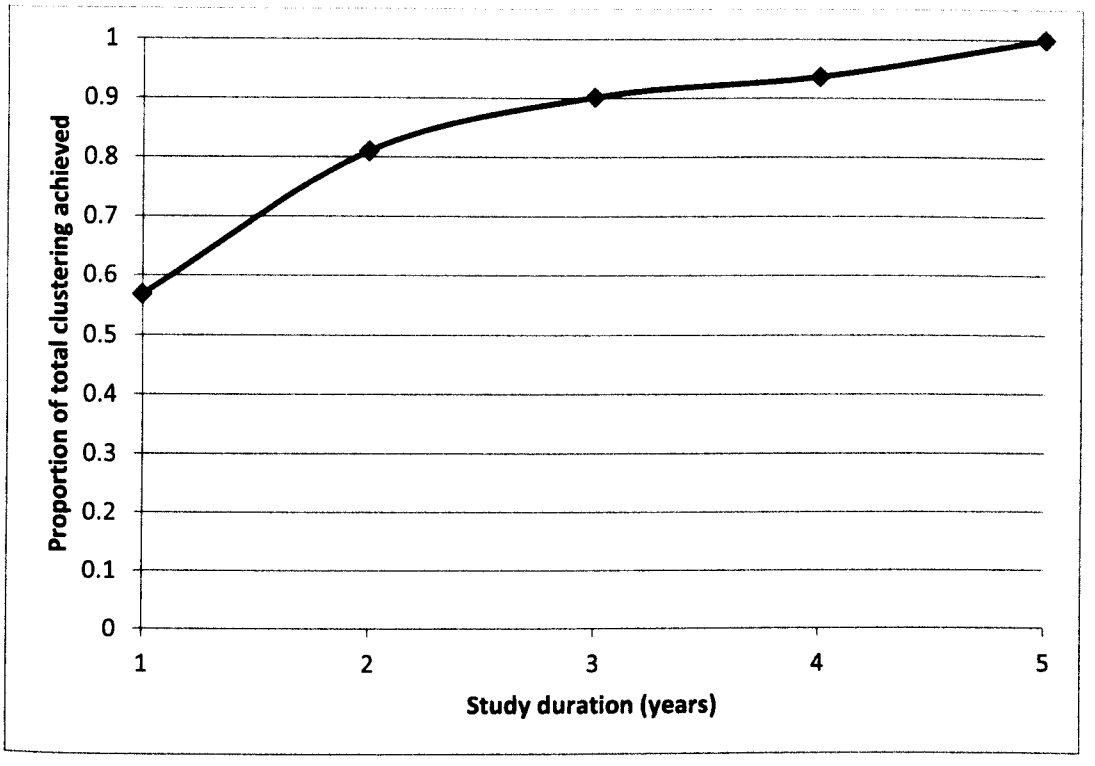


Figure 10-10: Ratio of clustering under various study durations (1 – 5 years) to clustering over the entire study period (five years), where clustering is defined using the n method using 24-locus VNTR.

## 10.18 Exclusion of Cases Due to Laboratory Contamination

In the present study, only the date of case report to ETS was available for all cases, not the date of receipt or processing of the clinical specimen in the laboratory. Therefore, no isolates were excluded from analyses due to suspected laboratory contamination.

Studies of laboratory contamination of tuberculosis cultures or isolates have resulted in a wide range of estimates for the proportion of cases affected by laboratory contamination, though a recent UK study found only 0.54 – 0.93% of cases with false positive results due to laboratory cross-contamination [333]. Although this is a serious problem for individual patients if resulting in misdiagnosis, contamination rates are low enough that laboratory contamination is unlikely to impact population-based molecular epidemiological studies such as the present study. Still, laboratory contamination could result in a slight increase in the proportion clustered in the study, though efforts to exclude cases due to laboratory contamination could alternatively result in legitimately clustered cases being excluded, slightly lowering the proportion clustered.

For comparison with other studies in the literature, it was observed that some molecular epidemiological studies have excluded cases due to suspected laboratory contamination [77, 321, 326, 329, 343-345], while many others have not [165, 273, 313, 332, 346-350]. One common method for excluding cases suspected of laboratory contamination is to exclude cases for which the following three criteria are met: 1) There is only a single positive culture for the case; 2) A test for acid-fast bacilli is smear-negative; and 3) Another case, processed in the laboratory on the same day, was also found culture-positive and had the same strain type profile as the suspected false positive case [77, 162, 326, 329, 351]. Similar methods have been used in other studies [164, 345].

## 10.19 Model Code and Published Algorithms

This appendix exhibits the source code and detailed algorithms that have been applied to the research questions in this thesis. The purpose and design of the model were described in Chapter 3. In addition, objective one of the thesis included making the model be fully and freely available so that others might understand its operation in detail, validate its applicability to particular problems, adapt it to new and different situations, and generally improve it for future use.

Accordingly, source code for the main program, with complete technical documentation, appears in this appendix. Included is all the code that implements the flow depicted in Figure 3-2, comprising approximately 2,500 lines. Not included is supporting code that is unnecessary for understanding the program, such as (1) input routines to gather parameters from commands and files, (2) output routines to display results as listings and graphs, and (3) standard subroutines such as those to generate random numbers from classical probability distributions. This code and all the supporting code, approximately 8,000 lines total, is downloadable free from a project website, www.cbs.umn.edu/modeling, and is also available free from the author upon request.

The source code included here is organized into 32 main sections, each containing, where applicable, (1) a number and title for the section, (2) an abstract describing the section, (3) conditions that must be established before code in the section is invoked, (4) conditions that will prevail when the code of the section has completed execution, (5) a description of the algorithm in a right-hand column, and (6) the code that accomplishes the algorithm in a left-hand column. The amount and style of the information presented is intended to supply that which is necessary and sufficient to validate the program's correctness, and which has been used in that way during development and testing of the program.

The code is a subset of the standard programming language C. The subset uses only those features necessary for efficient scientific programs. For example, indexed arrays rather than memory pointers are used to managing data within the program. This results in no loss of speed but affords an improvement in comprehensibility. The style of the coding is also meant to enhance readability and comprehensibility. For example, the local structure of the code is indicated by indentation, but with the block-defining

brackets omitted when unnecessary and placed unobtrusively when they are necessary, to reduce visual noise and clutter.

A number of general computer algorithms and techniques were developed as part of this thesis research, or were expanded and tested during this research. Five of those of which I am either auxiliary or primary author are now published in the computing literature. The papers describing these general algorithms are also necessary for fully understanding the code of the model, and accordingly they are also included, after the source code in this appendix, as follows.

1. Main source code of the model

2. Algorithm for managing groups of objects

3. Algorithm for managing schedules of future events

4. Algorithm for random numbers from any distribution

5. Method of organizing data read by the model

6. Method of multi-processing for model simulation

# Model for Tuberculosis in the UK

This individual-based model (IBM) is to simulate tuberculosis dynamics in the UK. This version of the model stores individual strain types for each infection to simulate strain type clustering patterns seen in disease cases. This version was developed for use in the West Midlands, a region with a population size of around five million people. One major characteristic of individuals is their region of birth, UK or non-UK. In the `SSAV` version of the model, the non-UK-born region of birth is divided into Sub-Saharan African born (SSA-born) and other non-UK born (ONUK-born) for some model parameters.

1. BACKGROUND ALGORITHMS. The core IBM algorithms, external to this simulation program, were written by Clarence Lehman (CL) in 2009. The method was first developed for simulation of HIV dynamics in the US. Beginning in January 2010, and with initial help from CL, Adrienne Keen (AK) adapted this IBM skeleton for modelling tuberculosis dynamics in the UK, first for fitting the model to tuberculosis notifications in England and Wales and then for simulating genotyping data from the West Midlands.

2. SUMMARY OF METHOD. This is an event-based simulation where continuous time is simulated directly. There is no arbitrary time step. Instead, events are processed one at a time, chronologically. The time variable `t` is time in years, with arbitrarily high resolution down to small fractions of a instant and all complexities and inaccuracies associated with multiple events during a finite time step vanish—such as undershooting zero when the sum of the rates times the width of the time step exceeds unity. Continuous time also allows all activities to occur in a single data array, rather than having to swap old and new arrays at each time step. Events are assumed to following probability distributions that vary through time and space.

States of the system never change spontaneously—all changes are induced by some other event in the system and usually scheduled in advance. The scheduled times are determined stochastically from functions whose characteristics may depend on the state of the individual and the environment at the time. An individual's age, sex, infection history, or any other considerations can be incorporated into the functions.

For example, death is scheduled at the time of birth, with the time chosen randomly from a life-span distribution for babies born in the simulated year. But the scheduled time of death is not immutable, nor are any other scheduled times in the system. If the individual develops disease, the scheduled time of death may be cancelled and a new time of death due to tuberculosis may be scheduled instead. At no time does the program visit an individual when it does not need to, and therein lies its speed.

Many future events may apply to each individual and are saved for that individual, but only the earliest among each individual's events enters a global "list of future events."

3. MAIN DATA STRUCTURES. Each individual is assigned a number 1 through `n` and recorded in a linear array `A` of structures `Indiv`. Each element of `A` defines the state of the correspondng individual.

For example, this would be defined as `struct A[indiv+3]`, the main array of individuals. Suppose there are 3 UK-born and 3 non-UK-born individuals, with a total maximum population size of 14 (`indiv`). Note, in the `SSAV` version of the model, SSAs are stored as non-UK born and it is not possible to tell from their ID number alone whether they are SSA-born or ONUK-born.

```
 n   A[n]                          A array index
-------------------------------------------------
 0  [Reserved]                (Null pointer for list)
 1  (Non-UK-born)
 2  (Non-UK-born)
 3  (Non-UK-born)                    immid-1
 4  [Empty]                          immid
 5  [Empty]
 6  [Empty]
 7  [Empty]                          maximm
-------------------(Imaginary separator, non-UK/UK born)
 8  (UK-born)                        maximm+1
 9  (UK-born)
10  (UK-born)                        ukbid-1
11  [Empty]                          ukbid
12  [Empty]
13  [Empty]
14  [Empty]                          indiv
15  External event, birth            indiv+1 or BIRTH
16  External event, immigration      indiv+2 or IMM
-------------------------------------------------
```

4. MODEL FITTING. Prior versions of the model were designed to work with an optimization algorithm for model fitting, currently found in `fit5.c`. In this version of the model, the fitting routine is not needed.

To run this program stand-alone, as opposed to inside the fitting routine, simply comment out the `=define main mainiac` and everything else will be handled automatically.

Below is a sample program call when it is running as an individual executable, not linked with the fitting routine. A different syntax is used for the call inside `fit5i.c`.

```
tb36gen  df=2.5 d1uk20=0.10 d2uk20=0.0003 d3uk20=0.05
```

When linked with the fitting routine, the model is called from the fitting routine, not as an independent executable, so that it is compatible with parallel runs using MPI commands. In this set up, the model accepts four variable disease risk parameters, `df`, `d1uk20[M]`, `d2uk20[M]`, and `d3uk20[M]`. See the function `Data` for information on these. Briefly, `df` is the factor by which UK-born disease risks are multiplied to obtain non-UK born disease risks. The other three parameters are UK-born disease risks for Primary, Reactivation, and Reinfection Disease respectively, in adult males (those aged 20 years and over). In this version of the model, disease risk are fixed for children under ten years of age, allowing for fewer variable parameters.

5. OTHER. Note that the following program substitutes the term `dec` for the C term `double`. It is short for "decimal", in contrast and parallel with "integer", saving valuable coding columns at the left of the line and helping data names line up.

The sequence of random numbers is specified on the command line with phrases like `randseq=0`, `randseq=1`, `randseq=-6`, `randseq=239702397623`, and so forth. Fixed sequences that are the same each time the program runs occur when `randseq` is 0 or greater. Each

integer gives a different sequence of random numbers. (Actually, it gives only a different starting point in a single long sequence of random numbers.) Positive integers, or zero, are typically used in testing because program results are precisely repeatable.

Arbitrary sequences that are different, with high probability, each time the program runs occur when **randseq** is negative. The date and time, measured to the nearest second, selects the starting sequence, then the negative value modifies that sequence. Thus if several instances of the program were started on separate processors at the same time, the first with **randseq=-1**, the second with **randseq=-2**, and so forth, each instance of the program is guaranteed a different random number sequence. Unlike the case with non-negative integers, however, the sequence be different each time the program runs, with very high probability.

The actual starting seed, incorporating the time of day if requested by a negative value of **randseq**, is stored in **rand0** and reported at the end of the run. That allows a run to be repeated exactly even if it was started with an arbitrary sequence.

6. NOTES ON GENETIC STRAINS. At model initialization, strains are randomly assigned from two different strain type distributions derived from empirical data, one for non-UK born and one for UK-born, saved in **sdimm** and **sduk**, respectively. Also, the total number of available strains in each distribution is **is0** and **is1**, respectively. These are set in define statements before the model is run to faciliate array initialization and contiguous number of strains. New mutant strains which appear as the simulation runs will be given strain IDs distinct from any in the **sdimm** or **sduk** distributions. These IDs will begin with integer **is0+is1** and will always be greater than or equal to this value. The variable **stid** holds the next available strain type ID for new mutants. The diagram below illustrates which strain IDs belong to which individuals in the simulation:

| StrainID | Available for | Generic Reference |
|---|---|---|
| 0 | (Not used) | |
| 1 | Initial non-UK, migrants | |
| 2 | Initial non-UK, migrants | |
| 3 | Initial non-UK, migrants | |
| 4 | Initial non-UK, migrants | is0 |
| 5 | Initial UK | is0+1 |
| 6 | Initial UK | |
| 7 | Initial UK | |
| 8 | Initial UK | is0+is1 |
| 9 | New mutant strain | is0+is1+1 |
| 10 | New mutant strain | |
| 11 | New mutant strain | |
| 14 | Next available new mutant | stid |

```
#include <stdio.h>
#include <stdlib.h>
#include <time.h>
#include "common.h"
#include "fileio.h"

#define PN (q1+1)              Number of elements in array N.
#define T0 1981                Start time of model, years.
#define T1 2012                End time of model, years. The simulation
                               ends before reaching this year.
#define TDATA 2007             First year of observed data, for reporting times.
#define SSAV   1               Switch model version depending on existence
```

```
                                    of separate SubSaharan African group,
                                    0=non-SSA, 1=SSA.
#define SUPER  1                    Flag for whether model is run on supercomputer,
                                    0=no, 1=yes (changes population sizes).
#define DPARAM 1                    Allows model to accept disease progression
                                    parameters (4 in this version), 0=no, 1=yes.
#define REC    1                    Flag for whether tallying recent transmissions.
#define BIRTH (indiv+1)             Index used for scheduling births.
#define IMM   (indiv+2)             Index for scheduling arrival of immigrants.
#define RT    (T1-T0)               Running time of model, calendar years.
#define NUK   0                     Array index for non-UK born.
#define UK    1                     Array index for UK-born.
#define HIV   2                     Array index for HIV+.
#define SSA   2                     Array index for SSA-born.
#define M     0                     Array index for males.
#define F     1                     Array index for females.
#define E     0.0000000001;         Small number added to some event times to
                                    ensure they happen in the future.
#define AC 122                      Age classes for mortality data.
#define LAT 5                       Years to Remote from recent (re)infection.
#define BY (2012-1870+1)            Number of birth cohorts for mortality data.
#define IS0 5000                    Number of strains for migrants.
#define IS1 1000                    Number of strains for UK-born at model
                                    initialization.
dec N[PN];                          Current number in each disease state.
dec N2[4][2][2][RT];                Population sizes in the model at end of year by
                                    age, sex, rob and year.
dec N3[4][2][2][RT];                Population sizes observed, which are compared
                                    with model population sizes and used to correct
                                    case numbers produced by the model.
int Np[4][2][2][RT];                Infected persons by age, sex, rob, year.
dec age1[2],age2[2],agec[2];        Accumulators for 1st and 2nd moments of age.
dec repc[4][2][2][2][RT];           Reported cases by age category, sex, rob,
                                    disease site and year.
dec repc2[4][2][2][2][RT][5];       Time/place of tranmission for reported cases,
                                    indexed as repc plus last index is 0 for
                                    total cases, 1 for recent/UK, 2 for older/UK
                                    3 for recent/NonUK and 4 for older/NonUK.
dec repc3[15000][7];                Data on typed cases, 0=age, 1=sex, 2=rob, 3=time
                                    of report, 4=place/time of infection, 5=strain ID,
                                    6=number of others with identical strain.
int ari[2][RT];                     Number of successful transmissions, by rob
                                    (UK/NUK) and year.
int clust[4][2][2][5];              Clustering data by age, sex, and rob, where
                                    last index gives cases which are: 0=unique and
                                    non-recent/non-UK,1=unique and recent/UK,
                                    2=clustered and non-recent/non-UK, 3=clustered
                                    and recent/UK, 4=total typed cases.
int deaths;                         Current number of deaths.
int events;                         Current number of events dispatched.
int immid;                          Next available ID number for immigrants.
int ukbid;                          Next available ID number for UK-born.
int stid;                           Next available ID for new strain types.
int repid;                          Next available ID for repc3 case report.
```

```
extern dec t;
```
Current time (Managed by `EventSchedule`).

```
dec pt;
```
Time of previous report.

```
dec t0 = T0;
```
Beginning time of simulation.

```
dec t1 = T1;
```
End time of simulation.

```
int tdata = TDATA;
```
First year of observed data.

```
int lup;
```
Year of last update to birth and immigration rates, which are sensitive to calendar year.

```
dec runid;
```
ID number for printing output files.

```
unsigned long startsec;
```
Starting clock time, seconds of Unix.

```
unsigned long rand0;
```
Starting random number seed.

```
struct Indiv *A;
```
State of each individual, including their characterisitics, saved event times, etc.

# 1. Parameters and control variables

## Population initialization

```
int maximm;
```
Maximum immigrants in pop'n at any time.

```
dec inf1981[121][3][2][9];
```
Cumulative probabilities of the 9 disease states for pop. initialization (by a,s,rob).

```
dec n1981[121][2][2];
```
Numbers in each age/sex/rob category at population initialization, 1981.

```
dec ssa1981[121][2];
```
Proportion SSA by age/sex category.

## Infection transmission

```
dec c[2][2];
```
Effective contacts per year per pulmonary case (smear+) by sex and region of birth.

```
dec pcc;
```
Probability effective contact is close contact (drawn from within own region of birth, UK or non-UK).

```
dec  s2[2];
```
Relative susceptibility to reinfection (s).

```
dec smear[121];
```
Proportion smear positive by age.

## Vaccination

```
dec  v1[2];
```
Efficacy of vaccine (rob).

```
dec  v2[2];
```
Portion vaccinated at designated age (rob).

```
dec  v3[2];
```
Average age of vaccination (rob).

## Disease progression

```
dec d1[2][3][121];
```
Proportion Recently Infected who progress to disease over first 5 years of infection (by sex,rob,age)

```
dec d3[2][3][121];
```
Proportion Reinfected who progress to disease over first 5 yrs of reinfection (by a,s,r).

```
dec drr[6];
```
Cumulative, relative risk of disease progression by year since infection, used with `d1` and `d3` (for first 5 years of infection).

```
dec B1[6];
```
Array of values for finding cumulative risk of in first five years of infection/reinfection, used with drr.

| | |
|---|---|
| `dec d2[2][3][AC+2];` | Proportion of those Remotely Infected who progress to disease, cumulative dsn by sex, rob (where r=0, 1, or 2. 2=HIV+ and SSA in SSA version of model), and age. |
| `dec A2[AC+2];` | Array of values for finding random time to disease for Remote Infection. |
| `dec ehiv;` | Factor by which non-UK born disease risks are multiplied for HIV+ SSA individuals. |
| `dec df;` | Factor by which UK-born disease progression rates are multiplied to get immigrant rates. |
| `dec d1uk10[2];` | Rates of disease progression for primary (1), |
| `dec d1uk20[2];` | Reactivation (2), and Reinfection (3) disease |
| `dec d2uk10[2];` | for those aged 0-10 (10) and 20+ (20) by sex. |
| `dec d2uk20[2];` | These help construct `d1`, `d2`, and `d3`. |
| `dec d3uk10[2];` | |
| `dec d3uk20[2];` | |
| `dec sdf1[2];` | Risk ratios for female:male disease |
| `dec sdf2[2];` | progression risks/rates (by age 0-10,20+). |
| `dec sdf3[2];` | |
| `dec presp;` | Proportion of all tuberculosis which is respiratory, for correcting disease risks in Vynn. and Fine to pulmonary disease risks in children. |
| `dec p1[121][2][2];` | Portion pulm -primary disease (a,s,rob) |
| `dec p2[121][2][2];` | Portion pulm -reactivation disease (a,s,rob) |
| `dec p3[121][2][2];` | Portion pulm -reinfection disease (a,s,rob) |
| `dec duk1p[2][2];` | Intermediate parameters for pulmonary-only |
| `dec duk2p[2][2];` | rates of disease progression. Used for |
| `dec duk3p[2][2];` | incorporating estimated rates from Vynnycky and Fine into rates for this model, which are combined pulmonary/non-pulmonary. Indexed by age (0-10yrs, 20+yrs) and sex. |
| `dec d1p[121][2][2];` | Intermediate param's for pulmonary-only rates |
| `dec d2p[121][2][2];` | of disease progression. Used to get overall |
| `dec d3p[121][2][2];` | rates using those estimated by Vynn. and Fine (which are pulmonary rates). Indexed by age, sex, and rob. Precursors to `d1`, `d2`, `d3`. |

### Disease recovery, indexed by sex

| | |
|---|---|
| `dec r3[2];` | Primary disease recovery rate |
| `dec r4[2];` | Reactivation disease recovery rate |
| `dec r5[2];` | Reinfection disease recovery rate |
| `dec r6[2];` | Primary non-pulmonary disease recovery rate |
| `dec r7[2];` | Reactivation non-pulm. disease recovery rate |
| `dec r8[2];` | Reinfection non-pulm. disease recovery rate |

### Mortality

| | |
|---|---|
| `dec A1[AC];` | Holds ages 0-121 which correspond to the cumulative probabilities in M1. |
| `dec M1[BY][2][AC];` | Cumulative probabilities of death by a given birth cohort, sex and age. |
| `dec cft[121][2][RT];` | Case fatality rate due to tuberculosis (a,type dis,y) |

Below are mortality rates used to generate lifetimes with exponential distribution. used for "reduction testing" of the model.

| | |
|---|---|
| `dec m1 [2][RT];` | Mortality of uninf/vacc/inf ind's (sex, y) |
| `dec m6 [2][RT];` | Mortality of primary disease (sex,y) |
| `dec m7 [2][RT];` | Mortality of reactivation disease (sex,y) |
| `dec m8 [2][RT];` | Mortality of reinfection disease (sex,y) |
| `dec m9 [2][RT];` | Mortality of primary non-pulm. disease (sex,y) |
| `dec m10[2][RT];` | Mortality of reactivated non-pulm. disease (s,y) |
| `dec m11[2][RT];` | Mortality of reinfection non-pulm. disease (s,y) |

## Birth and migration

| | |
|---|---|
| `dec bcy[RT];` | Births by calendar year. |
| `dec pmale[RT];` | Portion of newborns who are male by year. |
| `dec immig[RT];` | Total (uk+non-uk-born) immigrants by year. |
| `dec pimm[RT];` | Proportion of immigrants non-UK born by year. |
| `dec ssaim[RT];` | Proportion of non-UK born immigrants born in SubSaharan Africa by year. |
| `dec hivp[2][RT];` | HIV Prevalence in SubSaharan African born immigrants by sex,year. |
| `dec immsex[RT][3];` | Proportion immigrants who are male by yr and rob:0=non-UK,1=UK, 2=non-UK SSA. Note there are different input files for SSA and non-SSA version of the model. |
| `dec immage[RT][2][3][7];` | Cumulative proportion of immigrants (by yr, sex,rob) in age classes, for use with RandF. |
| `dec immageX[RT][2][3][6];` | Probabilities of 6 age classes (by yr,sex, rob) from ONS inflow data—as in **immsex**. Note there are two versions of the input file for this array. Precursor to **immage**. |
| `dec infimm[121][3][RT][9];` | Cumulative probabilities immigrants enter disease states, by age,rob and year. |
| `dec Ax[9];` | State variables which accompany **infimm**. |
| `dec ypb, ypi;` | Years per birth, years per immigrant. |
| `dec em[2][3];` | Annual emigration rate by sex, rob. |

## Strain type related

| | |
|---|---|
| `dec md;` | Mutations per year per strain type (diseased). |
| `dec mi;` | Mutations per year per strain (infected). |
| `dec is0 = IS0;` | Number of strains for migrants. |
| `dec is1 = IS1;` | Number of strains for UK-born. |
| `dec S0[IS0+1];` | Strain IDs to accompany strain distribution for migrants. |
| `dec S1[IS1+1];` | Strain IDs to accompany strain distribution UK-born at initialization. |
| `dec sdimm[IS0+1];` | Cumulative probabilities of strain types for non-UK born and migrants. |
| `dec sduk[IS1+1];` | Cumulative probabilities of strain types for UK-born at initialization. |
| `dec ptyped[2];` | Proportion of cases with strain type, by disease site, 0=non-pulm, 1=pulm. |

The following two recovery rates will not be used in version of model that defines Remote Infection as LAT years after most recent infection.

```
dec r1[2];                      Rate Recent Infection moves to remote (s)
dec r2[2];                      Rate Reinfection moves to remote (s)

dec proprep;                    Proportion of cases reported.

dec relativetime = 0;           Set for relative time reporting.
dec randseq = 0;                Random number sequence (set with randseq=N).
dec tgap    = 0.5;              Time between reports, years.
dec kernel  = 0;                Contagion kernel, 0=Panmictic, 1=Cauchy.
dec sigma   = 1;                Width of contagion kernel, where applicable.
```

```
struct IO fmt[] =               Format statements for input/output.
{ /*00*/ { (dec*)bcy,     {-'i',RT} },
  /*01*/ { (dec*)immig,   {-'i',RT} },
  /*02*/ { (dec*)pimm,    {-'i',RT} },
  /*03*/ { (dec*)ssaim,   {-'i',RT} },
  /*04*/ { (dec*)pmale,   {-'i',RT} },
  /*05*/ { (dec*)hivp,    {-'s',2,-'Y',RT}, {-'y',-'S'} },
  /*06*/ { (dec*)infimm,  {-'a',121,-'r',3,-'y',RT,-'q',9}, {-'R',0,SSAV+1,-'Y',-'Q',-'A'} },
  /*--*/ { (dec*)inf1981, {-'a',121,-'r',2,-'q',9},  {-'a',120,0,-'r',UK,UK,  -'q',1,7} },
  /*--*/ { (dec*)inf1981, {-'a',121,-'r',2,-'q',9},  {-'a',120,0,-'r',NUK,NUK,-'q',1,7} },
  /*09*/ { (dec*)ssa1981, {-'a',121,-'s',2}, {-'s',-'A'} },
  /*10*/ { (dec*)n1981,   {-'a',121,-'s',2,-'r',2},  {-'s',-'a',-'R',1,0,1} },
  /*11*/ { (dec*)immsex,  {-'i',RT,-'r',3}, {-'i',-'R',0,SSAV+1} },
  /*12*/ { (dec*)immageX, {-'i',RT,-'s',2,-'r',3,-'a',6}, {-'i',-'r',0,SSAV+1,-'s',-'a'} },
  /*13*/ { (dec*)immage,  {-'i',RT,-'s',2,-'r',3,-'a',7}, {-'i',-'R',0,SSAV+1,-'A',-'s'} },
  /*14*/ { (dec*)M1,      {-'i',BY, -'s',2,-'a',AC}, {-'s',-'i',-'A'} },
  /*15*/ { (dec*)cft,     {-'a',121,-'d',2,-'i',RT} },
  /*16*/ { (dec*)d1,      {-'s',2,-'r',3,-'a',121}, {-'s',-'r',0,1,-'A'} },
  /*17*/ { (dec*)d2,      {-'s',2,-'r',3,-'a',124}, {-'s',-'r',0,2,-'A'} },
  /*18*/ { (dec*)d3,      {-'s',2,-'r',3,-'a',121}, {-'s',-'r',0,1,-'A'} },
  /*19*/ { (dec*)inf1981, {-'a',121,-'s',2,-'r',3,-'q',9}, {-'r',-'s',-'A',120,0,-'Q',1,8} },
  /*20*/ { (dec*)smear,   {-'a',121} },
  /*21*/ { (dec*)N3,      {-'a',4,  -'s',2,-'r',2,-'i',RT} },
  /*22*/ { (dec*)repc,    {-'a',4, -'s',2, -'r',2, -'d',2, -'i', RT} },
  /*23*/ { (dec*)sdimm,   {-'i',IS0+1} },
  /*24*/ { (dec*)sduk,    {-'i',IS1+1} },
         { } };
```

## 2. Main initialization

This routine should be called each time the program is reused, to clear static variables for the next run. The function was added when `tb30i.c` was made into a function of the fitting routine, to implement parallel, replicate runs of the tuberculosis program. This would not be necessary if the program were called as independent executable, as before.

```
MainInit()
{ int i,j,k,l,m,n;

  for(i=0; i<PN; i++) N[i] = 0;
  for(i=0; i<2;  i++) age1[i] = age2[i] = agec[i] = 0;

  for(i=0; i<4;  i++)
  for(j=0; j<2;  j++)
  for(k=0; k<3;  k++)
  for(l=0; l<2;  l++)
  for(m=0; m<RT; m++)
  for(n=0; n<5;  n++)
  { repc[i][j][k][l][m]     = 0;
```

```
    repc2[i][j][k][l][m][n] = 0;
    ari[j][m]               = 0;
    N2[i][j][k][m]          = 0;
    N3[i][j][k][m]          = 0;
    Np[i][j][k][m]          = 0;
    clust[i][j][l][n]       = 0; }

  for(i=0; i<15000; i++)
  for(j=0; j<7; j++)
    repc3[i][j]=0;

  deaths = events = immid = ukbid = repid = stid = 0;
  t = pt = 0;
}
```

## 3. Main program

Due to an MPI bug not allowing **popen** and related routines to work, the tuberculosis program is defined as a function which returns an array of output (rather than stand-alone executable) for use with the fitting routine. A **define** statement is used to control whether tuberculosis program is a stand-alone executable or function within the fitting routine.

```
#define main dec *mainiac
```
Make main not the real main, for use with fitting routine. Comment out for independent executable.

```
#ifdef  main
static int fit5i = 1;
#else
static int fit5i = 0;
#endif
static int fitm  = 2;
```
Flag set when linked with fitting.

Flag set when not linked.

Flag set when fitting to rates, (0=numbers, 1=rates, 2=clust and 3=overall rates).

```
dec  out[1000]; int outi;
dec outn[1000]; int outni;
dec outc[1000]; int outci;
dec outo[1000]; int outoi;
```
Main output array (all).
Main output array (case numbers).
Main output array (clust).
Main output array (overall rates).

```
main(int argc, char *argv[])
{ int i, j, k, l, n, sid;

  startsec = time(NULL);
```
Retrieve the wall-clock time.

```
  if(fit5i==0) ErrorInit();
  MainInit();
  EventInit();
  FinalInit();
  ReportInit();
```
Trap system failures.
Start the main program.
Start the event queue.
Start the final reports.
Start the output reports.

```
  A = (struct Indiv *)
      calloc(indiv+3, sizeof(struct Indiv));
  if(A==0) Error(911.);
```
Allocate array of individuals. (Not static because of gcc bug restricting such arrays to 2GB.)

```
  if(SUPER) maximm = 900000;
  else      maximm = 500000;
```
Adjust **maximm** if not running on supercomputer.

```
  Data();                                   Read in appropriate data files
                                            and parameters, store to arrays.
  gparam(argc, argv);                       Collect parameters which have
                                            been changed on the command line.
  Param();                                  Update parameters/distributions
                                            affected by parameters changed.
  if(bcy[0]<=0.0001)                        Calculate years per birth and years
  { ypb = RT*100;                           per immigrant at t=t0 for
    printf("Births are zero!\n"); }         scheduling. If none should occur,
  else ypb = 1./bcy[0];                     make interval very large, so
  if(immig[0]<=0.0001)                      they never happen.
  { ypi = RT*100;
    printf("Immigrants are zero!\n"); }
  else ypi = 1./immig[0];
                                            Update time of last update for
  lup = t0;                                 parameters sensitive to
                                            calendar year.
  rand0 = abs(randseq);                     Start the random number sequence
  if(randseq>=0)  RandStart(rand0);         from a specified or an arbitrary
  else rand0 = RandStartArb(rand0);         place.
  EventStartTime(t0);                       Initialize the event queues.
  t = t0;                                   Set the starting time.
  stid = is0+is1;                           Set first available new strain
                                            strain ID for mutants.
  InitPop();                                Set up initial population.
  Report(argv[0]); pt = t;                  Report initial conditions.
  BirthG();                                 Start external event generators
  ImmigrateG();                             for birth and immigration.
  for(t=t0; t<t1; Dispatch())               Main loop: process events,
  { if(t-pt<tgap) continue;                 reporting results periodically.
    pt = t; Report(argv[0]); }
  Report(argv[0]);                          Get final report.
  Clust();                                  Get clustering statistics.
  Final();                                  Close processing and return to
  free(A);                                  caller.
  if(fit5i)                                 If linked with fitting routine,
    return fitm==0?outn: fitm==1?out:       return selected results.
           fitm==2?outc: outo;
  return 0;                                 Otherwise return a success code.
}
```

## 4. Dispatch next event

All events pass through this routine. It picks the earliest event in the list of pending events, sets the time to match that event, and performs the operations called for by that event. Typically that will result in other events being scheduled, to be seen in the future as they arrive at the top of the list of pending events.

**Entry:** The system is initialized with all events in the list ready for processing.
`t` contains the present time.
`t1` contains the ending time.

**Exit:** The next event has been processed and `events` incremented, if the events `time is less than t1`'.
`t` is advanced to the next event, which may be an unprocessed event at time greater than `t1`.

```
Dispatch()
{ int n; dec tw;

  tw = t;                                          Remember the previous time.
  n = EventNext(); if(t>t1) return;                Advance time to the next event.
  tstep(tw, t);                                    Record the size of the time step.
  events += 1;                                     Increment the events counter.

  switch(A[n].pending)                             Process the event.
  { case pVaccin:   Vaccination(n);  break;        [Vaccination]
    case pTransm:   Transmission(n); break;        [Transmission of an infection]
    case pRemote:   Remote(n);       break;        [Transition to latency]
    case pDisease:  Disease(n);      break;        [Progression to disease]
    case pDeath:    Death(n);        break;        [Death]
    case pMutate:   Mutate(n);       break;        [Strain type mutation]
    case pEmigrate: Emigrate(n);     break;        [Emigration]
    case pBirth:    BirthG();        break;        [Birth generator]
    case pImmig:    ImmigrateG();    break;        [Immigration generator]
    case pRep:      Rep(n);          break;        [Case report]

    default: Error2(921.2,                         [System error]
      "`A[",n,"].pending=",A[n].pending); }
}
```

## 5. Birth

This routine is dispatched when an individual is to be born. All newborns are Uninfected; exit from the Uninfected compartment is by vaccination to the Immune compartment, by infection to the Recent Infection compartment, and by emigration from the population or death.

**Entry:** `n` indexes an individual being born.
`b` contains the time of birth. Presently, this is the current time, though with some set up, it could be earlier than present (notably, `pmale` would have to be indexed differently).
`t` contains the current time.
`A[n].state` contains the present state of the record (can be any state, including 0, which is not a state but marks records not yet assigned).
`m1` contains the mortality rate for susceptible individuals, if applicable.
`em` contains the emigration rate.
`v1` contains vaccine efficacy.
`v2` contains the probability that an individual will be vaccinated.
`v3` contains the average age of vaccination.
`VTYPE` is zero if vaccinations are to match ODE conventions.
No event is scheduled for individual `n`.

**Exit:** `Birth` contains a status code.

0 The individual would die before the current time so no birth has been recorded and no event scheduled.

1 Entry `n` is initialized as a susceptible newborn and its first event is scheduled, either vaccination, emigration or death.

`A[n].state` marks a susceptible individual.

Counters in `N` are updated.

```
#define VTYPE 1                                   Vaccination type.

int Birth(int n, dec b)
{ int y, s, v, e; dec wd, we, wv;
  if(n<maximm+1) Error1(610.1, "n=",(dec)n);      Check for appropriate n, this
  if(n>indiv)    Error1(610.2, "n=",(dec)n);      routine does not allow immigrant
                                                  births or births to those with
                                                  index number greater than indiv.

  y = (int)t - (int)t0;                           Retrieve year index for arrays.
  A[n].sex = Rand()<pmale[y]? 0: 1;               Assign the newborn's sex.
  s = A[n].sex;

  wd = b+LifeDsn(s,t-b,m1[s][y]);                 Schedule a time of death and
  if(wd<t) Error(850.);                           check for errors.
  we = b+EmDsn(1,s,t-b,em[s][UK]);                Calculate time of emigration.

  A[n].tBirth    = b;                             Record the time of birth.
  A[n].tDeath    = wd;                            Record the time of death.
  A[n].tEmigrate = we;                            Record the time of emigration.
  A[n].rob       = 1;                             Set as born in UK.
  NewState(n, qU);                                Mark as Uninfected.
  A[n].tExit     = 0;                             Clear any other saved event
  A[n].tDisease  = 0;                             times or states.
  A[n].tTransm   = 0;
  A[n].tMutate   = 0;
  if(REC) A[n].inf = 1;
  if(SSAV) A[n].ssa = 0;
  A[n].strain = 0;

  v = 0; switch(VTYPE)                            Select the type of vaccination
  {                                               scheduling.
    case 0:
      wv = b+Expon(v1[UK]*v2[UK]/v3[UK]);         Generate a time for vaccination
      if(wv<wd && wv<we) v = 1;                   compatible with ODE models (for
      break;                                      testing).

    case 1:                                       Generate a vaccination sometime
      wv = b+v3[UK]+Rand();                       within the specified year if
      if(b<1993 && Rand()<(v1[UK]*v2[UK])         probabilities allow.
                && wv<wd && wv<we) v = 1;
      break;

    default: Error1(611., "",(dec)VTYPE); }       Improper vaccination type.

  if(v)                                           If vaccination occurs before
  { A[n].pending = pVaccin;                       death and emigration, schedule
    EventSchedule(n, wv);                         the vaccination.
    return 1; }

  if(we<wd)                                       Schedule emigration if that
```

```
  { A[n].pending = pEmigrate;                    is the earliest event.
    EventSchedule(n, we);
    return 1; }
  { A[n].pending = pDeath;                       Otherwise, schedule death.
    EventSchedule(n, wd);
    return 1; }
}
```

# 6. Immigration

This routine brings a new individual into the population from outside. The new individual assigned demographic and infection-related attributes according to appropriate probability distributions. They are then scheduled for their earliest event. All information stored for this individual is written over, in case their index number is being recycled from an individual leaving the study population through death or emigration.

**Entry:** `n` contains the index number of the new immigrant. The contents of the record is undefined.
`t` contains the current time.
`indiv` contains the highest index number for any individual.
`maximm` contains the highest index number for an immigrant.
`t0` contains begining time of the simulation.
`SSAV` contains model version, 0=non-SSA, 1=SSA
`ssaim` contains proportion of non-UK born immigrants from SSA by year.
`immsex[r]` contains the proportion of immigrants who are male, r=0, non-UK born; r=1, UK-born; r=2, SSA-born. Note in SSA and non-SSA versions of the model, `non-UK born` will be defined differently.
`hivp` contains the HIV prevalence, by sex and year, for SSA immigrants.
`infimm` contains cumulative probabilities immigrants enter disease states, by age, rob and calendar year.
`Ax` contains state variables which accompany `infimm`.
`M1` contains the mortality table for non-diseased (real runs only).
`A1` contains state variables which accompany `M1`.
`m1` contains the mortality rate (ODE validation only).
`em` contains the emigration rate.
`v1` contains vaccine efficacy.
`v2` contains the probability that an individual will be vaccinated.
`v3` contains the average age of vaccination.
No event is scheduled for individual `n`.

**Exit:** An event is scheduled for individual `n`.
`A[n].state` contains the disease state.
`A[n].tEntry` contains the time of entry to the new state.
`A[n].tBirth` contains the time of birth.
`A[n].tImm` contains the time of immigration.
`A[n].tDeath` contains the time of death.
`A[n].tEmigrate` contains the time of emigration.
`A[n].sex` contains the sex.

```
int Immigrate(int n)
{ int y,s,rob,rob2,ac,a,st; dec r,age,wd,we,wv,tinf;

  if(n>indiv) Error1(610.3, "n=",(dec)n);      Check for appropriate n.
  if(n<1)     Error1(610.4, "n=",(dec)n);
```

## Set up basic, uninfected immigrant

```
NewState(n,qU);                                  Assign to Uninfected state
                                                 to start with.
y = (int)t - (int)t0;                            Get array index for year.

A[n].tExit     = 0;                              Clear saved event
A[n].tDisease  = 0;                              times or states.
A[n].tTransm   = 0;
A[n].tMutate   = 0;
if(REC) A[n].inf = 1;
if(SSAV) A[n].ssa = 0;
A[n].strain = 0;

if(n<=maximm) A[n].rob = rob = 0;                Assign rob=0 to all non-UK born.
else          A[n].rob = rob = 1;                Assign rob=1 to UK-born.

s = 0;                                           Set sex as male to begin.

if(rob==0 && SSAV==1)                            If non-UK born and running
{ A[n].ssa = 0;                                  SSA version of model, check
  if(Rand()<ssaim[y])                            to see if SSA. If so, get their
  { A[n].ssa = 1;                                sex and HIV status.
    if(Rand()>immsex[y][SSA]) s = 1;             Get sex of SSA.
    if(Rand()<hivp[s][y]) A[n].ssa = 2; }        Get HIV status of SSA.
  else                                           If not SSA, leave as other non-
    if(Rand()>immsex[y][0]) s = 1; }             UK and assign sex.
else                                             Assign sex to non-UK-born in
  if(Rand()>immsex[y][rob]) s = 1;               non-SSAV model and UK-born.

A[n].sex = s;                                    Assign sex to record.

age=GetAge(n,s,rob);                             Assign age.
a = (int) age;                                   Save integer age.
rob2=rob;                                        Create rob2 (0,1 and 2)
if(SSAV && A[n].ssa) rob2=2;                     since here rob is only 0 or 1).

A[n].tBirth = t-age;                             Save time of birth based on age.

A[n].tDeath = wd                                 Assign time of death and check
            = t+LifeDsn(s,age,m1[s][y]);         death time is ok.
if(wd<A[n].tBirth+age) Error(612.1);

A[n].tEmigrate                                   Assign emigration time.
     = we
     = t+EmDsn(rob2,s,age,em[s][rob2]);
if(age<v3[rob] && Rand()<v1[rob]*v2[rob]         Determine if vaccination should
               && t<2005-(v3[rob]-age))          occur and assign vaccination
  wv = t+(v3[rob]-age)+Rand();                   time if so. If it should not
else wv = t+2*RT+Rand();                         occur, set to time which will
                                                 not happen in the model.
if(wv<wd && wv<we)                               Schedule vaccination if it is
{ A[n].pending = pVaccin;                        the earliest event.
  EventSchedule(n, wv); }

else if(wd<we)                                   Schedule death if it is the
{ A[n].tExit = wd;                               earliest event.
  A[n].pending = pDeath;
  EventSchedule(n, wd); }
```

```
  else                                          Otherwise schedule emigration if
  { A[n].tExit = we;                            it is the earliest event.
    A[n].pending = pEmigrate;
    EventSchedule(n, we); }

  A[n].tDisease  = 0;                           Clear time to disease.
  A[n].tTransm   = 0;                           Clear time to transmit.
  A[n].tMutate   = 0;                           Clear time of strain mutation.
```

### Assign disease state to immigrant and process accordingly

```
  st = 1+
    (int)RandF(Ax,infimm[a][rob2][y],9,1.0);    Get random disease state.
  if(st==1)                                     Do nothing if Uninfected.
    return 0;

  else if(st==2)                                Process Immune.
  { EventCancel(n);
    Vaccination(n);
    return 1; }

  else if(st==3)                                Process Recently Infected.
  { tinf = Rand()*5;                            Set random time since infection
    Infect(n,tinf,Strain(0),0);                 within the last five years.
    if(REC) A[n].inf = 3;                       Assign inf (time/place).
    return 2; }

  else if(st==4)                                Process Remotely infected.
  { EventCancel(n);
    A[n].strain = Strain(0);                    Assign infection strain.
    NewState(n,qD1);                            Put in temporary disease state
    Remote(n);                                  to facilitate re-scheduling of
                                                events in the Remote function.
    if(REC) A[n].inf = 4;                       Assign inf (time/place).
    return 3; }

  else if(st==5)                                Process Reinfected.
  { NewState(n,qI2);                            Put in Remote infection so
    tinf = Rand()*5;                            that Infect picks this up
    Infect(n,tinf,Strain(0),0);                 as reinfection, also draw
                                                random infection time.
    if(REC) A[n].inf = 3;                       Assign inf (time/place).
    return 4; }

  else if(st>5 && st<9)                         Process Primary, Reactivation,
  { EventCancel(n);                             Reinfection disease classes.
    A[n].strain = Strain(0);                    Assign infection strain.
    NewState(n,st-3);                           First put in correct infection
    Disease(n);                                 class and then send to Disease.
    if(REC)                                     Assign inf (time/place)
    { if(st==7) A[n].inf = 4;
      else      A[n].inf = 3; }
    return 5; }

  else
    Error(618.1);
    return 0;                                   (Will never reach this).
}
```

# 7. Vaccination

This routine is dispatched when an individual is scheduled for an effective vaccination. Ineffective vaccinations are already accounted for—they are never scheduled for this routine. Effective vaccinations are assumed to impart lifelong immunity; therefore individuals will never leave this state, except by dying or emigrating from the population.

**Entry:** `n` indexes an individual being born.
`t` contains the current time.
`A[n].state` contains the present state (always `qU`).
`A[n].sex` contains the individual's sex.
`A[n].tEmigrate` contains time of emigration.
`A[n].tDeath` contains time of death.
No event is scheduled for individual `n`.

**Exit:** `A[n].state` contains the new state, `qV`.
Counters in `N` are updated.

```
int Vaccination(int n)
{
  NewState(n, qV);                                  Change states.

  if(A[n].tEmigrate<A[n].tDeath)                    Schedule emigration if that is
  { A[n].pending = pEmigrate;                       the earliest event.
    EventSchedule(n, A[n].tEmigrate); }
  else                                              Otherwise, schedule death.
  { A[n].pending = pDeath;
    EventSchedule(n, A[n].tDeath); }

  return 0;
}
```

# 8. Infect a specified individual

Individuals receive infections from others via this routine. If the targeted individual is in the Uninfected or Remote Infection state, it acquires the infection and moves to the Recent Infection or Reinfection state. The individual is then scheduled to return to Remote Infection, develop disease, emigrate, die or have its strain type mutate, depending on probabilities of each and random chance. If the individual targeted for infection is in one of the other nine disease states, they are not susceptible to infection and the transmission event has no effect.

**Entry:** `n` indexes the individual to be infected.
`tinf` contains the time of infection, 5 or fewer years ago.
`strain` contains the strain ID number of infecting strain.
`type` differentiates infections at initialization or migration,
`type=0` or transmission during the simulation, `type=1`.
`t` contains the current time.
`A[n].state` contains the state of infection target.
`A[n].tEmigrate` contains the time of emigration of infection target.
`A[n].tDeath` contains the time of death for infection target.
`r1` and `r2` contain the latency rates for `qI1` and `qI3`.
`mi` contains the mutation rate for infection strains.
An event is still scheduled for individual `n`.

**Exit:** `Infect` describes the result.

0 The specified individual could not be infected or reinfected.
1 Return to remote infection is scheduled.
2 Disease is scheduled.
3 Death is scheduled.
4 Strain mutation is scheduled.
5 Emigration is scheduled.

`A[n].state` contains the new state, if applicable.
Entry `n` is infected (if `Infect` is nonzero).
Counters in `N` are updated.

```
int Infect(int n, dec tinf, int strain, int type)
{ int a,s,rob,y,q; dec d,r,wd,we,wdis,wr,wm;
  if(n>indiv||n<1)   Error1(610.3,"",n);      Check for appropriate n.
  if(strain>=stid)   Error1(616.0,"",strain);Check for appropriate strain ID.
  if(tinf>5||tinf<0) Error1(617.0,"",tinf);  Check for appropriate tinf.
  if(tinf==5) tinf=tinf-E;                    Correct tinf if equal to 5.

  a   = (int)(t-A[n].tBirth);                 Retrieve integer age.
  s   = A[n].sex;                             Retrieve sex.
  rob = A[n].rob;                             Retrieve region of birth.
  y   = (int)t - (int)t0;                     Get array index for year.

  switch(A[n].state)                          Determine the new state and
  {                                           its associated parameters.
    case qI2: r=r2[s]; q=qI3; break;
    case qU:  r=r1[s]; q=qI1; break;
    default:  return 0; }                     Avoid uninfectable states.
  EventCancel(n);                             Else cancel the pending event and
  NewState(n, q);                             mark this individual as infected.
  if(REC) A[n].inf  = 1;                      Save place of infection as UK
                                              (will be changed outside routine
                                              if infection is acquired abroad).
  if(type==1)                                 Increment ARI counter if infection
   ari[rob][y]++;                             occurred in the UK.

  A[n].strain = strain;                       Assign infecting strain type.

  wd = A[n].tDeath;                           Retrieve time of death.
  we = A[n].tEmigrate;                        Retrieve time of emigration.
  wr = t+LAT-tinf;                            After LAT years individual is
                                              defined as remotely infected.
  wdis = t+Tdis(n,a,s,rob,tinf)+E;            Calculate time to disease.
  if(wdis<=t) Error2(620.0,"t=",t, " wdis=",wdis);
  wm   = t+Expon(mi);                         Calculate strain mutation time.

  if(wd<we && wd<wr && wd<wdis && wd<wm)       If death is earliest event,
  { A[n].pending = pDeath;                    schedule the death and
    EventSchedule(n, wd);                     ignore everything else.
    return 3; }

  if(we<wr && we<wdis && we<wm)               If emigration is the earliest
  { A[n].pending = pEmigrate;                 event, schedule it and
    EventSchedule(n, we);                     ignore everything else.
    return 5; }

  if(wr<wdis && wr<wm)                        Otherwise, if transition to
  { A[n].pending = pRemote;                   remote infection is the
```

```
    EventSchedule(n, wr);
    A[n].tMutate = wm;
    return 1; }
```
schedule that, save mutation time, and ignore disease time.

```
  if(wm<wdis)
  { A[n].pending = pMutate;
    EventSchedule(n, wm);
    A[n].tDisease = wdis;
    A[n].tExit = wr;
    return 4; }
```
Otherwise, if mutation is the earliest event, schedule that and save time to disease and time to remote.

```
  { A[n].pending = pDisease;
    EventSchedule(n, wdis);
    return 2; }
}
```
Otherwise, schedule disease and ignore other times, as they will be recalculated at disease onset.

## 9. Enter compartment remote

This routine is dispatched when an infection becomes latent, entering the Latent Infection state (qI2). Recent Infection (qI1), Reinfection (qI3), and all disease compartments (qD1-qD6) can lead to this state. The Latent Infection state allows the possibility of progression to disease, mutation of strain type, death and emigration. Also, those with Latent Infection can be reinfected, although that is induced by a transmission event dispatched independently and not handled here.

**Entry:** `n` indexes the individual.
`t` contains the current time.
`A[n].state` contains the present state (can be `qI1`, `qI3`, `qD1-qD6`).
`A[n].tMutate` contains the strain mutation time.
`d2` contains the disease progression rate for `qI2`.
`m4` contains the mortality rate for `qI2`.
`mi` contains the mutation rate of strain types in those infected, but not diseased.
No event is scheduled for individual `n`.

**Exit:** `Remote` contains a status code:
2 Disease is scheduled.
3 Death is scheduled.
4 Mutation is scheduled.
5 Emigration is scheduled.
`A[n].state` represents Remote Infection (qI2).
`A[n].tDeath` is updated as necessary.
`A[n].tMutate` is updated as necessary.
Counters in `N` are updated.

```
int Remote(int n)
{ int y, a, s, rob, q; dec age, wdis, wd, we, wm;
```

```
  y   = (int)t - (int)t0;
```
Retrieve array index for year.

```
  age = t-A[n].tBirth;
```
Retrieve age.

```
  a   = (int) age;
```
Integer age.

```
  s   = A[n].sex;
```
Retrieve sex.

```
  rob = A[n].rob;
```
Retrieve region of birth.

```
  q = A[n].state;
```
Remember the previous state.

```
  NewState(n, qI2);                          Mark the individual as remote.
  if(REC)                                    Adjust inf to older infection
    if(A[n].inf==1 || A[n].inf==3)           if not already older.
      A[n].inf+=1;

  if(q>=qD1)                                 Establish a new time for strain
     A[n].tMutate = t+Expon(mi);             mutation if the prior state
                                             was disease.

  wdis = t+Tdis(n,a,s,rob,0);                Calculate time to disease.
  wd = A[n].tDeath;                          Retrieve time of death.
  we = A[n].tEmigrate;                       Retrieve time of emigration.
  wm = A[n].tMutate;                         Retrieve time of mutation.
  if(wd<wdis && wd<wm && wd<we)              If death is the earliest event,
  { A[n].pending = pDeath;                   schedule it.
    EventSchedule(n, wd);
    return 3; }

  if(wm<wdis && wm<we)                       Otherwise, if strain mutation
  { A[n].pending = pMutate;                  is the earliest event,
    EventSchedule(n, wm);                    schedule strain it and save
    A[n].tDisease = wdis;                    time to disease onset.
    return 4; }

  if(we<wdis)                                Otherwise, if emigration is
  { A[n].pending = pEmigrate;                the earliest event, schedule
    EventSchedule(n,we);                     it and ignore other times.
    return 5; }

  { A[n].pending = pDisease;                 Otherwise, schedule progression
    EventSchedule(n, wdis);                  to disease and ignore
    return 2; }                              everything else.
}
```

## 10. Disease

This routine is dispatched when an infection progresses to active disease. There are six distinct disease compartments, pulmonary and non-pulmonary compartments for Primary (qD1/qD4), Reactivation (qD2/qD5), and Reinfection (qD3/qD6). Individuals enter disease compartments from three infection compartments—Recent Infection, Latent Infection, and Reinfection—which determine the disease compartment they enter. This routine handles all transitions to disease.

Future events for diseased individuals include transmission of infection to others, recovery to Latent Infection, death, emigration, reporting of their disease case, or strain type mutation.

**Entry:** n indexes the individual progressing to disease.
t contains the current time.
A[n].state contains the state progressing to disease (can be qI1, qI2, or qI3).
A[n].tBirth contains the time of birth of the individual.
A[n].sex contains the sex of the individual.
A[n].rob contains region of birth of the individual.
A[n].tEmigrate contains the time of emigration.
A[n].tDeath contains the scheduled time of death.

`cft` contains the case fatality rate (actually a proportion).
`r3`, `r4`, `r5`,`r6`,`r7`, and `r8` contain recovery rates for `qD1`, `qD2`, `qD3`, `qD4`, `qD5`, and `qD6`, respectively.
`m6`, `m7`, `m8`, `m9`,`m10`, and `m11` contain mortality rates for the states named above in the ODE-compatible version of the model.
`p1`, `p2`, and `p3` contain the fraction of disease which becomes pulmonary, from sources `I1`, `I2`, and `I3`, respectively.
`c` contains the average number of new infections produced by this individual per year.
`proprep` contains the proportion of cases reported.
`md` contains the mutation rate for strains involved in disease.
No event is scheduled for individual `n`.

**Exit:** `Disease` contains a status code.
1 A transmission is scheduled.
2 Recovery is scheduled.
3 Death is scheduled.
4 Strain mutation is scheduled.
5 Emigration is scheduled.
6 Case report is scheduled.
`A[n].state` contains the new state.
`A[n].tDeath` contains a possibly updated time of death.
`A[n].tMutate` contains the new time of strain mutation, if applicable.
`A[n].tTransm` contains the next time of transmission, if applicable.
`A[n].tExit` contains the time of recovery to remote infection, or if would happen after death, the time of death.
`A[n].tRep` contains the time of disease case report, if applicable.
Counters in `N` are updated.

```
int Disease(int n)
{ int a, s, rob, y, ds, q; dec age, r, m, p, wm, we, wd, wr, wt, e, wrep;
  age = t-A[n].tBirth;                          Retrieve age.
  a   = (int)age;                               Calculate integer age.
  s   = A[n].sex;                               Retrieve sex.
  rob = A[n].rob;                               Retrieve region of birth, 0 or 1.
  y   = (int)t - (int)t0;                       Retrieve array index for year.

  switch(A[n].state)                            Determine the new state and its
  {                                             associated parameters.
    case qI1: r=r3[s]; m=m6[s][y]; p=p1[a][s][rob]; q=qD1; break;
    case qI2: r=r4[s]; m=m7[s][y]; p=p2[a][s][rob]; q=qD2; break;
    case qI3: r=r5[s]; m=m8[s][y]; p=p3[a][s][rob]; q=qD3; break;
    default:  Error(922.0); }

  if(Rand()>p)                                  If this should be non-pulmonary
  { switch(A[n].state)                          disease, update new state and
    {                                           associated parameters.
      case qI1: r=r6[s]; m= m9[s][y]; q=qD4; break;
      case qI2: r=r7[s]; m=m10[s][y]; q=qD5; break;
      case qI3: r=r8[s]; m=m11[s][y]; q=qD6; break;
      default:  Error(922.0); } }

  NewState(n, q);                               Mark the individual as diseased.

  wr = A[n].tExit = t+RecovDsn(s,age,r);        Establish time to remote.
  we = A[n].tEmigrate;                          Retrieve emigration time.
  wd = A[n].tDeath;                             Retrieve time of death.
```

```
A[n].tMutate = wm = t+Expon(md);
```
Establish new mutation time.

```
if(q>=qD4) ds = 0;
else       ds = 1;
```
Set disease site to non-pulmonary or pulmonary.

```
if(Rand()<cft[a][ds][y])
{ if(wr<wd && wr<we) e = wr;
  else if(wd<we)     e = wd;
  else               e = we;


  wd = t+0.99*(e-t);


  A[n].tDeath = wd; }
```
If person should die from disease, find earliest of death, emigration and recovery to assign death before these occur, assigning death time close to the end of disease duration.

Since disease death is before natural death time, replace it.

```
if(Rand()<proprep)
{ if(wr<wd && wr<we) e = wr;

  else if(wd<we)     e = wd;

  else               e = we;

  A[n].tRep = t+Rand()*(e-t); }
```
If case should be reported, find reporting time.
First find earliest of death, emigration and disease recovery to get range for reporting time.

Randomly assign reporting time.

```
else
  A[n].tRep = t+2*RT+Rand();
```
If case should not be reported, assign a reporting time beyond running time of model.

```
if(A[n].tRep==0) Error1(619., "n=",n);
wrep = A[n].tRep;
if(wd<wr)
  wr = wd;
```
Save reporting time.
If death would occur before recovery, give death precedence.

```
if(q<qD4 && Rand()<smear[a])
  wt = t+Expon(c[s][rob]);
else wt = t+2*RT + Rand();
A[n].tTransm = wt;
```
If this is pulmonary disease and smear positive, store time to transmit. Otherwise, set time to transmit which never happens.

```
if(wt<wr && wt<wm && wt<we && wt<wrep)
{ A[n].pending = pTransm;
  EventSchedule(n, wt);
  return 1; }
```
If transmission is earliest event, schedule it.

```
if(wrep<wr && wrep<wm && wrep<we)
{ A[n].pending = pRep;
  EventSchedule(n, wrep);
  return 6; }
```
If case report is earliest event, schedule it.

```
if(wr<wd && wr<wm && wr<we)
{ A[n].pending = pRemote;
  EventSchedule(n, wr);
  return 2; }
```
If recovery is the earliest event, schedule it.

```
if(wm<wd && wm<we)
{ A[n].pending = pMutate;
  EventSchedule(n, wm);
  return 4; }
```
If mutation is the earliest event, schedule it.

```
if(we<wd)
{ A[n].pending = pEmigrate;
```
If emigration is the earliest event, schedule it.

```
    EventSchedule(n,we);
    return 5; }
  { A[n].pending = pDeath;                    Otherwise schedule death.
    EventSchedule(n, wd);
    return 3; }
}
```

# 11. Transmission

Infectious individuals transmit infection via this routine. Another individual is selected to be infected, either randomly from within the same region of birth as the infectious individual (UK-born or non-UK born, even in SSA version of model), or randomly from the entire population. If the target individual is susceptible (Uninfected or Remote Infection states), infection is established and that individual is processed accordingly. If not, no transmission occurs.

After the transmission attempt, the infectious individual is then scheduled for another transmission, strain type mutation, recovery, case report, emigration or death.

**Entry:** `n` indexes the individual to transmit an infection.
`t` contains the current time.
`A[n].tExit` contains the time of recovery, or if that time is equal to or greater than the time of death, contains time of death.
`A[n].tDeath` contains the time of death.
`A[n].tMutate` contains the strain type mutation time.
`A[n].tEmigrate` contains the emigration time.
`A[n].rob` contains the region of birth, UK or non-UK.
`A[n].sex` contains the sex.
`A[n].strain` contains the infection strain ID number.
`pcc` contains the proportion of close contacts.
`maximm` contains the maximum ID number for immigrants.
Non-UK born individuals are indexed from 1 to (`immid-1`), a total of (`immid-1`) individuals.
UK-born individuals are indexed from (`maximm+1`) to (`ukbid-1`), a total of (`ukbid-maximm-1`) individuals.
No event is scheduled for individual `n`.

**Exit:** A new individual has been targeted for infection. If the infection takes hold, that individual is scheduled for strain mutation, disease, remote infection, emigration or death.
`Transmission` contains a status code.
1 Another infection is scheduled.
2 Recovery to remote infection is scheduled.
3 Death is scheduled.
4 Strain mutation is scheduled.
5 Emigration is scheduled.
6 Case report is scheduled.
Counters in `N` are updated.

```
#define SCHED(X,Y,Z)  { A[n].pending = X; EventSchedule(n,Y); return Z; }

int Transmission(int n)
{ int i, j, y, low, tot; dec age;
  static int v[] = { iTransm,iDeath,iEmigrate,iExit,iMutate,iRep, -1 };
```

```
  y = (int)t - (int)t0;                          Get year for array index.

  if(Rand()<pcc)                                 If targetting a "close contact",
  { if(A[n].rob)                                 obtain total number and lowest
    { low = maximm+1;                            ID number from individual's own
      tot = (ukbid-1) - low + 1; }               region of birth, for selection
    else                                         of random target for infection
    { low = 1;                                   within that region of birth.
      tot = immid-1 - low + 1; }

    do i=low+(int)(Rand()*tot);                  Find person other than self to
    while(i==n); }                               infect.

  else                                           If not a "close contact", choose
  { do                                           random person to infect from
    { tot = (immid-1) + (ukbid-maximm-1);        entire population, increment
      j = 1 + (int)(Rand()*tot);                 Adjust ID numbers for UK-born.
      if(j>=immid)
        i = j+(maximm+1-immid);
      else
        i = j;  }
    while (i==n); }                              Avoid infecting self.
  Infect(i,0,A[n].strain,1);                     Infect chosen individual.

  A[n].tTransm=t+Expon(c[A[n].sex][A[n].rob]);Establish time to transmit again.

  switch(i=Earliest(A[n].t, v))                  Schedule the earliest event.
  { case iRep:       SCHED(pRep,       A[n].tRep,       6);
    case iTransm:    SCHED(pTransm,    A[n].tTransm,    1);
    case iExit:      SCHED(pRemote,    A[n].tExit,      2);
    case iMutate:    SCHED(pMutate,    A[n].tMutate,    4);
    case iEmigrate: SCHED(pEmigrate, A[n].tEmigrate, 5);
    case iDeath:     SCHED(pDeath,     A[n].tDeath,     3);
    default:         Error1(922., "m=",(dec)i);          }

  return 0;                                      (Will never reach this.)
}
```

## 12. Mutation

This routine is dispatched when the strain type of an infected or diseased individual is scheduled for mutation. The mutation does not affect any other event. After mutation is processed, the individual is scheduled for their next event, which will depend on disease state.

**Entry:** `n` indexes an individual whose strain type is to mutate
`t` contains the current time.
`mi` contains the mutation rate for strains not in an active disease case (merely infection).
`md` contains the mutation rate for strains in a disease case.
`A[n].state` contains the current state (can be any infection or disease state).
`A[n].strain` contains the strain identification number of the current strain of infection or disease.
`A[n].tDeath` contains the saved time of death.
`A[n].tEmigrate` contains the saved time of emigration.

A[n].tExit contains the saved time to exit state.
A[n].tTransm contains the saved time to transmit again.
No event is scheduled for individual n.

**Exit:** The next event for individual n is scheduled.
A[n].tMutate contains the time of next scheduled strain mutation.
Mutation contains a status code.
1 Recovery to remote infection is scheduled.
2 Progression to disease is scheduled.
3 Death is scheduled.
4 Strain mutation is scheduled.
5 Emigration is scheduled.
6 Case report is scheduled.
Counters in N are updated.

```
Mutate(int n)
{ dec m, wm, wd, we, wdis, wr, wt, wrep;
  A[n].strain = stid++;                    Assign new, mutant strain type
                                           and update next available ID.

  if(A[n].state<=qI3) m = mi;              Determine appropriate mutation
  else m = md;                             rate and calculate time to
  wm = t+Expon(m);                         next mutation.

  wd   = A[n].tDeath;                      Get time of death.
  we   = A[n].tEmigrate;                   Get time of emigration.
  wdis = A[n].tDisease;                    Get time of disease.
  wr   = A[n].tExit;                       Get time to remote infection.

  if(A[n].state==qI2)                      Schedule events for remotely
  {                                        infected individuals (qI2).
    if(wd<we && wd<wdis && wd<wm)
    { A[n].pending = pDeath;               If death occurs first,
      EventSchedule(n, wd);                schedule it.
      return 3; }

    if(wm<we && wm<wdis)                   Otherwise, if strain mutation
    { A[n].pending = pMutate;              occurs first, schedule it.
      EventSchedule(n, wm);
      return 4; }

    if(wdis<we)                            Otherwise, if disease occurs
    { A[n].pending = pDisease;             first, schedule it.
      EventSchedule(n, wdis);
      return 2; }

    { A[n].pending = pEmigrate;            Otherwise, schedule emigration.
      EventSchedule(n, we);
      return 5; } }

  if(A[n].state<=qI3)                      Schedule events for the other
  {                                        infected classes (qI1, qI3).
    if(wd<wdis && wd<wr && wd<wm && wd<we)
    { A[n].pending = pDeath;               If death is earliest event,
      EventSchedule(n, wd);                schedule it.
      return 3; }

    if(wr<wdis && wr<wm && wr<we)          Otherwise, if transition to
    { A[n].pending = pRemote;              remote infection occurs
```

```
      EventSchedule(n, wr);                    first, schedule it and
      A[n].tMutate = wm;                       save mutation time.
      return 1; }

    if(wm<wdis && wm<we)                       Otherwise, if mutation is the
    { A[n].pending = pMutate;                  earliest event, schedule it.
      EventSchedule(n, wm);
      return 4; }

    if(wdis<we)
    { A[n].pending = pDisease;                 Otherwise, if disease onset
      EventSchedule(n, wdis);                  is earliest, schedule it.
      return 2; }

    { A[n].pending = pEmigrate;                Otherwise, schedule emigration.
      EventSchedule(n, we);
      return 5; } }

  {                                            Schedule events for diseased.
    wrep = A[n].tRep;                          Get time of case report.
    if(A[n].state<qD4)                         If this is pulmonary disease,
    { wt = A[n].tTransm;                       retrieve time for transmission
      if(wt<wd && wt<wr && wt<wm && wt<we && wt<wrep)
      { A[n].pending = pTransm;                and if it is the earliest
        EventSchedule(n, wt);                  event, schedule it and
        A[n].tMutate = wm;                     save mutation time.
        return 1; } }

    if(wrep<wd && wrep<wr && wrep<wm && wrep<we)
    { A[n].pending = pRep;                     If case report is the earliest
      EventSchedule(n, wrep);                  event, schedule it and save
      A[n].tMutate = wm;                       mutation time.
      return 6; }

    if(wr<wd && wr<wm && wr<we)                If recovery is the earliest
    { A[n].pending = pRemote;                  event, schedule it.
    EventSchedule(n, wr);
    return 2; }

    if(wm<wd && wm<we)                         If mutation is the earliest
    { A[n].pending = pMutate;                  event, schedule it.
    EventSchedule(n, wm);
    return 4; }

    if(wd<we)
    { A[n].pending = pDeath;                   If death is the earliest
    EventSchedule(n, wd);                      event, schedule it.
    return 3; }

    { A[n].pending = pEmigrate;                Otherwise, schedule emigration.
      EventSchedule(n, we);
      return 5; } }
}
```

# 13. Death

This routine is dispatched when individuals die. The routine logs dead individuals out of compartments as they leave the study population, to keep track of numbers of individuals in each compartment so that the array of individuals does not have to be scanned for that information. Also, the individuals index number is recycled with Transferso that array A' is always continuous.

**Entry:** n indexes an individual who has just died.
t contains the current time.
A[n].state contains the present state (can be any compartment).
A[n].tBirth contains the time of birth.
ukbid contains the next available index number for UK-born individuals.
No event is scheduled for individual n.

**Exit:** Either entry n is sent to the function Birth, to be initialized as a susceptible newborn and function returns 0 (DTYPE==0) or index number is recycled with Transfer (DTYPE==1), no birth is generated and function returns 1.
N[A[n].state] is decremented.
Counters age1, age2, and agec are updated.
Counters in N are updated.
deaths is incremented.

```
#define DTYPE 1
```

Allows for non-constant population size.

```
Death(int n)
{ int n2; dec age;

  deaths += 1;
  N[A[n].state]-=1;
  age = t-A[n].tBirth;
```

Increment the number of deaths.
Decrement the number in the state.
Compute the age at death.

```
  { age1[0] += age; age2[0] += age*age;
    agec[0] += 1; }
```

Accumulate statistics for mean age and its variance.

```
  if(DTYPE==0)
  { Birth(n, t);
    return 0; }
```

If population size is to be held constant, initiate a birth.

```
  if(A[n].rob)
  { n2 = ukbid-1; ukbid--; }

  else
  { n2 = immid-1; immid--; }
```

Avoid unoccupied index numbers in array A by transferring highest-numbered individual, n2, to index number n.

```
  Transfer(n, n2);
  return 1;
}
```

# 14. Emigration

This routine is dispatched when individuals migrate out of the study population. This routine logs the individual out of its compartment as they leave the population, maintaining numbers in each compartment so that the array of individuals never has to be scanned

for that information. Also, the individual's index number is recycled with `Transfer` so that array `A` is always continuous.

**Entry:** `n` indexes the individual.
`tli` contains the time of last immigration.
`A[n].state` contains the disease state.
`A[n].rob` contains the region of birth.
`immig[y]` contains the total number of immigrants each year.

**Exit:** `N[A[n].state]` is decremented.
`n` is recycled such that the highest-numbered individual within the same region of birth as `n` takes the index number `n` and array `A` remains continuous.

```
Emigrate(int n)
{ int n2;

  N[A[n].state] -= 1;                    Decrement state.

  if(A[n].rob)                           Use emigrant's region of birth
  { n2 = ukbid-1; ukbid--; }             to find highest index number of
                                         individual who will take over
  else                                   emigrant's index number, to
  { n2 = immid-1; immid--; }             prevent array from having gaps
                                         of unoccupied index numbers.
  Transfer(n, n2);
}
```

## 15. Immigration generator

This routine brings a new immigrant into the population and schedules the next immigrant's arrival. The routine can be thought of as an external immigration event generator. The routine uses a pseudo-individual (index number `IMM`) to schedule the immigrant arrivals at evenly spaced intervals.

**Entry:** `t` contains the current time.
`t0` contains the end time of model.
`pimm` contains the proportion of immigrants who are non-UK born.
`immid` contains the next available index number for non-UK born.
`ukbid` contains the next available index number for UK born.
`IMM` contains the index number for the pseudo-individual used to schedule external immigration events handled here.
`ypi` contains the years per immigration, re-calculated each year from data on immigrants per year (`immig`).

**Exit:** An immigrant is brought into the population and the next immigrant due is scheduled.

```
ImmigrateG()
{ int y, n;

  y = (int)t - (int)t0;                  Get integer year array index.

  if(Rand()<pimm[y]) { n = immid; immid++; } Determine whether immigrant will
  else               { n = ukbid; ukbid++; } be UK or non-UK born.

  Immigrate(n);                          Create immigrant.
```

```
  A[IMM].pending = pImmig;                Schedule next immigration.
  EventSchedule(IMM, t+ypi);
}
```

## 16. Birth generator

This routine initiates a birth and schedules the next birth, at regularly spaced intervals, acting as the external event generator for births. The routine uses a pseudo-individual (index number `BIRTH`) to schedule births at evenly spaced intervals.

**Entry:** `t` contains the current time.
`ukbid` contains the next available ID number for UK-born.
`A[BIRTH]` is the individual designated for scheduling births.
`ypb` contains years per birth, re-calculated each year from data on births per year (`births`).

**Exit:** A new individual is born.
The next birth is scheduled.
`ukbid` is incremented.

```
BirthG()
{
  Birth(ukbid,t); ukbid++;          Produce a birth and increment the next
                                    available index number for UK-born.
  A[BIRTH].pending = pBirth;        Schedule the next birth for ypb
  EventSchedule(BIRTH,t+ypb);       years into the future.
}
```

## 17. Change states

This routine logs individuals out of states as they leave them and into new states as they enter. It maintains counters of the numbers in each state so that the array of individuals never has to be scanned for that information.

**Entry:** `n` indexes the individual.
`q` contains the new state. This is a number greater than zero, in the range `q0` to `q1`.
`A[n].state` contains the old state, either 0 or in the range `q0` to `q1`. If 0, this record is not in use.
`A[n].bstate` includes the non-susceptible states visited thus far.

**Exit:** `A[n].state` contains the number of the new state (`q` on entry).
`A[n].bstate` incorporates the new state (`q` on entry).
`A[n].tEntry` contains the time of entry to the new state.
`N[u]` is decremented, where `u` represents the old state.
`N[v]` is incremented, where `v` represents the new state.

```
NewState(int n, int q)
{
  if(q>qU)                          Reduce the number in the old state unless
    N[A[n].state] -= 1;             the individual is entering Uninfected,
                                    which only happens at birth or immigration.
  if(N[A[n].state]<0)               Make sure the state has not become
```

```
    Error2(609.0, "q=",(dec)q,          negative.
                  " n=",(dec)n);

  A[n].state  = q;                      Change state.
  N[A[n].state] += 1;                   Increase the number in the new state.
}
```

## 18. Transfer

This routine transfer all information about an individual, including saved event times, to a new identification number. The routine then cancels the pending event for that index number and re-schedules it using the new index number. The routine is used to keep the array of individuals contiguous for each region of birth.

**Entry:** `n` is the new index number to be assigned, which has no event scheduled
`n2` is the current index number of the individual.
There is an event scheduled for `n2`.

**Exit:** `n` is the new index number of the individual.
The event scheduled for n2 is now re-scheduled under `n` and all other data from `n2` are transferred to `n`. `n2` no longer has an event scheduled and the index number is free to be used again.

```
Transfer(int n, int n2)
{
  if(n!=n2)
  { A[n] = A[n2]; EventRenumber(n, n2); }      Copy data and reschedule as n.
}
```

## 19. Add reported case

This routine keeps track of the number of reported cases by age, sex, place of birth, disease site and calendar year. After a case is reported, they are scheduled for their next event. For the version of the model with a genetic component, this routine also allows for a stain to be genotyped and reports genotyping and other data for those cases, including the strain type identifier, and the age, sex, rob, etc for the case.

**Entry:** `t` is the current time.
`t0` is model start time.
`t1` is the model end time.
`A[n].tBirth` contains the time of birth.
`A[n].sex` contains the sex of the individual.
`A[n].rob` contains the region of birth of the individual, 0=non-UK, 1=UK born.
`SSAV` is the version of the model running, 0=non-SSA, 1=SSA.
`repc` holds the numbers of reported cases.
`repc2` holds the numbers of cases by place and time of infection.
`ptyped` holds the proportion of cases with genotyping data, by rob.
`repc3` holds data on cases reported and genotyped.
`A[n].tDeath contains the individuals` time of death.
`A[n].tEmigrate` contains the individual's time of emigration.
`A[n].tExit` contains the individual's time to remote infection.
`A[n].tMutate` contains the individual's strain mutation time.

`A[n].ssa` holds country of birth in SSA version of model, 0=UK and non-UK other. 1=SSA, 2=SSA.HIV+
`A[n].tImm` contains the time of immigration to UK.
`A[n].tInfected` contains the time infection was acquired.
`A[n].inf` contains the time and place of infection 1=recent/UK, 2=older/UK, 3=recent/non-UK, 4=older/non-UK.
`A[n].strain` contains the strain ID for the case.
`A[n].state` contains the disease state of the individual.

**Exit:** `repc` contains an additional case, individual `n`.
`repc2` contains an additional case, if running `REC` version of the model.
`repc3` contains an additional case, if `n` was selected for genotyping.
`A[n].tRep` contains a time past the end of the simulation, so that individual `n` will not be reported again.
`Rep` contains a status code.
1 A transmission is scheduled.
2 Recovery is scheduled.
3 Death is scheduled.
4 Strain mutation is scheduled.
5 Emigration is scheduled.

```
Rep(int n)
{
  int s,r,y,acl,d; dec age,wt, wd, we, wr, wm, wrep;

  age = t-A[n].tBirth;                          Get age.
  if(age<15) acl=0;                             Find age class (classes which match
  else if(age<45) acl=1;                        notification rates).
  else if(age<65) acl=2;
  else acl=3;
  s = A[n].sex;                                 Get sex.
  r = A[n].rob;                                 Get region, 0=non-UK, 1=UK.
  y = (int)t - (int)t0;                         Get year for array index.
  if(A[n].state>=qD4) d=0;                      Get disease site (pulm/non-pulm)
  else d=1;                                     for arrray index.

  repc [acl][s][r][d][y] += 1;                  Increment reported cases.

  if(REC)                                       Increment reported cases for recent
  { repc2[acl][s][r][d][y][0]        += 1;      transmission stats, by inf.
    repc2[acl][s][r][d][y][A[n].inf] += 1; }

  if(Rand()<ptyped[r] && t>=2007)               If case is designated to be typed
  { repc3[repid][0] = age;                      and is within correct time window,
    repc3[repid][1] = s;                        store for genetic output.
    repc3[repid][2] = r;
    repc3[repid][3] = t;
    repc3[repid][4] = A[n].inf;
    repc3[repid][5] = A[n].strain;
    repid++;  }

  A[n].tRep = t1*2+Rand();                      Set reporting time distant enough
                                                that it cannot occur again.

  wd = A[n].tDeath;                             Get time of death.
  we = A[n].tEmigrate;                          Get time of emigration.
```

```
  wr = A[n].tExit;                          Get time to remote infection.
  wm = A[n].tMutate;                        Get strain mutation time.

  if(A[n].state<qD4)                        If this is pulmonary disease,
  { wt = A[n].tTransm;                      get time for transmission
    if(wt<wd && wt<we && wt<wr && wt<wm)     and if it occurs earliest,
    { A[n].pending = pTransm;               schedule it.
      EventSchedule(n, wt);
      return 1; } }

  if(wr<wd && wr<we && wr<wm)               If recovery occurs earliest,
  { A[n].pending = pRemote;                 schedule it.
    EventSchedule(n, wr);
    return 2; }

  if(wm<wd && wm<we)                        If mutation occurs earliest,
  { A[n].pending = pMutate;                 schedule it.
    EventSchedule(n, wm);
    return 4; }

  if(we<wd)                                 If emigration occurs earliest,
  { A[n].pending = pEmigrate;               schedule it.
    EventSchedule(n,we);
    return 5; }

  { A[n].pending = pDeath;                  Otherwise schedule death.
    EventSchedule(n, wd);
    return 3; }
}
```

## 20. Lifespan distribution

This routine assigns the number of remaining years to live for an individual, based on the present year, the individual's sex and age, and other factors. Various probability distributions may be selected. Exponentially distributed ages, with a constant chance of death in any year, are an option included for calibration of the model to results from an ordinary differential equation version of the model.

**Entry:** `sex` contains the individual's sex, 0=male, 1=female.
`age` contains the individual's present age, years and fractions thereof.
`mort` contains a mortality factor. For testing, this is the proportion of individuals who would die per year it deaths were strictly random (i.e., Poisson distributed in time).
`lifedsn` defines the lifespan distribution computation:
0 Exponential
1 Gompertz
2 Empirical life tables
`t` contains the present time.

**Exit:** `LifeDsn` contains the *remaining* life time (years until death) for the individual.

```
static int lifedsn = 2;                 Type of longevity distribution to be used.

dec LifeDsn(int sex, dec age, dec mort)
{ int yb, y, n; dec w;

  switch(lifedsn)
```

```
  {
    case 0: return Expon(mort);              Constant probability of death.
    case 2:                                  Empirical life tables.
    {  yb = (int)(t-age);                    Get year of birth.
       y  = yb-1870; if(y<0) y=0;            Get array index for birth year.
       w = RandF(A1, M1[y][sex], 122, age);
       return w; }
    default: Error(922.0); }                 Incorrect life span selection.

  return 0;                                  (Will never reach this.)
}
```

## 21. Emigration time distribution

This routine assigns the number of years until time of emigration from the study population for an individual, based on the present year, the individual's sex and age, and other factors in the condition of the individual. Exponentially distributed times, with a constant chance of emigration in each year, are included for calibration of the model to an ordinary differential equation version of the model.

**Entry:** `rob` contains the individual's region of birth, 0=non-UK, 1=UK
`sex` contains the individual's sex, 0=male, 1=female.
`age` contains the individual's present age, years.
`em` contains the emigration rate.
`emdsn` defines the emigration time distribution computation:
0 Exponential
2 Empirical migrant flow data.
`t` contains the present time.

**Exit:** `EmDsn` contains the *remaining* time in the UK for the individual in the updated version of the program. Note that many individuals will have dates of emigration beyond the running time of the model or beyond their own death date; these individuals will never emigrate.

```
static int emdsn = 0;                        Type of longevity distribution to be used.

dec EmDsn(int rob, int sex, dec age, dec em)
{ int yb, y, a, n; dec w;

  switch(emdsn)
  {
    case 0: return Expon(em);                Constant probability of
                                             emigration.
    default: Error(922.0);                   Incorrect life span selection.
  }
  return 0;                                  (Will never reach this.)
}
```

# 22. Recovery distribution

This routine assigns a time to remote infection based on the present year, the individual's sex and age, and other factors in the condition of the individual. Various probability distributions may be selected.

**Entry:** `sex` contains the individual's sex, 0=male, 1=female.
`age` contains the individual's present age, years.
`r` contains a recovery parameter describing the individual. For testing this represents the proportion that would recover per year if recovery were strictly random (i.e., Poisson distributed in time).
`t` contains the present time.

**Exit:** `RecovDsn` contains the time until recovery, in years.

```
static dec recovdsn = 0;      Type of recovery distribution to be used.
static dec rmu      = 0.0;    Centering of recovery distribution, years.
static dec rsigma   = 0.1;    Half-width of recovery distribution, years.

dec RecovDsn(int s, dec age, dec r)
{ dec w;

  switch((int)recovdsn)                              Select the type of recovery.
  { case 0: return Expon(r);                         (completely random)
    case 1: w = 0;                          break;   (completely fixed)
    case 2: w = Uniform(-rsigma,rsigma);    break;   (uniform variation)
    case 3: w = LogNormal(rmu, rsigma);     break;   (log-normal variation)
    case 4: w = Gauss(.0, rsigma);          break;   (truncated Gaussian variation)
    case 5: w = Cauchy(.0, rsigma);         break;   (truncated Cauchy variation)
    default: Error(922.); }

  return max(1e-9, w+1./r);                          Return the time for recovery,
}                                                    always after a slight delay.
```

# 23. Time to disease

This routine assigns a time to disease based on the individual's age, sex, region of birth, infection status (Recent Infection, Remote Infection, Reinfection), and in the SSAV version of the model, HIV status. Note that Recent Infection and Reinfection states are handled similarly, whereas Remote Infection is handled somewhat differently.

Notes on SSA version of model: Would like to get disease rates correct so things are comparable between the SSA and non-SSA models. I think it is best to leave them as a ratio (`ehiv`) and then when fitting model, the non-SSA model should fit higher disease risk for non-UK born than in the SSA model. In the SSA model they should be lower since a portion of non-UK born individuals will have elevated risk due to HIV.

**Entry:** `n` contains the individual's identifier.
`s` contains the individual's sex, 0=male, 1=female.
`a` contains the individual's integer age.
`rob` contains the individual's region of birth, 0=non-UK, 1=UK, 2=SSA
`tinf` contains the time since infection (years).
`d1` and `d3` contain the probability of progressing to disease for `qI1` and `qI3` over the first five years of infection.
`d2` contains the probability per year of progressing to disease for

qI2.
A[n].state contains the present state of the individual, qI1, qI2, qI3.
t contains the present time.
drr gives the relative risk of disease over the first five years of infection.
B1 contains the year of infection, for use with drr.
SSAV contains 1 if SSA version of model is to be run.

**Exit:** Tdis contains the number of years until disease development.

```
dec Tdis(int n, int a, int s, int rob, dec tinf)
{ dec d,age,w;
  if(SSAV && A[n].ssa==2)                  If running SSA version of model and
    rob=2;                                 individual is HIV+, adjust rob.

  switch(A[n].state)
  { case qI1:                              Process Recently Infection.
    { d = d1[s][rob][a]                    First calculate overall disease risk
                                           for tinf greater than 0 if applicable.

      if(Rand()>d)                         If disease should NOT occur, schedule
      { w = 2*RT+Rand();                   disease past the running time of
        return w; }                        model so that it never occurs.
      else
      { w = RandF(B1,drr,6,tinf);          If disease should occur, randomly
        return w; } }                      choose year, based on relative risk
                                           over five years.

    case qI3:                              Process Reinfection.
    { d = d3[s][rob][a]                    First calculate overall disease risk
                                           for tinf greater than 0 if applicable.

      if(Rand()>d)                         If disease should NOT occur, schedule
      { w = 2*RT+Rand();                   disease past the running time of
        return w; }                        model so that it never occurrs.
      else
      { w = RandF(B1,drr,6,tinf);          If disease should occur, get year
        return w; } }                      from relative risk over five years.

    case qI2:                              Process Remote Infection.
    { age = t-A[n].tBirth;
      w = RandF(A2,d2[s][rob],AC+2,age);
      return w;       }

    default: Error(922.);   }
  return 0;                                (Will never reach this.)
}
```

## 24. Get random age for immigrant

This routine assigns an age to an individual who is immigrating into the population. Age is randomly assigned based on probabilities of age classes from data. Probabilities of age classes are conditional on sex and region of birth.

**Entry:** n contains the individual's identifier.
s contains the individual's sex, 0=male 1=female.

`r` contains the individual's region of birth, 0=non-UK, 1=UK
`t` contains the present time.
`immage` contains cumulative probabilities of the age classes; to be compatible with future calls to `RandF`, the first cumulative probability is 0.
`SSAV` contains 1 if SSA version of model is to be run.

**Exit:** `GetAge` contains the age (in years) of the individual.

```
dec GetAge(int n, int s, int r)
{ dec rn,age; int y;

  rn = Rand();                          Get random number.

  if(SSAV)                              If running SSA version of model,
    if(A[n].ssa) r=2;                   check if SSA and adjust r (rob) if so.

  y = (int)t - (int)t0;                 Get array index for calendar year.

  if(rn<immage[y][s][r][1])             Assign a random age class within the
    return Rand()*15;                   correct age class, depending on the
  if(rn<immage[y][s][r][2])             random number draw and cumulative
    return Rand()*10+15;                probabilities of each age class specified
  if(rn<immage[y][s][r][3])             by immage.
    return Rand()*10+25;
  if(rn<immage[y][s][r][4])
    return Rand()*10+35;
  if(rn<immage[y][s][r][5])
    return Rand()*15+45;

  age=Expon(0.10)+60;                   For the age class 60+, add age of 60 plus
  if(age>=121) age=120+Rand();          draw from exponential distribution with
                                        mean of 10 years.
  return age;
}
```

## 25. Parameter changing

This function updates variables associated with parameters that change with each model run. This routine must come after any change to parameters, e.g. through the **gparam** function. Currently four disease risk parameters are varied during model fitting: `d1uk20[M]`, `d2uk20[M]`, `d3uk20[M]` and `df`. `d1uk20[M]` is the risk of developing Primary Disease for UK-born males aged 20 years above. `d2uk20[M]` is the annual risk of developing Reactivation Disease for UK-born males aged 20 years and above. `d3uk20[M]` is the risk of developing Reinfection Disease for UK-born males aged 20 years and above. `df` is the factor by which UK-born disease risks are multiplied to get non-UK born risks.

**Entry:** `drr` contains relative cumulative rates of disease progression by time since infection, up to five years.
`d1uk20[M]` contains the risk of developing Primary Disease for UK-born males aged 20 years above
`d2uk20[M]` contains the annual risk of developing Reactivation Disease for UK-born males aged 20 years and above.
`d3uk20[M]` contains the risk of developing Reinfection Disease for UK-born males aged 20 years and above.
`df` contains the factor by which UK disease risk are multiplied to get non-UK born disease risks.

`sdf1` contains the sex ratios of female:male disease risk (Primary Disease).
`sdf2` contains the sex ratios of female:male disease risk (Reactivation Disease).
`sdf3` contains the sex ratios of female:male disease risk (Reinfection Disease).
`presp` contains the proportion of disease respiratory for children.

**Exit:** `d1`contains Primary Disease risk by sex, rob (UK/nonUK/HIV) and age.
`d2`contains Reactivation Disease rate by sex, rob (UK/nonUK/HIV) and age.
`d3`contains Primary Disease risk by sex, rob (UK/nonUK/HIV) and age.

Notes: Disease progression risks are specified separately for Recent, Remote, and Reinfected individuals (`d1`, `d2`, `d3`), and also allowed to vary by sex, rob and age. Following Vynnycky and Fine, the age-dependency of risk is specified by two parameters—one for risk ages 0-10 (constant) and one for 20+ (constant). Risk between ages 10-20 is assumed to increase linearly from the rate at 10 to the rate at age 20.For those over 10 and under 20 overall disease progression rate is: `A0+(age-10)*((A20-A10)/10)`.

For `d1` and `d3`, these represent overall, cumulative risks of disease progression in the first five years of infection or reinfection for infection at a given age. The array `drr` specifies the cumulative relative risk over these five years and is used along with `d1/d2` to generate a time to disease, if applicable. For `d2` the disease progression risks are annual rates at a given age (and sex/rob) for disease progression. `d2` is converted to cumulative risk by age, and the cumulative distribution is used, along with current age of individual infected, to assign a time to disease.

Disease progression risks estimated by Vynnycky and Fine 1997 for white ethnic males are used to set UK-born males and UK-born female risk for those under ten years of age. For all ages, female risks are calculated by multiplying male risks by the risk ratios for sex, `sdf1`, `sdf2`, and `sdf3`. Risks for 20 year-olds are allowed to vary in the model (one for each disease types, so three parameters). For non-UK born risks, `df` is multiplied by the UK-born risks to get non-UK born risks. `df` is allowed to vary in the model. HIV-positive risks are further multiplied by `ehiv`. This means as the three UK-born risks and `df` are varied during model fitting, non-UK born disease progression risks would need to be re-generated each time the model is run.

One complication regards pulmonary versus overall disease risk. In Vynnycky and Fine 1997, risks are for respiratory disease. However, disease risk needed for the model is overall disease risk, pulmonary plus non-pulmonary. For those fixed in the model (UK-born under 10), the respiratory risk is corrected to equal overall (pulmonary + non-pulmonary) risk.

```
Param()
{ int a,s,r;

  dec ep = 0.00000000000001;          Check that ehiv and df are not
  if(ehiv<ep) ehiv= ep;               negative, making them ep, a small
  if(df<ep)   df  = ep;               positive number if not.

  mi=.106*md;                         Set infection mutation rate after
                                      disease mutation rate is read.

  if(DPARAM)                          For new way of varying disease risks:
  { if(d1uk10[M]<ep) d1uk10[M]=ep;    First check that all risks/rates are
    if(d1uk20[M]<ep) d1uk20[M]=ep;    positive (greater than a small positive
    if(d2uk10[M]<ep) d2uk10[M]=ep;    number).
    if(d2uk20[M]<ep) d2uk20[M]=ep;
    if(d3uk10[M]<ep) d3uk10[M]=ep;
    if(d3uk20[M]<ep) d3uk20[M]=ep;

    d1uk10[F] = d1uk10[M]*sdf1[0];    Also for new version of model, set
    d2uk10[F] = d2uk10[M]*sdf2[0];    females' values for disease progression,
```

```
  d3uk10[F] = d3uk10[M]*sdf3[0];
  d1uk20[F] = d1uk20[M]*sdf1[1];
  d2uk20[F] = d2uk20[M]*sdf2[1];
  d3uk20[F] = d3uk20[M]*sdf3[1];
```

relative to males, with sex ratios calculated from rates in Vynnycky and Fine.

```
  for(s=0; s<2; s++)
  { d1uk10[s] = d1uk10[s]/presp;
    d2uk10[s] = d2uk10[s]/presp;
    d3uk10[s] = d3uk10[s]/presp; }
```

Also for new version of model, set divide by **presp** to get overall (not respiratory) progression rates/risks.

```
  for(a=0; a<10; a++)
  for(s=0; s<2;  s++)
  { d1[s][UK][a] = d1uk10[s];
    d2[s][UK][a] = d2uk10[s];
    d3[s][UK][a] = d3uk10[s]; }
```

Now expand rates to all ages for by assuming constant risk/rate from age 0-10, linear increase from 10-20, and constant risk/rate from age 20+. Note, for **d1** and **d3** these are cumulative risks over first 5 yrs for (infected at age **a**) while for 'd2; these are annual rates of progression.

```
  for(a=10; a<20; a++)
  for(s=0;  s<2;  s++)
  { d1[s][UK][a] = d1uk10[s] + (a-10)*((d1uk20[s]-d1uk10[s])/10);
    d2[s][UK][a] = d2uk10[s] + (a-10)*((d2uk20[s]-d2uk10[s])/10);
    d3[s][UK][a] = d3uk10[s] + (a-10)*((d3uk20[s]-d3uk10[s])/10); }

  for(a=20; a<121; a++)
  for(s=0;  s<2;   s++)
  { d1[s][UK][a] = d1uk20[s];
    d2[s][UK][a] = d2uk20[s];
    d3[s][UK][a] = d3uk20[s]; } }

else
  for(a=0; a<121; a++)
  for(s=0; s<2;   s++)
  { d1[s][UK][a] = d1p[a][s][UK]/presp;
    d2[s][UK][a] = d2p[a][s][UK]/presp;
    d3[s][UK][a] = d3p[a][s][UK]/presp; }
```

For UK-born get overall disease risks (**d1** and **d3**) and annual risks (**d2**) from respiratory risks, taken from Vynn and Fine 1997 (divided by prop. of all tuberculosis which is respiratory (see **Data**). Indexed by sex, rob, and age.

```
for(a=0; a<121; a++)
for(s=0; s<2;   s++)
{ d1[s][NUK][a] = df*d1[s][UK][a];
  d2[s][NUK][a] = df*d2[s][UK][a];
  d3[s][NUK][a] = df*d3[s][UK][a]; }
```

For non-UK born: multiply UK-born risks by **df** to get non-UK born disease risks by sex, rob and age.

```
for(a=0; a<121; a++)
for(s=0; s<2;   s++)
{ if(d1[s][NUK][a]>1) d1[s][NUK][a] = 1;
  if(d2[s][NUK][a]>1) d2[s][NUK][a] = 1;
  if(d3[s][NUK][a]>1) d3[s][NUK][a] = 1;
```

Check that overall disease risks (**d1**, **d3**) are not above 1; also check annual rates (**d2**) are not above 1 for non-UK born.
}

```
if(SSAV)
{ for(a=0; a<121; a++)
  for(s=0; s<2;   s++)
  { d1[s][HIV][a] = ehiv*d1[s][NUK][a];
    d2[s][HIV][a] = ehiv*d2[s][NUK][a];
    d3[s][HIV][a] = ehiv*d3[s][NUK][a]; }
```

Get disease risks for HIV-positive SSAs by multiplying non-UK born risks/rates by factor **ehiv**.

```
  for(a=0; a<121; a++)
  for(s=0; s<2;   s++)
```

Make sure disease risks for HIV+ are not above 1.

```
    { if(d1[s][HIV][a]>1) d1[s][HIV][a]=1;
      if(d2[s][HIV][a]>1) d2[s][HIV][a]=1;
      if(d3[s][HIV][a]>1) d3[s][HIV][a]=1; } }
```

```
  for(s=0; s<2; s++)                    Fix end of array for d2 (longer).
  for(r=0; r<3; r++)
    d2[s][r][121] = d2[s][r][120];
```

```
  for(s=0; s<2; s++)                    Redefine d2 as cumulative probability of
  for(r=0; r<3; r++)                    disease progression; first translate risk
    d2[s][r][1] = d2[s][r][0];          in first year of life (d2[s][r][0]) as
                                        cumulative risk experienced before age 1
  for(a=2; a<=121; a++)                 (d2[s][r][1]). Then convert rest of array
  for(s=0; s<2; s++)                    to cumulative risks. Indexed by sex, rob
  for(r=0; r<3; r++)                    and age.
    d2[s][r][a] = d2[s][r][a-1]+(1-d2[s][r][a-1])*d2[s][r][a];
```

```
  for(s=0; s<2; s++)                    Make sure cumulative probability of disease
  for(r=0; r<3; r++)                    has not gone beyond 1.
  { if(d2[s][r][121]>1) Error(754.1); }
```

```
  for(s=0; s<2; s++)                    Finish cumulative distribution so that
  for(r=0; r<3; r++)                    cumulative risk before age 0 is 0 and
  { d2[s][r][0] = 0.0;                  those who should never progress to disease
    d2[s][r][122] = d2[s][r][121];      are assigned times far into the future so
    d2[s][r][123] = 1.0; }              that disease progression does not happen.
```

```
}
```

## 26. Initialize starting population

This function sets up the intial population by looping through matrix **n1981** which holds numbers in the initial population by age, sex, and rob. For the SSA version of model, **ssa1981** is used for the proportions of SubSaharan Africans among non-UK born in 1981 by age and sex.

Notes: In function **Param**, one could multiply 1981 (if they are proportions) by the initial population size, e.g. **initpop**. Assignments could also be made deterministic since the numbers are sufficiently large. This would need some planning for dealing with remainders and other numeric issues.

**Entry:** **n1981** contains numbers by age, sex, and rob for individuals in the population at initialization in 1981.
**ssa1981** contains the proportion of SSAs among non-UK born at population initialization.

**Exit:** The intial population is set up; each individual is assigned attributes and scheduled for exactly one event (other event times may be stored for an individual.

```
InitPop()
{ int a,s,i,n,st,rob; dec age,wd,we,wv,tinf;
  ukbid = maximm+1;                     Initialize ID numbers to
  immid = 1;                            correct values.
```

NOTE: This function could be made so that UK-born. non-UK born and SSA-born are initialized with one function, called three times. Also so that states are assigned and processed in a function.

```
  for(a=0; a<121; a++)                          First, initialize UK-born
  for(s=0; s<2;   s++)                          (rob=1) population for all age
  for(i=0; i<n1981[a][s][UK]; i++)              and sex categories.
  { n = ukbid; ukbid++;                         Take the next available ID.
    age = a+Rand();                             Assign age plus random bit.
    A[n].tBirth = t-age;                        Assign birth time from age.
    A[n].sex = s;                               Assign sex.
    A[n].rob = rob = UK;                        Set to UK-born.

    BasicInd(n,UK,age,s);                       Set up basic individual.
    DisState(n,UK,a); }                         Assign disease state and
                                                process accordingly.

  if(SSAV)                                      Process non-UK born in SSA
    for(a=0; a<121; a++)                        version of model, taking into
    for(s=0; s<2;   s++)                        account the proportion of
    for(i=0; i<n1981[a][s][NUK]; i++)           SSAs and their HIV
    { n = immid; immid++;                       status.
      age = a+Rand();
      A[n].tBirth = t-age;
      A[n].sex = s;
      A[n].rob = rob = NUK;
      if(Rand()<ssa1981[a][s])                  If SubSaharan African, indicate
      { A[n].ssa = 1;                           this and assign HIV status.
        rob=SSA;
        if(Rand()<hivp[s][0]) A[n].ssa=2; }

      BasicInd(n,NUK,age,s);                    Set up basic individual.
      DisState(n,rob,a); }                      Assign disease state and
                                                process accordingly.

  else                                          Process non-UK born for non-SSA
    for(a=0; a<121; a++)                        version of model.
    for(s=0; s<2; s++)
    for(i=0; i<n1981[a][s][NUK]; i++)
    { n=immid; immid++;
      age=a+Rand();
      A[n].tBirth = t-age;
      A[n].sex = s;
      A[n].rob = rob = NUK;

      BasicInd(n,NUK,age,s);                    Set up basic individual.

      DisState(n,NUK,a); }                      Assign disease state and
}                                               process accordingly.
```

## 27. Set up basic individual for population initialization

This function sets up a basic individual assigned to a death, emigration, or vaccination time. This is similar to birth but processes individuals of any age, sex, or region of birth (anyone being initialized when the model starts). The scheduled event may change if a

state other than Uninfected is assigned when diseaes state is assigned. This function is merely used to get the individual initialized.

**Entry:** A[n] individual is not scheduled for any event.
rob is 0 for non-UK born (ONUK and SSA), 1 for UK-born.
age is the age of the individual in years.
sex is 0 for males and 1 for females.

**Exit:** A[n] is in the Uninfected state and scheduled for its earliest event.

```
BasicInd(int n, int rob, dec age, int s)
{ dec wd, we, wv;

  NewState(n,qU);                                  Assign to Uninfected state.
  A[n].tDeath = wd = t+LifeDsn(s,age,m1[0][0]);    Assign time of death.
  if(wd<A[n].tBirth+age) Error(612.2);             Check death time.
  A[n].tEmigrate = we                              Assign time of emigration.
                 = t+EmDsn(rob,s,age,em[s][rob]);
  if(age<v3[rob] && Rand()<v1[rob]*v2[rob])        Calculate time to vaccination,
    wv = t+(v3[rob]-age)+Rand();                   set to time which never
  else wv = t+2*RT+Rand();                         happens if it should not occur.

  if(wv<wd && wv<we)                               If vaccination is the earliest
  { A[n].pending = pVaccin;                        event, schedule it.
    EventSchedule(n, wv); }

  else if(wd<we)                                   If death is the earliest
  { A[n].tExit = wd;                               event, schedule it.
    A[n].pending = pDeath;
    EventSchedule(n, wd); }
  else                                             Or, if emigration is the
  { A[n].tExit = we;                               earliest event, schedule that.
    A[n].pending = pEmigrate;
    EventSchedule(n, we); }
}
```

## 28. Assign disease state for initial member of population

This function assigns one of the eight disease states (8 instead of 11 because those with disease are not assigned to pulmonary or non-pulmonary until after being processed).

**Entry:** n is the individual's ID number.
A[n] is set up as basic Uninfected individual.
rob is the region of birth, 0=Non-UK, 1=UK, 2=SSA (if running SSAV version of model.
a is the integer age of the individual.
Ax contains the disease state, used for RandF along with inf1981.
inf1981 contains the probabilities of the different disease states, which are numbers in Ax.

**Exit:** A[n] has been assigned disease state (may remain unchanged) and processed accordingly.

```
DisState(int n, int rob, int a)
{ int st,str,r; dec tinf;
```

```
  st = 1+(int)RandF(Ax,inf1981[a][A[n].sex][rob],9,1.0); Get disease state.

  r=rob;                                          Get rob of 0/1 for assigning
  if(r==2) r=0;                                   place of infection in inf
                                                  and for assigning strain ID.

  switch(st)
  {
    case 2:                                       Send to vaccination.
      EventCancel(n);
      Vaccination(n);
      break;

    case 3:                                       Get the time of infection before
      tinf = Rand()*5;                            sending to Infect.
      Infect(n,tinf,Strain(r),0);
      if(REC)                                     Assign inf (time/place)
      { if(r) A[n].inf = 1;                       if REC version of model.
        else  A[n].inf = Rand()<.5? 1: 3; }
      break;

    case 4:                                       Before sending to Remote, put
      EventCancel(n);                             in temporary disease state to
      A[n].strain = Strain(r);                    facilitate re-scheduling of
      NewState(n,qD1);                            events in Remote function.
      Remote(n);
      if(REC)                                     Assign inf (time/place)
      { if(r) A[n].inf = 2;                       if REC version of model.
        else  A[n].inf = Rand()<.25? 2:4; }
      break;

    case 5:                                       Before sending to Infect, put
      NewState(n,qI2);                            in "Remote infection" state so
      tinf = Rand()*5;                            that this is correctly treated
      Infect(n,tinf,Strain(r),0);                 as Reinfection.
      if(REC)                                     Assign inf (time/place)
      { if(r) A[n].inf = 1;                       if REC version of model.
        else  A[n].inf = Rand()<.5? 1: 3; }
      break;

    case 6: case 7: case 8:                       Process disease states by first
      EventCancel(n);                             putting in correct infection
      A[n].strain=Strain(r);                      state and then sending to
      NewState(n,st-3);                           Disease.
      Disease(n);
      if(REC)                                     Assign inf (time/place)
      { if(r&&st==7) A[n].inf = 2;                if REC version of model.
        else if(r)   A[n].inf = 1;
        else if(st==7) A[n].inf = Rand()<.2? 2: 4;
        else           A[n].inf = Rand()<.4? 1: 3; }
      break;

    default:
      if(st!=1) Error(618.2); }
}
```

# 29. Choose strain identification number

This routine chooses a strain number at model initialization, or for immigrants at the time of immigration. Future versions of the model will expand tracking of strain type profile identifiers, which will require an expanded version of this routine.

**Entry:** `type` holds the type of strain number to be generated, where 0=Non-UK born at model initialization and migrants during the simulation, 1=UK-born at model initialization.
`S0` holds strain IDs for migrants, corresponding to probabilities held in `sdimm`.
`S1` holds strain IDs for UK-born at model initialization, which correspond to probabilities held in `sduk`.
`sdimm` holds cumulative probabilities of selection for strains in migrants, `S0`.
`sduk` holds cumulative probabilities of selection for strains in UK-born at model initialization, held in `S1`.
`is0` holds the number of non-UK born strains, for model initialization and for migrants upon entry to the UK.
`is1` holds the number of UK-born strains at model initialization.

**Exit:** A strain ID number is returned.

```
int Strain(int type)
{
  if(type==0)                                          Process Non-UK individuals, where
    return (int)RandF(S0,sdimm,is0+1,1);               strains are numbered 1 to is0.

  else                                                 Process UK-born, where strains are
    return (int)RandF(S1,sduk,is1+1,is0+1);            numbered is0+1 to is0+is1.
}
```

# 30. Cluster analysis

This function analyzes genetic typing data created during the simulation, checking whether cases have strains that match others in the population and computing cluster sizes for each strain. This routine is called once at the end of the simulation, so speed is not an issue.

**Entry:** `repc3` holds data on cases which are reported and genotyped; the first array dimension holds reporting identification numbers; the second dimension has 7 elements:
0 age
1 sex
2 birthplace
3 time of case report
4 place/time of infection (as in `A[n].inf`)
1 recent/UK
2 older/UK
3 recent/NonUK
4 older/NonUK
5 strain type of infection
6 set to zero.
`repid` holds the total number of reported cases.
`clust` is empty and set to zero, indexed as described its data

definition.

**Exit:** `repc3` is updated to include the number of individuals with the same strain in element 6 of the second array dimension.
`clust[ac][s][r][k]` counts the number of cases for each combination of
age class (0=0 14, 1=15 44, 2=45 64, 3=65),
sex (0=male, 1=female), and
rob (0=foreign, 1=UK) combination,
with the fourth array dimension counting where and when infected:
0 unique and non-recent/non-UK,
1 unique and recent/UK cases,
2 clustered and non-recent/non-UK,
3 clustered and recent/UK,
4 total of indexes 0 to 3.

```
int Clust()
{ int i,j,k,ac,s,r; dec age;

  for(i=0; i<repid; i++)                          For each case count the
  { for(j=0; j<repid; j++)                        number of other cases
      if (j!=i && repc3[i][5]==repc3[j][5])       with matching strains.
        repc3[i][6]++;

    age = repc3[i][0];                            Get age class, sex, and
    ac  = age<15?0: age<45?1: age<65?2: 3;        birthplace indexes.
    s   = repc3[i][1];
    r   = repc3[i][2];

    k = repc3[i][4]==1? 1: 0;                     Develop the array index for the
    if(repc3[i][6]>0) k += 2;                     fourth dimension, by recent/UK
                                                  and clustering.

    clust[ac][s][r][k]++;                         Increment cases by category and
    clust[ac][s][r][4]++; }                       accumulate the total.

  return 0;                                       Return to caller.
}
```

## 31. Check for monotonicity

This routine checks whether a table of cumulative probabilities is monotonically increasing and optionally whether it is bracketed by 0 and 1.

**Entry:** `p` is the table of cumulative probabilities.
`n` is the number of entries in the table.
`b` is set if the table should begin with 0 and end with 1.
`r1` and `r2` contain two numbers that may help to identify the location of the error. If such numbers will not help, then either or both contain zero.

**Exit:** The routine returns if the table appears to be correct. If not, an error message is issued and the routine never returns.

```
int monotone(dec p[], int n, int b, int r1, int r2)
{ int i;

  for(i=1; i<n; i++)                              Make sure the sequence never
    if(p[i-1]>p[i])                               decreases.
```

```
      Error3(621., "‘",r1, "‘ ",r2, "‘ ",i);
  if(b && (p[0]!=0||p[n-1]!=1))                    If requested. make sure it begins
    Error2(622., "‘",r1, "‘ ",r2);                 with 0 and ends with 1.
  return 0;                                        (Will never reach this).
}
```

## 32. Service routines

Various service routines, such as parameter reading and report writing, which are peripheral to the operation of the program, are included in file `service.c`. This becomes part of the main program but is kept separate for simplicity. The tables below define parameters that may vary each time the program is invoked, available to the parameter gathering routine `gparam`.

```
char *pntab[] =                                    Table of parameter names.
{ "s2[0]","s2[1]","c[0][0]","c[0][1]","c[1][0]", "c[1][1]",
  "v1[0]","v1[1]", "v2[0]", "v2[1]",  "v3[0]",   "v3[1]","ehiv",
  "r1[0]","r1[1]", "r2[0]", "r2[1]",  "r3[0]",   "r3[1]",
  "r4[0]","r4[1]", "r5[0]", "r5[1]",  "r6[0]",   "r6[1]",
  "r7[0]","r7[1]", "r8[0]", "r8[1]",   "df",     "runid",
  "d1uk20","d2uk20","d3uk20", "md",    "mi",
  "pmale[0]", "randseq", 0 };

dec *patab[] =                                     Table of parameter addresses.
{ &s2[0], &s2[1],  &c[0][0], &c[0][1],  &c[1][0], &c[1][1],
  &v1[0], &v1[1],  &v2[0],   &v2[1],    &v3[0],   &v3[1], &ehiv,
  &r1[0], &r1[1],  &r2[0],   &r2[1],    &r3[0],   &r3[1],
  &r4[0], &r4[1],  &r5[0],   &r5[1],    &r6[0],   &r6[1],
  &r7[0], &r7[1],  &r8[0],   &r8[1],    &df,      &runid,
  &d1uk20[0], &d2uk20[0], &d3uk20[0], &md,&mi,
  &pmale[0], &randseq, 0 };

#include "service.c"
```

Intn'l Conf. Scientific Computing, CSC12: 146–151, Proceedings, July 2012.

# Trading Space for Time: Constant-Speed Algorithms for Grouping Objects in Scientific Simulations

**Adrienne Keen**[1] **and Clarence Lehman**[2]
[1]London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK
[2]University of Minnesota, 123 Snyder Hall, 1474 Gortner Avenue, Saint Paul, MN 55108 USA

**Abstract**—*Microscale models that simulate discrete individuals commonly organize those individuals into groups of similar characteristics. During the simulation, individuals are frequently added to groups, removed from groups, and moved among groups. Moreover, individuals must be selected randomly based on probabilities that vary among the groups, and even that may vary within a group. Space and time limit the number of individuals, as the number desired can be large—up to the population of entire nations or more. Therefore algorithms for processing them should be efficient. Here we explain a set of Order-1 algorithms for managing groups. These algorithms run at a constant speed regardless of whether 100 individuals are included or 100 million. They use space–time tradeoffs recently made possible by large computer memories to bring the number of iterations per operation close to one. This increases the scale of problems that can be addressed by individual-based, agent-based, discrete-event, and other microscale simulations.*

**Keywords:** individual-based simulation, discrete event simulation, equation-free models, order-1 algorithms, memory–speed tradeoff

## 1. Introduction

One goal of the algorithms described here is to select individuals at random, from given groups, in the least possible time. Such operations can be needed billions of times during large-scale individual-based and other microscale simulations in science and industry [1] [2] [3]. For example, epidemiological models may need to select simulated individuals from given age groups, birthplaces, and susceptibility categories as targets of infection transmitted during the simulation.

Selection is easiest and fastest if the data structures for all members of a group occupy a contiguous block of memory, with no intervening gaps. Then selecting a random member is merely generating a random number between 1 and the number of members in the group, then indexing the corresponding member. If all individuals in the group have the same probability of being selected, that operation is clearly independent of the number of individuals in the group. In other words, it is of "Order-1," running at the same speed regardless of how many individuals are in a group. If the probabilities of being selected vary within the group, the algorithm remains Order-1, but with slightly more time required per selection.

During the course of the simulation, individuals will be added to groups, deleted from them, and moved between them. When an individual is deleted, for example, the gap in the group formed by that individual must be closed up. These addition–deletion operations occur frequently during the simulation, so ideally they should also be Order-1, as are the algorithms explained in this paper.

With high-speed algorithms for adding and deleting, individuals can be organized into groups even if random selection is not needed. For example, such organization can help tallying—keeping counts of those in different groups through time without scanning the array of individuals. Under the right circumstances, the algorithms can also be used within in many applications that require efficient priority queues [4].

Order-1 algorithms are rare, but they have been known since the early days of computer science [5]. They often rely on an abundance of computer memory to keep data structures sparse, and therefore were costly to apply in earlier days. Now, because allocating a gigabyte array is readily within the reach even of portable computers, new rules for memory use apply. Algorithms that trade additional memory for additional speed are possible and desirable.

Some groups may have myriad individuals and others only a few. The size of the groups may not be known in advance and may vary widely during the simulation. Therefore, it is not practical to allocate separate arrays large enough for each group, and dynamically allocating and reallocating memory would be unnecessarily inefficient. The algorithms described here eliminate the need for such reallocations, using space not needed by smaller groups to accommodate the needs of larger groups. The algorithms use buffer areas of allocated but unused memory to speed operations. Their running times become independent of the number of individuals and nearly independent of the number of groups.

## 2. Sample application

For an example of use of these algorithms, consider an epidemiological model having what is called "age dependent

mixing." Transmission of influenza, for example, is more likely between those in similar age groups, since individuals of similar ages spend much of their time in similar locales—such as day-cares, schools, business places, or assisted living facilities. These may have approximately uniform mixing within age groups, but reduced mixing between age groups [6]. An age-dependent mixing strategy can substitute for a more accurate but unknown contact network.

In an individual-based, discrete-event simulation of this type, when an infectious individual is about to infect another, the new individual must be selected efficiently from a list of perhaps tens of millions of individuals. Which group will receive the infection is determined in the program at large, outside the algorithms described here, via empirical or hypothetical probability distributions. Figure 1 is a hypothetical example of such a distribution. It represents the probability of transmission by age class from an infected four-year-old. The horizontal axis is the age of a susceptible individual, the vertical axis represents relative probability of receiving an infection from an infected four-year-old.

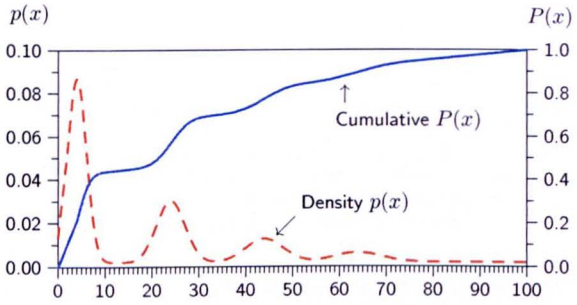


**Figure 1.** *Sample probability function for selecting groups, illustrated as the probability of transmission of infection from an infected four-year-old to each of 100 groups, here representing one-year human age classes. The probability of being selected depends on the group, according to an empirical probability distribution, determined or surmised. Left axis, probability density function, dashed curve. Right axis, corresponding cumulative probability distribution, solid curve.*

In this illustration, four-year olds can transmit infections to anyone but have the highest probability of transmitting to other four-year olds and to nearby ages, an increased probability of transmitting to those the age of their parents, peaking around 24 years old, and their grandparents, around 44 years old, and great-grandparents around 64 years old. Such distributions would be empirically estimated and groups to receive an infection would be selected many times during the simulation. This would be accomplished, for example, with a non-uniform random number generator that selects from arbitrary probability distributions [7] [8] [9], like the distribution in Figure 1. That is accomplished through the cumulative probability distribution by representing its inverse in an efficient way, as with piecewise-polynomial curves, then sampling from that inverse distribution with uniform random numbers (e.g., [10]). That process, accomplished outside these algorithms, determines which group of individuals will receive the infection. The algorithms described here are then applied to select an individual from the group. They are also applied to add individuals to groups, delete them, or move them between groups.

## 3. Algorithms and data structures

One array exists for individuals and two for groups (see appendix). $A[n]$ is a one-dimensional array of structures representing individuals, in order by group and contiguous within each group, but in no particular order within groups, and with possible gaps between groups. The relevant simulation data for each individual is carried in this array, which varies by application, but in our example would include such items as sex, birth date, birthplace, and geographic coordinates. Each individual may also carry information on its own probability of being selected within a group, in data element $A[n].v$. By convention, $A[1]$ is the first entry used, so that index 0 may be used as a null list index. This is the largest data structure, potentially containing tens or hundreds of millions of individuals and occupying gigabytes of memory.

$C[i]$ is a one-dimensional array of groups, identifying the lowest-numbered individual in $A[n]$ for each group, structured so that $C[0]$ is the index of the lowest numbered individual in the first group and that $C[j+1] - C[j] - E[j]$ is the number of individuals in group $j$. $E[i]$ is simply a one-dimensional array identifying the number of empty cells at the end of each group. Deletions increase the number of empty cells and additions draw from those empty cells. These are relatively small arrays, typically occupying only kilobytes each. They can be initialized by *GroupInit* (appendix), or by a custom routine. Global scalar variables are $ma$, the maximum number of individuals in $A$, and $nc$, the number of groups.

### 3.1 Selecting from a group

Because all individuals in a group occupy contiguous elements of array $A[n]$, selecting one at random simply involves generating a uniformly distributed random number between one and the number of individuals in the group, then selecting that individual after biasing the number to align with index numbers in $A[n]$ for the group. If all members of the group are equally likely, the process is complete. If probabilities of selection vary among members of the group, then the "sieve method" [11] applies within the group. That is implemented in Algorithm 1 (appendix). In effect, the sieve method tentatively selects a random member of the group and examines its probability of being selected, relative to others in the group. One additional uniform random number determines if the tentatively selected individual should remain unselected, based on the probability recorded

| | C[0] ↓ | | | C[1] ↓ | | | C[2] ↓ | | | | C[0] | C[1] | C[2] | E[0] | E[1] | E[2] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A[1] | A[2] | A[3] | A[4] | A[5] | A[6] | A[7] | A[8] | A[9] | A[10] | | | | | | |
| Initial | 0 | 0 | 0 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 1 | 4 | 7 | 3 | 3 | 4 |
| Add U | 0 | 0 | 0 | 1 | 1 | 1 | 2 U | 2 | 2 | 2 | 1 | 4 | 7 | 3 | 3 | 3 |
| Add O | 0 | 0 | 0 | 1 O | 1 | 1 | 2 U | 2 | 2 | 2 | 1 | 4 | 7 | 3 | 2 | 3 |
| Add P | 0 | 0 | 0 | 1 O | 1 P | 1 | 2 U | 2 | 2 | 2 | 1 | 4 | 7 | 3 | 1 | 3 |
| Add Z | 0 | 0 | 0 | 1 O | 1 P | 1 | 2 U | 2 Z | 2 | 2 | 1 | 4 | 7 | 3 | 1 | 2 |
| Add M | 0 | 0 | 0 | 1 O | 1 P | 1 M | 2 U | 2 Z | 2 | 2 | 1 | 4 | 7 | 3 | 0 | 2 |
| Add B | 0 B | 0 | 0 | 1 O | 1 P | 1 M | 2 U | 2 Z | 2 | 2 | 1 | 4 | 7 | 2 | 0 | 2 |
| Add W | 0 B | 0 | 0 | 1 O | 1 P | 1 M | 2 U | 2 Z | 2 W | 2 | 1 | 4 | 7 | 2 | 0 | 1 |
| Add S | 0 B | 0 | 1 M | 1 O | 1 P | 1 S | 2 U | 2 Z | 2 W | 2 | 1 | 3 | 7 | 1 | 0 | 1 |
| Add X | 0 B | 0 | 1 M | 1 O | 1 P | 1 S | 2 U | 2 Z | 2 W | 2 X | 1 | 3 | 7 | 1 | 0 | 0 |
| Add Y | 0 B | 1 S | 1 M | 1 O | 1 P | 2 X | 2 U | 2 Z | 2 W | 2 Y | 1 | 2 | 6 | 0 | 0 | 0 |
| Drop U | 0 B | 1 S | 1 M | 1 O | 1 P | 2 X | 2 Y | 2 Z | 2 W | 2 | 1 | 2 | 6 | 1 | 0 | 1 |
| | ↑ C[0] | ↑ C[1] | | | | ↑ C[2] | | | | | | | | | | |

**Figure 2.** *Illustration of addition and deletion. Each row shows the array of individuals $A[n]$, $1 \leq n \leq 10$, of groups $C[i]$, $0 \leq i \leq 2$, and of empty cells $E[i]$, same range on $i$. The top row is the initial empty state. The next 10 rows are individuals being added in random order by Algorithm 2. The bottom row is an individual being deleted by Algorithm 3. Numbers in the upper left of each cell of $A[n]$ indicate the group to which that cell is presently allocated, as defined by array $C[i]$. The letters in the centers of the cells of $A[n]$ represent distinct individuals assigned to the cells.*

for it in $A[n]$. If so, the individual is ignored and the selection process is repeated. The entire process is still Order-1 on the number of individuals, though with a higher coefficient.

## 3.2 Adding to a group

Adding a member to a group is a relatively simple but exacting process. If space is available at the end of the group, the new individual is simply placed there and the number of unused entries in the group, $E[i]$, is reduced by 1. If, on the other hand, the area allocated to the group is full, then one member from each of one or more neighboring groups must be shifted to make room. The external function *Transfer* is called to actually move the entries, since the larger program may have other lists that must be updated when an individual is moved.

This process is defined precisely in Algorithm 2 (appendix) and illustrated in Figure 2. The latter is a step-by-step example starting with an empty list of 10 individuals and filling it in random order. In the example there are 26 possible individuals, each with a fixed "name", 'A' through 'Z'. Each is assigned an initial group, which organizes the list, such that 'A' through 'J' initially belong to group 0, 'K' through 'T' initially belong to group 1, and 'U' through 'Z' initially belong to group 2. Individuals can be moved from group to group as the simulation proceeds. In practice, large numbers of individuals would be processed, not just ten.

Array $A[n]$ starts with 10 empty slots. The three groups, 0, 1, and 2, have an initial allocation of 3, 3, and 4 slots, respectively. Entries are added at the first available slot for their group, until that group is filled. Then entries cascade to the right or left, resting in the first available slot.

Individual 'U' is added first. It is initially in group 2, which begins at position $A[7]$. That is followed by 'O', 'P', 'Z', 'M', 'B', and 'W', all of which fall into empty slots pre-allocated in their initial groups. That situation is typical when the array is sparse, and in part leads to the algorithm's speed. However, when 'S' is to be added, the space allocated to its group (cells $A[4]$, $A[5]$, and $A[6]$) is full. To make room, group 2 could move down or group 1 could move up. This example shows the latter, and accomplishes that by shifting 'M' into the open cell at $A[3]$, changing $C[1]$ accordingly, and placing 'S' at the newly opened cell at $A[6]$. Note that 'S' could simply have been placed in position $A[3]$, rather than moving 'M' there. Various optimizations could be applied at the cost of a little additional complexity in the code, though the effects on overall timing would be minor.

Individual 'X', initially designated for group 2, falls immediately in the open cell at $A[10]$. Individual 'Y' is also of group 2, and the cells for that group are full. The algorithm moves 'S' to empty cell $A[2]$, then moves 'X' to $A[6]$, just vacated by 'S', and finally stores 'Y' in the vacated cell $A[10]$. Arrays $C$ and $E$ are updated in the process.

With this method, the maximum number of array elements moved is bounded by the number of groups, independent of the number of individuals. When sufficient memory is allocated to leave a fraction of the space free in $A[n]$, then typically no array elements must be moved. That makes the Order-1 coefficient as small as possible. It also makes it, for practical purposes, independent of the number of groups.

### 3.3 Deleting from a group

Deleting is simpler. The individual is removed and the member from the end of that group is moved into its place, which keeps the individuals in the group contiguous. For example, when individual 'U' is removed, that would leave a gap in the middle of group 2. Individual 'Y' at the end of the group is moved into its place, increasing by one the number $E[2]$ of empty slots at the end of the group. Deleting is always independent of the number of individuals and the number of groups.

Moving an individual from one group to another is merely deleting from one group and subsequently adding to another. As before, the external function *Transfer* is called to actually move the entries.

## 4. Timing

Timing tests of the above algorithms, starting with an empty array and building to one-hundred million individuals ($10^8$), averaged 1.04 seconds total on a 2.8 GHz processor, or 10.4 nanoseconds per addition. Individuals were added in random order, with all groups equally likely, into an array that had 15% more room than required. Selecting from 100 groups averaged 2.1 nanoseconds per selection. Deleting them all at the end averaged 6.8 nanoseconds each.

With sufficient space in array $A[n]$, as provided above, the algorithms become independent not only of the number of individuals but also of the number of groups. Keeping all else equal while increasing groups a thousand-fold, to 100,000 distinct groups, required no additional running time in the algorithm itself.

Nonetheless, large multi-gigabyte arrays such as these may exercise a processor's internal memory caches in various ways, and that can affect finer details of timing. In any case, the algorithms remain efficient for a very large number of groups.

## 5. Conclusions

The algorithms presented here can be incorporated into any individual-based or other microscale model, where they can speed simulations many orders of magnitude over alternative methods that are not Order-1. The methods applied in these algorithms are part of a large-scale simulation model developed by one of us (A.K.) for tuberculosis in the UK. Compilable copies of the code described here and related simulation algorithms are available free from the authors upon request.

## 6. Acknowledgements

We are grateful to Todd Lehman and Shelby Williams for reading the manuscript and helping with the presentation of the material, and to Richard Barnes for enlightening discussions on the application of these algorithms to priority queues. The project was supported in part by a resident fellowship grant to C.L. from the University of Minnesota's Institute on the Environment, by grants of computer time from the Minnesota Supercomputer Institute, Minneapolis, Minnesota, and by doctoral research funding to A.K. from the Modelling and Economics Unit at the Health Protection Agency, London.

## 7. Contributions

Both authors contributed equally. The algorithms began with an Order-1 method by A.K. for managing two groups. C.L. and A.K. worked jointly to incorporate any number of groups. A.K. proposed how to handle variable probabilities within groups. C.L. initially coded the algorithms and both authors contributed to the code and to the manuscript.

## References

[1] L. Gustafsson and M. Sternad, "Consistent micro, macro and state-based population modelling," *Mathematical Bioscience*, vol. 225, pp. 94–107, 2010.

[2] I. G. Kevrekidis and G. Samaey, "Equation-free multiscale computation: Algorithms and applications," *Annual Review of Physical Chemistry*, vol. 60, pp. 321–344, 2009.

[3] V. Grimm and S. F. Railsback, "Individual-based modeling and ecology," *Princeton University Press, New Jersey*, 2005.

[4] S. Edelkamp and S. Schrödl, "Heuristic search: Theory and applications," *Morgan Kaufmann*, p. 152, 2011.

[5] A. I. Dumey, "Indexing for rapid random access memory systems," *Computers and Automation*, vol. 6, pp. 6–9, 1956.

[6] J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S. Tomba, J. Wallinga, J. Heijne, M. Sadkowska-Todys, M. Rosinska, and W. J. Edmunds, "Social contacts and mixing patterns relevant to the spread of infectious diseases," *PLoS Medicine*, vol. 5, p. e74, 2008.

[7] C. Lehman and A. Keen, "Efficient pseudo-random numbers generated from any probability distribution," *Proceedings, International Conference on Modeling, Simulation, and Visualization Methods*, 2012.

[8] W. Hörmann and J. Leydold, "Continuous random variate generation by fast numerical inversion," *ACM Transactions on Modeling and Computer Simulation*, vol. 13, pp. 347–362, 2003.

[9] L. Devroye, "Non-uniform random variate generation," *Springer-Verlag, New York*, 1986.

[10] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, "Numerical recipes: The art of scientific computing, third edition," *Cambridge University Press, New York*, 2007.

[11] M.-T. Tsai, T.-C. Chern, J.-H. Chuang, C.-W. Hsueh, H.-S. Kuo, C.-J. Liau, S. Riley, B.-J. Shen, C.-H. Shen, D.-W. Wang, and T.-S. Hsu1, "Efficient simulation of the spatial transmission dynamics of influenza," *PLoS ONE*, vol. 5, p. e13292, 2010.

# 8. Appendix: Grouping Algorithms

To use the algorithms described in this paper, it is only necessary to understand the entry and exit conditions that appear at the beginning of each, not the code itself. Nonetheless, to allow complete evaluation of the algorithms, and to encourage further development of them, we present them as pseudo-code inspired by and simplified from the programming languages C, R, and Python. The algorithms are defined with sufficient precision that they can be tested, timed, or translated to other languages. Familiarity with a relatively few operators* and with the syntax of flow control (if, for, while, etc.), is sufficient to follow the algorithms. Text copies of this pseudo-code translated into operational C are available from the authors upon request, or from the associated website, 'www.cbs.umn.edu/modeling'.

**DATA STRUCTURES**

| | |
|---|---|
| $ma \equiv 100000000$ | Sample, maximum number of individuals in $A$. |
| $nc \equiv 10000$ | Sample, maximum number of groups in $C$. |
| **structure** *Individual* | Sample, data structure for individuals |
| **integer** $g$; | (Optional, group for this individual) |
| **real** $v$; | (Optional, relative probability of remaining unselected) |
| **real** *tbirth*; | (Sample, time of birth) |
| **integer** *rob*; | (Sample, region of birth). |
| **structure** *Individual* $A[ma+1]$; | Array of individuals. |
| **integer** $C[nc+2]$; | Array of beginning index for each group in $A$. |
| **integer** $E[nc+2]$; | Number of empty cells trailing each group. |

**Algorithm 1. SELECT RANDOM ELEMENT FROM GROUP**

**Upon entry to the algorithm,** **(1)** $k$ defines the group to be sampled. **(2)** $nc$ specifies the number of groups. **(3)** $C$ indexes the first individual in each group. **(4)** $E[k]$ contains the number of empty cells at the end of the group. **(5)** $A[i].v$ contains the probability of each individual in the group remaining unselected, relative to the individual least likely to remain unselected. All $A[i].v$ are zero if all individuals are equally likely. **(6)** *Rand* returns a uniformly distributed random number between 0 and 1, including 0 but not including 1. **At exit,** *GroupSelect* indexes the individual selected. If zero, the group is empty.

**integer** *GroupSelect*($k$) **integer** $k$; **integer** $h$, $n$;

| | |
|---|---|
| **if** $C[0] = 0$ **or** $k \geq nc$ : **return** 0; | 1. Guard against null cases. |
| $C[k+1] - C[k] - E[k] \rightarrow h$;<br>**if** $h \leq 0$ : **return** 0; | 2. Determine how many occupy the group. |
| **loop** until return :<br>$C[k] + h*Rand() \rightarrow n$;<br>**if** $A[n].v = 0$ : **return** $n$; | 3. Select an individual randomly and use it unless it has a probability of remaining unselected. |
| **if** $Rand() \geq A[n].v$ :<br>**return** $n$; | 4. Otherwise use it only in proportion to its relative probability. |

* *The pseudo-code given here is two-dimensional, as in the language Python, so that indentation completely defines the nested structure, with no need for bracketing characters such as '{' and '}'. Variables and function names are italicized and flow control and reserved words are bolded.*

*The assignment operator is represented either as '$\leftarrow$' or '$\rightarrow$', similar to assigments in R. The compound assignments '$a + 1 \rightarrow a \rightarrow b \rightarrow W[i][j]$' and '$W[i][j] \leftarrow b \leftarrow a \leftarrow a + 1$' are equivalent, first incrementing $a$ and placing the results back in $a$, then in $b$, and then in the $i, j$th element of the array $W$.*

*Using up-tick and down-tick operators to write '$\uparrow a$', '$\downarrow a$', '$a\uparrow$', and '$a\downarrow$' form pre- and post-increments by one, as in '`++a`', '`--a`', '`a++`', and '`a--`' of C.*

*Arrays are indexed as in the language C, starting with 0. Data types are* **'integer'** *and* **'real'**, *with the latter specifying floating point. Operator precedence is that of C, with assignments having lowest precedence. Logical operators such as* **'and'** *and* **'or'** *are preemptive, terminating a chain of logical operations as soon as the result is known. Permanent global assignments, as would be represented '`#define` $\alpha\ \beta$' in C, are rendered as '$\alpha \equiv \beta$'.*

**Algorithm 2. ADD ELEMENT TO GROUP**

**Upon entry to the algorithm,** **(1)** $A$ contains the list of individuals, ordered by group, with sufficient room for at least one more individual. **(2)** $C$ indexes the first individual in each group. **(3)** $E$ contains the number of empty cells in each group. **(4)** $k$ is the group for the individual to be added. **(5)** $ma$ contains the maximum number of individuals that may reside in $A$. **(6)** $nc$ contains the number of groups. **(7)** $Transfer(m, n)$ is an external function to move individuals from entry $m$ to entry $n$, including updating of any external information. **At exit,** **(1)** *GroupAdd* returns the index of an available entry in $A$ where the individual is to be added. If zero, none can be added. **(2)** $A$, $C$, and $E$ are updated to include space for the new individual.

**integer** *GroupAdd*($k$) **integer** $k$; **integer** $i$, $m$, $n$, $d$;

| | |
|---|---|
| **if** $C[0] = 0$ : **return** 0; | 1. Guard against null cases. |
| $0 \rightarrow d$;<br>**while** $k - d \geq 0$ **or** $k + d + 1 \leq nc$ :<br>$k - d \rightarrow i$; **if** $i \geq 0$ **and** $E[i] > 0$ : **exit loop**;<br>$k + d + 1 \rightarrow i$; **if** $i \leq nc$ **and** $E[i] > 0$ : **exit loop**;<br>$-1 \rightarrow i$; $\uparrow d$;<br>**if** $i < 0$ : **return** 0; | 2. Search forward and backward simultaneously for the nearest group with an empty slot, returning with failure if the array is full. |
| **while** $i \geq k$ :<br>$\downarrow E[i]$; $C[i + 1] - E[i] - 1 \rightarrow m$;<br>**if** $i = k$ : **return** $m$;<br>$C[i] \rightarrow n$; **if** $m \neq n$ : *Transfer*($n$, $m$);<br>$\uparrow E[i - 1]$; $\uparrow C[i]$; $\downarrow i$; | 3. If there is a slot at the present location, use it, or if forward, cascade it back to the current location. |
| **while** $i \leq k$ :<br>$\downarrow E[i]$; $C[i + 1] - 1 \rightarrow m$;<br>**if** $i = k$ : **return** $m$;<br>$C[i + 2] - 1 \rightarrow n$; **if** $m \neq n$ : *Transfer*($n$, $m$);<br>$\uparrow E[i + 1]$; $\downarrow C[i + 1]$; $\uparrow i$; | 4. Otherwise, if there is a slot earlier in the list of groups, cascade it forward to the current location. |

**Algorithm 3. DELETE ELEMENT FROM GROUP**

**Upon entry to the algorithm,** **(1)** $k$ is the group for the individual to be deleted. **(2)** $n$ indexes the individual being deleted, whose entry is ready for reuse. **(3)** $A$ contains the list of individuals, ordered by group. **(4)** $C$ indexes the first individual in each group. **(5)** $E$ contains the number of empty cells in each group. **(6)** $nc$ contains the number of classes. **(7)** $Transfer(m, n)$ is an external function to move individuals from entry $m$ to entry $n$, including updating of any external information. **At exit,** **(1)** *GroupDelete* returns zero if the operation failed. **(2)** $A$, $C$, and $E$ are updated to exclude the deleted individual.

**integer** *GroupDelete*($k$, $n$) **integer** $k$, $n$; **integer** $m$;

| | |
|---|---|
| **if** $C[0] = 0$ **or** $k > nc$ **or** $C[k + 1] - C[k] \leq 0$ : **return** 0; | 1. Guard against null cases. |
| $\uparrow E[k]$; $C[k + 1] - E[k] \rightarrow m$;<br>**if** $n \neq m$ : *Transfer*($m$, $n$); | 2. Transfer the last occupied entry to the deleted slot. |
| **return** 1; | 3. Return with success. |

**Algorithm 4. INITIALIZE ALL GROUPS**

**Upon entry to the algorithm,** **(1)** $nc$ is the number of groups. **(2)** $ma$ is the maximum number of individuals. **At exit,** **(1)** $C[i]$ indexes the location for the first individual of group $i$. **(2)** $E[i]$ contains the number of entries initially in group $i$. Groups are of equal size to within the limits of integer arithmetic.

*GroupInit*() **integer** $i$, $k$, $n$, $r$;

| | |
|---|---|
| $ma/nc \rightarrow n$; $ma - nc*n \rightarrow r$; | 1. Compute the group size and remainder. |
| $1 \rightarrow k$;<br>**for** $i$ **from** 0 **to** $nc$ :<br>$n \rightarrow E[i]$; **if** $i \geq nc - r$ : $\uparrow E[i]$;<br>$k \rightarrow C[i]$; $k + E[i] \rightarrow k$; | 2. Initialize the location and size of each group, distributing the remainder across higher-numbered groups. |
| $0 \rightarrow E[nc]$; | 3. Close the list of groups. |

# Trading Space for Time: Constant-Speed Algorithms for Managing Future Events in Scientific Simulations

**Clarence Lehman[1], Adrienne Keen[2], and Richard Barnes[1]**
[1]University of Minnesota, 123 Snyder Hall, Saint Paul, MN 55108, USA
[2]London School of Hygiene and Tropical Medicine, Keppel St., London WC1E 7HT, UK

*"It is a mistake to try to look too far ahead. The chain of destiny can only be grasped one link at a time."*
—Winston Churchill

**Abstract**— *Given vast increases in computing capacity, applications in science and engineering that were formerly interpreted with ordinary or partial differential equations, or by integro-partial differential equations, can now be understood through microscale modeling. Interactions among individual particles—be they molecules, viruses, or individual humans—are modeled directly, rather than first abstracting the interactions into mathematical equations and then simulating the equations. One approach to microscale modeling involves scheduling all events into the future, wherever that is possible. With sufficient space-for-time tradeoffs, this considerably improves the speed of the simulation, but requires scheduling algorithms of high efficiency. In this paper we describe our variation on calendar queues and their usage, presenting detailed algorithms, intuitive explanations of the methods, and notes from our experiences applying them in large-scale simulations. Results can be useful to scientists in ecology, epidemiology, economics, and other disciplines that employ microscale modeling.*

**Keywords:** microscale modeling, discrete event simulation, calendar queues, pending events set, space-time tradeoff

## 1. Introduction

The obvious approach to model a large number of discrete interacting entities, hereinafter called "individuals," is to emulate what is done to model continuous systems with differential equations. That is, select a small time step $\Delta t$, compute how the system will change during the interval $\Delta t$, update the system with those changes, then advance to the next time step. In ordinary differential equations, as the time step shrinks, the dynamics of the simulated system converge to the correct behavior. This is "macroscale modeling," following Euler's method or its many variations [1]. With a model of 100 compartments, representing, for example, 100 age classes in a human population, relatively few dynamical variables must be examined and updated in each time step.

The same approach works with microscale modeling, though with difficulties. At each time step, each individual is examined to determine what interactions will occur during that time step. The difficulty with this approach is twofold. First, each individual acts as a dynamical variable, so there can be many millions or hundreds of millions of variables to be examined and updated in each time step. Moreover, as the time step shrinks to assure convergence, it becomes exceedingly unlikely that anything will happen to a given individual during the time step. Therefore, in contrast with its macroscale counterpart, that approach to microscale modeling spends most of its time checking and finding nothing to do.

Inspiration for a faster approach comes from an alternative method of solving differential equations. Instead of determining what will happen during the present small time step, an algorithm can determine at what time in the future the next event will occur. This can be determined reliably for the very next event, and the precise process for doing so is called Gillespie's method [2]. It is the complement of the standard method.[1]

Despite certain epistemological difficulties about projecting the future that are beyond the scope of the present paper, hinted at in Churchill's statement above, Gillespie's method can be extended to determine possible times for all future events in many dynamical systems of scientific interest—or at least all events that control the fate of the system. But the number of future events can be large, with many events per individual, and the number of individuals in the simulation may be tens or hundreds of millions or more.

Fortunately, algorithms are known that are extremely efficient at handling schedules of future events. Discovered by Randy Brown in 1988 [3], these are called "calendar queues" or "pending event sets," and have been undergoing successive refinements ever since (e.g. [4] [5] [6] [7]). They have the desirable—and remarkable—property that their speed is independent of the number of events scheduled.

[1]In simulating $f(x) = dx/dt \approx \Delta x/\Delta t$, a small time step $\Delta t$ can be established, such as 0.01 seconds, and the change in population (or other simulated quantity) can be estimated as $\Delta x \approx f(x)\Delta t$. That is Euler's method. Alternatively, a change in quantity $\Delta x$ can be specified (such as a population growth of one individual) and the time for that to occur can be estimated as $\Delta t \approx \Delta x/f(x)$. That is Gillespie's method. Thus the mathematics for the two are complementary.

Adding an event, canceling one, or finding the next event about to occur is the same whether the schedule contains 100 events or 100 million. That is, they are "Order 1" algorithms.

In this paper we present our adaptation of calendar queues to large-scale individual-based modeling in epidemiology. Lessons should be applicable to areas including ecology, economics, and other physical and natural sciences. We attempt to make our presentation intuitive for access by scientists and other readers outside computer science. The goals of this paper are to (1) review the idea of space-for-time trade-offs that have become widely useable and applicable to other algorithms (e.g. [8]), (2) explain our variation on calendar queues and their incorporation in microscale simulations, and (3) present our algorithms in full detail for use and adaptation by others.

## 2. Space for time

The persistent increase in random access computer memories has carried algorithms through a "phase change," wherein a slow continuous advance in memory sizes has resulted in a rapid, almost abrupt, change in some of the rules for constructing algorithms for scientific programs. If it will speed processing, computer algorithms can now afford to allocate hundreds of millions of bytes of empty space—even if that space will never be used. This is a "space-for-time tradeoff." With large memories now available, such allocation is no longer wasting memory. On the contrary, leaving memory unused, or leaving it applied to insignificant purposes, is wasting it.

A basic space–time tradeoff arises with numerical keys. Suppose we have 10,000 items, each identified by a distinct six-digit "key," and with keys randomly distributed among values from '000000' to '999999'. Suppose each of the 10,000 items occupies 100 memory cells (e.g. 100 bytes). Stored contiguously, this will require $10^4 \times 10^2 = 10^6$ memory cells. In such a compact arrangement, searching can be relatively fast if the entries are kept in numeric order[2]. However, in this case adding and deleting will be slow, averaging $N$ or more accesses to keep the list contiguous and in order. On the other hand, if the entries are left in random order, adding and deleting will be fast, 1 to 3 accesses only,[3] but searching will be slow, sequentially checking each entry until the right one is encountered. The point is, this minimal-space approach inevitably results in algorithms that are slow in one respect or another.

An alternative is to "waste" memory by allocating one slot in memory for each of the million entries possible. Now to search for a specific six-digit key, say key '314159', the algorithm merely goes directly to the 314,159th entry of the table. Only one access to the memory array is thus needed to retrieve, and the same is needed to add or delete. With 10,000 active entries, this space–time tradeoff speeds the algorithm 5,000 fold. However, it comes at the expense of 100-million memory cells, about one-tenth of a gigabyte. Such cavalier abandon in the use of memory would have been unthinkable until recently, but if speed is the utmost criterion, then allocating an extra 1/10 GB to accomplish a multi-thousand-fold increase in speed is the clear and proper choice.

This approach extends to larger keys through the method of "hash coding," which is directly related to calendar queues. Hash coding is an Order-1 algorithm known at least since Arnold Dumey in 1956 [9]. The key may be an individual's first, last, and middle name, for which the space required for direct access would be astronomical, beyond the power of any computer presently foreseeable. Even if the key was only a nine-digit social security number, such as 123-45-6789, providing one direct-access entry for all possible social security numbers would be prohibitively large.

The simplest solution merely extracts the rightmost six digits of the social security number and indexes an array of a million entries with those six digits. Of course, as many as 1,000 individuals may share the same last six digits of their social security numbers, so "collisions" can occur. But with only 10,000 entries of a million active, and assuming all possible social security numbers are equally likely, each entry in the array has only a 0.01 chance of being occupied, so the chance that two or more individuals will occupy the same cell is very small. Nonetheless, the possibility of collisions must be provided for, and a variety of practical methods have been devised [10]. Once that is done, locating an individual by social security number, or indeed by first, last, and middle name, can be accomplished in one access, or arbitrarily close to one access, with a sufficiently large space-for-time tradeoff.

Dumey's scheme [9] was to use a modulus operation by considering the key to be a large number, dividing it by the size of the memory array (number of entries in the array), then discarding the quotient and using the remainder to index the array—as in the social security example. In that case, the rightmost six digits were equivalent to the remainder after division by one million. Essentially the same underlying scheme is applied in calendar queues, dividing the scheduled time by the size of the memory array (one year's worth of minutes in the intuitive example to follow), and using the remainder to index the array. Therefore, the same space-for-time tradeoffs that make hash-coded accesses maximally fast also can make calendar queues, properly programmed, maximally fast for managing large numbers of future events.

## 3. Future events

Having emphasized the value of spending memory to buy time, we must also say that it is pointless to spend memory

[2]For instance, by using a binary search algorithm, which is of Order $\log_2 N$ accesses, where $N$ is the number of items in the list

[3]New entries can be added at the end in 1 access; deleted entries can be swapped with the entry at the end in 3 accesses.

when it does not buy time. The more events that are scheduled at once, the greater the amount of memory that is needed to handle them efficiently, in direct proportion to the number scheduled. Also, the more that the number of events scheduled vary during the simulation, the more frequently the data structures should be optimized by "resizing" [3].

Therefore, to help keep the scheduling algorithms efficient, our microscale simulation programs withhold all but one event per individual from them. Characteristics of individuals are maintained in a large array of data structures, $A[n]$, indexed by individual number $n$, which ranges from one to some maximum value. This array includes data of two types: (1) information about the individual, such as, in a model of human events, date of birth, sex, geographic location, and so forth, and (2) a list of all future events relevant to that individual. This large array is not processed nor examined by the scheduling routines described in this paper.

Only the earliest among the events pending for each individual is entered into the global schedule, with the data structure $A[n]$ holding the rest. Such withholding of information has several benefits: (1) the number of events managed by these algorithms is considerably reduced, (2) the number of events that must be canceled and rescheduled is reduced, and (3) the size of the scheduling data structures are predictable, with precisely one event per individual. This partly obviates the need for the scheduling algorithms to maintain separate lists for near, intermediate, and far future events, as in some variations of calendar queues [11], and also eliminates the need for time-consuming "resizing" operations [3].

# 4. Intuitive view

We want to (1) schedule new events, (2) cancel existing events, and (3) notify a dispatcher as the time for each event arrives—all three with maximal efficiency. The coding details can be subtle, but the overall operation is not. It can be understood intuitively through a physical analogy.

Assume, for a specific illustration, that half a million events are to be scheduled over the next five years, and that they appear more-or-less randomly throughout that period. Suppose that each event has a ticket with (1) a unique event number and (2) a scheduled time, represented at least to the nearest second, but possibly much finer.

Now consider a series of pigeon-hole bins to contain the tickets, one bin representing each minute of an entire year. The first bin represents the first minute after midnight on New Year's Day, the second bin represents the second minute, and so forth to the last bin, which represents the last minute on December 31st. That is 366 days × 24 hours/day × 60 minutes/hour = 527,040 bins total, each labeled with the month, day, hour, and minute that it represents. Each bin also has a flag that can be lowered or raised according to whether the tickets in the bin are known to be in chronological order. We assumed half a million events to be scheduled, less than one event per bin on average.

## 4.1 Creating a new event

Events are created as the simulation proceeds, each associated with a particular individual and with a precisely assigned time, usually stochastically assigned. In an ecological model these may represent a time of birth or death, in an epidemiological model they may represent the time of onset of a disease, or the time for transmission to another individual. In any case, new events arise frequently during the simulation. The procedure for scheduling a new event is quite easy:

1. Go to the bin representing the month, day, hour, and minute for the event. Although the year, second, and any fraction of a second are not used to select a bin, they are later used to place events in precise chronological order.
2. Drop the event's ticket on top of the others in the bin.
3. Raise the flag on the bin to indicate that its tickets may no longer be in chronological order.

That required only a single operation, regardless of how many events were in the bin. We take it to be important merely to drop the ticket atop others in the bin, as above, rather than trying to sort it into place among other tickets in the bin. Earlier implementations of calendar queues [3] keep all bins always sorted, but that can be disabling if a large number of events accumulate in any bin. Such accumulation can occur during testing or simulation.

## 4.2 Canceling an existing event

Once scheduled, events may occasionally have to be canceled. For example, in an epidemiological model, a healthy individual may become the target of an infection. Whatever the next event in their life was, it may have to be rescheduled as the simulated individual progresses toward disease and infectiousness. Therefore, the existing event will be canceled and the earliest of other future events for the individual will be scheduled instead. In the physical analogy, that requires three steps:

1. Go to the bin representing the month, day, hour, and minute for the event. As before, ignore the year, second and any fraction of a second.
2. Flip through them to find the ticket for the event in question.
3. Destroy that ticket.

That required one operation for every ticket in the bin, but on average there is only one ticket in the bin. Canceling an event can be slow if the events cluster badly, because of the need to flip through the tickets in the bin. But canceling is not a usual operation. The two common operations are adding events (described above) and dispatching them (described next).

### 4.3 Dispatching the next event

The simulation proceeds stepwise by locating the earliest among all events in the schedule, removing it, then processing it. This is efficient, but it involves several steps:

1. Go to the bin representing the current day, hour, and minute.
2. If the flag on the bin is raised, arrange the tickets in chronological order and lower the flag.
3. Leave any tickets for future years in the bin.
4. Process any tickets from this year, day, hour, and minute, each to be handled precisely in sequence as the scheduled second and fraction of a second arrives.
5. If any tickets for the current bin arrive while the bin is being handled, put them in their proper position among the other tickets.

This required only one operation for each ticket, plus one or two more per ticket to order them chronologically before dispatching the contents of the bin. Again, on average there is only one ticket in the bin.

This method intentionally does not keep tickets in the bins ordered, using instead "just-in-time sorting." Usually this will make little difference, since the bins are intentionally designed to be nearly empty. However, as described earlier, if unexpected clustering occurs, this just-in-time sorting will be much faster than keeping the contents of all bins in order each time an event is added.

Within a simulation program using these scheduling algorithms, the individual associated with the ticket being handled will have other pending events in its entry of data structure $A[n]$. The simulation program will then pass the earliest of these to the scheduling algorithms, through a call to *EventSchedule*.

The discussion above shows how the algorithms achieve their speed—by maintaining at least as many bins as there are tickets. If there were sixty times as many tickets—thirty million—the same speed of operation could be maintained simply by increasing the number of bins by sixty, to one bin for each second.

## 5. Applications

The algorithms described here have been applied and tested in a large-scale multi-compartmental epidemiological model of tuberculosis transmission developed by one of us (A.K.). That model runs with upwards of $6 \times 10^7$ individuals (60 million), representing the entire population of the UK, on multiple parallel processors for parameter fitting by simulated annealing. Each individual has many events pending, including, for example, scheduled times of death, emigration, onset of disease for recently infected individuals, next transmission for infectious individuals, potential vaccination for juveniles, and so forth.

In this epidemiological model, typical runs spanned 30 simulated years and used 75 million bins occupied by 60 million individuals. Each run consumed about 80 seconds on a 2.8 GHz processor, using a little over 6 GB of memory on each of 30 to 50 parallel processors. The average time increment between scheduled events was 14 simulated seconds, with a standard deviation of 12 seconds. The minimum was less than a simulated microsecond, whenever stochastic events appeared by chance close together in time. The maximum time increment was 53 simulated seconds. Thus the time steps are very small compared with a corresponding macroscale model.

In simplified timing tests on the same processor, outside of the operation of the epidemiological model, a list with $6 \times 10^7$ individuals needed 30 nanoseconds on average to schedule each new event, 18 nanoseconds to cancel an event, and 12 nanoseconds to dispatch each event when its time arrived. This was near-ideal conditions, with new events arising in sequence in a way that minimized clustering in the schedule. Expanding the number of individuals by a factor of more than 16, to $10^9$ individuals (one billion) required exactly the same amount of time per operation—within small bounds of statistical error—demonstrating the Order-1 behavior of the algorithms.

On the other hand, events arising in random order needed 90 nanoseconds to schedule each new event into a list of 60 million and 180 nanoseconds into a list of one billion. The three to six-fold increase can be attributed to interactions with internal memory caches. Such caches grow less useful as memory accesses become less localized.

## 6. Algorithmic details

The intuitive picture sketched above converts directly into the algorithms displayed in the appendix. As implemented in the algorithms, the bins need not correspond to standard time units such as minutes, but can be any values.

A simulation begins by adding one or more events, typically one event per individual, and ends either at a predetermined time or when the last event has been dispatched. Array $A[n]$ would be established earlier with a collection of pending events for each individual $n$. The main simulation program would be structured as follows:

**[1]** *ProgramInit*();
**[2]** **loop for all** $n$ **in** $A[n]$ :
    *EventSchedule*$(n,$ *earliest*$(n))$;
**[3]** **loop for** $t$ **from** 0 **to** $t_{max}$:
    *Process*$($*EventNext*$())$;
**[4]** **exit**;

In step 1 above, *ProgramInit* sets the initial conditions for the program, including allocating all individuals that will start the simulation and all future events that are known for each. Step 2 moves through all individuals, selects the earliest event for each (*earliest*$(n)$), and schedules each event by calling *EventSchedule*. With all events to start the simulation scheduled, step 3 repeatedly asks for the next chronological event by calling *EventNext* and passing the

number of that event to *Process*. In turn, *Process* will call upon *EventSchedule* and possibly *EventCancel* and *EventRenumber* while carrying out the simulation. *ProgramInit*, *Process*, and *earliest*, as well as array $A[n]$, are written as part of the simulation program. The rest are scheduler algorithms detailed in the appendix.

The two main data structures organizing the earliest event for each individual are (1) a circular array of integers $Q[h]$, each heading a linked list in $P[n]$ of events scheduled for time bin $h$, and (2) an array of integers $P[n]$, each continuing the linked list from $Q[n]$. The number of entries in $P[n]$ must equal the number in external array $A[n]$, and like $A[n]$, $P[n]$ is indexed by individual number. But the number of time bins $Q[h]$ may be smaller or larger than the number of individuals. The size of $Q[n]$ is a matter of optimization. It is typical to have one time bin for each event that could be scheduled, meaning each bin will represent a single event on average. A space–time tradeoff occurs because optimal allocation leaves about one-third of the bins empty.[4]

Each bin $Q[h]$ represents many related times, all equal modulo the width of the series of time bins, $Qw$. The width $Qw$ of all bins combined is also a matter of optimization. If it is much too large, events will tend to cluster near the bin being dispatched. If it is much too small, events will tend to spread out, with most bins containing events that are for the more distant future. A suitable value for $Qw$ can be found by knowledge of the system being analysed, or by experimental trials to find a good speed of operation.

For speed of addition, the lists of events in $P[i]$ are not maintained in any particular order, but each bin is sorted chronologically before it is dispatched. Any sorting algorithm used should have (1) best performance when the list is already partially sorted, e.g. Order $N$, important because lists will remain partially sorted from earlier passes, (2) high-speed when sorting only 1 and 2 entries, which are the most common, and (3) good worst-case performance, e.g. Order $N \log_2 N$. The sorting routine presented as Algorithm 5 in the appendix has these properties.

## 7. Conclusions

The algorithms presented here can be incorporated into any individual-based or other microscale model, where they can speed simulations many orders of magnitude over alternative methods that are not Order-1.

They are part of a large-scale simulation model developed by one of us (A.K.) for tuberculosis in the UK. Sixty million individuals thus can be handled by allocating less than a gigabyte of random access memory—within the reach even of portable computers. In practice, these algorithms should be able to schedule, cancel, and dispatch up to $10^7$ or more events per second with 60 million or more pending events maintained in the queue. Therefore, they should not become a bottleneck in the simulation as a whole.

Compilable copies of the code described here and related simulation algorithms are available free from the authors upon request.

[4]Under random distribution, $1/e = 37\%$ will be empty. That can be shown to be optimal for overall speed if all bin operations are equally fast.

## 8. Acknowledgements

We thank Fred Lehman, Holly MacCormick, Peter Hawthorne, Ben Kerr, Celia Hemmerich, and Shelby Williams for discussions and encouragement. The project was supported in part by a resident fellowship grant to C.L. from the UMN Institute on the Environment, by grants of computer time from the Minnesota Supercomputer Institute, and by doctoral research funding to A.K. from the Modelling and Economics Unit at the Health Protection Agency, London.

## 9. Contributions

C.L. considered the notion of stochastically prescheduling the earliest future event for each individual and developed an initial application. A.K. expanded the approach for her large-scale tuberculosis model, leading to refinements in the algorithms. A.K. conceived a grouping method to work in concert and make these algorithms practical for multi-compartment simulation models [8]. R.B. participated in the evaluation, applications to other areas, and the literature review. C.L. coded the algorithms and A.K. tested them in large-scale operation. All authors contributed to the manuscript.

## References

[1] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, "Numerical recipes: The art of scientific computing, third edition," *Cambridge University Press, New York*, 2007.

[2] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *Journal of Physical Chemistry*, vol. 81, pp. 2340–2361, 1977.

[3] R. Brown, "Calendar queues: a fast 0(1) priority queue implementation for the simulation event set problem," *Commun. ACM*, vol. 31, no. 10, pp. 1220–1227, Oct. 1988.

[4] R. Rönngren, J. Riboe, and R. Ayani, "Lazy queue: an efficient implementation of the pending-event set," *SIGSIM Simul. Dig.*, vol. 21, no. 3, pp. 194–204, Apr. 1991.

[5] G. A. Davidson, "Calendar p's and q's," *Communications of the ACM*, vol. 32, pp. 1241–1242, 1989.

[6] T. Hui and I. Thng, "Felt: A far future event list structure optimized for calendar queues," *Simulation*, vol. 78, no. 6, pp. 343–361, 2002.

[7] G. Yan and S. Eidenbenz, "Sluggish calendar queues for network simulation," in *Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, 2006. MASCOTS 2006. 14th IEEE International Symposium on*, 2006, pp. 127–136.

[8] A. Keen and C. Lehman, "Trading space for time: Constant-speed algorithms for grouping objects in scientific simulations," *Proceedings, International Conference on Scientific Computing.*, pp. 146–151, 2012.

[9] A. I. Dumey, "Indexing for rapid random access memory systems," *Computers and Automation*, vol. 6, pp. 6–9, 1956.

[10] D. E. Knuth, "The art of computer programming, volume 3: Sorting and searching, second edition," *Addison-Wesley, Reading, MA*, 1998.

[11] R. Goh and I. Thng, "Mlist: An efficient pending event set structure for discrete event simulation," *International Journal of Simulation-Systems, Science & Technology*, vol. 4, no. 5-6, pp. 66–77, 2003.

# 10. Appendix

To use the algorithms described in this paper, it is only necessary to understand the entry and exit conditions that appear at the beginning of each, not the code itself. Nonetheless, to allow complete evaluation of the algorithms, and to encourage further development of them, we present them as pseudo-code inspired by and simplified from the programming languages C, Python, and R. The algorithms are defined with sufficient precision that they can be run, tested, timed, modified, or translated to other languages. Familiarity with a relatively few operators* and with the syntax of flow control (if, for, while, etc.), is sufficient to follow the algorithms. *WarnMsg* and *ExitMsg* display error messages and the latter terminates the program. Not all functions return values. Text copies of this pseudo-code translated into operational C are available from the authors upon request, or from the associated website `www.cbs.umn.edu/modeling`.

**PROGRAM PARAMETERS**

| | |
|---|---|
| $TN \equiv (100000000)$ | Example, maximum number of time bins. |
| $PN \equiv (100000003)$ | Example, maximum number of forward indexes to time bins. |
| $TW \equiv 20$ | Example, time width of all bins combined (for optimization). |

**INTERNAL DATA STRUCTURES**

| | | | |
|---|---|---|---|
| $PZ \equiv -1$ | | | Marker for empty bins. |
| **real** | $T[PN]$ | $\leftarrow 0$; | Time for each scheduled event. |
| **integer** | $P[PN]$ | $\leftarrow PZ$; | Forward indexes within bins, ending with zero. |
| **integer** | $Q[TN]$ | $\leftarrow 0$; | First index for the bin, with zero for empty bins, negative for unsorted bins. |
| **real** | $Qw$ | $\leftarrow TW$; | Interval of time represented for each cycle in $Q$. |
| **integer** | $Qn$ | $\leftarrow TN$; | Number of elements in $Q$. |
| **integer** | $Qi$ | $\leftarrow 0$; | Index of the immediate time bin. |
| **integer** | $Qe$ | $\leftarrow 0$; | Number of events in all bins. |
| **real** | $Qt0$ | $\leftarrow 0$; | Earliest time representable this cycle in $Q$. |
| **real** | $Qt1$ | $\leftarrow TW$; | Earliest time beyond this cycle in $Q$. |
| **real** | $t$ | $\leftarrow 0$; | Current time, last dispatched event. |

**Algorithm 1. SCHEDULE A NEW EVENT**

**Upon entry to the algorithm, (1)** $n$ contains the number (starting with 1) of a new event. **(2)** $te$ contains the time at which the new event will occur. **(3)** $P[n]$ indicates that the event is unscheduled (equal to $PZ$). **(4)** The scheduling data structures are prepared as described above. **At exit, (1)** the event has been scheduled, to occur when the proper time arrives. **(2)** $T[n]$ records the time $te$ of the event. **(3)** $P[n]$ links the event with others in its time bin.

*EventSchedule*($n$, $te$) **integer** $n$, **real** $te$; **integer** $i$; **real** $tr$;

| | |
|---|---|
| **if** $n < 1$ **or** $n \geq PN$: *ExitMsg*(3);<br>**if** $P[n] \neq PZ$: *ExitMsg*(4);<br>**if** $te < t$: *ExitMsg*(5); | 1. Check the index and make sure an event is not already scheduled and is not in the past. |
| $te \rightarrow T[n]$; | 2. Record the time of the new event. |
| $(te - Qt0)/Qw \rightarrow tr$; $tr - (int)tr \rightarrow tr$;<br>$tr * Qn \rightarrow i$; | 3. Convert the time to a bin number. |
| $abs(Q[i]) \rightarrow P[n]$, $-n \rightarrow Q[i]$, $\uparrow Qe$; | 4. Add the event to the list for that bin and increment the number of events. |

* *The pseudo-code given here is two-dimensional, as in the language Python, so that indentation completely defines the nested structure, with no need for bracketing characters such as '{' and '}'. Variables and function names are italicized and flow control and reserved words are bolded.*

*The assignment operator is represented either as '$\leftarrow$' or '$\rightarrow$', similar to assignments in R. The compound assignments '$a + 1 \rightarrow a \rightarrow b \rightarrow W[i][j]$' and '$W[i][j] \leftarrow b \leftarrow a \leftarrow a + 1$' are equivalent, first incrementing $a$ and placing the results back in $a$, then in $b$, and then in the $i, j$th element of the array $W$.*

*The expression structure '$c\,?\,u : v$', where $c$ is a condition, $u$ is an if-expression, and $v$ is an else-expression, follows that of C. Using up-tick and down-tick operators to write '$\uparrow a$', '$\downarrow a$', '$a\uparrow$', and '$a\downarrow$' form pre- and post-increments by one, as in '`++a`', '`--a`', '`a++`', and '`a--`' of C.*

*Arrays are indexed as in the language C, starting with 0. Data types are* **'integer'** *and* **'real'**, *with the latter specifying floating point. Operator precedence is that of C, with assignments having lowest precedence. Logical operators such as* **'and'** *and* **'or'** *are preemptive, terminating a chain of logical operations as soon as the result is known. Permanent global assignments, as would be represented '`#define` $\alpha\ \beta$' in C, are rendered as '$\alpha \equiv \beta$'.*

**Algorithm 2. CANCEL AN EXISTING EVENT**

**Upon entry to the algorithm, (1)** $n$ contains the number (starting with 1) of the event to be cancelled. **(2)** $T[n]$ contains the scheduled time of the event. **(3)** the scheduling data structures are prepared as described above. **At exit,** the event has been removed from the list.

```
EventCancel(n)  integer n; integer i, j, jp; real tr;
  if n < 1 or n ≥ PN: ExitMsg(6);
  if P[n] = PZ: ExitMsg(7);                                   1. Check the index and make sure an
                                                                 event is scheduled.
  (T[n] − Qt0)/Qw → tr; tr − (int)tr → tr;
  tr*Qn → i;                                                  2. Convert the time to a bin number,
                                                                 modulo the duration of the cycle.
  if subcancel(n, i): return;
  (i − 1 + Qn) mod Qn → i; if subcancel(n, i): return;        3. Remove it from its normal bin
  (i + 2 + Qn) mod Qn → i; if subcancel(n, i): return;           or from an adjacent bin above or
                                                                 below (due to rounding error).
  ExitMsg(8);                                                 4. If the specified event was not in
                                                                 the list, signal an error.

integer subcancel(n, i)  integer n, i; integer j, jp;
  0 → jp, abs(Q[i]) → j;
  loop while j > 0:                                           1. Scan the list of pending events in
    if j = n:                                                    this bin and remove the specified
      if jp > 0: P[j] → P[jp];                                   event. (The average number of events
      else Q[i] > 0?P[j]: −P[j] → Q[i];                          in non-empty bins is about 1.5)
      PZ → P[j]; if ↓Qe < 0: ExitMsg(9);
      return 1;
  j → jp, P[j] → j;
  return 0;
```

**Algorithm 3. DISPATCH THE NEXT EVENT**

**Upon entry to the algorithm, (1)** $T$ contains the time for each scheduled event. **(2)** The scheduling data structures are prepared as described above. **At exit, (1)** $EventNext$ contains the number of the next event. If zero, no events are scheduled. **(2)** $t$ contains the time of the next event, if $NextEvent$ is not zero.

```
integer EventNext()  integer j, n;
  loop while Qe > 0:
    loop while Qi < Qn:                                       1. Advance to the next non-empty
      Q[Qi] → j; if j = 0: ↑Qi; repeat loop;                     bin.

      if j < 0:                                               2. Sort the bin if it may be necessary
        sort(P, −j, 0, order) → Q[Qi] → j;                       (usually sorts 1 or 2).

      if T[j] < Qt1:                                          3. If the event belongs to this pass,
        if P[j] = PZ: ExitMsg(2);                                remove it, decrement the number of
        P[j] → Q[Qi], PZ → P[j], ↓Qe;                            events, advance the time, and return
        T[j] → t; return j;                                      its index.

      ↑Qi;                                                    4. Advance to the next bin and repeat.

    0 → Qi, Qt0 + Qw → Qt0, Qt0 + Qw → Qt1;                   5. Circle back to the first bin.

  return 0;                                                   6. Signal completion of all events.
```

**Algorithm 4. RENUMBER AN EVENT**

**Upon entry to the algorithm, (1)** $n$ contains the new index number, which has no event scheduled. **(2)** $m$ contains the current index number of the event. **At exit, (1)** $n$ is the new index number. **(2)** The event originally scheduled as $m$ is re-scheduled as $n$. Event $m$ no longer has an event scheduled and the index is free to be reused.

```
EventRenumber(n, m)  integer n, m;
  if n < 1 or n ≥ PN: ExitMsg(10);                            1. Check the indexes and make sure
  if m < 1 or m ≥ PN: ExitMsg(11);                               they are in range.

  if n ≠ m:
    T[m] → T[n];                                              2. Transfer the time.
    EventCancel(m);                                           3. Cancel the old number.
    EventSchedule(n, T[n]);                                   4. Reschedule as the new number.
```

**Algorithm 5. SORTING**

**Upon entry to the algorithm,** (1) *list* points to an array of forward indexes. *list*[0] is unused. (2) $p$ indexes the first element of the list, which ends with a zero. (3) $n$ contains the number of items in the list, if known. If zero, the number of items is not known and *sort* should count. (4) $c$ compares two list elements $u$ and $v$. It returns negative, zero, or positive when $u < v$, $u = v$, and $u > v$, respectively. **At exit,** *sort* indexes the first element in the sorted list, which ends with a zero. The original ordering is preserved for entries that are equal.

```
integer *P, pc, pr, m, (*order)(int, int);

integer sort(list, p, n, c)  integer list[], p, n, (*c)(int, int); integer i;
```

```
c → order, list → P;
```
1. Record calling parameters.

```
if n = 0: p → i; loop while i > 0: P[i] → i, n↑;
if n = 0 or p = 0: return 0;
if n = 1: return p;
```
2. Count the number of elements and return empty and single-element lists immediately.

```
if n = 2:
  if order(p, P[p]) ≤ 0: return p;
  P[p] → i, p → P[i], 0 → P[p];
  return i;
```
3. If the list contains only two elements, sort it by inspection.

```
p → pc; return isort(n);
```
4. Otherwise sort the full list.

**Partition into sorted sublists. Upon entry,** (1) $n$ defines the minimum number of elements to be sorted. (2) $P$ is the list of forward indexes. (3) $pc$ indexes the first element of the list. (4) *order* compares two list elements. **At exit,** (1) *isort* indexes the first element in the sorted list, which ends with a zero index. (2) $m$ defines the number of elements which were actually sorted, greater than or equal to its value on entry. (3) $pc$ indexes the element following the last element sorted. If the entire list has been sorted, $pc$ is null.

```
integer isort(n)  integer n; integer wp1, wp2, m1;
if n ≤ 1:
  if pc = 0: return 0;
  pc → wp1, 0 → m;
```
1. If a single element is requested, initialize variables and check for error in count.

```
  loop : pc → pr, P[pc] → pc, m + 1 → m;
    if pc = 0: return wp1;
    if order(pr, pc) > 0: exit loop;
  0 → P[pr]; return wp1;
```
2. Then scan forward in the list to find the longest list that is already in order and return that list.

```
isort(n/2) → wp1;
if n ≤ m: return wp1;
```
3. If multiple elements are requested, sort the first part of the list and return if enough was sorted.

```
m → m1, isort(n − m) → wp2, m + m1 → m;
return imerge(wp1, wp2);
```
4. If it was not, then sort what remains and merge the two sublists.

**Merge sublists. Upon entry,** (1) $P$ is the list of forward indexes. (2) $p$ and $q$ index the first element of a sorted primary and secondary list, respectively. (3) *order* compares two list elements. **At exit,** *imerge* indexes the first element of the list merged in order. In case of equal entries, those from the primary list appear first.

```
integer imerge(p, q)  integer p, q; integer pb, v;
if p = 0: return q; if q = 0: return p;
```
1. Handle empty lists.

```
if order(p, q) > 0: q → pb, 1 → v;
else p → pb, 3 → v;
loop while v > 0:
```
2. Save the beginning of the list and select the proper routine.

```
  loop while v = 1: q → pr, P[q] → q;
    if q = 0: −2 → v;
    else if order(p, q) ≤ 0: 2 → v;
  if v = 2: p → P[pr];
```
3. Scan for a secondary element greater than or equal to the current primary element and mend the secondary list.

```
  loop while v ≥ 2: p → pr, P[p] → p;
    if p = 0: −1 → v;
    else if order(p, q) > 0: 1 → v;
  if v = 1: q → P[pr];
```
4. Scan for a primary element greater than the current secondary element, mend the primary list, and repeat.

```
if v < −1: p → P[pr]; else q → P[pr];
return pb;
```
5. Attach any remaining elements and return the merged list.

*Intn'l Conf. Modeling, Simulation, & Visualization, MSV12: 121–127, Proceedings, July 2012*

# Efficient Pseudo-Random Numbers Generated from Any Probability Distribution

**Clarence Lehman**[1] **and Adrienne Keen**[2]
[1]University of Minnesota, 123 Snyder Hall, 1474 Gortner Avenue, Saint Paul, MN 55108, USA
[2]London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK

*"Truth is much too complicated to allow anything but approximations."*
—John von Neumann, 1947

**Abstract**—*Microscale simulations and other applications in science, engineering, and commerce need an abundance of pseudo-random numbers drawn from non-classical probability distributions, including empirical distributions that may be incompletely known. Discrete-event simulations that assign random times to future events have further requirements, including numbers drawn from subsets of distributions to help establish initial conditions, or to deal with events that are partially complete. Fast methods are known for generating pseudo-random numbers accurately from arbitrary probability distributions, but those methods do not combine the full range of necessary algorithms outlined here. In this paper we provide techniques and computer code for practical high-speed generation of pseudo-random numbers from any continuous, discontinuous, or discrete probability distribution, reducing the need for approximation by standard probability functions. The techniques are designed for the kinds of scientific simulations presently emerging.*

**Keywords:** non-uniform random numbers, uniform random numbers, pseudo-random numbers, probability distributions, numerical simulations

## 1. Introduction

Computer generated pseudo-random numbers are needed at every step in stochastic simulations—as well as to establish representative sets of initial conditions in deterministic simulations, to draw samples for statistical bootstrapping and other operations, to identify uncertainty in models by varying parameters, to randomize experimental designs, to develop test cases for commercial software, and for many other applications in science, industry, and art. They can arise in such quantities as to become a significant part of the total time for the computation itself.

For example, an emerging application arises in discrete-event simulation [1], where stochastically assigned times of future events must be determined in advance. When a simulated individual is born, the future time of death may be assigned from empirical probabilistic "life tables" for the year, geographic location, and other conditions of the individual being simulated. Initial conditions for the simulation can start with an empirically or hypothetically derived "age distribution." Subsets of the life tables (sub-distributions) are sampled to determine how long each individual will live, based on initial conditions. Moreover, when new individuals may enter the population as immigrants, sub-distribution sampling is also needed to assign future times of death. Examples and code that follows will illustrate these points.

In what follows we shall omit the prefix "pseudo" in "pseudo-random", it being understood that repeatable sequences of numbers generated by deterministic computer algorithms can be at best apparently random. The starting point for random number generation from a desired distribution is random number generation from the standard uniform distribution. That is, random numbers greater than equal to zero and less than one, all drawn with equal likelihood and with no correlation between any prescribed pair of numbers. This is difficult to achieve in practice, but much theory and effort have been dedicated to the problem, and a number of acceptable algorithms are known (e.g. [2]).

The question is, given a standard uniform random number, how can that be accurately converted to a random number from an arbitrary distribution? Answers were found in the earliest days of computers, as early as John von Neumann in 1945, and some of those methods are still in use [3]. The early methods, of necessity, used essentially no computer memory. However, with the vast computer memories now available, not using available memory is wasting it, and time-for-space tradeoffs that support arbitrary distributions have been published recently [3].

In this paper we (1) provide further background on the kinds of probability distributions needed in random number generation, (2) introduce a new aspect we call "sub-distribution sampling," (3) provide detailed algorithms for generating numbers from the kinds of distributions that arise in practice, and (4) compare processor time required of selected methods.

## 2. Density and cumulative functions

The "probability density function" is the most familiar representation, but it is the corresponding "cumulative prob-

ability distribution" that is employed for generating random numbers. In what follows, for brevity we shall write "density function" for the probability density function, "cumulative function" for the cumulative probability distribution. Also, we shall write "inverse function" or "inverse cumulative function" for the inverse of the cumulative probability distribution.

The vertical axis of a density function is the "probability density" that a corresponding value on the horizontal axis will be drawn by chance. This is not the actual probability of it being drawn, since in many distributions, the probability of any precise value being drawn is 0, amongst the infinite set of possible values. The probability density can take on any value from zero to infinity. If one considers the probability that a single random number drawn from the distribution will fall between two specified values, such as between 1.49 and 1.51, that probability is equal to the average value of the density function over that interval, multiplied by the width of the interval. Or what is equivalent, it is the area beneath the density function from the left to the right endpoint of the interval. Thus the area under the entire density function becomes 1. Modes of the distribution correspond to peaks in the density function, while the median and the mean are not immediately visible.

The cumulative function carries the same information as the corresponding density function, but in a different form. The vertical axis is the probability that a number less than or equal to the corresponding value on the horizontal axis will be drawn from the distribution. The vertical axis of a cumulative function is thus constrained to a range of 0 to 1. While the density function can have peaks and valleys, the cumulative function is either level or increasing. The cumulative function is one degree smoother than the corresponding density function, since it is its integral. The median of the distribution corresponds to the half-way point on the vertical axis ($y = 0.5$) and modes of the distribution correspond to places of maximum slope. The mean is not immediately visible in the cumulative function.

See Figures 1A through 1D for examples of four cumulative functions shown above their corresponding density functions.

## 3. The inverse cumulative technique

Using the inverse cumulative function to generate random numbers from any desired distribution is a long-recognized approach [4], and the idea is straightforward. Start with a uniform random number $U$ between 0 and 1, locate that number on the vertical axis of the cumulative function, and find which value on the horizontal axis corresponds. That value is the desired random number. See the arrows in Figures 1A through 1D, where $P = 0.25$ on the vertical axis maps to $-1$, 4.25, 1.0638, and 2, respectively, on the horizontal axes of the various cumulative distributions. Note that domains from which a random number generator may never select random values correspond to perfectly level stretches in the cumulative function (Figure 1B and 1D), and that discrete random numbers correspond to vertical jumps (Figure 1D). Also, as Devroye points out [4], using the same value of $U$ like this for multiple distributions, or using correlated values $f(U \pm \rho_i)$, with $\rho_i$ being a small random variate, draws correlated random numbers from multiple distributions. Likewise, using uniform random numbers equal to $U$ and $f(1-U \pm \rho_i)$ will generate negatively correlated numbers from any distribution, or pair of distributions. Both kinds of correlation can be useful in applications.

The inverse technique is simple graphically but not necessarily numerically, for it involves computing the inverse function. A few classical distributions, such as the exponential, Cauchy, and Pareto, have inverse functions that can be written in terms of elementary functions [4] and therefore computed directly. In general, however, inverting an arbitrary cumulative distribution is computationally difficult, in which case generating the corresponding random numbers is slow.

Hörmann and Leydold [5] explained how to improve the speed by computing the inverse function only once when the simulation begins, then approximating it by interpolation as the simulation proceeds. This can be done by computing a series of $x, y$ pairs from the cumulative function, then exchanging $x$ and $y$ and fitting the $y, x$ pairs to obtain the inverse. Approximating the inverse in this way is reasonably fast and can be made as accurate as desired for many distributions.

In this paper we exhibit a variation that is simpler yet still fast, and that supports the full set of functions required. We approximate not the inverse cumulative function, but the cumulative function itself, with quadratic pieces that join smoothly, without sharp corners, where one piece ends and the next begins, or which approximate smoothness as accurately as we need. We can then invert the function piecewise as the simulation runs, since inverting quadratic functions involves only the quadratic formula. We use this because it directly supports "sub-distribution sampling," which is useful in general discrete-event simulations [1] and other micro-scale computations. It also supports efficient random numbers from discrete density functions (e.g. Poisson), step functions (e.g. empirical histograms), or piecewise linear (e.g. empirical function estimates). See Figure 1 for examples. The corresponding algorithms (appendix) are short and relatively simple.

The algorithms we present can be made to reproduce known distributions as accurately as desired, although in many cases such accuracy is superfluous. Classical distributions may be used because they are known and available, even though they may not closely approximate the distributions of interest. For instance, waiting times in an individual-based computer model may be selected from an exponential distribution that is used largely as a convenience, for correspondence with differential equation models. More-

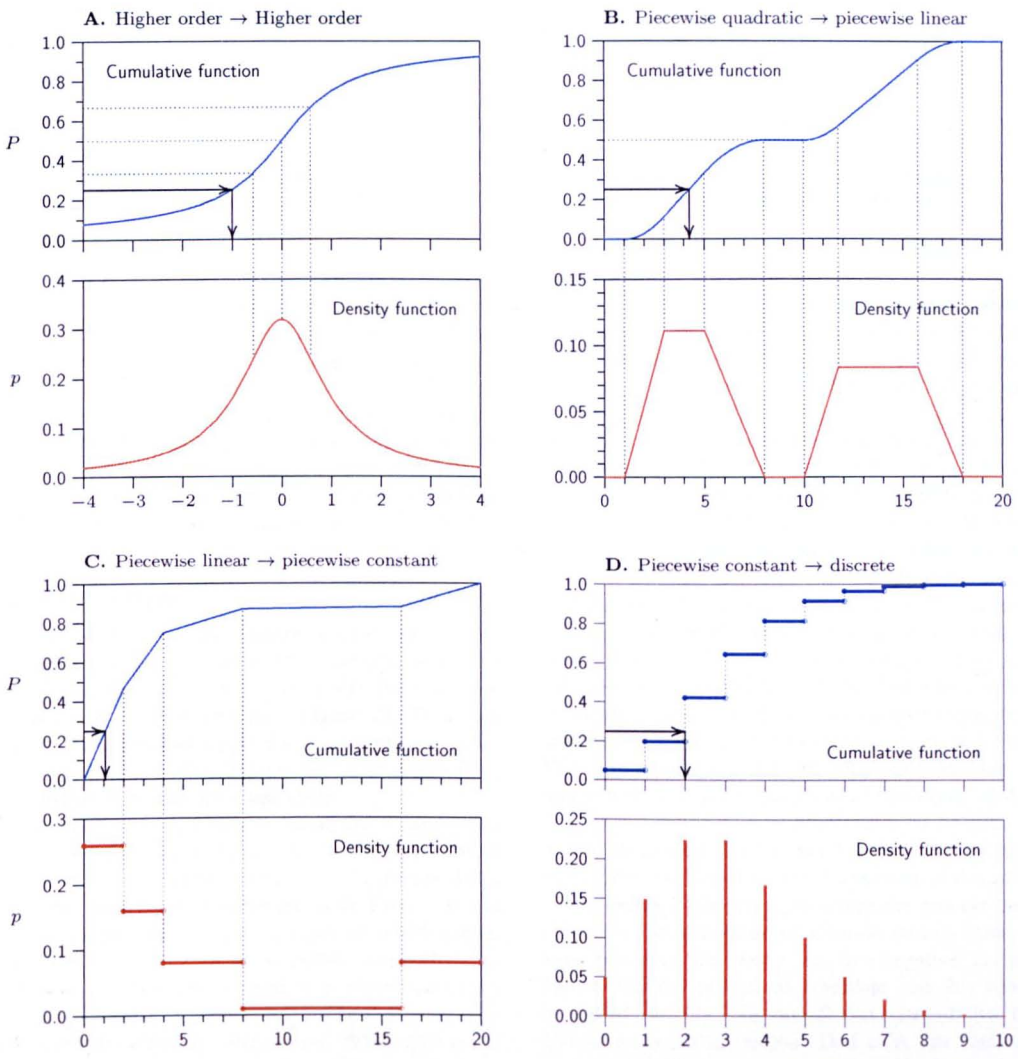


**Figure 1. Random number generation at four levels of continuity.** *Horizontal axes represent values of random variables. The vertical axis for cumulative functions represents the probability that the random variable is less than or equal to the corresponding value $x$ on the horizontal-axis, and for density functions represents the average probability of a random number being in an arbitrarily small surrounding interval.* **(A)** *A standard Cauchy distribution, which does not converge to a mean. That renders it difficult to replicate purely with finite approximations. It represents the distribution of slopes that are associated with random angles. Other classical distributions, such as normal, lognormal, exponential, and chi-square, which do converge to means, are also higher order like this.* **(B)** *A hypothetical empirical distribution of two lobes developed from a piecewise linear density function, with each lobe equally likely. See Figure 2 for data structures of this example. Multi-modal distributions like this arise, for example, in carbon-14 dating, though those functions are typically more complicated than this illustration.* **(C)** *A hypothetical empirical distribution of failure rates of machine parts. Probability of failure is relatively high for new parts, drops to a minimum at intermediate ages, then rises again when parts get older.* **(D)** *A Poisson distribution with mean of 3. This discrete distribution represents frequencies of co-occurrence of random events with delta-function spikes.*

| $i$: | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| $X[i]$: | 0 | 1 | 3 | 5 | 8 | 10 | 11.75 | 15.75 | 18 | 20 |
| $Y[i]$: | 0 | 0 | .1111 | .3333 | .5000 | .5000 | .5729 | .9063 | 1 | 1 |
| $Q[i]$: | 0 | 0 | 1/9 | 1/9 | 0 | 0 | 1/12 | 1/12 | 0 | 0 |

**Figure 2. Data structures.** *Three one-dimensional arrays, $X[i]$, $Y[i]$, and $Q[i]$ carry the $x$ values, the cumulative $y$ values, and optionally the density $y$ values, respectively, for a probability function. This example corresponds to the distribution of Figure 1B. In practice, arrays are often much larger than this illustration.*

over, a simulation may benefit from using data directly, such as lifetime survivorship data available for multiple years and geographic areas, rather than approximating the data with standard distributions, then handling those distributions exactly. On top of all this, many empirical distributions are poorly known. Important distributions, such as the duration of infectiousness for certain asymptomatic diseases, may not be known to more than one digit of accuracy. Therefore, we are not recommending these algorithms as general solutions for all cases, but they can be particularly useful in practical cases where even the density function may be imperfectly known.

## 4. Data structures

Each cumulative function and, where needed, each corresponding density function, is defined by a set of two or three matched one-dimensional arrays that define the functions at selected points in their domains (Figure 2). These are constructed by the user and supplied to the algorithms, either to approximate classical distributions but more typically to represent distributions derived empirically.

The first array is $X[i]$, which contains the $x$ values that cover the range of random values to be generated. With entries in one-to-one correspondence, $Y[i]$ contains the $y$ values for the cumulative distribution, with $Y[0] = 0$ and $Y[i_{\max}] = 1$. Optionally, $Q[i]$ can be supplied, which carries the derivatives of $Y[i]$. That is, it carries the probability density function. When $Q[i]$ is used, it is piecewise linear, making $Y[i]$ piecewise quadratic. With a sufficient number of points, this can model any distribution. When $Q[i]$ is not used, $Y[i]$ is taken to be piecewise linear or piecewise constant and the implied $Q[i]$ is a piecewise constant histogram. That corresponds to many empirical distributions, such as life tables.

## 5. Algorithms

The appendix defines four algorithms sufficient to accomplish the goals of this paper:

1. *Cinverse*: Evaluates inverse cumulative functions.
2. *Cforward*: Evaluates general functions.
3. *Cdiscrete*: Evaluates discrete functions.
4. *Cintegral*: Prepares cumulative functions.

Algorithm 1, *Cinverse*, accepts a probability, typically as a uniform random number, plus a starting point $g$, and returns a corresponding random number for the cumulative distribution defined by $X$, $Y$, and optionally $Q$. It calls *Cforward* and *Cdiscrete* to accomplish its work. The starting point $g$ is a minimum value for the random number returned. It is typically set to the minimum value of the distribution, $X[0]$, but may be a greater value. In that case numbers are selected from the remainder of the distribution, starting at $g$, transformed as $Z(x) = \big(Y(x+g) - Y(g)\big)/\big(1 - Y(g)\big)$. This we call "sub-distribution sampling," which has such uses as initializing a population with individuals of various ages before the simulation begins, or handling immigrants to a population of random ages and projecting their remaining lifetimes. Note that "memoryless" distributions (the exponentials) are invariant under this transformation.

Algorithm 2, *Cforward*, determines the value $y$ of a function at a specific point $x$, here used to obtain the value of the cumulative function at a random value $x$. It calls *Cdiscrete* to accomplish its work. The function is defined by tables $X$ and $Y$ of values for corresponding points, and optionally a table $Q$ of derivatives of $Y$ at each point in $X$. Values in tables $X$ and $Y$ are increasing. Passing $Y[i]$ as a function of $X[i]$ processes forward functions, while passing $X[i]$ as a function of $Y[i]$ processes their inverses. ($Y$ need not be increasing if $Q$ is not supplied, though processing non-cumulative functions is not a purpose of this subroutine.)

Algorithm 3, *Cdiscrete*, is where the process begins. For cumulative distributions of discrete density functions that have precisely one entry per non-negative integer, as in Figure 1D, the process is complete and this routine may be called directly. For non-discrete cumulative functions, *Cinverse* is called instead. That calls this routine to start and then handles any necessary inverse linear or quadratic interpolation. This routine is nothing more than a recursive binary search that processes ordered tables of $n$ entries in time proportional to $\log_2 n$.

Algorithm 4, *Cintegral*, creates a cumulative function on demand, given an array of points $X[i]$ and a density function $Q[i]$. It integrates using quadratic interpolation. It is not needed for linear interpolation or discrete distributions, where the cumulative distribution is simply the sum of the density function values, as in Figure 1D.

## 6. Timing

In timing tests, drawing $10^7$ random values from a Poisson distribution with a mean of 3, as in Figure 1D, averaged 0.89 seconds on a 2.4 GHz processor for both a standard

iterative method [6] and for this method. For means greater than 3, the standard iterative method was slower. Generating numbers from discrete distributions like the Poisson is exact and is particularly fast because no interpolation is needed. In a continuous case, drawing $10^7$ random values from a lognormal distribution required 1.14 seconds overall with the standard Box–Muller method and 1.07 seconds with this method, on the same processor. The 6% improvement in speed is not significant, but it is significant that a general method like this is competitive with the speed of classical custom methods. Timing tests showed that the recursive binary-search of Algorithm 3 could be speeded by about 10% by using an equivalent iterative algorithm, at the cost of a little greater complexity.

## 7. Discussion

Devroye [4] listed six factors for assessing general random number generation, (1) speed, (2) initialization time, (3) memory requirements, (4) portability, (5) generality, and (6) simplicity/readability. He pointed out that the sixth factor is the most neglected.

We find that his first factor, speed, is as important now as then. Despite the enormous increase in computer speeds, computers now labor under proportionally longer simulations. The second, initialization time, is less important, since it typically vanishes into the time for the simulation itself. Moreover, with large memories spaces that can be allocated, cumulative distributions may be precomputed and read from files, so initialization times become essentially zero. The third factor, memory, is largely irrelevant now—except insofar as it increases complexity—thanks to the then-incomprehensible rise in computer memory sizes. The fourth factor, portability, is now easily had with careful coding, for which countless examples exist. Therefore, speed, generality, and simplicity remain as important factors.

The algorithms we present satisfy the simplicity factor well. They require under 60 lines of computer code altogether and are completely exhibited here, with code and accompanying meta-code. They are also fast, slightly outperforming even well-established methods like the Box–Muller algorithm for drawing numbers for standard distributions like the lognormal. Finally they are general, written to handle any continuous or discrete distributions for which the density function or cumulative function is known.

Modest increases in speed could be obtained in these algorithms at the cost of complexity, trading space for time by storing the inverse cumulative function as a direct-access table, with one entry per lattice point in the probability space $Y$. This would eliminate the binary search, though that is not slow. It would work if the slope of the cumulative function never gets very close to zero. A more modest increase in speed, whenever sub-distribution sampling applies, could result from rescaling the random number differently so that the entire table need not be searched, but only the part covering value $g$ and above. That could eliminate a few calls of binary-search recursion if $g$ was large.

Generality could be increased even further. As written, the algorithms handle all the kinds of distributions shown in Figure 1, but not mixtures of those types—for example, density functions that are discrete in some parts of the domain and continuous in other parts. Such functions are compatible with the algorithms detailed here and the algorithms could be extended to accommodate them, but at the cost of a little complexity, should the need for such hybrid distributions ever arise.

## 8. Conclusions

The algorithms presented here can be incorporated wherever efficient random numbers drawn from arbitrary distributions are needed. These algorithms have been successfully used in a large-scale simulation model developed by one of us (A.K.) for tuberculosis in the UK. Compilable copies of the code described here and related simulation algorithms are available free from the authors upon request.

## 9. Acknowledgements

We are grateful to Todd Lehman and Lori Thomson for discussions and help with the presentation, and to Todd Lehman for timing comparisons of binary-searching iteratively and recursively. This project was supported in part by a resident fellowship grant to C.L. from the UMN Institute on the Environment, by grants of computer time from the Minnesota Supercomputer Institute, and by doctoral research funding to A.K. from the Modelling and Economics Unit at the Health Protection Agency, London.

## 10. Contributions

A.K. conceived the approach to sub-distribution sampling, which was necessary in her discrete-event simulations, and which inspired this effort. C.L. extended the ideas to piecewise quadratic functions and coded the resulting algorithms. A.K. tested and applied the techniques in a large-scale individual-based model for tuberculosis. Both authors contributed to the manuscript.

## References

[1] J. Banks, J. S. C. II, B. L. Nelson, and D. M. Nicol, "Discrete-event system simulation, fourth edition," *Pearson Prentice Hall, New Jersey*, 2005.

[2] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, "Numerical recipes: The art of scientific computing, third edition," *Cambridge University Press, New York*, 2007.

[3] W. Hörmann, J. Leydold, and G. Derflinger, "Automatic nonuniform random variate generation," *Springer-Verlag, Berlin*, 2004.

[4] L. Devroye, "Non-uniform random variate generation," *Springer-Verlag, New York*, 1986.

[5] W. Hörmann and J. Leydold, "Continuous random variate generation by fast numerical inversion," *ACM Transactions on Modeling and Computer Simulation*, vol. 13, pp. 347–362, 2003.

[6] D. Knuth, "Seminumerical algorithms, volume 2, third edition," *Addison-Wesley, Reading, MA*, 1997.

# 11. Appendix

To use the algorithms described in this paper, it is only necessary to understand the entry and exit conditions that appear at the beginning of each, not the code itself. Nonetheless, to allow complete evaluation of the algorithms, and to encourage further development of them, we present them as pseudo-code inspired by and simplified from the programming languages C, R, and Python.

The algorithms are defined with sufficient precision that they can be run, tested, timed, modified, or translated to other languages. Familiarity with a relatively few operators* and with the syntax of flow control (if, for, while, etc.), is sufficient to follow the algorithms. Text copies of this pseudo-code translated into operational C are available from the authors upon request, or from the associated website www.cbs.umn.edu/modeling.

The algorithms assume a uniform random number generator *Rand*, which returns values in the range of 0 to 1, including 0 but not including 1, as is typical for uniform random number generators. *WarnMsg* and *ExitMsg* display error messages and the latter terminates the program.

---

**Algorithm 1. Evaluate inverse cumulative function, with sub-distribution sampling.**

**Upon entry to the algorithm, (1)** $k$ describes the piecewise order of the function: 0=constant, 1=linear, 2=quadratic. **(2)** $y$ contains a value between 0 and 1, representing a probability. **(3)** $g$ is given value in the range $X[0]$ to $X[n-1]$, inclusive. **(4)** $n$ is the number of entries in tables $X$, $Y$, and $Q$. **(5)** $X$ is a table of strictly increasing values in the set of numbers to be generated. **(6)** $Y$ is a table of probabilities, each being the probability that a value will be less than or equal to the corresponding value in $X$. **(7)** $Q$ is a table of probability densities, in effect the derivative of $Y$ at every point in $X$. **At exit,** *Cinverse* returns the value from the given distribution corresponding to probability $y$, starting at value $g$. Note that if $g > 0$, this is the value from the rescaled distribution.

```
real Cinverse(k, y, g, n, X, Y, Q)  integer k, n; real y, g, X[], Y[], Q[];
  integer i; real r, s, d, h, a, b, c, p, w;

  if X[0] > g or X[n − 1] < g: ExitMsg(1);                          1. Check the bounds of both tables.
  if Y[0] ≠ 0 or Y[n − 1] ≠ 1: ExitMsg(2);

  if y < Y[0]: return X[0];                                         2. Handle variables outside the
  if y > Y[n − 1]: return X[n − 1];                                    normal range.

  y → r; g − X[0] → d;                                              3. Rescale the probability value if only
  if d: Cforward(k, g, 0, n − 1, X, Y, Q) → p,                         part of the distribution is to be
    p + r*(1 − p) → r;                                                 sampled.

  Cdiscrete(Y, 0, n, r) → i;                                        4. Bracket the probability value and
  if k = 0: return X[i] − d;                                           return if it is piecewise constant.

  X[i + 1] − X[i] → w;                                              5. If this is piecewise quadratic,
  if k = 2 and Q: (Q[i + 1] − Q[i])/(2*w) → a,                         generate the coefficients of the
    Q[i] → b, Y[i] − r → c;                                            quadratic equation, ax² + bx + c.
  else 0 → a;

  if a: b*b − 4*a*c → s; if s < 0: ExitMsg(3);                      6. If the equation actually has a quadratic
    sqrt(s) → s, (−b + s)/(2*a) → h;                                   term, invert it using the positive root
    if h < 0 or h > w: ExitMsg(4);                                     of the quadratic formula.

  else                                                              7. If it is only linear, invert it
    Y[i + 1] − Y[i] → s;                                               with linear interpolation.
    if s: (r − Y[i])/s → s; else 1 → s;
    s*w → h;

  return X[i] + h − d;                                              8. Return the result.
```

* *The pseudo-code given here is two-dimensional, as in the language Python, so that indentation completely defines the nested structure, with no need for bracketing characters such as '{' and '}'. Variables and function names are italicized and flow control and reserved words are bolded.*

*The assignment operator is represented either as '←' or '→', similar to assigments in R. The compound assignments 'a + 1 → a → b → W[i][j]' and 'W[i][j] ← b ← a ← a + 1' are equivalent, first incrementing a and placing the results back in a, then in b, and then in the i, jth element of the array W.*

*The expression structure 'c? u : v', where c is a condition, u is an if-expression, and v is an else-expression, follows that of C. Using up-tick and down-tick operators to write '↑a', '↓a', 'a↑', and 'a↓' form pre- and post-increments by one, as in '++a', '--a', 'a++', and 'a--' of C.*

*Arrays are indexed as in the language C, starting with 0. Data types are* **'integer'** *and* **'real'***, with the latter specifying floating point. Operator precedence is that of C, with assignments having lowest precedence. Logical operators such as* **'and'** *and* **'or'** *are preemptive, terminating a chain of logical operations as soon as the result is known. Permanent global assignments, as would be represented '#define α β' in C, are rendered as 'α ≡ β'.*

**Algorithm 2. Evaluate general function.**

**Upon entry to the algorithm,** (1) $k$ describes the piecewise order of the function: 0=constant, 1=linear, 2=quadratic. (2) $x$ specifies the independent variable. (3) $i0$ and $i1$ define the first and last entries, respectively, in tables $X$, $Y$, and $Q$. (4) $X$ and $Y$ define the independent and dependent variables, respectively. (5) $Q$ defines the derivative of the function, if $k$ is 2. Otherwise $Q$ is null. **At exit,** $Cforward$ returns the value of the function at point $x$. If $x$ is below or above the range defined in table $X$, the minimum or maximum value, respectively, in table $Y$ is returned.

**real** *Cforward*($k$, $x$, $i0$, $i1$, $X$, $Y$, $Q$) **integer** $k$, $i0$, $i1$; **real** $x$, $X[\,]$, $Y[\,]$, $Q[\,]$;
**integer** $i$; **real** $h$, $s$, $u$, $w$;

| | |
|---|---|
| **if** $k > 1$ **and** $Q = 0$: *ExitMsg*(5); | 1. Check for certain invalid calls. |
| **if** $x < X[i0]$: **return** $Y[i0]$;<br>**if** $x > X[i1]$: **return** $Y[i1]$; | 2. Handle variables outside of the normal range. |
| *Cdiscrete*($X$, $i0$, $i1 - i0 + 1$, $x$) $\rightarrow i$;<br>**if** $k = 0$: **return** $Y[i + 1]$; | 3. Bracket the independent variable and return if piecewise constant. |
| $X[i + 1] - X[i] \rightarrow w$, $x - X[i] \rightarrow u$; | 4. Compute $x$-width and displacement. |
| **if** $k = 2$:<br>$(Q[i + 1] - Q[i])/w \rightarrow s$,<br>$u*(Q[i] + u*s/2) \rightarrow h$; | 5. If a derivative is supplied, compute the $y$-value with quadratic interpolation. |
| **else**<br>**if** $w$: $u/w \rightarrow w$; **else** $1 \rightarrow w$;<br>$w*(Y[i + 1] - Y[i]) \rightarrow h$; | 6. Otherwise interpolate linearly. |
| **return** $Y[i] + h$; | 7. Return the computed $y$-value. |

**Algorithm 3. Evaluate discrete function.**

**Upon entry to the algorithm,** (1) $T$ addresses a strictly increasing table of two or more values. (2) $b$ indexes the beginning entry to be examined in $T$. (3) $n$ defines the number of entries to be examined in $T$, at least 2. (4) $v$ specifies the value to be located in $T$, with $T[b] \le v \le T[b + n - 1]$. **At exit,** $Cdiscrete$ indexes the local pair of table entries containing $v$, such that $T[loc] \le v \le T[loc + 1]$.

**integer** *Cdiscrete*($T$, $b$, $n$, $v$) **integer** $b$, $n$; **real** $T[\,]$, $v$; **integer** $m$;
$(n + 1)/2 - 1 \rightarrow m$;
**return** $m \le 0$? $b$: $v < T[b + m]$? *Cdiscrete*($T$, $b$, $m + 1$, $v$): *Cdiscrete*($T$, $b + m$, $n - m$, $v$);

**Algorithm 4. Prepare cumulative function.**

**Upon entry to the algorithm,** (1) $n$ is the number of entries in tables $X$, $Y$, and $Q$. (2) $X$ is a table of strictly increasing values in the set of random numbers to be generated. (3) $Y$ is a table to receive the cumulative function associated with the corresponding values in $X$. (4) $Q$ is a table representing the piecewise linear density function associated with the corresponding values in $X$. **At exit,** (1) $Cintegral$ returns the number of entries in $Y$ up to and including the first entry that saturates at 1. (2) $Y$ contains the piecewise quadratic cumulative distribution function associated with the corresponding values in $X$.

**integer** *Cintegral*($n$, $X$, $Y$, $Q$) **integer** $n$; **real** $X[\,]$, $Y[\,]$, $Q[\,]$;
**integer** $i$, $m$; **real** $w$;

| | |
|---|---|
| **for** $i$ **from** 1 **to** $n - 1$:<br>**if** $X[i - 1] \ge X[i]$: *ExitMsg*(6); | 1. Make sure the domain is strictly increasing. |
| $0 \rightarrow Y[0]$, $1 \rightarrow m$;<br>**for** $i$ **from** 1 **to** $n - 1$:<br>$Y[i - 1] \rightarrow Y[i]$, $X[i] - X[i - 1] \rightarrow w$;<br>$(Q[i - 1] + (Q[i] - Q[i - 1])/2)*w + Y[i] \rightarrow Y[i]$;<br>**if** $Y[i] > 1$: *WarnMsg*(7), $1 \rightarrow Y[i]$;<br>**if** $Y[i] < 1$: $i + 1 \rightarrow m$; | 2. Integrate the probability density function to obtain the cumulative distribution function. |
| **if** $Y[n - 1] \ne 1$: *WarnMsg*(8);<br>**return** $m + 1$; | 3. Make sure it adds and return the number of operational entries. |

Intn'l Conf. Information & Knowledge Engineering, IKE12: 47–53, Proceedings, July 2012

# The Centinel Data Format: Reliably Communicating through Time and Place

**Clarence Lehman[1], Shelby Williams[2], and Adrienne Keen[3]**
[1]University of Minnesota, 123 Snyder Hall, 1474 Gortner Avenue, Saint Paul, MN 55108, USA
[2]University of Minnesota, 100 Ecology, 1987 Upper Buford Circle, Saint Paul, MN 55108, USA
[3]London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK

*"A library book lasts as long as a house, for hundreds of years."*
—Thomas Jefferson, 1821

**Abstract**—*A common experience among scientists and engineers is storing and sharing data, the capacity for which has advanced immensely since laboratory notebooks were only paper and ink. However, since that time, the sustainability of data has decreased. Even though our digital data should be safer and more secure than ever, a continuing cascade of obsolescence in computer media and software can actually make it less so. Here we outline an ensemble of free tools and techniques that we call "Centinel," designed to manage, communicate, and archive digital datasets. Rather than embedding error-correcting codes as part of the computer media, Centinel exposes them and places them with the data and metadata. Thus even printed copies of the data form reliable storage media that can last indefinitely without intervening attention. Centinel complements standard methods for data sustainability, such as data migration. Unified approaches, as we outline here, benefit reliability and longevity of data.*

**Keywords:** database, data archive, data longevity, data reliability, error correcting codes

## 1. Introduction

In 1815 began one of the largest scientific data collection projects ever launched [1]. Legions of surveyors walked regularly spaced transects along 2,500,000,000 meters of the Louisiana Territory, recording the biological species, geographic locations, and diameters of selected trees near periodic sample points—plus other information on soils, vegetation, and boundaries of wetlands. For almost a century the survey continued. Now, another century after the last data were recorded, the results form one of the most visible efforts ever, organizing the rural landscape into square sections along those transects. The results also form one of the best preserved and widely available datasets ever. Think of which present datasets, in your personal experience, are guaranteed to be extant and usable well into the 22nd century.

A large part of the reason the survey data survived was that it was recorded on paper and protected at many different governmental sites. In the meantime, technology changed immensely. Computers emerged and increased in capacity so relentlessly that the Library of Alexandria's ancient charge of organizing and cataloging all human knowledge began to draw within reach. Global access to digital data can make that knowledge available to all. Large-scale private enterprises are aiming at this goal, but individuals in academia and industry are established sources of knowledge and therefore have a special role in achieving this.

Here we are addressing that role—of scientists, engineers, and others who collect empirical data, share it, and want to preserve it for the future. In this report we explain how digital computer techniques of today combine naturally with paper methods of prior centuries to create a form of digital storage that can reliably persist into future centuries and improve electronic processing today.

## 2. What Centinel is and is not

The general topic that Centinel addresses has been long discussed (e.g., [2] [3] [4] [5] [6] [7] [8] [9]) and a complete solution is not yet available. Centinel combines the words "century" and "sentinel," guarding data for extended periods. One goal for Centinel is to ensure that the digital data it encodes will be accessible in a century or more, without the need for care and intermediate steps by humans. A second goal is to protect data over a shorter term, from the time of initial creation to the time of final processing. Centinel works by (1) keeping all metadata with the data, (2) protecting data with line-by-line error correcting codes, (3) providing a format easily readable by humans as well as computers and scanners, (4) supporting a reliable digital format that works on any media, including paper and verbal communications, to protect data from unintentional alteration, and (5) supplying an extensible, self-defining format with accompanying tools that help computer programmers know that the data entering their programs are correct. Centinel is an approach to data management, but also a set of basic computer utilities for writing, reading, editing, separating, joining, ordering, and aligning data. It avoids structures that are error prone

```
6674844762232577  Keyword SpAbbr: Abbreviations for species names. Abbreviations contain the
0629561874138616    first three letters of the genus name followed by the first three letters
0211050455008008    of the species name. The full species names are recorded with their
5515307245627135    abbreviations in table "species codes" at the end of the chapter.
5915322805104717  Keyword Date: Date species was collected. Format year-month-day.
1453182442695072  Keyword CollID: Unique code assigned to species sample collected.
1382423906566782  Keyword Cover: Estimated canopy cover, in percent. Dashes indicate missing
5953391885352618    data. (See "methods" at the end of the chapter.)
0748783303437946  Keyword HtMax: Maximum height, in meters. Dashes indicate missing data.
0229302812296440    (See "methods" at the end of the chapter.)
0602554115737437  Keyword HtMin: Minimum height, in meters. Dashes indicate missing data.
0229302812296440    (See "methods" at the end of the chapter.)
0000000000000000
1976160343505769  :Site :Code     :SpAbbr  :Date       :CollID   :Cover :HtMax :HtMin
4554847814214755  :1600 :P1600D04 :Abibal  :1989-08-21 :AMB00555 :    - :    5 :    5
2645745581124348  :1600 :P1600D01 :Abibal  :1989-08-21 :AMB00604 :    2 :    1 :    1
1076375677295808  :1600 :R1600EA  :Abibal  :1989-08-24 :AMB00666 :    3 :    1 :    1
2000445884315808  :1600 :R1600EA  :Abibal  :1989-08-24 :AMB00668 :    5 :    6 :    6
0582355170295008  :1600 :R1600EA  :Abibal  :1991-08-05 :AMB01719 :    2 :    2 :    2
1485325476235008  :1600 :R1600EA  :Abibal  :1991-08-05 :AMB01722 :    4 :    6 :    6
4100414960104041  :1600 :R1600EA  :Acerub  :1991-08-05 :AMB01503 :    2 :    2 :    2
5773084583093978  :1600 :P1600B01 :Agrsca  :1989-08-25 :AMB00456 :    3 :    2 :    2
4766066289426272  :1600 :P1600D01 :Amerot  :1991-06-17 :AMB01439 :    2 :    2 :    2
```

**Figure 1.** *Excerpt of a sample Centinel data file from a large ecological database, with metadata above and error correcting codes called "centinels" at left. Here colons separate columns rather than vertical bars. In the Centinel structure, error detection and correction stays with the data rather than with the computer medium.*

and supports good data management practices, for example as outlined in [10] and [11].

Centinel is not intended to substitute for large-scale interactive databases undergoing continual manipulation, such as in PostgreSQL, MySQL, or Access. It is, however, a good format for long and medium-term retention of such databases, as Centinel format can be readily exported from them through simple utility programs, and conversely, imported through conventional means or by scanning. Nor is Centinel intended as a complete solution to the problem of storing all data at national and international scales (e.g. [12] [13]), but rather as a solution for individual research and development groups to help maintain their data.

The Centinel format shown in Figure 1 supports the movement of data through place and time. A dataset documented sufficiently with complete descriptions as its metadata, and protected with error correcting "centinels," can be transmitted to another researcher in a distant place without separate documentation and time spent explaining the data, or equivalently it can be transmitted forward to another researcher in the distant future. In other words, it can be archived. Instead of error detecting and correcting codes being applied to the storage media, as is the common method today, codes in Centinel are applied to the data themselves, and stay with the data through all media changes. That simple but unusual characteristic fills a gap in existing data methods and provides confidence in the data across distant places and times. Multiple printed copies of the data can be stored throughout the world and scanned with optical character recognition in the remote future. The centinels, checked automatically against the scanned results, are the essential link to data reliability.

As in some other databases, Centinel has multiple equivalent formats, which we call "singular," "columnar," and "mixed." Long lines of data in singular format can extend onto new lines, indented as in Figure 1. Here is a simpler file in singular format:

```
Class:  1
ID:     123
Age:    21
Region: SSA

Class:  1
ID:     47
Age:    7
Region: UK

Class:  2
ID:     723
Age:    70
Region: US
```

Below are the same data in columnar format:

```
| Class | ID   | Age | Region
| 1     | 123  | 21  | SSA
| 1     | 477  |  7  | UK
| 2     | 723  | 70  | US
```

And below is mixed format:

```
Class: 1
| ID   | Age | Region
| 123  | 21  | SSA
| 477  |  7  | UK

Class: 2
| 723  | 70  | US
```

These formats are interchangeable. The choice is a matter of space, readability, and ease of processing. All software written to handle Centinel data should process the three formats equally.

Printed copies of data with error-correcting centinels need not be limited to small data sets. For example, the genome of the fruit fly (*Drosophila melanogaster*), represented with one base-64 symbol for each of its 47 million codons, would require approximately 6000 pages—not absolutely prohibitive to print for an important, expensive dataset. By comparison, the King James Bible is 4.3 million characters, about one-tenth of this genome, and more than one copy of that work has been printed.

## 3. How Centinel works

Centinel protects data when they are complete and ready to be archived. But it can also be used when the data are first entered, to guard against accidental modifications of datasets undergoing incremental change.

To explain how Centinel works, we must consider what it means for data to be digital. Two properties are essential. First, the data must be represented by "symbols" that have only a finite number of states. Second, the shapes of any two distinct symbols must be separated by a sufficient gap, so that a symbol for one datum does not, except very rarely, degrade into a different symbol for a different datum. Symbols can take various forms—binary 0 and 1 encoded electronically in computer memories are one example of digital data. The Arabic numerals 0–9 printed on paper are another. With these ideas in mind, Figure 2 shows analog versus digital representations of a function, $y = f(x)$.

An analog form on paper could take the form of a graph, Figure 2A. The value on the vertical axis varies smoothly, and can be read to reasonable accuracy with a ruler and a careful eye. However, each time the graph is copied, its accuracy diminishes. The curve becomes successively blurred, the right side may get slightly skewed with respect to the left, and so forth. In contrast, the entire curve in digital form is defined by coefficients, Figure 2B. When this digital version is copied by re-typesetting, it will not degrade, for the individual symbols will be recognized for what they are and reproduced intact. A new font may even change '$x$' to 'x', but the meaning of the symbol will remain.

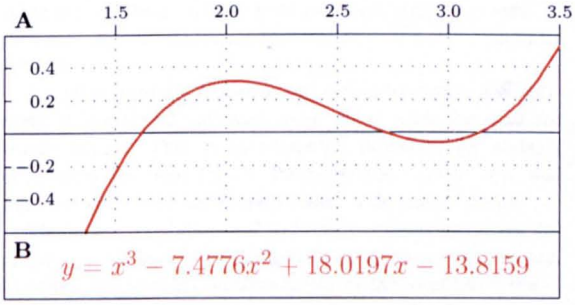


**Figure 2.** *Non-electronic analog and digital data for the same curve. Printed copies of the digital data (B) will not degrade over time as will the analog version (A) of the same data.*

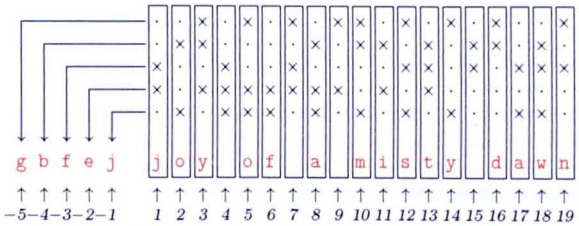


**Figure 3.** *Error-correcting "centinels" (left) for a 19-character message (right). Each centinel covers a distinct combination of columns, such that any unmatched centinels identify which column is in error and how to correct it. (See code in the appendix for details.)*

Thus digital data are not at all restricted to electronic media, but paper can carry digital data as well, and has done so for millennia. Moreover, some of the most common digital information read by computers today is recorded directly on paper, plastic, metal, and other substrates. The ubiquitous bar code is a case in point, though bar codes are not human-readable as Centinel-protected data are.

A significant separation between symbols in appearance or physical state keeps unavoidable small degradation in information from changing the message, because one symbol does not easily degrade into another. However, separation of symbols is not enough. For highest reliability, error correcting codes must be applied to the digital data to prevent rare alterations of one symbol into another from changing the message, except with negligibly small probability.

Centinel uses a "Hamming code" for arbitrary symbols, a generalization of the original code [14] for binary digits. Such codes we call "centinels," and they appear at the left of each line, at the end of each printed page, and at the end of each file. They can correct any single-symbol error in a line and detect any two-symbol errors. In addition, with high probability they detect multiple-symbol errors, including errors in the centinels themselves.

Each symbol is assigned a small integer and the integers for a given subset of columns are summed. The sum, modulo the number of symbols, is translated back to a symbol, as in columns $-1$ to $-5$ of Figure 3. This is repeated for carefully chosen subsets of columns which allow errors to be located and corrected. Then the results are translated to decimal form, as in Figure 1, to mask the actual random combinations of symbols, which by happenstance can spell out any word.

Complete details are in the Centinel algorithms (appendix). These details are part of the metadata and should be included with archived data.

## 4. Comparison with other approaches

A standard approach to data archiving is a rigorous effort of continually transferring data from old media and old software to new, before the old media and software become completely obsolete—keeping the data "alive" so to speak.

That is called "migration" [12]. It is a practical, well-tested method, though it can be labor intensive and susceptible to catastrophic failure.

Successful migration requires a central discipline maintained over long periods. Any lapse in the chain of migration will result in the complete loss of data. Successful migration will be practical for large, well funded data sets. However, for many small data sets, discipline and funding can easily lapse over long periods of time.

Timing is key, as migration must take place while (1) machines that can read the media still exist, (2) programs encoding the information are still operational, and (3) the media and the information stored on it have not deteriorated.

It follows that the best chance of success in data preservation will be for (1) media that require no advanced or specialized machinery to read them, (2) formats that require no complex computer programs to process them, or at worst require the simplest programs that can be described completely in a few pages of text, as in the Centinel algorithm (appendix), and (3) media and encoding methods that will themselves last a century or more. Centinel allows data preservation with a single migration.

A second method is called "encapsulation." Fully successful migration to new media will be worthless if the software that accesses the data ceases to exist. For example, an organization producing software may go out of existence and no other organization may support the old format. This has happened repeatedly in the history of computing. Encapsulation aims to include with the data all software that accesses the data, in a form that can be translated to future machinery. That is, of course, easiest when the corresponding software is as limited as possible.

Two other methods proposed for data archiving are "emulation" and "technology-preservation." In emulation, the complete hardware and software architectures to retrieve the data are migrated forward with the data and "emulated" on the future system. That practice was widespread and successful among mainframe computers in the 1960s, where one generation of computers would emulate the hardware of the generation before. But as computers become increasingly complex in their architecture and operating software, it becomes difficult to make this practical into the indefinite future.

In technology-preservation, the actual hardware and software is preserved, museum-style, along with the data for future access. This is problematic, however, for today's computers are built for the moment, not built to last, and may not even boot up properly after a decade of disuse.

Therefore, emulation and technology-preservation are not related to Centinel, but migration and encapsulation are. Centinel implements encapsulation in the simplest form—under 100 lines of code (appendix)—and with a single migration, creates digital documents that last as long as possible—up to a century or more.

## 5. Suggestions

In conclusion, we offer the following: (1) To keep electronic data safe, prepare early for archiving. (2) Archive data in the simplest formats possible. (3) Document data to the highest standards. (4) Associate documentation directly with the data it describes, ideally in the same file. (5) Keep multiple copies in separate locations. (6) Regularly convert working files from proprietary databases to archival format. (7) Keep printed copies of critical data, with Centinel-like guard symbols and documentation for future recovery.

For full details and utility programs supporting this project, see www.cbs.umn.edu/centinel.

## 6. Acknowledgements

We thank Eville Gorham, Jan Janssens, Todd Lehman, Eric Lind, Richard McGehee, David Tilman, Richard Barnes, and all others who lent help and encouragement during this ongoing project. This project was supported in part by a National Science Foundation LTER grant to David Tilman and by a University of Minnesota database grant to Eville Gorham.

## References

[1] L. A. Schulte and D. J. Mladenoff, "The original US public land survey records, their use and limitations in reconstructing presettlement vegetation," *Journal of Forestry*, vol. 99, pp. 5–10, 2001.

[2] J. Rothenberg, "Ensuring the longevity of digital documents," *Scientific American*, vol. 272, pp. 42–47, 1995.

[3] A. Waugh, R. Wilkinson, B. Hills, and J. Dell'oro, "Preserving digital information forever," *Proceedings of the Fifth ACM Conference on Digital Libraries*, pp. 175–184, 2000.

[4] D. Butler, "The future of electronic scientific literature," *Nature*, vol. 413, pp. 1–3, 2001.

[5] C. Tristam, "Data extinction," *Technology Review*, vol. 105, pp. 37–42, 2002.

[6] K.-H. Lee, O. Slattery, T. Lu, R. McCrary, and Victor, "The state of the art and practice in digital preservation," *Journal of Research of the National Institute of Standards and Technology*, vol. 107, pp. 93–106, 2002.

[7] S. Ong, "Worm storage is not enough," *IBM Systems Journal*, vol. 46, pp. 363–369, 2007.

[8] U. Duerig, "High density multi-level recording for archival data preservation," *Applied Physics Letters*, vol. 99, p. 023110, 2011.

[9] J. Marberg, "Towards SIRF: Self-contained information retention format," *Proceedings of the Annual International Systems and Storage Conference, Haifa, Israel*, 2011.

[10] E. T. Borer, E. W. Seabloom, M. B. Jones, and M. Schildhauer, "Some simple guidelines for effective data management," *Bulletin of the Ecological Society of America*, vol. 90, pp. 205–214, 2009.

[11] M. C. Whitlock, "Data archiving in ecology and evolution: Best practices," *Trends in Ecology and Evolution*, vol. 26, pp. 61–65, 2011.

[12] S. Rabinovici-Cohen, M. E. Factor, D. Naor, L. Ramati, P. Reshef, S. Ronen, J. Satran, and D. L. Giaretta, "Preservation datastores: New storage paradigm for preservation environments," *IBM Journal of Research and Development*, vol. 52, pp. 389–399, 2008.

[13] H. Heslop, S. Davis, and A. Wilson, "An approach to the preservation of digital records," *National Archives of Australia. Link at http:// www. naa. gov. au/ recordkeeping/ er/ digital_ preservation/ summary. html or http:// www. naa. gov. au*, 2000.

[14] R. W. Hamming, "Error detecting and error correcting codes," *The Bell System Technical Journal*, vol. 26, pp. 147–160, 1950.

[15] B. Kernighan and D. Ritchie, "The C programming language," *PrenticeHall, Englewood Cliffs, NJ*, 1978.

## 7. Appendix: The Centinel algorithm

The complete algorithm that encapsulates Centinel files is given here in a subset K&R C [15]. The material below, together with Kernighan and Ritchie's book, should allow the algorithm to be transcribed into future programming languages and the data to be extracted from Centinel files as long as the printed form is extant.

The algorithm adds an error-correcting code to each line of a text-based file, another to each page, and a third to the entire file. Each output line begins with a decimal error correcting code guarding that line, and also guarding the error correcting code itself, then the text of the line. In printed form another decimal code guards the entire page and a third guards the entire file.

In computing the error correcting code, leading and trailing white space is skipped, multiple blanks count as a single blank, and end-of-line codes are not counted. The code at the beginning of the line is not counted either. The assignment between symbols and numbers is specified in array $s$ below, where 'a' is number 1, 'b' is number 2, 'A' is number 27, and so forth. Any similar assignment could be substituted.

In the algorithms below, flow control and reserved words are bolded, variables and function names are italicized, and certain operations such as '<=', '>=', '!=', and '==' are displayed in a mathematical form as '$\leq$', '$\geq$', '$\neq$', and '$\equiv$', respectively.

---

**DATA STRUCTURES**

```
#define C      256      1. Maximum character code plus 1.
#define L      120      2. Maximum data length, excluding guard symbols.
#define G      8        3. Number of guard symbols.
#define COL    9        4. Number of symbols columns displayed on the page.
#define PAGEL  50       5. Number of lines per page.
#define IDENT  127      6. Identity symbol.

char s[] =                                   7. Character set available for present application.
  "_abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMNOPQRSTUVWXYZ0123456789"
  " .,;:!?+-*/\\=\"'()[]{}<>^&%|";

int nchar;                                   8. Maximum number of characters in present application.
char seq[C];                                 9. Sequence number for each symbol in the set.
char f[C][C];                               10. Modulo sum and difference tables.
char ptn[L][G + 1];                         11. Pattern of guard symbols for each position.
int pagef = PAGEL;                          12. Number of lines on first page.
int pages = PAGEL;                          13. Number of lines per subsequent page.

int ipage = 0;                              14. Page index.
int ifile = 0;                              15. File index.
char in[L + 1];                             16. Input line.

char line[L + 1], page[L + 1], file[L + 1]; 17. Current line, page, and file.
char guard[G + 1];                          18. Guard symbols, individual characters.
```

---

**END OF PAGE**

**Upon entry to the algorithm, (1)** *page* contains a list of symbols representing the current page. **(2)** *ipage* indexes the next entry for the page. **(3)** $a$ is set if a blank line should follow the code, indicating end of page. (This is not used on the last page of the file, because the code for the entire file follows immediately.) **At exit, (1)** Guard symbols for the page are displayed. **(2)** *guard* is destroyed. **(3)** *ipage* is set to zero.

```
seqpage(a) int a;
{
  if (ipage ≡ 0) return;
  page[ipage] = 0; ecc(guard, page);
  seqn(guard, "; ", ""); if (a) printf("\n");
  ipage = 0; }
```

**MAIN PROGRAM**

```
main(argc, argv) int argc; char *argv[];
{ char c; int i, j, k;

  if (argc > 1)
  { pagef = atoi(argv[1]);
    if (pagef < 2 || pagef > 100) pagef = PAGEL;
    pages = pagef; }

  if (argc > 2)
  { pagef = atoi(argv[2]);
    if (pagef < 2 || pagef > 100) pagef = PAGEL; }

  s[0] = IDENT;
  for (i = 0; s[i]; i = i + 1) seq[s[i]] = i;
  nchar = i;

  for (i = 0; i < C; i = i + 1)
  for (j = 0; j < C; j = j + 1)
    f[i][j] = IDENT;

  for (i = 0; s[i]; i = i + 1)
  for (j = 0; s[j]; j = j + 1)
  { k = i + j; if (k ≥ nchar) k = k - nchar;
    f[s[i]][s[j]] = s[k]; }

  for (i = 3; i ≤ 7; i = i + 2) colgen(i, G - 1);
  ipage = 0; ifile = 0;

  while (fgets(in, L, stdin))
  { i = strlen(in);
    if (in[i - 1] ≡' \n') in[i - 1] = 0;

    line[0] =' -';
    for (i = j = 0; in[i]; i++)
    { c = in[i]; if (seq[c] ≡ 0) c =' ';
      if (line[j] ≡' ' && c ≡' ') continue;
      line[++j] = c; }
    line[++j] = 0;

    ecc(guard, line + 1); seqn(guard, "", in);

    page[ipage++] = guard[G - 1];
    if (ipage ≥ pagef) seqpage(1), pagef = pages;

    file[ifile++] = guard[G - 1];
    if (ifile ≥ L) ifile = ifile - 1; }

  seqpage(0); file[ifile] = 0;
  ecc(guard, file); seqn(guard, ".", "");
  ifile = 0; }
```

1. If an entry parameter has been supplied, take it to be the page length.
2. Determine the number of symbols in the set while developing a list of sequence numbers.
3. Clear the modulo addition table.
4. Construct tables mapping all symbol pairs to corresponding sums.
5. Generate odd guard patterns.
6. Compute the error-correcting code for the line.
7. Compress multiple blanks from the input line.
8. Compute the ECC guard symbols.
9. If this is the end of the page, prepare a code for the entire
10. If this is the end of the page, prepare a code for the entire
11. At the end of the file, prepare a code for the entire file.

**COMPUTE CENTINELS**

**Upon entry to the algorithm, (1)** *gs* points to an area of length $G + 1$ to receive the results. **(2)** *line* points to the line. **(3)** $G$ defines the number of guard digits to be computed. **(4)** *ptn* defines which line positions contribute to which guard digits. **(5)** *f* contains the modulo-addition table for all symbols. **At exit,** *gs* contains the guard symbols for the line.

```
ecc(gs, line) char *gs, *line;
{ int i, j;

  for (i = 0; i < G; i = i + 1) gs[i] = IDENT;

  for (i = 0; i < G; i = i + 1)
  for (j = 0; line[j]; j = j + 1)
    if (ptn[j][i] ≡' X')
      gs[i] = f[gs[i]][line[j]];
  gs[G] = 0; }
```

1. Clear all the guard symbols.
2. Generate each guard symbols.
3. following the table that shows which line positions contribute to which guard symbols.

**CONVERT CENTINELS TO INTEGERS**

**Upon entry to the algorithm, (1)** *gs* contains the guard symbols. **(2)** *sep* contains a separator character. **(3)** *sym* contains the string of symbols. **At exit,** *gn* contains the corresponding integer sequence numbers.

```
seqn(gs, sep, sym) char *gs, *sep, *sym;
{ int i;
  for (i = 0; i < G; i = i + 1)
    printf("%02d", seq[gs[i]]);
  printf("%s%s\n", sep, sym); }
```

1. Display the sequence numbers for the guard symbols.
2. Display the full line.

**GENERATE PERMUTATIONS**

**Upon entry to the algorithm, (1)** $n$ defines the number of guard symbols to be marked. **(2)** $k$ defines the position for the initial mark. **(3)** $l$ defines the column number on the line, starting with 0. **(4)** *ptn* contains an area to receive the permutations. **(5)** $w$ contains a work area for generating the permutations. **At exit, (1)** All permutations have been generated. **(2)** $l$ is advanced by the number of combinations generated. **(3)** $ptn[0..l]$ contains the permutations generated thus far. **(4)** $w$ contains the most recent permutation generated.

```
colgen(n, k) int n, k;
{ static char w[G + 1] = ""; static int l = 0; int i;
  if (w[0] ≡ 0)
    for (i = 0; i < G; i = i + 1) w[i] =' -';
  if (n > 0) for (i = k; i ≥ n − 1; i = i − 1)
  { w[i] =' X';
    colgen(n − 1, i − 1);
    w[i] =' -'; }
  else if (l < L)
  { for (i = 0; i < G; i = i + 1)
      ptn[l][i] = w[i];
    l = l + 1; } }
```

1. On the first call, establish a null pattern in the array.
2. Mark the guard symbol for each possible position and generate all permutations within that position.
3. If there are no deeper permutations, save the current permutation and advance the column number.

*Intn'l Conf. Parallel & Distributed Processing, PDPTA12(1): 90–94, Proceedings, July 2012*

# Symmetry and Simplicity in Simulation: Reducing Complexity in Alternate Parallel–Serial Processing

**Clarence Lehman**[1] **and Adrienne Keen**[2]
[1]University of Minnesota, 123 Snyder Hall, 1474 Gortner Avenue, Saint Paul, MN 55108, USA
[2]London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK

**Abstract**— *Certain simulations are characterized by alternating periods of "expansion" and "contraction." For example, simulated populations of migratory birds may congregate in a central geographic location for overwintering, handled by a single processor, then fan out to dispersed locations for local ecological interactions during the rest of the year, handled by one processor per location. In another application, the difficult problem of fitting parameters to large-scale stochastic simulation models may fan out to numerous processors computing independent stochastic trajectories from the same initial conditions, then "contract" to allow a new, more likely set of parameters to be estimated from the computed distribution of independent trajectories. The contraction is commonly handled by a designated "master processor." In this paper, we point out a simpler, completely symmetric algorithm in which all processors act identically and no processor is designated master. We have used it for applications in simulated annealing and exhibit it here in a standard MPI (Message Passing Interface) environment.*

**Keywords:** parallel processing, symmetric multiprocessing, parallel–serial simulation, parameter fitting, individual-based modeling

## 1. Introduction

The goal of this paper is to demonstrate a symmetric technique for coordinating multiple processors and contrast that technique with a more usual master–subordinate technique. We consider the case where multiple processors calculate results independently of one another for an extended length of time, from seconds to minutes or more. The processors then pool their results before partitioning the calculations and expanding to multiple independent processors again. This repeats through multiple expansion–contraction phases until the computation converges to some result. We exhibit the algorithms in detail and illustrate them within an application of parameter fitting.

## 2. Algorithms

We assume each processor accepts a data structure as input and returns the results of its calculations in the same or another data structure. For simplicity here, we represent this data structure as an array of double-precision floating-point numbers, $local[j]$, though it could take any form. In addition, an array of these data structures, $global[i][j]$, has one row per processor. In the symmetric algorithm, all processors use this array, but in the master–subordinate algorithm, only the master uses it. Any processor can be the master, but here we make it the one numbered 0. The processor number is placed in an integer variable named *cproc* by Function 1 of the appendix. Algorithms in the appendix encapsulate the MPI environment [1] and provide a degree of system independence.

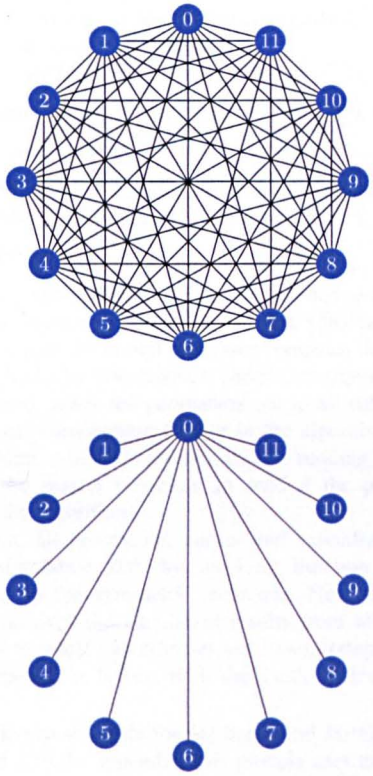


**Figure 1.** *Communications among a dozen processors in the symmetric technique, top, and the master–subordinate technique, bottom. The standard technique on the bottom might seem simpler, because of fewer interconnections, but asymmetries actually make it more complex.*

We present the algorithms in a stylized form of the language C, as an alternative to pseudocode, so that they define the interface precisely, and so they can be compiled, run, and modified. In the displayed algorithms, flow control and reserved words are bolded, variables and function names are italicized, and certain operations such as '<=', '>=', '!=', and '==' are displayed in a mathematical form as '$\leq$', '$\geq$', '$\neq$', and '$\equiv$', respectively.

## 2.1 Symmetric technique

In applications of parameter fitting, computing a new set of parameters is a global step, needing input from all the processors together. Therefore, it seems natural to assign that step to a master processor. However, such assignment is unnecessary. If all processors share information equally, then every processor can compute the new parameters for itself, using the same algorithm that would be used by the master processor. No time is lost, for all subordinate processors would be waiting for the master processor anyway. No chance of error arises, for all processors are executing the same code. And a notable simplification results, cutting the number of lines of code needed almost four-fold (see discussion below).

For the symmetric case, the program begins by invoking function *MPBegin* and ends by invoking *MPEnd*, defined in the appendix (steps 1 and 4, respectively, in the algorithm below). Variables *t* and *tmax* are integers recording the current and the maximum times, respectively. The main loop has five lines.

```
[1] MPBegin();
[2] for (t = 0; t ≤ tmax; t++)
      { converge = NewParameters(global);
        if (converge) break;
[3]     Simulate(local);
        MPCommon(global, local, W); }
[4] MPEnd();
```

Steps 1 and 4 above merely begin and end multiprocessing operations. Step 2 loops through the procedure, with each processor invoking function *NewParameters* to handle data from all the processors (array *global*), for example computing a new set of parameters by whatever technique is desired—gradient descent, simulated annealing, genetic search, and so forth. That function returns an indication of whether the process has converged, and if convergence is detected, the program terminates the main loop.

Step 3 invokes function *Simulate*, which carries out the current processor's simulation task, getting its parameters from array *local* and returning its results in the same array. Finally each processor communicates its results to all other processors and symmetrically receives all results back by invoking function *MPCommon* before repeating. $W$ in the example is the width of arrays *global* and *local*.

In addition to the code for *MPBegin* and *MPEnd* (Functions 1 and 2 in the appendix), this process uses only the five lines of code of Function 3 in the appendix. The amount of data communicated is $WN(N-1)$ elements, where $W$ is the number of elements per processor and $N$ is the number of processors. Processors pass messages only once per iteration.

## 2.2 Master–subordinate technique

When the global process is assigned to a master processor, the program also begins by invoking function *MPBegin* and ends by invoking *MPEnd*, defined in the appendix (steps 1 and 8, respectively, in the algorithm below). The variable *cproc* defines the processor number, which was not needed in the symmetric case. As before, *t* and *tmax* are integers recording the current and the maximum times, respectively. The main loop has nine lines.

```
[1] MPBegin();
[2] for (t = 0; t ≤ tmax; t++)
[3] { if (cproc ≡ 0)
      { converge = NewParameters(global);
        if (converge) break;
        MPMasterSend(global, W); }
[4]   else MPSubordinateReceive(local, W);
[5]   Simulate(local);
[6]   if (cproc ≡ 0) MPMasterReceive(global, W);
[7]   else MPSubordinateSend(local, W); }
[8] MPEnd();
```

At the beginning of the loop, *cproc* is tested to determine whether the master processor is running (processor number 0). If so, then the master processor computes the new parameters, checks for convergence, and if convergence has not been achieved, sends the parameters out to all subordinates for additional computation (step 3 in the algorithm). If, on the other hand, a subordinate processor is running, it merely waits for the master processor to send it the parameters (step 4 in the algorithm).

After that, all processors, master and subordinate alike, run one simulation task by invoking function *Simulate* (step 5), as in the symmetric technique. Next the master processor receives the simulation results from all subordinates (step 6) while subordinates send them (step 7). Then the loop repeats. As before, $W$ is the width of arrays *global* and *local*.

In addition to the code for *MPBegin* and *MPEnd* (Functions 1 and 2 in the appendix) this process uses the twenty-eight lines of code of Functions 4a, 4b, 5a, and 5b in the appendix. The amount of data communicated is $2W(N-1)$ elements, where $W$ is the number of elements per processor and $N$ is the number of processors. Both master and subordinate processors pass messages twice per iteration.

## 3. Discussion

The symmetric version needs no conditional if–else statements to determine which processor is running. It makes half the number of calls to the message passing interface per iteration, which simplifies the communications. It is somewhat less error prone, not only because of its greater simplicity, but also because multiple processors can be automatically checking each others work, detecting such mistakes as uninitialized variables that may behave differently under different conditions. It is shorter. The symmetric version uses 5 lines within its loop in the algorithm above and calls upon the 5 lines of Function 3 in the appendix, for a total of 10 lines in the loop. Functions 4a, 4b, 5a, and 5b in the appendix do not exist in the symmetric version. The master–subordinate version, in contrast, uses 9 lines within its loop and calls upon the additional 28 lines of Functions 4a, 4b, 5a, and 5b in the appendix, for a total of 37 lines in the loop.

The code to support symmetric multiprocessing is thus almost four times as compact. This savings can become compounded in the application code, because that code does not have to differentiate between master and subordinate communications. In parts of the code not connected with inter-processor communication, such as printing, one processor may still have to act as master. Yet since each level of reduction in complexity can be significant in a large program, this technique is preferred, all other things being equal.

One thing not equal is energy consumption. With all processors computing during the contraction phase, more heat will be generated and more energy consumed. Computations performed during the contraction phase will typically be short and simple compared with long and complicated simulations in the expansion phase, so this will be negligible. But if it is not, then a master–subordinate approach might be preferred.

Another thing not equal is the amount of information communicated among processors. The master–subordinate technique has only $N-1$ communication paths, where $N$ is the number of processors, whereas the symmetric technique has $\frac{1}{2}N(N-1)$ paths (Figure 1). Though this can be a large difference, it can also be insignificant in many applications. If the computation step is seconds or minutes or more, as it often will be, the microseconds or milliseconds dedicated to communications will vanish into the rest of the computation.

## 4. Conclusions

The symmetric version is simple. It is a viable way of communicating among multiple processors that can be incorporated into any expansion–contraction simulation programs, or related kinds of simulations. We have used it successfully in a large-scale simulation model developed by one of us (A.K.) for human tuberculosis in the UK. Compilable copies of the code described here and related simulation algorithms are available free from the authors upon request.

## 5. Acknowledgements

We are grateful to Shuxia Zhang and David Porter for helpful discussions of the message passing interface, to Mark Nelson for generous guidance to make things work, to Todd Lehman for reading the manuscript and helping with presentation of the code, and for Richard Barnes for help of many kinds.

## 6. Contributions

C.L. and A.K. worked together on various techniques for controlling the multiprocessors on a large-scale simulation carried out by A.K. C.L. conceived and coded the symmetric technique and A.K. built it into the simulation program and tested it in actual use. Both authors reviewed the code and contributed to the manuscript.

## 7. Funding

This project was supported in part by grants of computer time from the Minnesota Supercomputer Institute, Minneapolis, Minnesota, and by doctoral research funding to A.K. from the Modelling and Economics Unit at the Health Protection Agency, London.

## References

[1] M. Snir, S. Otto, S. Huss-Lederman, D. Walker, and J. Dongarra, "MPI: The complete reference," *MIT Press, Cambridge, MA*, 2012.

# 8. Appendix

This appendix defines the precise connections with the "Message Passing Interface," MPI [1]. Prototypes for functions and constants are defined in a file "mpi.h", which must be included at the top of the code. Then each process must call Function 1 before beginning. That function initializes communications, determines the number of processors that are participating, and assigns a number to the current processor. Furthermore, each process calls Function 2 after its work is complete, just before ceasing operations.

Function 3 is used in both symmetric and master–subordinate techniques. In this example it assembles an *nproc* by $w$ array of data from all processors, where *nproc* is the number of processors and $w$ is the number of data elements shared by each processor. It is typically called at the end of each processing step to synchronize all processors and put them all in a common data state. Processing is delayed until all processors have called this function. Therefore, note that all processors must call at corresponding points in the cycle or operations could deadlock.

Functions 4a, 4b, 5a, and 5b are additional algorithms needed for master–subordinate processing. Function 4a sends data from the master processor, numbered 0, to all processors allocated, including itself. Data to be sent reside in the *nproc* by $w$ array of data. Function 4a is typically called at the beginning of each processing step to synchronize all processors and give each processor the data it needs to carry out the next step. The master process must call this function and all others must call the companion function *MPSubordinateReceive*, Function 4b, at corresponding points in the cycle. Function 4b receives data from the master processor, resulting from that processor's call to 4a.

Function 5a receives data from subordinate processors, whose numbers are greater than 0. Data are assembled in the *nproc* by $w$ array. Function 5a is typically called by the master processor at the end of each processing step to receive results back from all processors and compute the data to begin the next step. The master process must call this function and all others must call the companion function *MPSubordinateSend*, 5b, at corresponding points in the cycle. Function 5b sends data back to the master processor, numbered 0, to satisfy its call to *MPMasterReceive*, Function 5a.

---

**Function 1.**

**Upon entry to the algorithm,** no conditions are significant. **At exit, (1)** multiprocessing operations have commenced. **(2)** *nproc* contains the total number of processors allocated. **(3)** *cproc* contains the number of the current processor, in the range 0 to $nproc - 1$.

```
int MPBegin()
{ static int argc; static char **argv; int n;
  MPI_Init(&argc, &argv);                          1. Initialize message processing.
  MPI_Comm_rank(MPI_COMM_WORLD, &cproc);           2. Determine this processor's number.
  MPI_Comm_size(MPI_COMM_WORLD, &nproc);           3. Determine the number of processors.
  return 0; }                                      4. Return to caller.
```

---

**Function 2.**

**Upon entry to the algorithm,** multiprocessing operations have closed. **At exit,** the main program may itself exit.

```
int MPEnd() { MPI_Finalize(); return 0; }
```

---

**Function 3.**

**Upon entry to the algorithm, (1)** *local* is a vector of $w$ data elements (double precision floating point) that are this processor's contribution to the global data set. **(2)** *global* is a *nproc* by $w$ matrix to receive the values of *local* from all processors. **(3)** $w$ contains the width of *local* and *global*. **At exit,** *global*[$n$] contains a copy of the contents of *local* from each processor $n$, where $n$ ranges from 0 to $nproc - 1$. In particular, the *local* of this processsor passed on entry is in row *global*[*cproc*].

```
int MPCommon(double global[][], double local[], int w)
{ MPI_Allgather(local, w, MPI_DOUBLE,
                global, w, MPI_DOUBLE,
                MPI_COMM_WORLD);
  return 0; }
```

**Function 4a.**

**Upon entry to the algorithm, (1)** *cproc* is 0. **(2)** *global* is a *nproc* by $w$ matrix containing values to be sent to each processor $n$ in row *global*[$n$]. **(3)** $w$ contains the width of *global*[$i$]. **At exit,** *global*[$n$] has been sent to each processor $n$.

```
int MPMasterSend(double *global, int w)
{ double *temp =
    (double*)malloc(w*sizeof(double));
  MPI_Scatter(global, w, MPI_DOUBLE,
              temp, w, MPI_DOUBLE,
              0, MPI_COMM_WORLD);
  free(temp);
  return 0; }
```

1. Allocate a temporary area to receive the master's data back from itself.
2. Send data from *global*[$n$] to each processor $n$.
3. Release the temporary area.
4. return to caller.

**Function 4b.**

**Upon entry to the algorithm, (1)** *cproc* is not 0. **(2)** *local* is vector of $w$ elements to receive data from the master processor. **At exit,** *local* contains the data received.

```
int MPSubordinateReceive(double local[], int w)
{ MPI_Scatter((double*)0, 0, MPI_DOUBLE,
              local, w, MPI_DOUBLE,
              0, MPI_COMM_WORLD);
  return 0; }
```

**Function 5a.**

**Upon entry to the algorithm, (1)** *cproc* is 0. **(2)** *global*[*nproc*][$w$] is an area to receive values for each processor $n$ in row *global*[$n$]. **(3)** *global*[0] contains any results from the master processor, to be sent back to itself. **(4)** $w$ contains the width of *global*[$i$]. **At exit,** *global*[$n$] contains the results from each processor $n$, except that *global*[0] is unchanged.

```
int MPMasterReceive(double global[][], int w)
{ int i;
  double *temp =
    (double*)malloc(w*sizeof(double));
  for (i = 0; i < w; i++) temp[i] = global[i];
  MPI_Gather(temp, w, MPI_DOUBLE,
             global, w, MPI_DOUBLE,
             0, MPI_COMM_WORLD);
  free(temp);
  return 0; }
```

1. Allocate a temporary area to receive the master's data back from itself.
2. Send data from *global*[$n$] to each processor $n$.
3. Release the temporary area and return to caller.

**Function 5b.**

**Upon entry to the algorithm, (1)** *cproc* is not 0. **(2)** *local* is vector of $w$ elements to send to the master processor. **At exit,** the data have been sent.

```
int MPSubordinateSend(double local[][], int w)
{ MPI_Gather(local, w, MPI_DOUBLE,
             (double*)0, 0, MPI_DOUBLE,
             0, MPI_COMM_WORLD);
  return 0; }
```

1. Send data to processor 0.
2. Return to caller.