Modelling BCG vaccination in the UK

What is the impact of changing policy?

By

Samuel Abbott

Bristol Medical School: Population Health Sciences
University of Bristol

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Doctor of Philosophy in the Faculty of Health Sciences.

April 2019

Word count: 36, 969

Abstract

Bacillus Calmette–Guérin (BCG) remains the only licensed vaccine against Tuberculosis. In 2005, England changed from universal vaccination of school-age children to targeted vaccination of high-risk neonates. Little work has been done to assess the impact of this policy change. This thesis evaluates the impact of this change.

Whilst the characteristics of Tuberculosis in England have been reported elsewhere, little attention has been given to the role of BCG. Consequently, I explored and combined, the available data sources. Reporting on data quality issues, trends in incidence rates and differences in outcomes stratified by BCG status.

Prior to the change in policy, several studies were carried out to assess the impact. The strength of this evidence has not been reassessed. I recreated one such study and found that there was a greater impact than previously thought.

Determining the benefits of being BCG vaccinated is necessary to properly assess the impact of the policy change. I evaluated the evidence that vaccination may improve outcomes for Tuberculosis cases in England and found that there was some evidence of an association between vaccination and reduced mortality.

Surveillance data can help assess whether changes in vaccination policy have influenced incidence rates. I used surveillance data to determine whether those at school-age, or neonates, were directly affected by the policy change. I found the policy change was associated with increased notifications in the UK born but this was outweighed by a reduction in notifications in the non-UK born.

Statistical modelling is restricted by the available data. Incorporating non-linear and indirect effects can also be complex. Therefore, I developed a dynamic model of Tuberculosis, fitting to available data, and forecast the long term and indirect effects of the policy change. I found that school-age vaccination *placeholder* Tuberculosis incidence rates compared with neonatal or no vaccination.

Acknowledgements

Thank you to Ellen Brooks-Pollock and Hannah Christensen, for supervising this thesis and for providing guidance and support. I would also like to Matthew Hickman for providing additional supervisory support.

This thesis could not have been completed with out the support of the Tuberculosis section at Public Health England. Of particular help were: Maeve K Lalor, Dominik Zenner, and Colin Campbell. Mary Ramsay, also at Public Health England, provided invaluable insights.

Nicky Welton provided support with the statistical methodology used in Chapter 7 - thank you.

This work was funded by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Evaluation of Interventions at University of Bristol in partnership with Public Health England. Thank you for this opportunity.

Thank you to my friends and family for providing support over 4 years.

Finally, thank you to my partner, Venexia Walker, without whom this would definitely have gone unfinished.

Contributiorship statements

Abbott S. *getTBinR: an R package for accessing and summarising the World Health Organisation Tuberculosis data*, Journal of Open Source Software, 2019, 4(34), 1260., doi: `https://doi.org/10.21105/joss.01260`

SA conceived, designed, and undertook the work. SA wrote the paper and approved the work for publication.

Abbott S, Christensen H, Lalor MK, Zenner D,Campbell C, Ramsay M, Brooks-Pollock E. *Exploring the effects of BCG vaccination in patients diagnosed with tuberculosis: observational study using the Enhanced Tuberculosis Surveillance system*, doi: `https://doi.org/10.1101/366476`

SA, HC, and EBP conceived and designed the work. SA undertook the analysis with advice from all other authors. All authors contributed to the interpretation of the data. SA wrote the first draft of the paper and all authors contributed to subsequent drafts. All authors approve the work for publication and agree to be accountable for the work.

Abbott S, Christensen H, Welton NJ, Brooks-Pollock E. *Estimating the effect of the 2005 change in BCG policy in England: A retrospective cohort study*, doi: `https://doi.org/10.1101/567511`

SA, HC, and EBP conceived and designed the work. NJW provided guidance on the statistical methods used. SA undertook the analysis with advice from all other authors. All authors contributed to the interpretation of the data. SA wrote the first draft of the paper and all authors contributed to subsequent drafts. All authors approve the work for publication and agree to be accountable for the work.

Signed :

Signed:

Date: *placeholder*

Declaration

I declare that the work in this thesis was carried out in accordance with the requirements of the University of Bristol's Regulations and Code of Practice for Research Degree Programs and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is my own work. Work done in collaboration with, or with the assistance of other, is indicated as such. Any views expressed in this thesis are those of the author.

Signed:

Date: *placeholder*

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Tuberculosis (TB) is one of the oldest human diseases, with recorded cases in ancient Egypt, renaissance Europe, and in the modern day across the globe. It is thought that roughly one-third of the world's population has been infected with TB, with 1% of the world's population being infected annually. However, the vast majority of these cases will never develop active disease. This reservoir of disease presents a challenge for control and eradication as, even if transmission can be halted, new cases will still occur for many years to come. While many people might consider TB to be a problem of the past in England, in 2017 there were 5,900 notified cases, the majority of which occurred in vulnerable communities; where incidence rates can be as much as 15 times higher than in the general population.[1] Globally TB remains the leading cause of death from infectious disease.[2]

The BCG vaccine was developed in 1921 and was introduced to the UK in 1953. Globally, it has been shown to offer variable protection that may reduce over time.[2] However, there is strong evidence that BCG offers high levels of protection for children, and more generally within the UK born population.[3] It remains the only TB vaccine with over 100 million doses given globally each year. Serious side effects are very rare but scarring commonly occurs at the site of injection. In 2005, the UK withdrew the universal BCG program for those at school age and introduced a targeted program of vaccination for babies that were deemed to be at high risk.[4] This was motivated by several years of declining transmission, the evidence of high levels of protection in children and a belief that other control measures would be more cost-effective.[4] Since this change in policy, declining incidence appears to support this decision.[1] However, due to TB's complex dynamics, the long-term effects are difficult to predict.

The availability of data is revolutionising the way we view the world; in few other areas has this revolution been felt more than in public health. In 2000, Public Health England (PHE) launched a routine surveillance system for TB, which records demographic, clinical, and microbiological information on all notified cases.[1] This data-set allows us to study the details of TB epidemiology in England more easily than ever before. Whilst this information would present much of interest by itself, by combining it with other data-sets we can adjust for the changing demographics of the English population to study the trends in TB over time.

This thesis explores the impact of changing BCG policy in England, with the aim of inform-

ing global efforts to control TB. I makes use of the detailed PHE routine TB surveillance data to explore the current epidemiology of TB in England. I then use statistical models that make use of this data to explore the impact from the 2005 change in BCG policy. Finally I develop, and fit, a detailed mechanistic model of TB and BCG vaccination in England in order to forecast the ongoing impact of the change in policy versus multiple alternative scenarios. The remainder of this chapter outlines: the theoretical framework used in this thesis; the aims and objects that were used to motivate this thesis; the chapter structure of this thesis; and the output from this thesis.

1.1 Theoretical framework

This thesis uses three main techniques to explore the impact of BCG vaccination on TB in England. These are: data exploration and visualisation, statistical modelling, and mechanistic modelling. Each of these techniques is outlined in the following sections.

1.1.1 Data exploration and visualisation

Data visualisation is often discounted in favour of more complex statistical or mechanistic approaches. However, as an exploratory tool it can be used to generate hypotheses that can then be evaluated using more complex techniques. It can also be used to visualise results from more complex methods that can then be used as a form of validation.

In this thesis, data visualisation is used extensively in Chapter 2 to explore the epidemiology of TB globally and in Chapter 4 to explore the epidemiology of TB in England. Chapter 4 also uses visualisation to generate many of the hypotheses that are then explored in further detail throughout the rest of this thesis. The remaining thesis chapters use data visualisation to explore the data used in the analysis and the results from the analysis.

1.1.2 Statistical modelling

At the most basic level a statistical model is a set of assumptions that outline the generative process of some sample data.[**???**] These assumptions describe a set of probability distributions, that approximate the population distribution from which the data has been sampled. Statistical models are usually specified using mathematical equations that relate one or more random variables to non-random variables. An example of this is a linear regression which maps a series of variables, using a linear relationship, to generate a numeric outcome variable. Statistical models can be used to represent complex multivariate relationships that would not be possible to visualise. They can also be used to test alternative scenarios without altering the underlying data, see Chapter 7 for an example of this.

In this thesis, a variety of statistical models are used to explore complex multivariate relationships. Uses cases include: adjusting for confounding variables when estimating the relationship between BCG vaccination and TB outcomes (Chapter **??**); and estimating the impact on incidence rates from ending the BCG schools scheme after accounting for various confounders (Chapter 7).

1.1.3 Mechanistic modelling

Mechanistic mathematical models provide an assumption based framework for understanding the transmission of infectious diseases.[5] Mechanistic models can be used to simplify complex real-world systems, whilst retaining a linkage to real-world processes.[6] This is unlike statistical models, which instead focus on modelling the underlying structure of the data generally without reference to the real-world processes. There are multiple mechanistic modelling approaches the most common being compartmental based models and individual based models.[5,6] Both of these approaches can be represented as a deterministic or stochastic - the latter includes a random component whilst the former does not. Recently, mechanistic models have been combined with statistical models to account for the fact that events may be only partially observed.[7] Mechanistic models have an advantage over statistical models in that they can more easily be used to compare alternative scenarios over long periods of time for which observed data does not exist.

In this thesis, a partially observed stochastic compartmental model of Tuberculosis is developed that models demographic processes such as ageing, births and deaths, as well as vaccination and TB treatment (Chapter 8). Compartmental infectious disease models generally operate by separating a given population into a series of groups, most commonly susceptible, infectious and recovered populations.[5,6] Movements between these groups are then modeled using a series of differential equations. Transmission is modeled using mass action,[5,6] where infected cases are assumed to randomly interact with susceptible individuals. Additional detail can be added to this model by stratifying the population further and adding additional parameters to modify the degree of mixing between populations. See [5,6] for a theoretical introduction to infectious disease models and [7] for implementation details using R.

1.2 Aims and objectives of the thesis

1.2.1 Aim

To understand the impact of BCG vaccination on the epidemiology of tuberculosis in England, and to use this understanding to forecast the future effects of current and historic BCG vaccination policy.

1.2.2 Objectives

- To describe the current epidemiology of Tuberculosis in England, in the context of global Tuberculosis epidemiology.
- To assess some of the statistical modelling evidence used to justify the 2005 change in BCG vaccination policy in the UK.
- To assess whether there is evidence in routinely-collected surveillance data that BCG vaccination impacts outcomes for TB cases in England.
- To assess the effects of the 2005 change in vaccination policy on those eligible for vaccination.
- To develop a parsimonious transmission dynamic model of tuberculosis which captures current, and historic. vaccination policy and reflects our current understanding of TB epidemiology in England.
- To fit this model using all available data sources.

- To investigate the effectiveness of universal school-age vs. universal neonatal vs. no vaccination using the previously developed transmission dynamic model.

1.3 Chapter overview

- **Chapter 2:** Background information is given on Tuberculosis and the BCG vaccine. This information helps motivate future chapters and is useful for non-subject area experts.
- **Chapter 3:** `getTBinR`, an R package that facilitates downloading Tuberculosis relevant data from the World Health Organization and provides functionality for visualising the downloaded data, is introduced. The motivation and context for this package as part of the wider thesis is also outlined.
- **Chapter 4:** This chapter describes the epidemiology of tuberculosis in England, using routine national data-sets. Focusing on: the impact of missing data; the mechanisms underlying that missing date; seasonal trends; the role of age; UK birth status; BCG status; trends in TB incidence over time; and TB outcomes in England using case rates. These data are used in all subsequent chapters in this thesis.
- **Chapter 5:** This chapter recreates a simulation based statistical model that was used as part of the decision making process that led to the 2005 change in BCG vaccination policy. It extends the previously implemented model by fully capturing all parameter and model uncertainty, and updating the underlying data. It then estimates the impact in real-terms of the change in policy using this updated model.
- **Chapter 6:** This chapter uses regression analysis to explore the evidence that BCG vaccination is associated with positive outcomes for active TB cases in England. Any evidence that this is the case may strengthen the case for extending BCG vaccination coverage.
- **Chapter 7:** This chapter uses multilevel statistical models to assess the effects of the 2005 change in BCG vaccination policy on the populations targeted by each vaccination scheme.
- **Chapter 8:** In this chapter a mechanistic model of TB and BCG vaccination in England is developed. The model structure is fully justified based on the known epidemiology of TB in England. Model parameters are given prior distributions based on routine surveillance data (Chapter 4), the published literature, and assumptions based on expert knowledge where no other source exists.
- **Chapter 9:** In this chapter the model developed in the previous chapter is fitted to the routine surveillance data (Chapter 4) using bayesian methods. Multiple scenarios are considered, with the results evaluated using their DIC.
- **Chapter 10:** In this chapter the model, developed and fitted in the previous chapters, is used to forecast the impact of universal BCG vaccination at school-age vs. universal vaccination of neonates vs. no vaccination from 2005 on-wards. The results from this forecasting are compared to the actual notifications recorded from 2005 to 2015, with the divergence from the observed data discussed. The ongoing impact of each policy is then discussed through to 2030.

- **Chapter 11:** Results from all previous Chapters are summarised and discussed as a whole. The strengths and weaknesses of the analysis in this thesis are outlined. Further work is suggested. Future work is outlined.

1.4 Thesis output

This thesis has produced both traditional and non-traditional output. The traditional output includes peer reviewed papers, preprints and talks at academic conferences. The non-traditional output includes open source research software, software for improving the academic workflow, dashboards for exposing relevant data, dashboards for exploring the modelling methods used in this thesis, and an educational dashboard for teaching some the benefits of vaccination. These outputs are detailed in the following section.

1.4.1 Peer reviewed papers

- Abbott S. *getTBinR: an R package for accessing and summarising the World Health Organisation Tuberculosis data*, Journal of Open Source Software, 2019, 4(34), 1260., doi: `https://doi.org/10.21105/joss.01260`

1.4.2 Papers under review

- Abbott S, Christensen H, Lalor MK, Zenner D,Campbell C, Ramsay M, Brooks-Pollock E. *Exploring the effects of BCG vaccination in patients diagnosed with tuberculosis: observational study using the Enhanced Tuberculosis Surveillance system*, doi: `https://doi.org/10.1101/366476`

- Abbott S, Christensen H, Welton NJ, Brooks-Pollock E. *Estimating the effect of the 2005 change in BCG policy in England: A retrospective cohort study*, doi: `https://doi.org/10.1101/567511`

1.4.3 Software

Packages

- `getTBinR`: The `getTBinR` R package facilitates downloading the most up-to-date version of multiple Tuberculosis relevant data sources from the World Health Organization, along with the accompanying data dictionaries. It also contains functions to allow easy exploration of the data via searching data dictionaries, summarising key metrics on a regional and global level, and visualising the data in a variety of highly customisable ways. See Chapter 3 for further details. Install from CRAN with `install.packages("getTBinR")` or install the development version from GitHub with `devtools::install_github("seabbs/getTBinR")`. Link: `https://www.samabbott.co.uk/getTBinR/`

- `tbinenglanddataclean`: An R package that contains the functions and documentation required to reproduce all data import and munging used in this thesis. This package provides a workflow to facilitate reproducing all analysis in this thesis and to expedite the work of others using data from

the Enhanced Surveillance System (Chapter 4). Available from GitHub using `devtools::install_github("seabbs/tbinenglanddataclean")`. Link: `https://www.samabbott.co.uk/tbinenglanddataclean/`

- **`idmodelr`**: An R package that contains a library infectious disease models as well as modelling utilities. It provides tooling that includes: example SEI/SEIR/SHLIR/SHLITR model code, a model solving wrapper; a model summary function; a scenario analysis function). Used by the Explore infectious disease model dashboard (`http://seabbs.co.uk/shiny/exploreidmodels/`) for all functionality. Available from GitHub using `devtools::install_github("seabbs/idmodelr")`. Link: `https://www.samabbott.co.uk/idmodelr/`

- **`prettypublisher`**: An R package that improves the R based reproducible research workflow. It provides tooling that includes: paper and figure referencing; effect size reporting; percentage reporting; P value reporting; and produces a table ready for further word processing. Used throughout this thesis. Available from GitHub using `devtools::install_github("seabbs/prettypublisher")`. Link: `https://www.samabbott.co.uk/prettypublisher/`

Interactive tools

- **Explore global Tuberculosis**: Developed to showcase `geTBinR` (`https://www.samabbott.co.uk/getTBinR/`) package functionality. This dashboard allows the interactive exploration of WHO TB data. It can also be used to generate a static, country level, report on TB epidemiology. Link: `http://seabbs.co.uk/shiny/ExploreGlobalTB/`

- **Explore Tuberculosis in England and Wales**: Developed to allow public Public Health England Tuberculosis Notification data to be explored interactively. Key interventions are highlighted and link to trends in Tuberculosis notifications. This app is used in its static form in Chapter 2. Link: `http://seabbs.co.uk/shiny/TB_England_Wales/`

- **Explore infectious disease models**: Developed to be used within a modelling short course at the University of Bristol (`https://github.com/bristolmathmodellers/biddmodellingcourse`). This dashboard allows the user to simulate and compare a variety of compartmental infectious disease models. All model code is surfaced in an easily view-able format to allow for users to develop their own models. Link: `http://seabbs.co.uk/shiny/exploreidmodels/`

- **Introduction to Tuberculosis models**: Developed to allow simple Tuberculosis models to be explored in an interactive session. Inspired by practicals from the Introduction to Tuberculosis, run by TB MAC (`http://tb-mac.org/`) at the 2017 Union conference. Link: `http://seabbs.co.uk/shiny/intro_to_tb_models/`

- **The pebble game**: Developed as a learning aid to help a general audience understand the impact of vaccination on infectious disease dynamics. Used at Green Man 2016 as part of a week of outreach work and subsequently developed further. Link: `http://seabbs.co.uk/shiny/thepebblegame/`

1.4.4 Talks

- **Assessing the Evidence for Universal Bacillus Calmette-Guèrin (BCG) Vaccination in England** - Research and Applied Epidemiology Scientific Conference 2016, Warwick, United Kingdom. Received best abstract from an early career researcher. Link: `https://www.samabbott.co.uk/talk/phe-applied_epi-2016/`
- **Beneficial effects of BCG vaccination in outcomes for patients with active TB: observational study using the Enhanced Tuberculosis surveillance system 2000-2014** - Research and Applied Epidemiology Scientific Conference 2017, Warwick, United Kingdom. Received best PhD student abstract. Link: `https://www.samabbott.co.uk/talk/phe-applied-epi-2017/`
- **Beneficial effects of BCG vaccination in outcomes for patients diagnosed with TB: observational study using the Enhanced Tuberculosis surveillance system 2009-2015** - 48th Union World Conference on Lung Health. Link: `https://www.samabbott.co.uk/talk/union-2017/`
- **Estimating the effect of the 2005 UK BCG vaccination policy change: A retrospective cohort study using the Enhanced Tuberculosis Surveillance system, 2000-2015** - Research and Applied Epidemiology Scientific Conference 2018, Warwick, United Kingdom. Link: `https://www.samabbott.co.uk/talk/phe-applied-epi-2018/`

1.5 Summary

- This chapter provides an introduction to Tuberculosis and the BCG vaccine. It then motivates the remainder of this thesis.
- An outline of the theoretical framework used throughout this thesis is given.
- The aims and objectives of this thesis are detailed.
- An overview of the chapters is provided.
- Finally the dissemination of this work so far is summarised, broken down into peer reviewed output, preprints, software output and talks given at academic conferences.

Chapter 2

Background

2.1 Tuberculosis

Tuberculosis (TB) is thought to infect over 1.7 billion people globally, of which 5-15% will develop active TB in their lifetime.[2] Of this number around 10% are likely to die from TB or TB related causes. TB is preventable and curable, but the majority of cases occur in less economically developed countries and are never diagnosed.[2] In the following section, the natural history of Tuberculosis, the risk factors, the treatment, the global impact, and the impact in England and Wales are explored.

2.1.1 Natural history of Tuberculosis

TB is primarily a respiratory disease (pulmonary TB) caused by the bacterium Mycobacterium tuberculosis, although it can also affect other parts of the body (extra-pulmonary TB). TB spreads via airborne droplets that are expelled when individuals with active pulmonary TB cough. After an infection with TB, 5-10% of individuals develop primary disease within 1-2 years of exposure. Children are more likely to develop active disease and to develop it more quickly than adults.[2] The majority of individuals then enter a latent stage in which they passively carry TB mycobacterium. Reactivation of bacilli can then occur many years later due to a loss of immune control.[8]

Both active and latent TB cases represent a range of diverse individual states. Pulmonary cases are typically responsible for the vast majority of transmission.[9] Latent cases may have completely cleared the bacterium or be asymptotically carrying reproducing active TB bacterium.[8] Adolescents have the highest risk of developing active TB, usually in the form of pulmonary TB.[2] The risk of developing pulmonary TB, versus extra-pulmonary TB, varies with age. For instance, younger children are more likely to develop pulmonary TB.[2]

The most common symptoms are a chronic cough with sputum containing blood, fever, night sweats and weight loss. Infectiousness, mortality and likelihood of developing various types of TB vary with age.

2.1.2 Known risk factors

TB has been associated with several risk factors, the most common of which is HIV. HIV increases the rate of activation by 20-fold and TB is the most common cause of AIDS-related death.[10] Increased risk of TB can also be the result of other medical conditions, such as diabetes, or lifestyle and environmental factors. These include smoking, low socioeconomic status, high density living, homelessness, incarceration, and drug use.[11–13]

2.1.3 Treatments

Treatment for TB consists of a six month course of multiple antibiotics (see Table 2.1). These usually consist of isoniazid, rifampicin, pyrazinamide and ethambutol (known as first line drugs). If the disease is resistant to treatment with the first line drugs then second line drugs such as aminoglycosides, fluoroquinolones, and cycloserine are employed. The side effects for these drugs are generally far more severe and the treatment regime is longer, typically 12-24 months. The World Health Organisation now recommends the use of the Directly Observed Treatment short-course (DOTS), which focuses on 5 action points.[14] These are:

1. Political commitment with increased and sustained financing
2. Case detection through quality-assured bacteriology
3. Standardized treatment with supervision and patient support
4. An effective drug supply and management system
5. Monitoring and evaluation system and impact measurement

Table 2.1: A timeline of interventions against TB. Antibiotics used to treat TB are commonly given together, with those with the fewest side effects given first. Second line antibiotics are then used if the initial treatment fails or tests show the strain is multiply drug resistant. BCG - Bacillus Calmette–Guérin; TB – Tuberculosis; MRSA - Methicillin-resistant Staphylococcus aureus; DOTS - Directly Observed Treatment Short-course

Year	Intervention	Type	Line	Detail
1921	BCG	Vaccination		The first use of the Bacillus Calmette–Guerin (BCG) vaccine in humans, it remains the only vaccine against Tuberculosis (TB). Efficacy has been shown to vary depending on latitude (0-8010-15 years after vaccination.
1944	Streptomycin	Antibiotic	Second	The first antibiotic and the first bacterial agent against TB.
1944	4-Aminosalicylic acid	Antibiotic	Second	The second antiobiotic to be developed. Due to lower potency than other antibiotics it is not considered a first line treatment.
1952	Isoniazid	Antibiotic	First	Used against both active and latent TB, it may also be given as a prophylatic therapy.
1952	Cycloserine	Antibiotic	Second	An antibiotic with severe side effects such as kidney failure and neurological conditions, which is therefore restricted for use against multiple drug resistant Tuberculosis.
1952	Pyrazinamide	Antibiotic	First	First discovered in 1936, it was first used against TB in 1952. Although showing no effect in-vitro it was shown to be effective in treating TB in mice. Used only for treating TB and never on it's own.
1953	School age BCG	Vaccination		After a successful trial, which showed high effectiveness for the vaccine, BCG was introduced in the UK for those at school leaving age as peak incidence was then in young, working-adults.
1962	Ethambutol	Antibiotic	First	Believed to work by interfering with TB bacteria's metabolism. There are some concerns that it may not be safe to give during pregancy, as it may lead to vision loss in the baby.
1971	Rifampicin	Antibiotic	First	Taken daily for at least a period of 6 months, if given alone resistance develops quickly. It may also be used in the treatment of MRSA amongst other diseases.
1995	DOTS	Strategy		Directly Observed Treatment, Short-Course (DOTS) is introdued by the World Health Organisation as a control strategy for TB. The intermittent, supervised system aims to eliminate drug default.
2005	Neonatal high risk BCG	Vaccination		Due to a continued decline in TB incidence rates in the indigenous UK population, the BCG programme was refocused as risk-based. This meant vaccinating high risk neonates rather than those most likely to transmit TB.

Table 2.1: A timeline of interventions against TB. Antibiotics used to treat TB are commonly given together, with those with the fewest side effects given first. Second line antibiotics are then used if the initial treatment fails or tests show the strain is multiply drug resistant. BCG - Bacillus Calmette–Guérin; TB – Tuberculosis; MRSA - Methicillin-resistant Staphylococcus aureus; DOTS - Directly Observed Treatment Short-course *(continued)*

Year	Intervention	Type	Line	Detail
2012	Bedaquiline	Antibiotic	Second	The first new antiobiotic for use against TB in 40 years, reserved for use against multiple drug resistant TB. Approved via a fast track process, higher mortality in those that recieve the antibiotic has caused significant concern.

2.1.4 Global TB

TB is a global disease with an estimated 10.4 million new cases in 2016, of which 4.3 million were missed by health systems.[2] Global incidence rates have decreased year on year since the early 2000's, with an average year on year decrease of 2.9%. However, global TB incidence remains above 134 per 100,000 population (Figure 2.1). On a regional level, incidence rates vary with Africa and South-East Asia having a greater concentration of cases. In the Eastern Mediterranean incidence rates have remained relatively stable over the last 10 years.

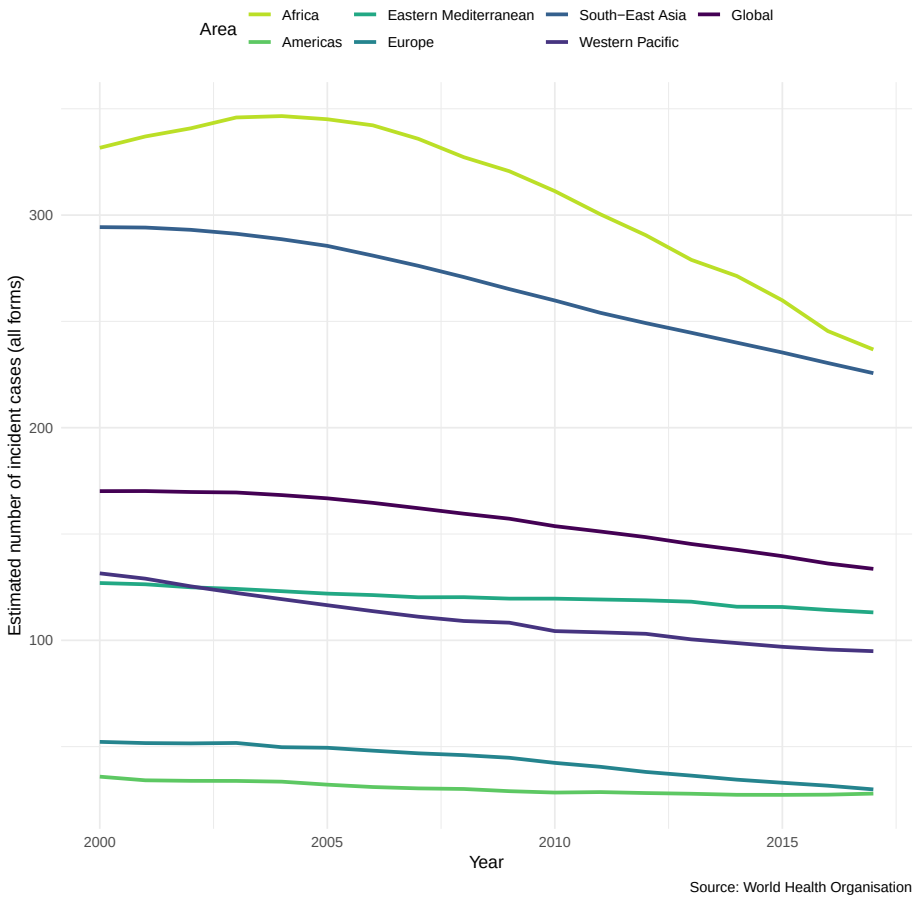


Figure 2.1: Tuberculosis incidence rates (per 100,000) by region and globally from 2000 until 2017. Globally incidence rates have been declining since the early 2000's but this decline varies with region. Plot produced using getTBinR

Regional incidence rates only tell part of the story as TB incidence rates vary significantly within regions. Six countries: India, Indonesia, China, Nigeria, Pakistan, and South Africa account for 60% of new cases. India, Indonesia, and Nigeria comprise nearly half of the gap between incident and notified cases.[14] Figure 2.2 shows both regional similarities and countries, like Mongolia, that stand out as having higher TB incidence rates than surrounding countries.

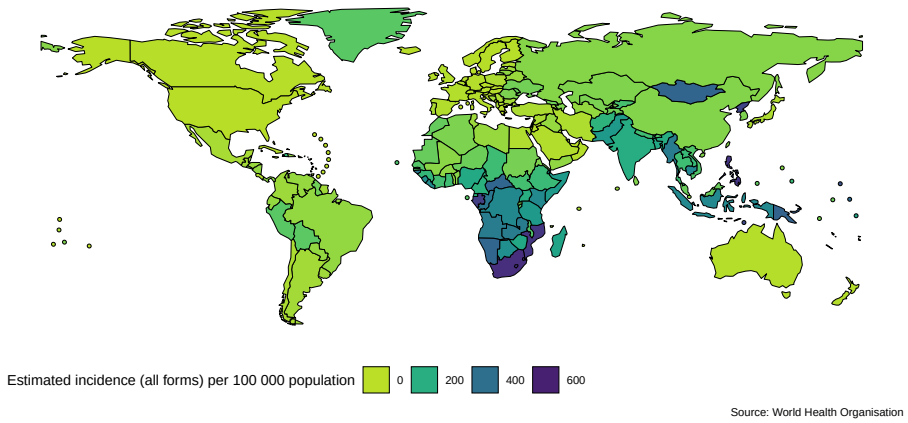


Figure 2.2: Global map of country level Tuberculosis incidence rates (per 100,000 population) in 2017. Note the clustering of countries with high incidence rates in southern and central Africa and southern Asia. Map produced using getTBinR

TB remains one of the top 10 causes death worldwide, leading to 1.7 million deaths in 2016 alone.[2] Whilst the absolute number of deaths due to TB has fallen since 2000, the average global rate of decline in TB mortality rates was only 2.9% between 2000-2016. However, unlike the trend observed for incidence rates, the year-on-year decline of TB mortality rates has remained consistent in all regions (Figure 2.3). Several regions, including Africa and Europe, have seen TB mortality rates fall to below 50% of those in 2000.

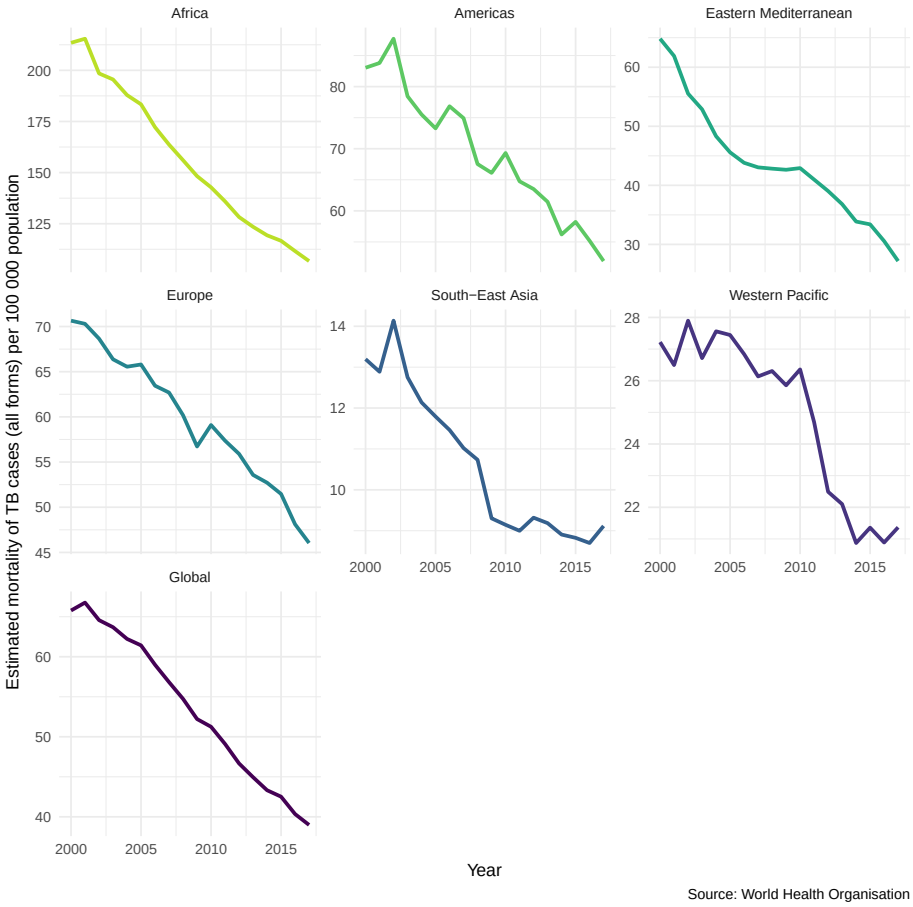


Figure 2.3: Tuberculosis (TB) mortality rates (per 100,000) by region and globally from 2000 until 2017. Mortality rates from TB have been falling in all regions since 2000. Plot produced using getTBinR

There is an ongoing global co-epidemic of HIV and TB, with people living with HIV accounting for 1.4 million TB cases in 2016. 22% of deaths from TB were in those living with HIV. Whilst this is a global problem, it is a particular issue in sub-Saharan Africa with over 60% (95% CI: 55%-64%) of incidence TB cases in South Africa also having HIV (Figure 2.4). This compares to a global mean of 9.1% (95% CI: 6.0%-13.0%) and a mean of 26.7% (95% CI: 17.4%-38.1%) in Africa.

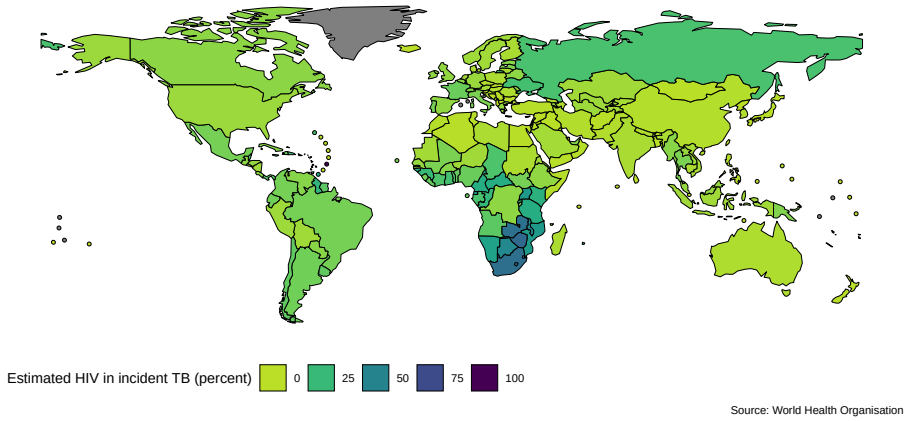


Figure 2.4: Global map of estimated HIV in incidence TB (percent) in 2017. Note that high percentage of TB cases with HIV in sub-saharan Africa. Map produced using getTBinR

Multi-drug-resistant TB (MDR-TB), which is defined as being resistant to at least isoniazid and rifampin, made up 4.6% of all new TB cases in 2015 (480,000). It can be acquired both through treatment failure and through transmission. Treatment requires the use of second line antibiotics, which often have more severe side effects and are more likely to fail, with only 52% successfully treated globally compared to 83% for drug susceptible TB.[14] As for HIV co-infection, drug resistance is globally heterogeneous with some regions, like countries in the former USSR, having a much higher proportion of drug resistant cases. Figure 2.5 shows the country level proportion of cases with at least rifampicin resistance and highlights the higher level of rifampicin resistance in countries formerly in the USSR. 88% of rifampicin cases in Russia in 2017 also had MDR-TB, which is comparable to the global median of 83% (Table 2.2). Across all regions Europe had the highest median percentage of rifampicin cases with MDR-TB (100%), with Africa having the lowest (67%). This variation may indicate areas in which drug resistance is developing (regions with low proportions of MDR to rifampicin resistance), from regions in which MDR-TB results from transmission or importation (regions with high proportions of MDR to rifampicin resistance).

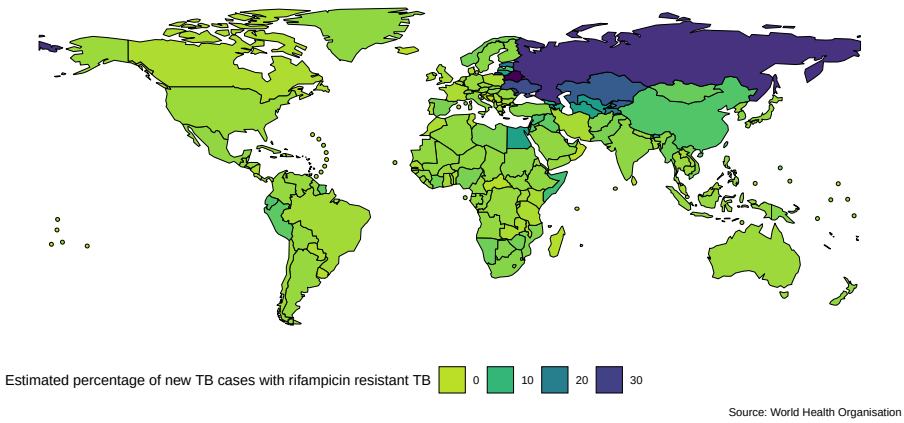


Figure 2.5: Global map of the estimated percentage of new Tuberculosis cases with rifampicin resistance (percent) in 2017. Not that a far higher percentage of TB cases have rifampicin resistance in the former Soviet Union that in the rest of the world. Map produced using getTBinR (see refgetTBinR for details)

Table 2.2: Percentage (%) of rifampicin resistant Tuberculosis (TB) cases that have multi-drug resistant TB in Russia and regional medians, with interquartile ranges.

Area	Median percentage of RR* cases with MDR** (95% IQR***)
Russian Federation	88.00 (NA to NA)
Africa	67.00 (39.45 to 100.00)
Americas	83.00 (56.25 to 99.25)
Eastern Mediterranean	79.00 (21.15 to 100.00)
Europe	100.00 (69.60 to 100.00)
Global	83.00 (40.12 to 100.00)
South-East Asia	91.00 (36.50 to 99.50)
Western Pacific	78.00 (66.50 to 100.00)

* Rifampicin resistance
** Multi-Drug Resistant TB
*** Interquartile Range

All statistics that are not referenced in this section were generated using `getTBinR` - see Chapter 3 for further details.

2.1.5 TB in the England and Wales

TB Notifications

TB incidence in England and Wales has decreased dramatically from a century ago (Figure 2.6, or see `http://www.seabbs.co.uk/shiny/TB_England_Wales` for an interactive dashboard). However, in the past several decades, incidence rates first stabilised and have since increased since their lowest point in the 1990's. In 2000 there were 6044 notified TB cases in England, increasing to a maximum of 8280 notified TB cases in 2011. Since then, notifications have declined year on year.[15] Figure 2.6 includes the interventions discussed above (Table 2.1) and indicates that the introduction of several antibiotics and BCG vaccination in the 1950s may have led to an extended decrease in incidence.

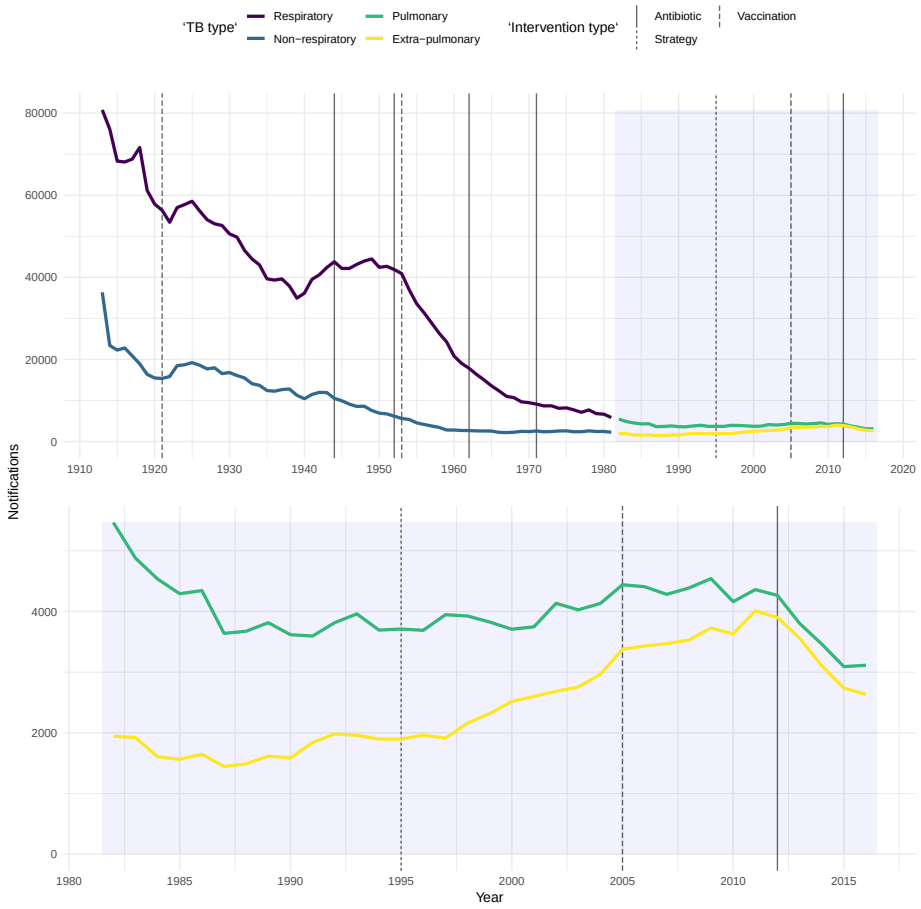


Figure 2.6: TB notifications in England and Wales from 1913 to 2017, stratified initially by respiratory/non-respiratory status and from 1982 by pulmonary/non-pulmonary TB. Interventions are highlighted with vertical lines, with linetype denoting the type of intervention, more information on each intervention is available in the corresponding table. Figure produced using tbinenglanddataclean (https://www.samabbott.co.uk/tbinenglanddataclean/)

Heterogeneity of TB

TB incidence in England and Wales is highly heterogeneous with over 70% of cases occurring in the non-UK born population. Incidence rates in the non-UK born (49.4 per 100,000, in 2016) are 15 times higher than in the UK born population (3.2 (95% CI 3.0-3.3) per 100,000, in 2016).[1] The age distribution of cases in the UK born and non-UK born populations differ, with the UK born population having a relatively uniform distribution. Meanwhile, the non-UK born have higher incidence rates in those aged 80 years and older (69.3 per 100,000 in 2016), those aged 75 to 79 years (62.9 per 100,000 in 2016) and those aged 25-29 years old (61.6 per 100,000 in 2016) [1]. In the non-UK born, the majority of cases occur amongst those who have lived in the UK for at least 6 years (63%) - this has increased year on year since 2010 (49%).[1] However, in 2016, 23.3% (420/1,800) of non-UK born cases had traveled outside the UK, with the majority returning to their country of origin. Incidence rates in the UK born are between 3 and 14 times higher in non-White ethnic groups compared to the White ethnic groups.[1]

The majority of cases occur in urban areas. London alone accounts for 39% of cases, with an incidence rate of 25.1 (per 100,000; 95% CI 24.1-26.2; in 2016).[1] England has few cases of MDR-TB cases, with only 68 cases recorded in 2016. Similarly the number of co-infections with HIV is low with only 3.8% of cases in 2015 having HIV - the majority of these cases were born in countries with high HIV prevalence. In 2016, 11.1% of TB cases in 2016 had at least one social risk factor, compared with 11.7% in 2015.[1] In general cases with social risk factors are more likely to have drug resistant TB, worse TB outcomes, and to be lost to follow up. [1] Amongst cases who were of working age in 2016, with a known occupation; 35.2% (1,491/4,240) were not in education or employment, 10.2% (432) were either studying or working in education; and 7.1% (304) were healthcare workers.[1]

TB Transmission

As TB incidence rates alone cannot be used to assess current TB transmission due to reactivation of those latently infected, the incidence rate in UK born children (0-14 years old) is used as a proxy for transmission. Incidence rates in UK born children have fallen 47% from 3.4 per 100,000 in 2008 to 1.8 per 100,000 in 2016.[1] This indicates that TB transmission has fallen in the last decade. However, BCG vaccination was introduced for those neonates at high risk of TB in 2005, which may partly be responsible for the observed reduction in incidence rates.

Strain typing or whole genome sequencing is used to establish case clustering. This can be used to rule out transmission between cases but does not necessarily confirm transmission. Approximately 60% of cases cluster with at least one other case, and whilst this varies year on year, the fluctuations appear to be small (approximately 1-2%).[1] Therefore interpreting any trend in TB transmission from the current strain typing data is difficult. Between 2010 and 2016, the median cluster size was 3 cases (range 2-244). In these clusters, 74.4% (2,141/2,878) consisted of less than 5 cases and only 8.8% of clusters had more than 10 cases [1]. UK born cases were more likely to cluster than non-UK born cases (71.1%, 4,200/5,910 vs. 56.1%, 10,166/18,121).[1]

Pulmonary Vs. Extra-Pulmonary TB

Figure 2.7 shows that since the 1980s the proportion of extra-pulmonary TB compared to pulmonary TB has increased from 26.2% (1944/7410) in 1982 to 45.8% (2634/5748) in 2016. This may be attributed to the age distribution of TB cases changing as different age groups are more likely to progress to pulmonary vs extra-pulmonary TB. It may also be related to the increase of non-UK born cases as a higher proportion of non-UK born cases have extra-pulmonary disease only (51.4%, 2,103/4,089, in 2016), compared to UK born cases (31.9%, 467/1,465, in 2016).[1] For more details on TB in England, see Chapter 4 and the Public Health England 2018 TB report from which the summary data discussed above was extracted.[1]

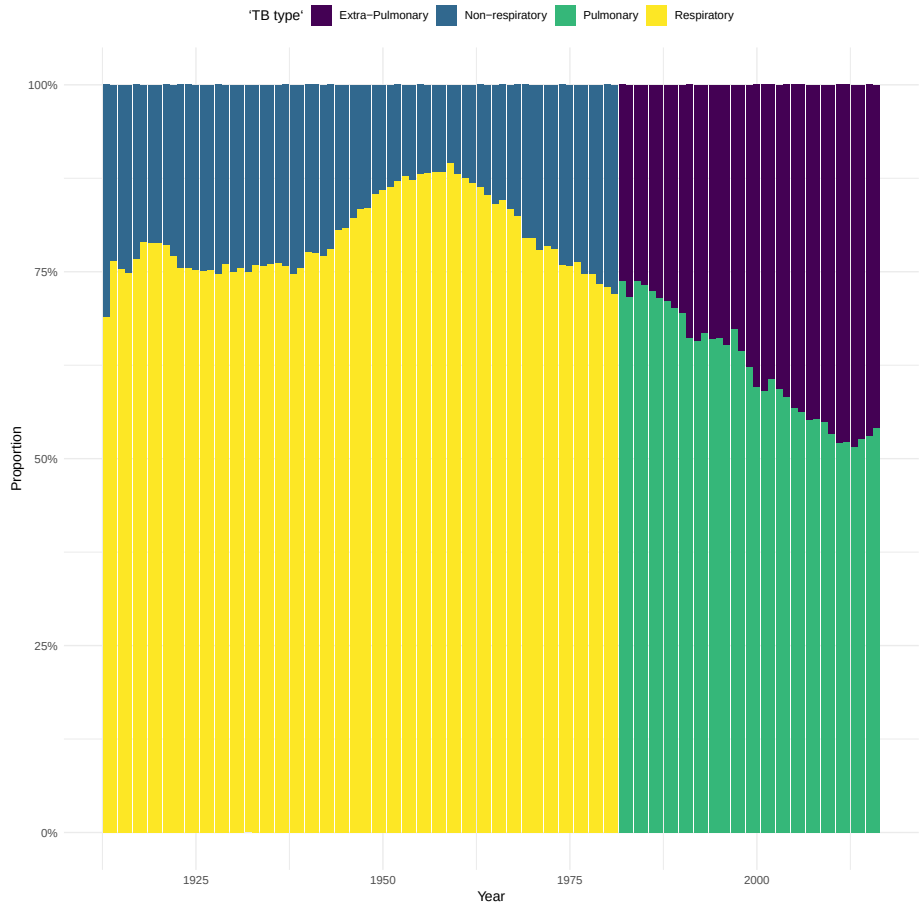


Figure 2.7: From 1913 until 1981 the figure shows the proportion respiratory vs. non-respiratory cases and from 1982 it shows the proportion of pulmonary vs. non-pulmonary TB. Figure produced using tbinenglanddataclean (https://www.samabbott.co.uk/tbinenglanddataclean/)

2.2 The Bacillus Calmette–Guérin Vaccine

The Bacillus Calmette–Guérin (BCG) vaccine was first given to humans in 1921 and remains the only licensed vaccine for TB.[16] The BCG vaccine is a live vaccine and was developed

by weakening a strain of Mycobacterium bovis, which is commonly found in cows, over a period of 13 years.[2] Serious side effects are rare, although a small scar at the injection site is common. This section details the action, effectiveness, duration of protection, effects and usage of the BCG vaccine.

2.2.1 Vaccine action

The BCG primarily acts by directly preventing the development of active, symptomatic disease. However, there is some evidence to suggest that the BCG vaccine also provides partial protection against initial infection.[3] There is no evidence that BCG vaccination post infection with TB provides protection from developing active TB disease.[2]

2.2.2 Vaccine effectiveness

The effectiveness of the vaccine is impacted by the age at which it is given, the latitude of the individual, and the period of time that has lapsed since vaccination. Multiple randomized control studies (RCTs) have been conducted on BCG efficacy. It has consistently been shown to be highly protective in children.[17] Efficacy in adults ranges from 0% to 78%,[19] with an MRC trial in England finding that BCG was 78% effective.[20]

A meta-analysis of RCTs indicated that increased protection is associated with distance from the equator.[19] One hypothesis for this is that there is a greater density of mycobacterium near the equator that may mask, or block, the protection offered by the BCG vaccine.[21] Recently it has been found that much of this latitude effect may be due to stringency in tuberculin skin testing (TST), with lower stringency near the equator.[2] TST screening tests for the presence of Tuberculosis infection but may give a false positive if the subject has been exposed to other mycobacteria or the BCG vaccine. Reduced stringency would lead to a greater number of TB positive individuals being vaccinated. These individuals would then receive no protection from the vaccine and would lead to a reduced estimate of the effectiveness of the vaccine overall.

2.2.3 Duration of protection

The effectiveness of the BCG vaccine has been shown to reduce over time.[22] However, there is good evidence that protection can last up to 10 years, with limited evidence of protection beyond 15 years.[22] Although, a recent study found that protection from active TB may extend later into life in England.[23] There is little evidence to suggest that re-vaccination boosts the protection offered by initial vaccination.[2]

The limited duration of protection has informed vaccination policy globally.[4] In countries where the BCG vaccination has been shown to be effective when given later in life, vaccination at school-age results in high levels of BCG effectiveness in young adults. As young adults are typically responsible for large amounts of TB transmission, this is likely to reduce TB incidence rates. Vaccination of neonates, on the other hand, provides protection against TB early in life. TB outcomes can be very poor at this time, but early life vaccination can lead to lower levels of protection later in life when transmission is more likely. This results in a trade-off with BCG vaccination of neonates being more effective in low effectiveness settings and in settings with lower TB incidence rates, whereas school-age vaccination is potentially more effective in settings with high BCG effectiveness and higher TB incidence rates.

2.2.4 Additional effects of BCG vaccination

Until recently little attention has been given to any additional effects of BCG vaccination.[24,25] However, there is now some evidence that BCG vaccination induces innate immune responses that may provide non-specific protection[26] and reduce all-cause neonatal mortality.[27,28] Additionally, BCG vaccination may improve outcomes for individuals with active TB disease. TB patients with BCG scars have been found to respond better to treatment with earlier sputum smear conversion.[29] There is also evidence to support an association between BCG vaccination and reduced TB[22] mortality. The evidence for additional effects of BCG vaccination on outcomes in individuals with notified TB, in England, is explored further in Chapter **??**.

In addition to its effect on TB outcomes, the BCG vaccine has also been found to be effective at preventing leprosy (with an RR of 0.45 (95% CI: 0.34-0.56)), with some evidence that this protection was stronger in those vaccinated before 15 years of age.[2] Additionally, there is some evidence that the BCG vaccine can provide protection against Non-Tuberculosis mycobacteria infections, with an estimated effectiveness of 50%.[2]

2.2.5 Usage Globally

The BCG vaccine is one of the mostly widely-used vaccines worldwide, with approximately 100m doses given annually.[30] However, due to the variable estimates of BCG efficacy, vaccination has been controversial since its development. The World Health Organisation (WHO) recommends vaccination for all neonates as early as possible after birth in high burden settings. Vaccination in low burden settings is dependent on the country specific epidemiology of TB.[2,31] This recommendation is based on the strong evidence that the BCG is highly protective in children,[17,18] whilst it's effectiveness has been shown to vary with latitude when given later in life.[32] Historically, different strategies have been utilized worldwide. This includes universal vaccination of those at most risk of on-wards transmission and high-risk group vaccination targeting either neonates or children.[22]

In addition, BCG vaccination policies have differed by the number of doses given, the method of application (although most countries now use the intradermal route), and the strain type used.[4] Policies have also changed over time within countries due to changes in evidence, global best practice, TB incidence rates and HIV incidence. This means that in order to understand the current impact of BCG vaccination in a population it is important to know the both the current vaccination policy but also historic vaccination policies.

As of 2011, among 180 countries with available data, 157 countries recommended universal BCG vaccination. The remaining 23 countries had either never implemented a universal programme or have switched to targeted vaccination of high risk individuals.[4] Most countries began universal programmes between the 1940s and 1980s due to high levels of TB incidence and strong evidence the effectiveness of the BCG vaccine.[20] In the last 20 years 49 of these countries reported changing their vaccination programme with 27 countries reporting major changes in the last 10 years.[4] Globally, in countries that have BCG vaccination policies in place, coverage is estimated to be between 70% and 100%.

2.2.6 Usage in England

In England, universal school-aged vaccination was introduced after a MRC trial in the 1950s estimated BCG's effectiveness at 80% in the white UK born population.[20] In 2005, the UK shifted from this strategy to targeted vaccination in neonates deemed at high risk.[22] This change was a reflection of current WHO vaccination policy,[31] falling TB incidence rates, an increasing proportion of TB cases occurring in the non-UK born,[1] and modelling evidence that suggested stopping the BCG schools scheme would have minimal long term effects on incidence rates.[33] The impact of this change in policy is explored throughout this thesis but in particular in Chapter 5, Chapter 7 and Chapter 10.

Since 2015, BCG vaccination has been included in the Cover Of Vaccination Evaluated Rapidly (COVER) programme, allowing coverage to be estimated in areas of England with universal vaccination (implemented due to high incidence rates based on WHO guidelines). Coverage for areas in England implementing targeted vaccination remains unknown. In London current coverage estimates are made by Local Authority and range from 5.3% to 92.1%.[1] These estimates may not be reliable as COVER has only recently begun to include returns for BCG, meaning that data quality maybe poor. Prior to the switch to targeted neonatal vaccination coverage in those at school leaving age was thought to be approximately 75%.[33]

2.2.7 Replacement vaccines

Multiple replacement vaccines are currently in clinical trials.[4,34] Vaccine candidates include both live and sub-unit vaccines. Many of these candidate vaccines serve as a boosts to the BCG vaccine, with the BCG vaccine being administered prior to the candidate vaccine.[4] Several BCG replacements are also being trialed, both based on alternative methods of attenuating TB mycobacteria and using other approaches.[34] However, in the short-to-midterm it is unlikely that a new vaccine will replace the BCG vaccine. This means that it's optimal usage is as important as ever.

2.3 Summary

- This chapter provides an overview of the natural history, risk factors, treatment, global epidemiology, and epidemiology in England and Wales of Tuberculosis.
- This chapter also details the action, effectiveness, duration of protection, effects and usage of the BCG vaccine. The only licensed vaccine for Tuberculosis.
- Motivation is given for the remaining chapters in this thesis.

Chapter 3

getTBinR: an R package for accessing and summarising the World Health Organization Tuberculosis data

3.1 Introduction

Developing tools for rapidly accessing and exploring data sets benefits the public health research community by enabling multiple data sets to be combined in a consistent manner, increasing the visibility of key data sets, and providing a framework that can be used to explore key questions of interest. Tooling also reduces the barriers to entry, allowing non-specialists to explore data sets that would otherwise be inaccessible. This widening of access may also lead to new insights and wider interest for key public health issues.

`getTBinR` is an R package [35] to facilitate working with the data [36] collected by the World Health Organization (WHO) on the country level epidemiology of Tuberculosis (TB). All data is freely available from the WHO and the package code is archived on Zenodo [37] and Github. The aim of `getTBinR` is to allow researchers, and other interested individuals, to quickly and easily gain access to a detailed TB data set and to start using it to derive key insights. It provides a consistent set of tools that can be used to rapidly evaluate hypotheses on a widely used data set before they are explored further using more complex methods or more detailed data. The functions provided in this package were developed to have sensible defaults to allow those new to the field to quickly gain key insights but also allow sufficient customisation so that experienced users may rapidly prototype new ideas.

The data sourced by `getBTinR` is collected by the WHO, via member governments, and used to compile the yearly global TB report [36]. Data collection encompasses TB incidence, TB mortality rates, the age distribution of TB cases, the proportion of drug resistant cases, case detection rates, and treatment rates. For a complete description of the data and data collection process, see [36]. These data are used by the WHO, governmental organisations and researchers to summarise country level TB epidemiology, as well as the wider global and regional picture.

The `getTBinR` package facilitates downloading the most up-to-date version of multiple TB relevant data sources from the WHO, along with the accompanying data dictionaries. It also contains functions to allow easy exploration of the data via searching data dictionaries, summarising key metrics on a regional and global level, and visualising the data in a variety of highly customisable ways. In addition, it provides both a dashboard and an automated country level report that enables the global, regional, and country level picture to be quickly summarised. An example of a potential use of the package is to explore estimates of the TB case fatality ratio [38]. In a few lines of code, using only built in package tooling, large regional differences can be discovered. Further insights can then be found by linking to other publicly available data sets or using a model based analysis. See `https://www.samabbott.co.uk/getTBinR/` for documentation. The remainder of this chapter outlines the key functionality of `getTBinR v0.5.7` and summarises the current impact.

3.2 Installation

`getTBinR` has been released to the Comprehensive R Archive Network (CRAN) and can therefore can be installed with the following code,

```
install.packages("getTBinR")
```

As `getTBinR` is under active development, the development version can be installed from GitHub with the following (provided the `devtools` package is installed),

```
# install.packages("devtools")
devtools::install_github("seabbs/getTBinR")
```

3.3 Data extraction and variable look-up

`getTBinR` provides a single user facing function for data extraction, `get_tb_burden`. This function downloads multiple data sets from the WHO, cleans variables names where required, and finally joins all data sets together. To reduce unnecessary downloads, and improve performance, downloads are cached automatically. `get_tb_burden` is called by all other package functions allowing for a seamless user experience. `get_data_dict` has similar functionality to `get_tb_burden` but extracts data dictionaries rather than the underlying data. It is called by the majority of the package functions in order to provide intelligent labels.

To improve the user experience, and to facilitate intelligent labeling, `getTBinR` provides a search function for the data dictionary (`search_data_dict`). This function is able to search, using fuzzy matching, for variables, variable descriptions, and key phrases. The code below gives an example search for `country` and `e_inc_100k` (TB incidence rate) variables, along with an accompanying search for variables referencing mortality.

```
search_data_dict(var = c("country","e_inc_100k"),
                 def = c("mortality"), verbose = FALSE)
```

```
# A tibble: 11 x 4
   variable_name     dataset      code_list definition
   <chr>             <chr>        <chr>     <chr>
```

```
 1 country           Country ide~ ""        Country or territory name
 2 e_inc_100k        Estimates    ""        Estimated incidence (all forms~
 3 e_mort_100k       Estimates    ""        Estimated mortality of TB case~
 4 e_mort_100k_hi    Estimates    ""        Estimated mortality of TB case~
 5 e_mort_100k_lo    Estimates    ""        Estimated mortality of TB case~
 6 e_mort_exc_tbhiv~ Estimates    ""        Estimated mortality of TB case~
 7 e_mort_exc_tbhiv~ Estimates    ""        Estimated mortality of TB case~
 8 e_mort_exc_tbhiv~ Estimates    ""        Estimated mortality of TB case~
 9 e_mort_tbhiv_100k Estimates    ""        Estimated mortality of TB case~
10 e_mort_tbhiv_100~ Estimates    ""        Estimated mortality of TB case~
11 e_mort_tbhiv_100~ Estimates    ""        Estimated mortality of TB case~
```

3.4 Data visualisation

`getTBInR` implements a range of functions to allow rapid development of complex visuals, with minimal R knowledge. All functions make use of cached data so that no data needs to be provided and can automatically match variables to variable names. Additionally, all visualisation functions have a common user interface, allowing for knowledge transfer between functions. As all visualisation functions return `ggplot2` objects (a commonly used R graphing library), user modification is readily supported.

Functionality that is common to all plotting functions is the ability to: plot data for a given list of countries; fuzzy match country names; plot data for a given metric present in the data; compute percentage changes from raw metric values; look up the supplied metric to see if the data dictionary contains an appropriate name; plot data over a user supplied range of years; facet over a user supplied variable; implement a user supplied transform (i.e log scaling); modify the colour palette used; and switch to comparable interactive graphics (using the `plotly` package). In addition to this, `plot_tb_burden` and `plot_tb_burden_summary` can incorporate confidence intervals. By default this is done by searching the data provided for matching variables. Function specific functionality is outlined below.

3.4.1 Mapping TB burden metrics

The `map_tb_burden` function makes use of an inbuilt, country level, shapefile (a geospatial vector data format) to produce a global or regional map of the metric supplied. Figure 3.1 gives a global overview of country level TB incidence rate. The use of a map here allows for the identification of spatial patterns that would be difficult to distinguish using other plot types. Figure 3.1 was produced with the following code,

```
map_tb_burden(metric = "e_inc_100k", verbose = FALSE)
```

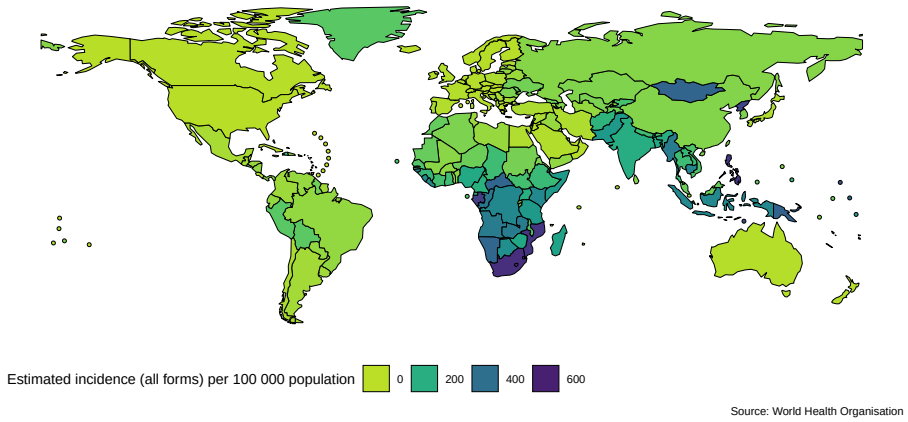


Figure 3.1: Map of global TB incidence rates in 2017 as generated by 'getTBinR'. Visualising the data with a map allows for spatial trends to be rapidly explored.

3.4.2 Plotting an overview for a given TB metric

The `plot_tb_burden_overview` function returns a dot plot that allows the trend over time of a metric to be plotted in a simplified way. Figure 3.2 shows incidence rates, by country, in Europe from 2000 to 2017. The dot plot format allows us to identify common trends across countries, after ranking for incidence rate. A more traditional line plot of the same data would be difficult to interpret due to the large number of countries. Figure 3.2 was produced with the following code,

```
plot_tb_burden_overview(metric = "e_inc_100k",
                        countries = "United Kingdom",
                        compare_to_region = TRUE,
                        interactive = FALSE,
                        verbose = FALSE)
```

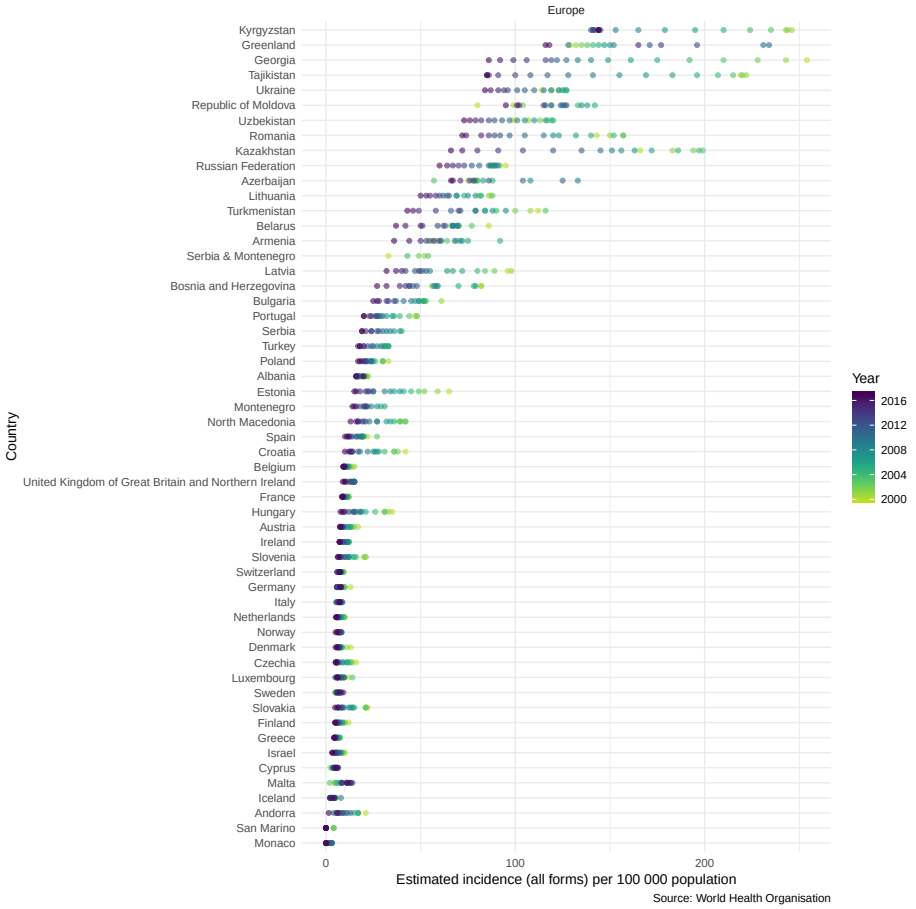


Figure 3.2: Dot plot showing trends over time in TB incidence rates in Europe ordered by TB incidence rates in 2017.

3.4.3 Plotting a comparision between country, regional and global metric values

The `plot_tb_burden_summary` function plots a regional, global, or custom overview of the supplied metric and can also include country level data for comparison. It can make use of a range of summary methods including: the country level mean, country level median, and summarised rates and proportions. Rates and proportions can be weighted with a user supplied variable but the package default is to use the summarised population. Confidence intervals are recomputed using a bootstrapping method where appropriate so that country level uncertainty is properly incorporated into the summarised metrics. The data can also be smoothed using a locally weighted regression to provide trend lines. Figure 3.3 shows a regional summary of TB incidence rates produced using `plot_tb_burden_summary`. This plot allows regional trends to be identified and compared against the global trend. Figure 3.3 was produced with the following code,

```
plot_tb_burden_summary(conf = NULL, metric_label = "e_inc_100k",
                       verbose = FALSE)
```

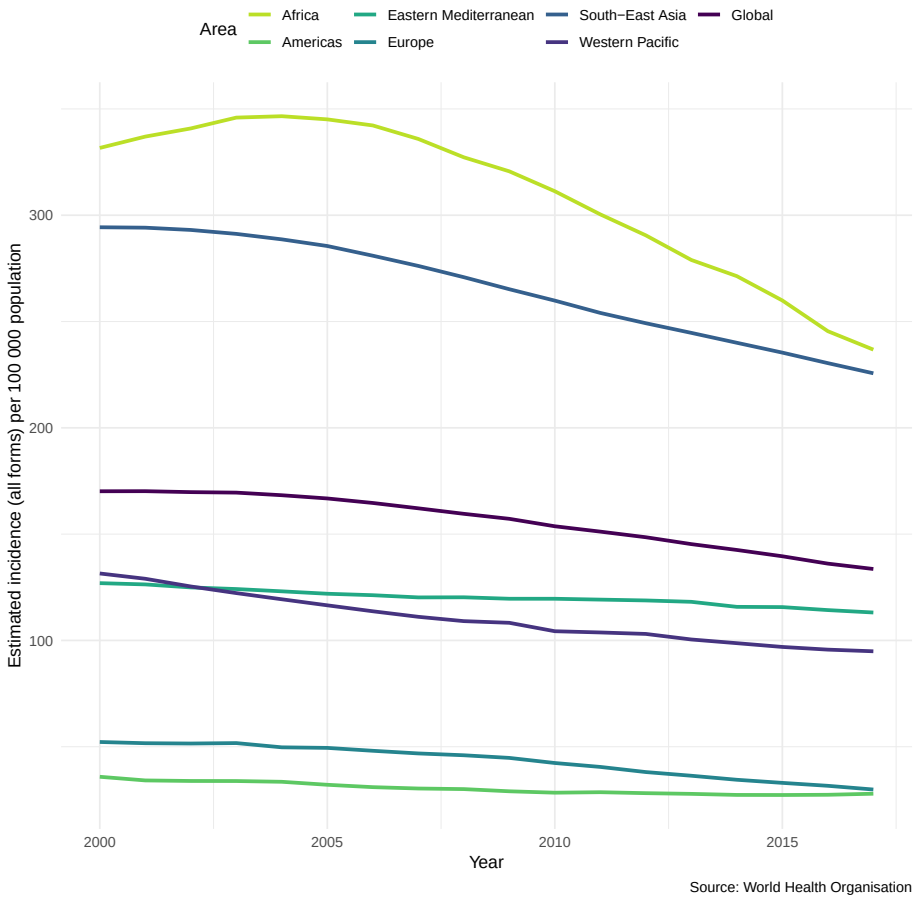


Figure 3.3: TB incidence by region and globally as computed and visualised by 'getTBinR'. Confidence intervals have been disabled in order to avoid obscuring the dominant trends.

3.5 Plotting country level trends for a given metric

The `plot_tb_burden` function plots TB trends at a country level using a simple, unaggregated, line plot. This allows for trends identified with the more complex plotting functions outlined above to be examined in more detail. Figure 3.4 shows the trend over time in TB incidence rates in the United Kingdom, along with confidence intervals. Unlike the plots above the focus on a single country allows changes over time to be more easily understood. Figure 3.4 was produced with the following code,

```
plot_tb_burden(metric = "e_inc_100k",
               countries = "United Kingdom",
               verbose = FALSE)
```

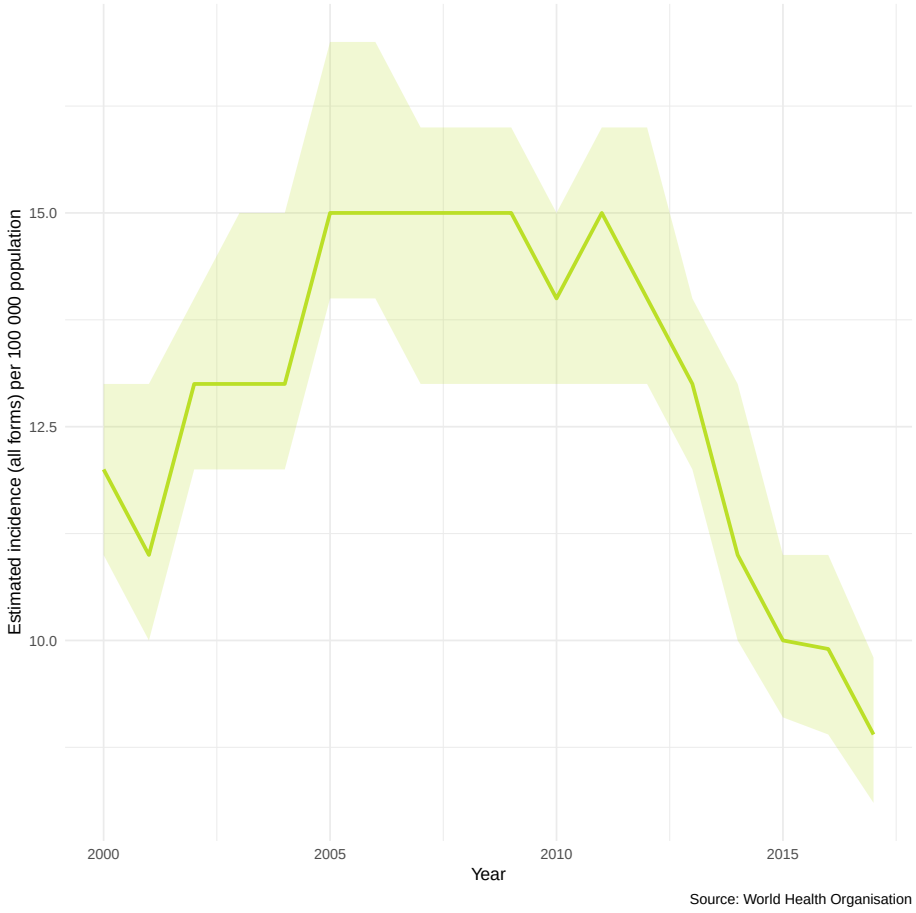


Figure 3.4: TB incidence rates over time, with confidence intervals, in the UK. As produced by 'getTBinR'.

3.6 Data summarisation

The same summarisation functionality outlined in 3.4.3 is also available in a separate function, `summarise_tb_burden`, which can be used to generate summarised data sets for further analysis or visualisation. All non-plotting functions outlined for `plot_tb_burden_summary` are implemented here. The code below summarises TB incidence rates in the UK, in Europe, and globally.

```
summarise_tb_burden(metric = "e_inc_num",
                    stat = "rate",
                    countries = "United Kingdom",
                    compare_to_world = TRUE,
                    compare_to_region = TRUE,
                    verbose = FALSE)
```

```
# A tibble: 144 x 5
   area                           year e_inc_num e_inc_num_lo e_inc_num_hi
   <fct>                         <int>     <dbl>        <dbl>        <dbl>
 1 United Kingdom of Great Brita~  2000      11.9         10.7         13.1
```

```
 2 United Kingdom of Great Brita~  2001      11.5          10.3          12.7
 3 United Kingdom of Great Brita~  2002      13.1          11.8          14.3
 4 United Kingdom of Great Brita~  2003      13.4          12.1          14.8
 5 United Kingdom of Great Brita~  2004      13.2          11.9          14.5
 6 United Kingdom of Great Brita~  2005      15.3          13.8          16.6
 7 United Kingdom of Great Brita~  2006      15.3          13.8          16.4
 8 United Kingdom of Great Brita~  2007      14.7          13.2          16.1
 9 United Kingdom of Great Brita~  2008      15.0          13.5          16.1
10 United Kingdom of Great Brita~  2009      14.5          13.1          15.9
# ... with 134 more rows
```

3.7 Dashboard

To explore the package functionality in an interactive session, or to investigate TB without having to code extensively in R, a shiny dashboard has been built into the package. This can either be used locally using,

```
run_tb_dashboard()
```

Or accessed online (`http://www.seabbs.co.uk/shiny/ExploreGlobalTB`). Any metric in the WHO TB data can be explored, with country selection using the built in map, and animation possible by year. The shiny app can also be used to generate the country level reports discussed in the next section. Figure 3.5 shows a screenshot of the dashboard, with South Africa selected as the country of interest.

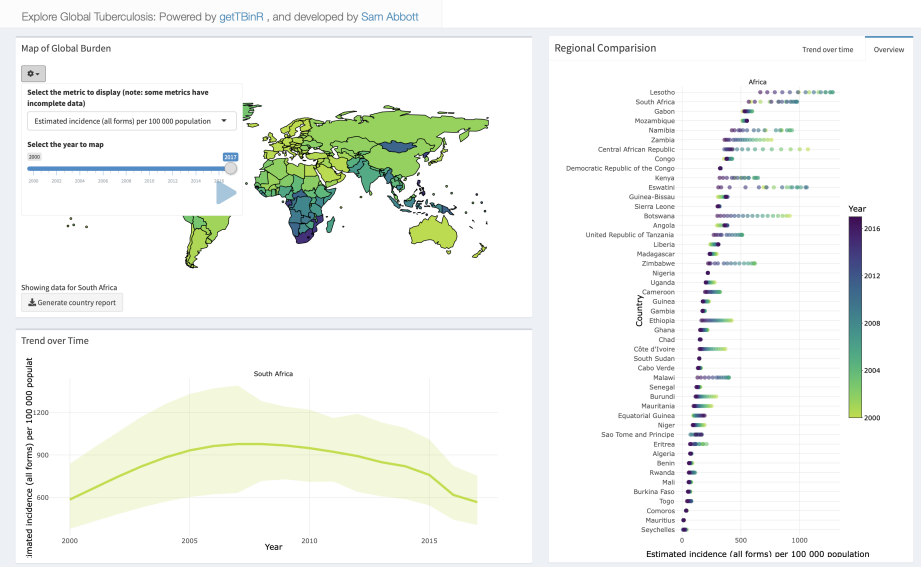


Figure 3.5: Snapshot of the built in package dashboard.

3.8 Country report

An automated country level report has also been built into `getTBinR`. This summarises key TB metrics and provides regional and global rankings. The most commonly required plots

are also produced, including the trend in TB incidence rates, proportion of cases that lead to death, and the proportion of cases with MDR TB. The report can be generated with the following code,

```
## Code saves report into your current working directory
render_country_report(country = "United Kingdom", save_dir = ".")
```

Figure 3.6 shows a screenshot of the start of the report for the United Kingdom. Note the automated reporting of country rankings in the text, along with summary metrics of interest.

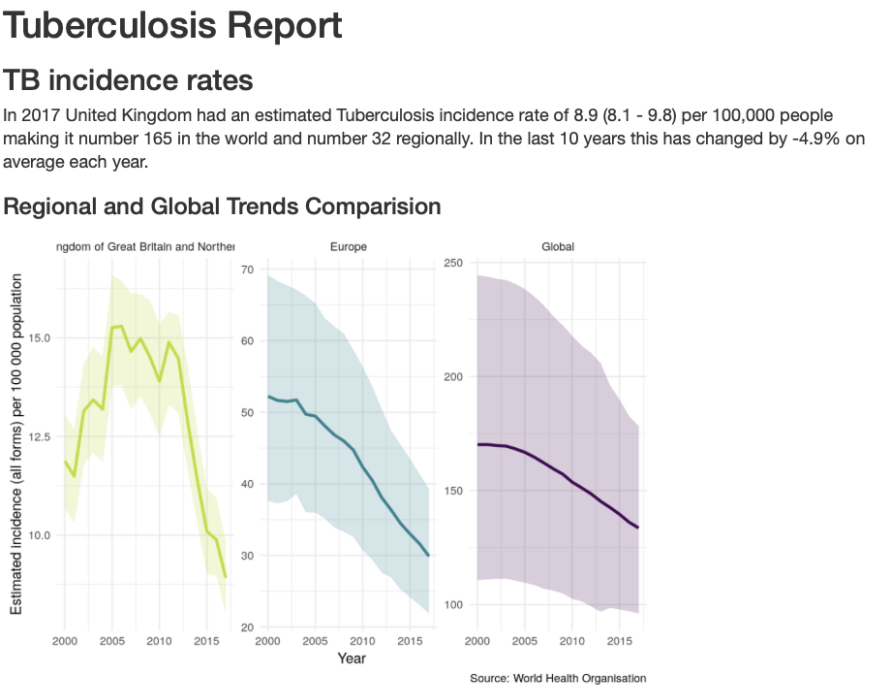


Figure 3.6: Screenshot of the start of the built in package summary report, for the United Kingdom.

3.9 Package Infrastructure

`getTBinR` has been developed using R package best practices and as such is thoroughly tested using an automated testing suite that runs against Linux, Windows and MacOS environments. Package documentation is supplied in a searchable website (`https://www.samabbott.co.uk/getTBinR/`) and a development environment can be launched with a single button press (`https://mybinder.org/v2/gh/seabbs/getTBinR/master?urlpath=rstudio`). Use cases for the package have been outlined using multiple case studies, see the package documentation for details.

3.10 Summary

- In this chapter I have introduced `getTBinR` an R package for accessing, summarising and visualising the WHO TB surveillance data set used to compile the yearly WHO

global TB report.

- I have outlined the need for data access packages in general - more specifically explaining the purpose of `getTBinR`, detailing the package functionality and outlining the package infrastructure used.
- As of the 1st of February 2019, `getTBinR` has been released on CRAN for over a year. It has been downloaded over 7000 times, has no outstanding issues, and has received multiple updates greatly expanding the functionality available. The standalone dashboard hosted online (`http://www.seabbs.co.uk/shiny/ExploreGlobalTB/`) has had over 2000 unique users.

Chapter 4

The epidemiology of tuberculosis, and the role of BCG vaccination, in England

4.1 Introduction

Whilst the characteristics of tuberculosis in England have been reported elsewhere,[1] and key risk factors such as non-UK birth status have been identified,[39] little attention has been given to the role of BCG vaccination. In particular, there is little information available regarding the demographics of vaccinated versus unvaccinated cases and the impact of BCG vaccination on TB outcomes in England has not been explored. There has also only been limited reporting of the age distribution, and trends over time, in incidence rates stratified by UK birth status.

In this chapter I explore the epidemiology of TB in England using routine data-sets, with a particular focus on the impact of missing data, the mechanisms underlying that missing date, seasonal trends, the role of age, UK birth status and BCG status. I have also estimated incidence rates, stratified by UK birth status and age, which I then used to identify trends in TB incidence over time. Finally I report TB outcomes in England using case rates, again stratified by BCG status and UK birth status. These data will then be used throughout this thesis to explore the impact of BCG vaccination on TB outcomes (Chapter 6), to estimate the direct impact of the 2005 change in BCG vaccination policy (Chapter 7), to parameterise a dynamic TB transmission model (Chapter 8) and to fit the same model to data (Chapter **??**).

4.2 Data sources

4.2.1 Enhanced tuberculosis surveillance system

Background

The enhanced tuberculosis surveillance system (ETS) is maintained by Public Health England (PHE) and collects demographic, clinical, and microbiological data on all notified cases in England. Notification is required by law, with health service providers having to inform

PHE of all confirmed TB cases.[1] Data collection began in 2000 and was expanded, with additional variables, with the launch of a web based system in 2008.[40] It is updated annually with de-notifications, late notifications and other updates. A descriptive analysis of tuberculosis epidemiology in England is published each year, which reports on data collection, cleaning, and trends in TB incidence at both a national, and sub-national level.[1]

Data extraction and management

Data on all notifications from the Enhanced Tuberculosis Surveillance (ETS) system from 2000 to 2015 were obtained from PHE. Data fields included: notification date, age, Public Health England centre, occupation, ethnic group, UK birth status, years since entry to the UK, date of symptom onset, date of presentation, date of diagnosis, date of treatment start, date of treatment end, date of death, Pulmonary TB status, culture status, sputum smear status, drug resistance, BCG vaccination status, year of vaccination, outcome at 12 months, overall outcome, and cause of death. Invalid entries were replaced with missing values unless otherwise noted, with character variables stored as factors using their most common entry as the baseline. Notifications from Scotland, Northern Ireland and Wales were dropped from the data-set. Several variables were created, or modified, for use in further analysis, Table 4.1 summarises these variables. The code used for data cleaning is available as an R package (`https://www.samabbott.co.uk/tbinenglanddataclean/reference/clean_munge_ets_2016.html`).

Table 4.1: Variables derived or modified from the Enhanced Tuberculosis Surveillance system for use in the analyses throughout this thesis.

Created/modified variable	Description
Years since BCG	Derived using year of vaccination and year of notification. Categorised into ≤ 10 and 11+ due to the evidence of waning protection for the BCG vaccine.[22]
Age at BCG	Derived using year of vaccination and age at vaccination. Categorised into < 1, 1 to x < 12, 12 to x < 16 and ≤ 16 to capture historic vaccination policy.[41]
Successful treatment	For cases that had a recorded date of starting treatment, with their outcome recorded at the latest available follow up. Those that completed treatment are defined as successfully treated; treatment failure is defined as those that stopped treatment, were lost to follow up, those that died during follow up from TB, those that died during follow up were TB contributed to their death, and those who were still on treatment. Those that were not evaluated were treated as missing.
Mortality	Assessed via follow up at 12 and 24 months; mortality is defined as cases with an overall outcome of died, and survival is defined as those that completed treatment, were still on treatment, and stopped treatment. Those that were lost to follow up, or not evaluated were treated as missing

Created/modified variable	Description
TB mortality	For cases with an overall outcome of died, and whose cause of death was known to be TB or to be related to TB. Those that were known to have not died, or who were known to have died from a cause other than from TB were defined to have not died from TB.
Death due to TB	Death due to TB is defined as those that died directly from TB, or where TB had contributed to their death with death not due to TB being cases that died from any other cause. Conditioned on all-cause mortality, for cases with a known cause of death.

Structure of the ETS

The ETS is in a wide format with each notification having a single row, and with each unique variable having a single column. This structure means that the progression of tuberculosis in each individual is captured by a series of dates rather than as a series of events. As notifications are not linked to a unique patient I.D it is possible that individuals are duplicated within the ETS, with multiple notifications. These recurrent notifications have been flagged within the data extract by the TB section at PHE. The majority of variables are factors, with a significant minority of numeric and date variables.

Data completeness

Missing data can take several forms, data that is missing completely at random (MCAR), data that is missing at random (MAR) and data that is missing not at random (MNAR).@Sterne2009a Data that is MAR is missing with a mechanism that is conditional on observed variables, whilst MNAR is missing with a mechanism that is conditional on variables that are not observed. Data that is MAR, and MNAR may lead to biases when analysing the data, however it is not possible to deduce from the observed data what the mechanism driving missing data is. Therefore, it is necessary to account for these potential biases during the analysis stage. This is possible using a variety of methods such as scenario analysis accounting for the 'best' and 'worst' case scenario's, and multiple imputation of missing data using additional variables in the data set to inform the imputation model.@Sterne2009a

As the ETS is aggregated across England, from a variety of sources, some level of missing data is inevitable. This takes two forms; under-reporting of notified cases, of which there is some evidence in the literature,[42] and data missing for a notified case. The former is particularly problematic as apart from using comparative studies the structure of those that are not notified is unknown. For variables that are missing data within the data-set it is possible to calculate the proportion of missing data (Figure 4.1, Table 4.2) but care must be taken to account for nested variables such as date of death and year of BCG vaccination. This can be done by assuming the value for the top level variable when it is known that the variable is not truly missing. An example of this is using overall outcome for data of death when notifications are known to have not died. Doing this shows high completeness for common demographic variables such as sex, age, ethnic group and UK birth status. More problematically, BCG status and year of BCG status have a high percentage missing, even

after accounting for the introduction of national collection of these variables in 2008. Socio-economic status (as national quintiles) was not collected until 2010 but after this point is highly complete. Comparing pre 2009 and post 2008 in Table 4.2 (and by inspecting Figure 4.1) there are also issues of changing completeness over time,[1,43] if this is not accounted for than it may lead to spurious trends. Figure 4.1 also indicates that there are multiple groups of variables that share a common pattern of missing data.

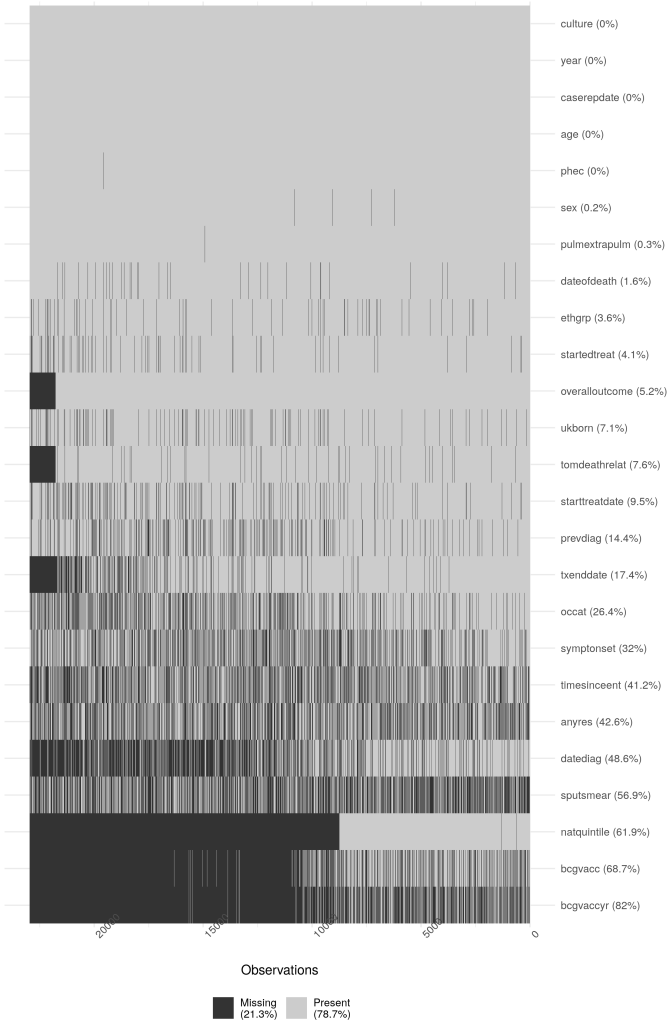


Figure 4.1: Summary plot of missing data in the extract of the Enhanced Tuberculosis Surveillance data used in this thesis. Due to the large size of the dataset, the data has been sub-sampled with only 20% of the data shown in this figure. Notifications have been ordered by date of notification from left to right. The following subset of variables are shown; year (year), sex (sex), age (age), Public Health England Centre (phec), Occupation (occat), Ethnic group (ethgrp), UK birth status (ukborn), Time since entry (timesinceent), date of symptom onset (symptonset), date of diagnosis (datediag), started treatment (startedtreat), date of starting treatment (starttreatdate), treatment end date (txenddate), pulmonary or extra-pulmonary TB (pulmextrapulm), culture (culture), sputum smear status (sputsmear), drug resistance (anyres), previous diagnosis (prevdiag), BCG status(bcgvacc), Year of BCG vaccination (bcgvaccyr), overall outcome (overalloutcome), cause of death (tomdeathrelate), socio-economic status quintiles (natquintile), and date of death (dateofdeath). Nested variables have been accounted for (i.e data of death has had an entry added for cases that are known to have not died), so that true missingness for all variables is estimated.

Table 4.2: Breakdown of missing data from the ETS prior to the web based system (pre 2009) and post (post 2008) by variable, ordered by the percentage missing for a subset of variables. The following subset of variables are shown; year (year), sex (sex), age (age), Public Health England Centre (phec), Occupation (occat), Ethnic group (ethgrp), UK birth status (ukborn), Time since entry (timesinceent), date of symptom onset (symptonset), date of diagnosis (datediag), started treatment (startedtreat), date of starting treatment (starttreatdate), treatment end date (txenddate), pulmonary or extra-pulmonary TB (pulmextrapulm), culture (culture), sputum smear status (sputsmear), drug resistance (anyres), previous diagnosis (prevdiag), BCG status(bcgvacc), Year of BCG vaccination (bcgvaccyr), overall outcome (overalloutcome), cause of death (tomdeathrelate), socio-economic status quintiles (natquintile), and date of death (dateofdeath). Nested variables have been accounted for (i.e data of death has had an entry added for cases that are known to have not died), so that true missingness for all variables is estimated.

	Pre 2009		Post 2008	
Variable	Missing (N)	Missing (%)	Missing (N)	Missing (%)
natquintile	63175	100.0	8120	15.7
bcgvaccyr	62479	98.9	31421	60.8
bcgvacc	61916	98.0	17133	33.2
datediag	45557	72.1	10303	19.9
sputsmear	32912	52.1	32094	62.1
timesinceent	29084	46.0	18670	36.2
anyres	27485	43.5	20995	40.7
occat	24870	39.4	5513	10.7
symptonset	23937	37.9	12829	24.8
txenddate	18711	29.6	1137	2.2
prevdiag	13204	20.9	3148	6.1
starttreatdate	9151	14.5	2127	4.1
tomdeathrelat	7539	11.9	1191	2.3
ukborn	6230	9.9	1825	3.5
overalloutcome	6044	9.6	0	0.0
startedtreat	4242	6.7	602	1.2
ethgrp	2811	4.4	1229	2.4
dateofdeath	1235	2.0	357	0.7
pulmextrapulm	177	0.3	213	0.4
sex	101	0.2	110	0.2
phec	32	0.1	0	0.0
age	25	0.0	0	0.0
caserepdate	0	0.0	0	0.0
year	0	0.0	0	0.0
culture	0	0.0	0	0.0

For nested variables, with rare outcomes, assuming the top level variable value can mask the underlying amount of missing data. An alternative approach is to filter the data for the top level variable required for the nested variable to be defined and to then compute the proportion of these notifications that are missing the data for the outcome of interest. For the date of starting treatment this approach leads to an estimate of 5.9% (6434/108410) being missing, which is more complete than previously estimated. For cases that are known to have completed treatment 16.5% (13804/83891) are missing a date for the end of treatment. In notifications that are known to have died, 26.6% (1592/5976) were missing the date of death and 44.9% (2686/5976) were missing the cause of death. In any analysis were these variables are used the missing data for these variables will need to be carefully adjusted for. In particular, if cause of death is used it must be clearly stated that it is highly missing and results based on this variable should be properly caveated.

Drivers of Variable completeness

As previously discussed, missing data may be MAR or MNAR, which may introduce biases into any analyses based on these data. This is of particular importance for variables that have high levels of missingness, as any introduced bias is likely to have a greater impact on the overall results, and for variables that are used extensively in analyses later in this thesis. Unfortunately MNAR data cannot be detected, so bias from this source cannot be discounted. However, it is possible to detect MAR mechanisms from observed variables that would not necessarily be included in a model used for analysis. For this reason, in the following section, I explore variables associated with data being missing for several key variables including: BCG status, year of BCG vaccination, date of death, cause of death, date of symptom onset, date of diagnosis, date of starting treatment and date of ending treatment. All of these variables were shown to have high levels of missing data in the previous section and they will all be used extensively throughout this thesis.

In order to explore the drivers of missing data I have reformulated the problem as a logistic regression for each variable of interest, with the outcome being data completeness (complete/missing). This allows variables that are hypothesised to be related to missing data to be adjusted for and their independent impact on data completeness to be measured. Unlike classic approaches to missing data, such as multiple imputation by chained regression (MICE),[44] this is not an imputation but insteaqd allows associations between observed varibales and missing outcome data to be estimated. The details of the approach are discussed below.

Method In order to reformulate missing data as a logistic regression the following steps must be taken,

1. For the variable of interest create a new temporary binary variable, called data status, that is “Missing” when the variable of interest is missing and “Complete” when it is not. Specify “Complete” as the baseline.

2. For nested variables exclude notifications that do not have the top level outcome required by the variable of interest. An example of this is excluding cases that did not die, or have a missing overall outcome, when investigating TB mortality.

3. Specify the hypothesised drivers of missingness for the variable of interest. These should be variables with a reasonable hypothesis for how they would drive missingness

in the variable of interest. They must also be relatively complete as this approach does not impute missing confounder data.

4. Fit a logistic regression model; with the temporary data status variable as the outcome, adjusting for the hypothesised drivers of missingness.
5. Exponentiate the returned coefficients, and confidence intervals, so that they represent Odds Ratios (ORs).
6. Refit the model, dropping each variable in turn and then comparing the updated model with the full model using a likelihood ratio test.
7. Interpret the results, using the estimated size of the effect, the width of the confidence intervals and the size of the Wald and likelihood ratio test p values to determine which variables are related to missingness for the variable of interest. Evidence should be interpreted on a spectrum, rather than using arbitrary significance cut-offs.[45] To avoid issues of multiple testing the level of evidence should be weighted based on the number of variables adjusted for and the number of outcomes explored.

For all outcomes considered I adjusted for the same set of demographic variables that where both highly complete and also plausibly linked to missingness for all outcomes considered. They are; year, sex, age (grouped as 0-14 year old's, 15-65 year old's and 65+), ethnic group, UK birth status and socio-economic status (national quintiles). For socio-economic group 1 indicates the most deprived quintile. Complete case analysis has been used, with the dataset limited to notifications from 2010 and on-wards as socio-economic status was not collected prior to this.

BCG status It is clear that BCG status is missing with a MAR mechanism for the variables considered (Table 4.3). BCG data being missing is strongly associated with year of notification, sex age, ethnic group, and socio-economic status. It appears that after adjusting for other variables data completeness increased from 2010 until 2012 but has since showed no clear trend. Men appear to be more likely than women to have a missing BCG status, with the non-UK born also being more likely than the UK born to be missing BCG status. The proportion of those missing BCG status increases with age, with those aged 65+ being over 4 times more likely to be missing BCG status than those aged 0-14 years old. There is also evidence to suggest that notifications in the lowest socio-economic group are more likely to have a missing BCG status but there was no clear evidence of a trend across socio-economic quintiles. The White ethnic group was more likely to have a missing BCG status than any other ethnic group.

Table 4.3: Results from a logistic regression model with data completeness (Complete/Missing) for BCG vaccination as an outcome, adjusted for; year, sex, age (grouped as 0-14 year olds, 15-65 year olds and 65+), ethnic group, UK birth status and socio-economic status (national quintiles). For socio-economic group 1 indicates the most deprived quintile. Notifications from 2010 onwards were included as socio-economic status was not collected before this. Complete case analysis weas used. Odds ratios shown are adjusted for all explanatory variables. The model indicates that BCG status is missing at random for the variables considered.

Variable	Category	Missing (N)	Notifications (41659)	Odds Ratio	P value (Wald)	P value (LRT)
Year	2010	31.3% (2235)	7143			1.27e-08
	2011	29.8% (2319)	7781	0.93 (0.87, 1.00)	0.0606	
	2012	27.9% (2164)	7755	0.85 (0.79, 0.91)	6.32e-06	
	2013	27.1% (1907)	7034	0.79 (0.74, 0.86)	1.01e-09	
	2014	30.1% (1907)	6327	0.91 (0.85, 0.98)	0.017	
	2015	29.7% (1668)	5619	0.89 (0.82, 0.96)	0.00348	
Sex	Female	27.4% (4847)	17664			8.74e-11
	Male	30.6% (7353)	23995	1.16 (1.11, 1.21)	9.49e-11	
Age	0-14	13.1% (235)	1793			1.67e-157
	15-44	26.0% (6557)	25235	2.10 (1.82, 2.43)	5.83e-24	
	45-64	32.8% (2964)	9026	2.84 (2.45, 3.30)	3.16e-43	
	65+	43.6% (2444)	5605	4.42 (3.80, 5.15)	1.43e-81	
Ethnic group	White	35.4% (2959)	8359			2.15e-41
	Black-Caribbean	24.6% (228)	928	0.62 (0.52, 0.72)	3.96e-09	
	Black-African	27.3% (1966)	7204	0.73 (0.67, 0.80)	1.34e-12	
	Black-Other	24.1% (89)	369	0.65 (0.51, 0.83)	0.000717	
	Indian	25.9% (2805)	10848	0.62 (0.58, 0.68)	5.2e-31	
	Pakistani	33.2% (2258)	6806	0.89 (0.82, 0.97)	0.00569	
	Bangladeshi	27.9% (469)	1680	0.71 (0.62, 0.80)	1.03e-07	
	Chinese	33.6% (166)	494	0.88 (0.72, 1.07)	0.202	
	Mixed / Other	25.3% (1260)	4971	0.65 (0.59, 0.71)	4.93e-20	
UK birth status	Non-UK Born	29.5% (9104)	30880			7.2e-18
	UK Born	28.7% (3096)	10779	0.75 (0.70, 0.80)	1.3e-17	
Socio-economic status	1	30.7% (4948)	16131			4.88e-08
	2	26.8% (3383)	12621	0.84 (0.80, 0.89)	3.58e-10	
	3	29.2% (1905)	6530	0.92 (0.86, 0.98)	0.0117	
	4	30.1% (1142)	3796	0.91 (0.84, 0.99)	0.0264	
	5	31.8% (822)	2581	0.94 (0.85, 1.03)	0.174	

Year of BCG vaccination As for BCG status, year of BCG vaccination is also clearly missing with MAR mechanisms for the variables considered (Table 4.4). As for BCG status men were more likely to have a missing year of BCG vaccination as were the non-UK born. Older notifications were again more likely to have missing data, with those aged 65+ being more than 2 times more likely to have a missing year of vaccination. However, unlike BCG vaccination status, year of notification shows a clear trend of increasing data completeness from 2010 until 2015. Additionally, for year of BCG vaccination the White ethnic group is more likely to have complete data than any other ethnic group, with those of Black-Caribbean descent being over 3 times more likely to have a missing year of BCG vaccination. Socio-economic status is highly associated with year of vaccination being missing but there is little clear evidence of a trend. The second, and third, poorest quintiles were more likely to have a missing year of vaccination. Whilst the richest, and second richest quintiles were

less likely to have a missing year of vaccination.

Table 4.4: Results from a logistic regression model with data completeness (Complete/Missing) for year of BCG vaccination as an outcome, adjusted for; year, sex, age (grouped as 0-14 year olds, 15-65 year olds and 65+), ethnic group, UK birth status and socio-economic status (national quintiles).For socio-economic group 1 indicates the most deprived quintile. Notifications from 2010 onwards were included as socio-economic status was not collected before this. Complete case analysis weas used. Odds ratios shown are adjusted for all explanatory variables. The model indicates that year of BCG vaccination is missing at random for the variables considered.

Variable	Category	Missing (N)	Notifications (20835)	Odds Ratio	P value (Wald)	P value (LRT)
Year	2010	61.0% (2090)	3424			2.03e-07
	2011	59.6% (2304)	3869	0.93 (0.84, 1.03)	0.149	
	2012	56.2% (2216)	3945	0.82 (0.75, 0.91)	7.74e-05	
	2013	55.7% (2025)	3638	0.82 (0.74, 0.90)	5.17e-05	
	2014	56.6% (1776)	3138	0.86 (0.77, 0.95)	0.00279	
	2015	54.2% (1530)	2821	0.75 (0.67, 0.83)	5.8e-08	
Sex	Female	55.5% (5089)	9174			6.9e-06
	Male	58.8% (6852)	11661	1.14 (1.08, 1.21)	6.89e-06	
Age	0-14	43.9% (488)	1111			3.94e-14
	15-44	58.3% (8216)	14102	1.54 (1.34, 1.76)	3.44e-10	
	45-64	57.6% (2526)	4388	1.66 (1.44, 1.93)	6.62e-12	
	65+	57.6% (711)	1234	2.02 (1.69, 2.42)	1.5e-14	
Ethnic group	White	44.2% (1370)	3102			5.94e-82
	Black-Caribbean	77.5% (371)	479	3.91 (3.12, 4.95)	3.56e-31	
	Black-African	65.2% (2524)	3870	1.83 (1.63, 2.05)	5.44e-25	
	Black-Other	72.0% (154)	214	2.89 (2.12, 3.99)	4.01e-11	
	Indian	56.1% (3516)	6267	1.17 (1.06, 1.30)	0.00247	
	Pakistani	51.6% (1583)	3066	1.09 (0.97, 1.22)	0.136	
	Bangladeshi	73.1% (583)	797	2.67 (2.23, 3.20)	4.74e-26	
	Chinese	58.2% (142)	244	1.43 (1.09, 1.89)	0.0111	
	Mixed / Other	60.7% (1698)	2796	1.50 (1.33, 1.69)	3.32e-11	
UK birth status	Non-UK Born	61.1% (9665)	15808			4.35e-28
	UK Born	45.3% (2276)	5027	0.64 (0.59, 0.69)	4.37e-28	
Socio-economic status	1	55.4% (4221)	7615			2.2e-124
	2	66.3% (4463)	6729	1.60 (1.49, 1.72)	3.9e-39	
	3	59.4% (2019)	3401	1.22 (1.12, 1.33)	5.5e-06	
	4	45.3% (838)	1848	0.71 (0.64, 0.79)	4.25e-10	
	5	32.2% (400)	1242	0.41 (0.36, 0.47)	1.43e-39	

Date of death For date of death there is some evidence that data is missing with an MAR mechanism for ethnic group and socio-economic status, with little evidence for any other association (Table 4.5). These associations should be interpreted carefully due to the strength of the evidence when compared to the number of tests conducted, there is a high likelihood of a type 1 error. Whilst the confidence intervals were wide for all ethnic groups there was some weak indication that the White ethnic group were more likely to have a complete date of death than other ethnic groups. Similarly, those in the lowest socio-economic group were somewhat more likely to have a complete date of death than other quintiles. The reduction in the levels of evidence found for case of death may be linked to the reduction in power for this outcome, as mortality is a rare outcome.

Table 4.5: Results from a logistic regression model with data completeness (Complete/Missing) for date of death as an outcome, adjusted for; year, sex, age (grouped as 0-14 year olds, 15-65 year olds and 65+), ethnic group, UK birth status and socio-economic status (national quintiles). For socio-economic group 1 indicates the most deprived quintile. Notifications from 2010 onwards were included as socio-economic status was not collected before this. Complete case analysis weas used. Odds ratios shown are adjusted for all explanatory variables. The model indicates that there is some evidence that date of death is missing at random for ethnic group, with weaker evidence for all other variables.

Variable	Category	Missing (N)	Notifications (1883)	Odds Ratio	P value (Wald)	P value (LRT)
Year	2010	16.6% (53)	320			0.0876
	2011	15.9% (52)	327	0.95 (0.62, 1.46)	0.818	
	2012	14.5% (51)	351	0.81 (0.53, 1.25)	0.342	
	2013	13.5% (42)	312	0.73 (0.46, 1.14)	0.163	
	2014	9.5% (30)	317	0.52 (0.32, 0.84)	0.0081	
	2015	13.3% (34)	256	0.69 (0.43, 1.11)	0.133	
Sex	Female	14.8% (97)	657			0.609
	Male	13.5% (165)	1226	0.93 (0.70, 1.23)	0.608	
Age	0-14	10.0% (1)	10			0.929
	15-44	15.7% (31)	198	1.90 (0.32, 36.43)	0.556	
	45-64	14.6% (68)	465	1.92 (0.33, 36.42)	0.549	
	65+	13.4% (162)	1210	1.95 (0.34, 37.04)	0.536	
Ethnic group	White	11.1% (102)	920			0.00373
	Black-Caribbean	21.7% (10)	46	1.58 (0.67, 3.51)	0.274	
	Black-African	20.1% (27)	134	1.49 (0.76, 2.94)	0.251	
	Black-Other	20.0% (1)	5	1.59 (0.08, 11.72)	0.687	
	Indian	17.4% (64)	367	1.08 (0.62, 1.92)	0.789	
	Pakistani	8.0% (20)	249	0.50 (0.25, 0.99)	0.0483	
	Bangladeshi	22.7% (10)	44	1.65 (0.67, 3.87)	0.261	
	Chinese	14.3% (3)	21	0.89 (0.19, 3.00)	0.864	
	Mixed / Other	25.8% (25)	97	1.99 (1.01, 3.92)	0.0462	
UK birth status	Non-UK Born	16.6% (167)	1004			0.133
	UK Born	10.8% (95)	879	0.67 (0.40, 1.14)	0.128	
Socio-economic status	1	11.4% (79)	695			0.0265
	2	18.3% (86)	470	1.67 (1.19, 2.35)	0.0033	
	3	16.2% (48)	296	1.49 (0.99, 2.22)	0.0548	
	4	12.7% (30)	237	1.21 (0.75, 1.90)	0.429	
	5	10.3% (19)	185	0.95 (0.54, 1.62)	0.866	

Cause of death For cause of death there is less evidence of an MAR mechanism, with little evidence of an association for year, sex, age, or socio-economic group (Table 4.6). There was, however, strong evidence of an association with ethnic group and very weak evidence of an association with UK birth status. The White ethnic group was less likely to have an incomplete cause of death when compared to the majority of other identified ethnic groups but there was evidence to suggest that cause of death was more likely to be missing in those identifying as being of Black-Caribbean, Black-Other, Indian and Bangladeshi descent. The confidence intervals for these estimates were wide, indicating that these estimates may not be reliable. There was again some weak evidence to suggest that the UK born were more likely to be missing a cause of death than the non-UK born, which reverses the trend observed in the other variables explored. The reduction in the levels of evidence found for

case of death may be linked to the reduction in power for this outcome, as mortality is a rare outcome.

Table 4.6: Results from a logistic regression model with data completeness (Complete/Missing) for cause of death as an outcome, adjusted for; year, sex, age (grouped as 0-14 year olds, 15-65 year olds and 65+), ethnic group, UK birth status and socio-economic status (national quintiles). For socio-economic group 1 indicates the most deprived quintile. Notifications from 2010 onwards were included as socio-economic status was not collected before this. Complete case analysis weas used. Odds ratios shown are adjusted for all explanatory variables. The model indicates that cause of death is missing at random for ethnic group and UK birth status, with little evidence for any other variables

Variable	Category	Missing (N)	Notifications (1883)	Odds Ratio	P value (Wald)	P value (LRT)
Year	2010	45.0% (144)	320			0.724
	2011	45.6% (149)	327	1.03 (0.75, 1.41)	0.85	
	2012	45.3% (159)	351	1.02 (0.75, 1.39)	0.905	
	2013	43.9% (137)	312	0.99 (0.72, 1.37)	0.954	
	2014	44.8% (142)	317	0.96 (0.70, 1.32)	0.793	
	2015	38.7% (99)	256	0.80 (0.57, 1.12)	0.196	
Sex	Female	44.7% (294)	657			0.628
	Male	43.7% (536)	1226	0.95 (0.78, 1.16)	0.628	
Age	0-14	50.0% (5)	10			0.116
	15-44	35.4% (70)	198	0.64 (0.17, 2.48)	0.509	
	45-64	43.0% (200)	465	0.90 (0.24, 3.44)	0.874	
	65+	45.9% (555)	1210	0.96 (0.25, 3.67)	0.957	
Ethnic group	White	48.2% (443)	920			0.000704
	Black-Caribbean	21.7% (10)	46	0.40 (0.18, 0.82)	0.0173	
	Black-African	45.5% (61)	134	1.41 (0.85, 2.36)	0.183	
	Black-Other	20.0% (1)	5	0.41 (0.02, 2.87)	0.428	
	Indian	35.7% (131)	367	0.83 (0.55, 1.27)	0.388	
	Pakistani	49.4% (123)	249	1.47 (0.95, 2.29)	0.0857	
	Bangladeshi	27.3% (12)	44	0.60 (0.27, 1.26)	0.189	
	Chinese	52.4% (11)	21	1.64 (0.64, 4.23)	0.302	
	Mixed / Other	39.2% (38)	97	1.00 (0.58, 1.72)	0.991	
UK birth status	Non-UK Born	40.1% (403)	1004			0.072
	UK Born	48.6% (427)	879	1.41 (0.97, 2.07)	0.073	
Socio-economic status	1	43.7% (304)	695			0.345
	2	40.0% (188)	470	0.93 (0.72, 1.18)	0.54	
	3	42.9% (127)	296	0.98 (0.74, 1.31)	0.916	
	4	49.8% (118)	237	1.24 (0.91, 1.69)	0.172	
	5	50.3% (93)	185	1.21 (0.86, 1.71)	0.262	

Date of symptom onset For date of symptom onset there was strong evidence of an MNAR mechanism for all variables considered, except for sex (Table 4.7). As found previously, the likelihood of date of symptom onset being missing reduced with year of notification. Children (0-14 years old) were more likely to have a missing date of symptom onset than any other age group as were those in any socio-economic quintile when compared to the poorest group. UK born cases were more likely to have a complete date of symptom onset than non-UK born cases, with the White ethnic group being more likely to have a missing date of symptom onset than most other ethnic groups.

Table 4.7: Results from a logistic regression model with data completeness (Complete/Missing) for date of symptom onset as an outcome, adjusted for; year, sex, age (grouped as 0-14 year olds, 15-65 year olds and 65+), ethnic group, UK birth status and socio-economic status (national quintiles).For socio-economic group 1 indicates the most deprived quintile. Notifications from 2010 onwards were included as socio-economic status was not collected before this. Complete case analysis weas used. Odds ratios shown are adjusted for all explanatory variables. The model indicates that date of symptom onset is missing not at random for the variables for all variables considered, except for sex.

Variable	Category	Missing (N)	Notifications (41659)	Odds Ratio	P value (Wald)	P value (LRT)
Year	2010	34.0% (2426)	7143			0
	2011	30.1% (2339)	7781	0.83 (0.78, 0.89)	2.03e-07	
	2012	24.2% (1878)	7755	0.61 (0.57, 0.66)	2.24e-40	
	2013	17.5% (1233)	7034	0.41 (0.38, 0.44)	8.22e-108	
	2014	11.8% (744)	6327	0.25 (0.23, 0.28)	1.52e-188	
	2015	6.9% (390)	5619	0.14 (0.13, 0.16)	8.88e-245	
Sex	Female	22.0% (3894)	17664			0.93
	Male	21.3% (5116)	23995	1.00 (0.95, 1.05)	0.93	
Age	0-14	38.1% (684)	1793			3.59e-73
	15-44	20.5% (5182)	25235	0.35 (0.31, 0.39)	6.74e-77	
	45-64	20.7% (1870)	9026	0.37 (0.33, 0.42)	2.18e-58	
	65+	22.7% (1274)	5605	0.43 (0.38, 0.49)	2.31e-39	
Ethnic group	White	20.9% (1749)	8359			3.98e-09
	Black-Caribbean	23.1% (214)	928	1.04 (0.88, 1.23)	0.658	
	Black-African	23.0% (1654)	7204	0.89 (0.80, 0.98)	0.0179	
	Black-Other	18.7% (69)	369	0.79 (0.60, 1.04)	0.106	
	Indian	22.2% (2404)	10848	0.86 (0.79, 0.94)	0.00119	
	Pakistani	19.2% (1305)	6806	0.75 (0.68, 0.83)	5.56e-09	
	Bangladeshi	23.9% (401)	1680	1.05 (0.91, 1.20)	0.524	
	Chinese	18.8% (93)	494	0.74 (0.58, 0.94)	0.016	
	Mixed / Other	22.6% (1121)	4971	0.93 (0.83, 1.03)	0.152	
UK birth status	Non-UK Born	21.9% (6774)	30880			5.44e-12
	UK Born	20.7% (2236)	10779	0.77 (0.71, 0.83)	7.6e-12	
Socio-economic status	1	19.9% (3218)	16131			5e-17
	2	22.9% (2888)	12621	1.22 (1.15, 1.29)	9.51e-11	
	3	24.2% (1578)	6530	1.33 (1.24, 1.43)	5.79e-15	
	4	22.0% (837)	3796	1.20 (1.09, 1.31)	8.72e-05	
	5	18.9% (489)	2581	1.00 (0.89, 1.12)	0.991	

Date of diagnosis For date of diagnosis there was again strong evidence for an MAR mechanism for all variables considered, except for sex for which there was very weak evidence (Table 4.8). Increasing completeness was found for year of notification as seen previously, as was an increased likelihood of missing data in males and the non-UK born. The White ethnic group was less likely to be missing data on the data of diagnosis as compared to the majority of other ethnic groups, as were the poorest socio-economic group compared to all other socio-economic quintiles. Children (0-14 years old) were again more likely to be missing data than adults in any age group.

Table 4.8: Results from a logistic regression model with data completeness (Complete/Missing) for date of diagnosis onset as an outcome, adjusted for; year, sex, age (grouped as 0-14 year olds, 15-65 year olds and 65+), ethnic group, UK birth status and socio-economic status (national quintiles).For socio-economic group 1 indicates the most deprived quintile. Notifications from 2010 onwards were included as socio-economic status was not collected before this. Complete case analysis weas used. Odds ratios shown are adjusted for all explanatory variables. The model indicates that date of diagnosis is missing at random for the variables for all variables considered, except for sex.

Variable	Category	Missing (N)	Notifications (41659)	Odds Ratio	P value (Wald)	P value (LRT)
Year	2010	26.9% (1918)	7143			1.65e-283
	2011	22.3% (1736)	7781	0.78 (0.72, 0.84)	6.72e-11	
	2012	18.8% (1458)	7755	0.63 (0.58, 0.68)	4.31e-31	
	2013	12.9% (909)	7034	0.41 (0.37, 0.44)	1.46e-89	
	2014	10.4% (659)	6327	0.32 (0.29, 0.35)	1.05e-118	
	2015	7.4% (415)	5619	0.22 (0.19, 0.24)	1.63e-154	
Sex	Female	16.9% (2984)	17664			0.0296
	Male	17.1% (4111)	23995	1.06 (1.01, 1.12)	0.0298	
Age	0-14	19.4% (348)	1793			0.000164
	15-44	17.8% (4504)	25235	0.76 (0.67, 0.87)	3.85e-05	
	45-64	15.9% (1434)	9026	0.73 (0.64, 0.84)	1.54e-05	
	65+	14.4% (809)	5605	0.72 (0.62, 0.84)	1.84e-05	
Ethnic group	White	12.5% (1043)	8359			2.91e-67
	Black-Caribbean	25.2% (234)	928	2.21 (1.87, 2.61)	2.58e-20	
	Black-African	21.9% (1577)	7204	1.49 (1.34, 1.66)	5.45e-13	
	Black-Other	17.9% (66)	369	1.32 (0.98, 1.74)	0.0587	
	Indian	18.0% (1957)	10848	1.09 (0.99, 1.21)	0.0858	
	Pakistani	11.8% (805)	6806	0.75 (0.67, 0.84)	8.98e-07	
	Bangladeshi	21.5% (361)	1680	1.57 (1.35, 1.82)	2.63e-09	
	Chinese	13.4% (66)	494	0.82 (0.61, 1.07)	0.153	
	Mixed / Other	19.8% (986)	4971	1.32 (1.18, 1.48)	1.84e-06	
UK birth status	Non-UK Born	18.4% (5696)	30880			6.07e-16
	UK Born	13.0% (1399)	10779	0.71 (0.65, 0.77)	1.23e-15	
Socio-economic status	1	14.4% (2317)	16131			1.05e-45
	2	19.6% (2469)	12621	1.48 (1.39, 1.58)	1.7e-32	
	3	20.3% (1325)	6530	1.62 (1.50, 1.75)	8.07e-34	
	4	17.0% (645)	3796	1.37 (1.24, 1.52)	4.77e-10	
	5	13.1% (339)	2581	1.07 (0.94, 1.21)	0.313	

Date of starting treatment For date of starting treatment there is little evidence that missing data is associated with any variable considered, except for year of notification (Table 4.9). Variable completeness improved year on year, with a 96% drop in missing data in 2015 compared to 2010.

Table 4.9: Results from a logistic regression model with data completeness (Complete/Missing) for date of starting treatment as an outcome, adjusted for; year, sex, age (grouped as 0-14 year olds, 15-65 year olds and 65+), ethnic group, UK birth status and socio-economic status (national quintiles).For socio-economic group 1 indicates the most deprived quintile. Notifications from 2010 onwards were included as socio-economic status was not collected before this. Complete case analysis weas used. Odds ratios shown are adjusted for all explanatory variables. There is little evidence that the missing data for the date of starting treatment is associated with any variable considered, except for year of notification.

Variable	Category	Missing (N)	Notifications (40977)	Odds Ratio	P value (Wald)	P value (LRT)
Year	2010	3.5% (244)	7020			2.4e-70
	2011	3.2% (242)	7655	0.91 (0.76, 1.08)	0.281	
	2012	2.5% (187)	7628	0.69 (0.57, 0.84)	0.000211	
	2013	2.2% (154)	6923	0.63 (0.51, 0.77)	8.29e-06	
	2014	0.8% (51)	6239	0.23 (0.17, 0.31)	1.29e-21	
	2015	0.1% (8)	5512	0.04 (0.02, 0.08)	3.62e-19	
Sex	Female	2.2% (383)	17439			0.83
	Male	2.1% (503)	23538	0.99 (0.86, 1.13)	0.83	
Age	0-14	3.0% (54)	1783			0.157
	15-44	2.2% (539)	25000	0.72 (0.53, 0.98)	0.0303	
	45-64	2.0% (180)	8896	0.68 (0.49, 0.95)	0.0209	
	65+	2.1% (113)	5298	0.69 (0.49, 0.99)	0.042	
Ethnic group	White	2.3% (182)	8055			0.423
	Black-Caribbean	2.2% (20)	916	0.89 (0.54, 1.39)	0.626	
	Black-African	1.9% (139)	7140	0.73 (0.55, 0.96)	0.025	
	Black-Other	3.0% (11)	368	1.33 (0.67, 2.38)	0.379	
	Indian	2.1% (230)	10707	0.86 (0.67, 1.10)	0.23	
	Pakistani	2.4% (158)	6721	0.92 (0.72, 1.19)	0.536	
	Bangladeshi	2.2% (37)	1665	0.88 (0.59, 1.29)	0.526	
	Chinese	1.7% (8)	483	0.68 (0.30, 1.33)	0.307	
	Mixed / Other	2.1% (101)	4922	0.86 (0.64, 1.15)	0.308	
UK birth status	Non-UK Born	2.1% (646)	30481			0.763
	UK Born	2.3% (240)	10496	0.97 (0.79, 1.18)	0.763	
Socio-economic status	1	2.3% (364)	15884			0.517
	2	2.1% (263)	12422	0.92 (0.78, 1.08)	0.32	
	3	2.0% (131)	6435	0.89 (0.72, 1.09)	0.282	
	4	1.9% (70)	3712	0.83 (0.63, 1.07)	0.168	
	5	2.3% (58)	2524	1.04 (0.77, 1.37)	0.804	

Date of ending treatment For date of ending treatment there is evidence that missing data is associated with year of notification and weaker evidence of an association with ethnic group and socio-economic status, with little evidence for any other variable. As found previously, variable completeness increased over time. There was some evidence that poorest socio-economic group was more likely to be missing the date of ending treatment but the evidence for this was mixed. The White ethnic group was more somewhat likely to be missing date of treatment ending than most other ethnic groups.

Table 4.10: Results from a logistic regression model with data completeness (Complete/Missing) for date of starting treatment as an outcome, adjusted for; year, sex, age (grouped as 0-14 year olds, 15-65 year olds and 65+), ethnic group, UK birth status and socio-economic status (national quintiles). For socio-economic group 1 indicates the most deprived quintile. Notifications from 2010 onwards were included as socio-economic status was not collected before this. Complete case analysis weas used. Odds ratios shown are adjusted for all explanatory variables. There is little evidence that the missing data for the date of starting treatment is associated with any variable considered, except for year of notification.

Variable	Category	Missing (N)	Notifications (33606)	Odds Ratio	P value (Wald)	P value (LRT)
Year	2010	2.9% (182)	6171			2.52e-14
	2011	2.6% (177)	6855	0.88 (0.71, 1.08)	0.223	
	2012	2.4% (164)	6882	0.80 (0.64, 0.99)	0.0379	
	2013	1.5% (97)	6298	0.51 (0.39, 0.65)	8.39e-08	
	2014	1.2% (66)	5341	0.40 (0.30, 0.53)	2.2e-10	
	2015	1.4% (28)	2059	0.45 (0.30, 0.66)	0.00011	
Sex	Female	2.1% (311)	14630			0.859
	Male	2.1% (403)	18976	1.01 (0.87, 1.18)	0.86	
Age	0-14	2.7% (44)	1617			0.711
	15-44	2.0% (419)	21027	0.83 (0.60, 1.18)	0.282	
	45-64	2.3% (165)	7272	0.88 (0.62, 1.27)	0.479	
	65+	2.3% (86)	3690	0.83 (0.56, 1.23)	0.338	
Ethnic group	White	2.9% (176)	6076			0.00931
	Black-Caribbean	2.8% (21)	753	1.01 (0.62, 1.57)	0.972	
	Black-African	1.9% (114)	6071	0.69 (0.52, 0.93)	0.0162	
	Black-Other	2.3% (7)	306	0.88 (0.37, 1.78)	0.751	
	Indian	1.7% (150)	8842	0.66 (0.51, 0.87)	0.00317	
	Pakistani	2.5% (140)	5668	0.94 (0.72, 1.22)	0.63	
	Bangladeshi	1.3% (18)	1409	0.48 (0.28, 0.78)	0.00533	
	Chinese	2.8% (11)	396	1.09 (0.54, 1.99)	0.787	
	Mixed / Other	1.9% (77)	4085	0.75 (0.54, 1.02)	0.0724	
UK birth status	Non-UK Born	1.9% (480)	25174			0.153
	UK Born	2.8% (234)	8432	1.17 (0.94, 1.45)	0.151	
Socio-economic status	1	2.4% (308)	13080			0.000621
	2	1.7% (170)	10266	0.72 (0.60, 0.87)	0.000888	
	3	1.9% (100)	5265	0.82 (0.65, 1.03)	0.0917	
	4	2.8% (84)	2994	1.19 (0.92, 1.52)	0.178	
	5	2.6% (52)	2001	1.07 (0.78, 1.44)	0.681	

Biases in the ETS

Routine observational data-sets are subject to numerous potential biases, such as selection bias, recall bias, measurement bias, and unmeasured confounding.[46] Additionally as the data has not been collected with a specific analysis in mind there maybe issues with the specificity of variables. The ETS system is likely to suffer from all of the above biases to some extent, which must be accounted for as far as possbile, and explicitly stated at every level of analysis. The most important consideration is that the ETS system is unlikely to be representative of the general population as it contains only notified TB cases that occurred in England during the study period, research questions must therefore be either limited to active tuberculosis patients, or when extended to the general population the differing

population demographics must be accounted for. If this is not done then any results may be due to selection bias. Additionally, multiple variables may suffer from misclassification bias, including BCG status which can be assessed via vaccination record, the presence of a scar, or case recall: this may lead to spurious associations.[47] Validation studies would be required to account for this, which is beyond the scope of this thesis.

Date variables in the ETS

For analyses that aim to reproduce temporal trends in TB incidence, such as dynamic modelling studies, it is important to understand which variables represent the most accurate date of contact with the health system and more generally on what scale date variables can be considered reliable. In the ETS extract used in this thesis there are several date variables that encode useful information including; the date of notification, the date of symptom onset, the date of diagnosis, the date of starting treatment, the date of completing treatment, and the date of death. In the following section I explore these variables using counts and proportions aggregated to then nearest year, month and day.

As seen in the previous section (Section 4.2.1), many of these variables have a large proportion of missing data, with date of notification and date of starting treatment having the least amount of missing data. The date of notification represents the simplest variable to use to represent when a case can be defined to have occurred as it is complete for all records. Unfortunately, cases may be notified at any stage of active TB, from initially becoming symptomatic to post-mortem diagnosis and notification. Despite this limitation, date of notification can be used as a baseline on which to judge other date variables and some of these limitations may be mitigated by aggregating data by month or by year. Figure 4.2 a.) shows the number of TB notifications by year and Figure 4.2 b.) shows the number of TB notifications by month. These figures indicate that aggregating by year, rather than month, reduce the level of noise in the estimates and makes the trend over time easier to identify. This is an acceptable approximation if inference is being drawn on the scale of years. For shorter term processes, such as the duration of treatment which is generally considered to take approximately 6 months (Chapter 2), aggregating by year would reduce the accuracy of the estimated parameter. There is some evidence of a seasonal trend in notifications (Figure 4.2 c.)), with a higher proportion of cases notified in the May, June and July than in the rest of the year. This seasonality would have to be accounted for if conducting analysis on a monthly scale and date of notification was being used as the date of first contact with the health system. There is little evidence that date of notification varies by day of the month (Figure 4.2 d.)).

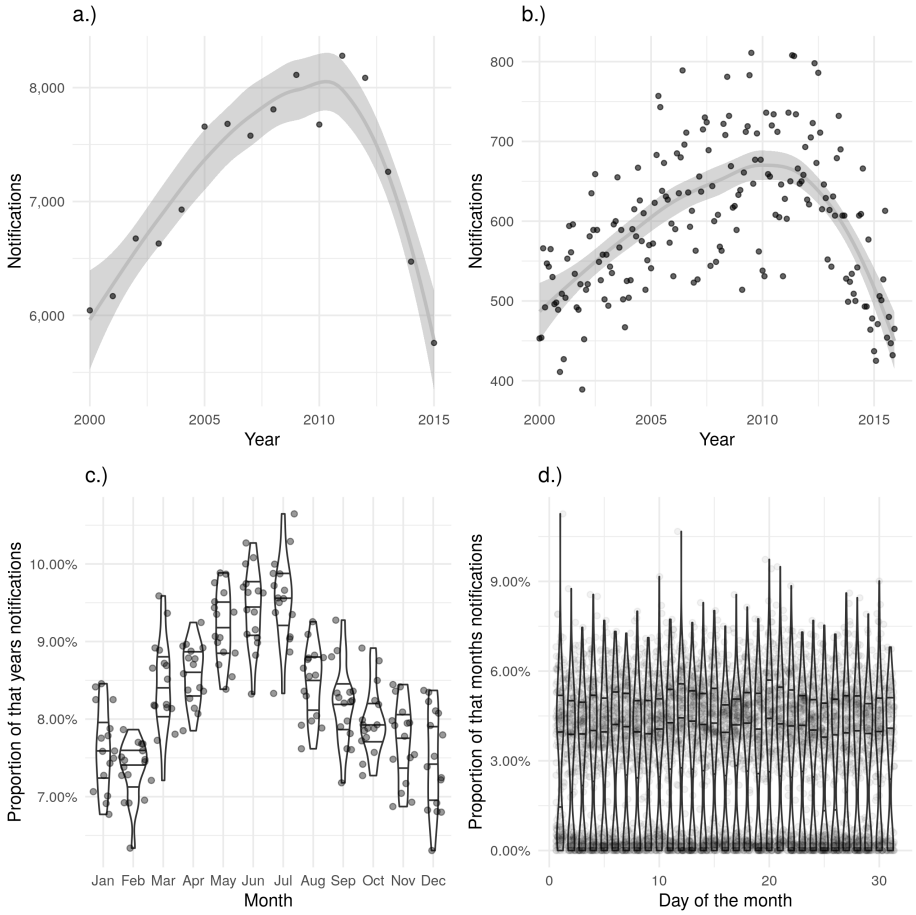


Figure 4.2: a.) and .b) show notifications over time by date of notification in the ETS, with a.) aggregated by year and b.) aggregated by month. A trendline has been produced using a locally weighted regression model. Both of these plots show the same overall trend, but b.) contains a large amount of apparent noise. c.) Shows the proportion of cases notified in a given month for each year, with some evidence of a seasonal trend. d.) Shows the proportion of cases notified on a given day for each month, there is little evidence of between day variation in cases notified.

An alternative measure is to use the date of symptom onset, this represents the closest approximation to the date when a case became infectious. Unfortunately there are multiple issues with this measure, the first of which being is that it is only 68.0% (78054/114820) complete across the data extract. Additionally, completeness changes with time, with 65.7% (3969/6044) in 2000, 60.4% (4720/7809) in 2008, and 87.7% (5677/6472) in 2014 this could lead to spurious trends in the number of cases. Perhaps most importantly the date of symptom onset is highly susceptible to recall bias with the majority of cases becoming symptomatic on the first of each month (Figure 4.3 d.)), with some evidence that a greater number of cases occur in January than would be expected (Figure 4.3 c.)). Another possible measure of the number of cases is the date of diagnosis, this should be a more reliable variable than the date of symptom onset, as it does not rely on the recall of the case. However it is only 51.3% (58960/114820) complete across the dataset, with strong evidence of increasing

completeness going from 11.6% (699/6044) in 2000, 52.2% (4078/7809) in 2008, and 89.4% (5786/6472) in 2014. This trend would be hard to properly account for in any analysis and therefore this variable should not be used as a primary measure.

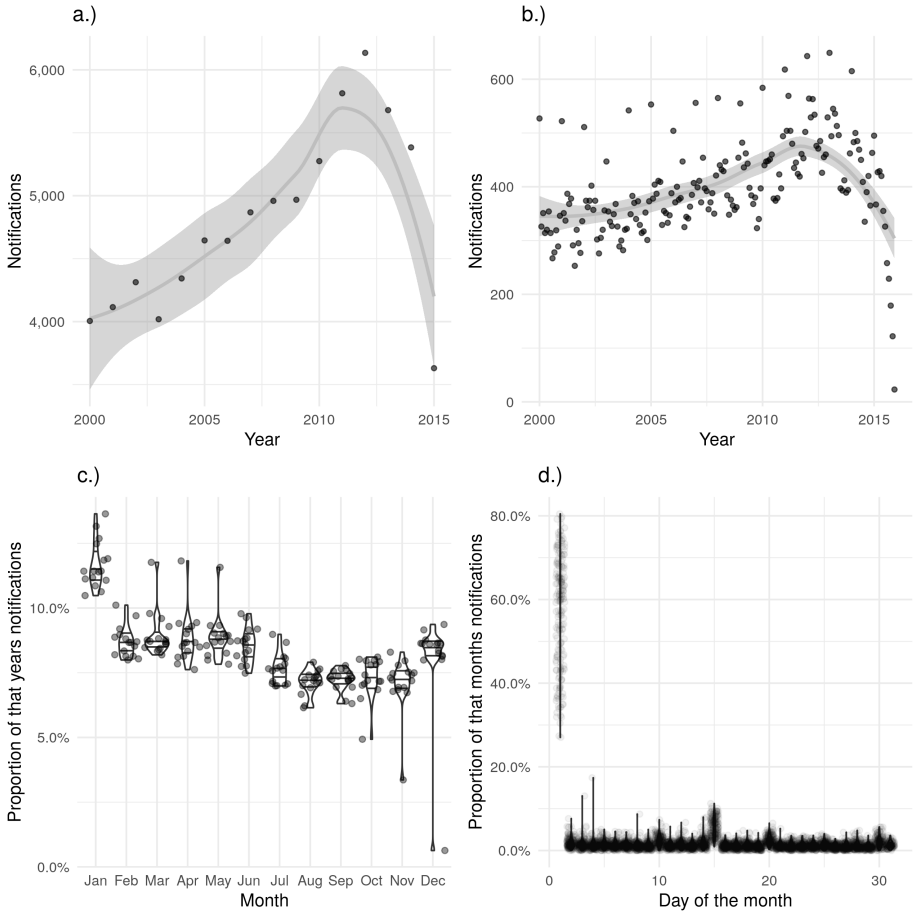


Figure 4.3: a.) and .b) show notifications over time by date of symptom onset in the ETS, with a.) aggregated by year and b.) aggregated by month. A trendline has been produced using a locally weighted regression model. Both of these plots show the same overall trend, but b.) contains a large amount of apparent noise. c.) Shows the proportion of cases notified in a given month for each year, with some evidence of a seasonal trend and a higher proportion of cases reporting symptoms starting in January than would be expected. d.) Shows the proportion of cases notified on a given day for each month, with a much higher proportion of cases reproting symptoms on the first of the month than would be expected. On both the scale of months and years there is some evidence of recall bias, with the first month, or first day, reporting higher proportions of cases than would be expected.

The date of starting treatment should be a more reliable contact date as it records an official contact with the health system. Indeed it was 75.7% (4464/5899) complete in 2000 which increased year on year to 98.8% (5612/5680) complete in 2015. This increasing completeness may lead to a temporal bias if not properly adjusted for when evaluating the date of starting

treatment over time. As for the data of notification there is some evidence of a seasonal trend for date of starting treatment (Figure 4.4 c.)), with a peak of cases starting treatment in May, June and July. However, this seasonal trend is difficult to identify when cases starting treatment are visualised by month over time (Figure 4.4 b.)). Unlike the date of symptom onset there is little evidence of recall bias by month, or by day (Figure 4.4 c.) and d.)).

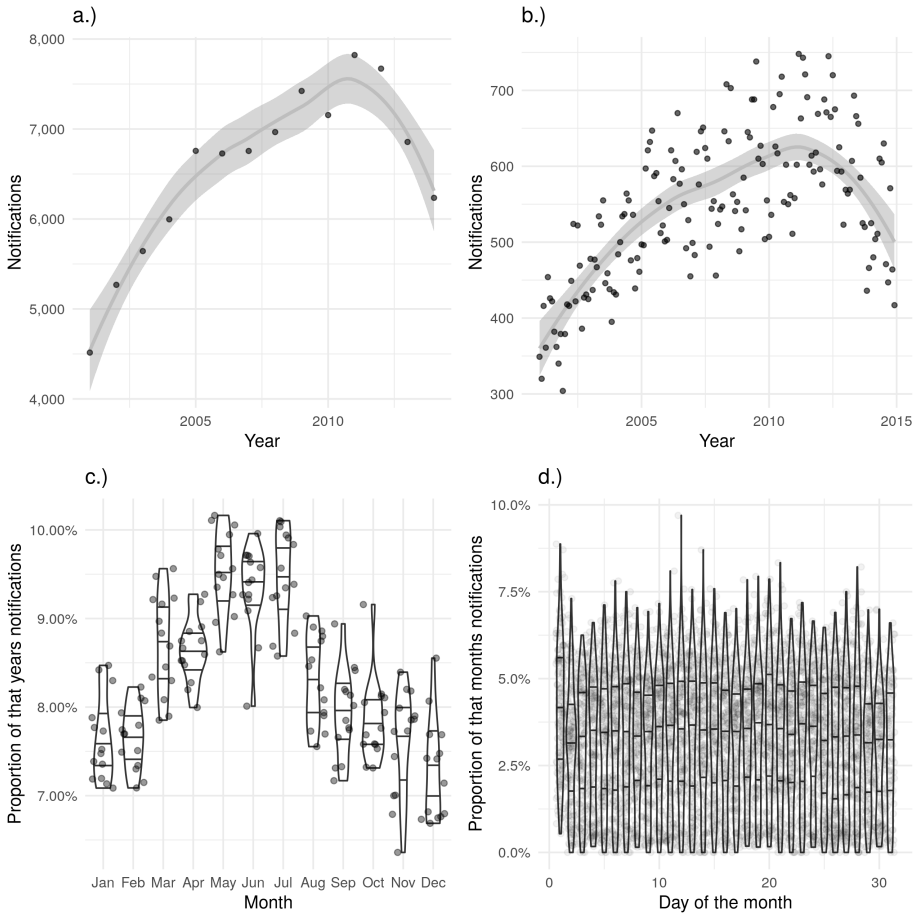


Figure 4.4: a.) and .b) show notifications over time by date of starting treatment in the ETS, with a.) aggregated by year and b.) aggregated by month. A trendline has been produced using a locally weighted regression model. Both of these plots show the same overall trend, but b.) contains a large amount of apparent noise. c.) Shows the proportion of cases starting treatment in a given month for each year, with some evidence of a seasonal trend. d.) Shows the proportion of cases starting treatment on a given day for each month, with little evidence of between day variation. Data is only shown from 2001 until 2015 and prior to 2001 this variable was not recorded and it is not complete for 2015.

The date of ending treatment does not appear to display similar seasonality (Figure 4.5 c.)). This maybe because treatment time varies between individuals and this dilutes the seasonality observed for the date of starting treatment. As noted previously there was some

evidence of recall bias when the proportion of those ending treatment was examined on a day of the month basis, with a larger proportion ending treatment on the first of the month than on any other day (Figure 4.5 d.)). The date of ending treatment was not recorded in 2000, or 2001, and was highly missing for the first several years after collection began (45.4% (2593/5712) complete in 2002 and 58.7% (3475/5921) complete in 2003). From 2009 it was over 90% complete, reaching 97.7% (5359/5486) complete in 2013. As for the other data variables discussed this increasing completeness over time may lead to a bias if not accounted for in future analyses.

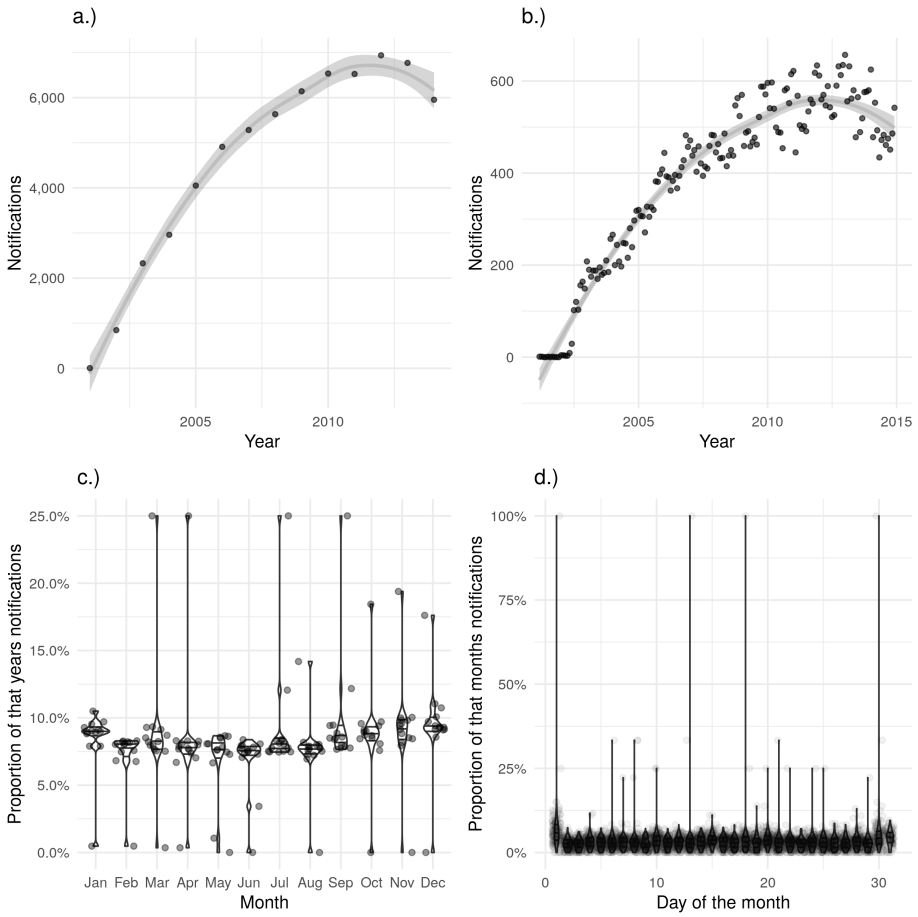


Figure 4.5: a.) and .b) show notifications over time by date of treatment ending in the ETS, with a.) aggregated by year and b.) aggregated by month. A trendline has been produced using a locally weighted regression model. Both of these plots show the same overall trend, but b.) contains a large amount of apparent noise. c.) Shows the proportion of cases finishing treatment in a given month for each year, with little evidence of a seasonal trend. d.) Shows the proportion of cases finishing treatment on a given day for each month, with a much higher proportion of cases finishing treatment on the first of the month than would be expected. On the scale of months there is some evidence of recall bias, with the first day reporting higher proportions of cases than would be expected. Data is only shown from 2001 until 2015 and prior to 2001 this variable was not recorded and it is not complete for 2015.

Finally date of death displays little evidence of seasonal variation or recall bias (Figure 4.6 c.) and d.)) but has a strong temporal trend for data completeness, with a year on year increase. Data was not collected in 2000 and was only 11.8% (199/1689) complete in 2001, data completeness remained below 20% until 2005 when it increased to 38.3% (353/921). This can be seen as a discontinuity when deaths are aggregated by year and plotted (Figure 4.6 a.)). Missing data also masks a drop in notified cases that died, which fell from 1451 in 2005 to 921 in 2006. In comparison, only 352 cases in 2005 and 353 cases in 2006 had a date

of death. Data completeness has remained below 80% with increases in data completeness decreasing year on year.

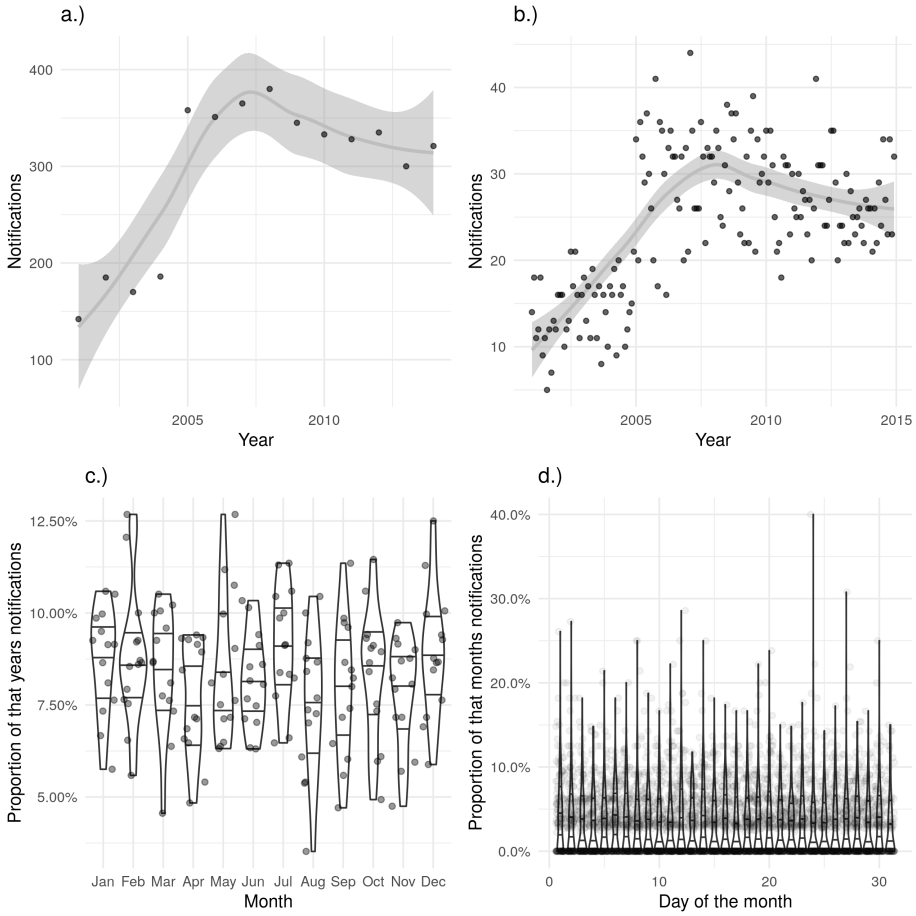


Figure 4.6: a.) and .b) show notifications over time by date of death in the ETS, with a.) aggregated by year and b.) aggregated by month. A trendline has been produced using a locally weighted regression model. Both of these plots show the same overall trend, but b.) contains a large amount of apparent noise. c.) Shows the proportion of cases who died in a given month for each year, with no evidence of a seasonal trend. d.) Shows the proportion of cases who died on a given day for each month, with little evidence of between day variation. Data is only shown from 2001 until 2015 and prior to 2001 this variable was not recorded and it is not complete for 2015.

4.2.2 Demographic data

Background

Demographic data used in this thesis is drawn from two main sources; mid-year resident populations, by single year of age, downloaded from the Office for National Statistics (ONS) website for 2000 to 2015 and population estimates from the yearly April to June Labour Force Survey (LFS) stratified by single year of age and UK birth status. The LFS is a study

of the employment circumstances of the UK population and provides the official measures of employment and unemployment in the UK. It also records other details such as ethnicity and UK birth status which may be used, along with population weightings, to estimate the UK and non-UK born population.

Data management

The mid-year population estimates were transformed from wide format into tidy data,[48] with the population estimates from 2000 being reformatted to match those from 2001 onwards. Data from the Labour force survey was available by year, so each dataset was separately imported into R.[49] Reporting practices have changed with time so the appropriate variables for age, country of origin, country of birth, and survey weight were extracted from each yearly extract, standardised, and combined into a single tidy dataset. The LFS data was then aggregated, accounting for survey weight, by year, age, and UK birth status to provide yearly estimates of the UK born/Non-UK born demographics by age. Finally 5 year age groups were defined using the single year of age.

Data structure, completeness, and biases.

Both the mid-year ONS population estimates,[50] and the LFS are assessed for performance and quality elsewhere.[51,52] However both have several failing that it is important to note, as they could introduce bias in future analysis. Whilst the ONS mid-year, and LFS estimates compare well when aggregated by age (Figure 4.7) there is more disagreement when they are broken down by 5 year age groups (Figure 4.8). For those at working age both data sources are comparable (with approximately a 1% difference across all years), however for children, young adults, and those who are 85+ the LFS underestimates the total population. This is particularly the case for older adults with between a 5% and 20% discrepancy for those aged 85-89 and a 25% to 45% discrepancy between those aged 90+. This could be problematic as these age groups often have the most severe outcomes to tuberculosis infection. A pragmatic approach to this is to exclude those aged 90+ from future analysis as results for this age group will be subject to large amounts of uncertainty which will be difficult to directly incorporate into the results.

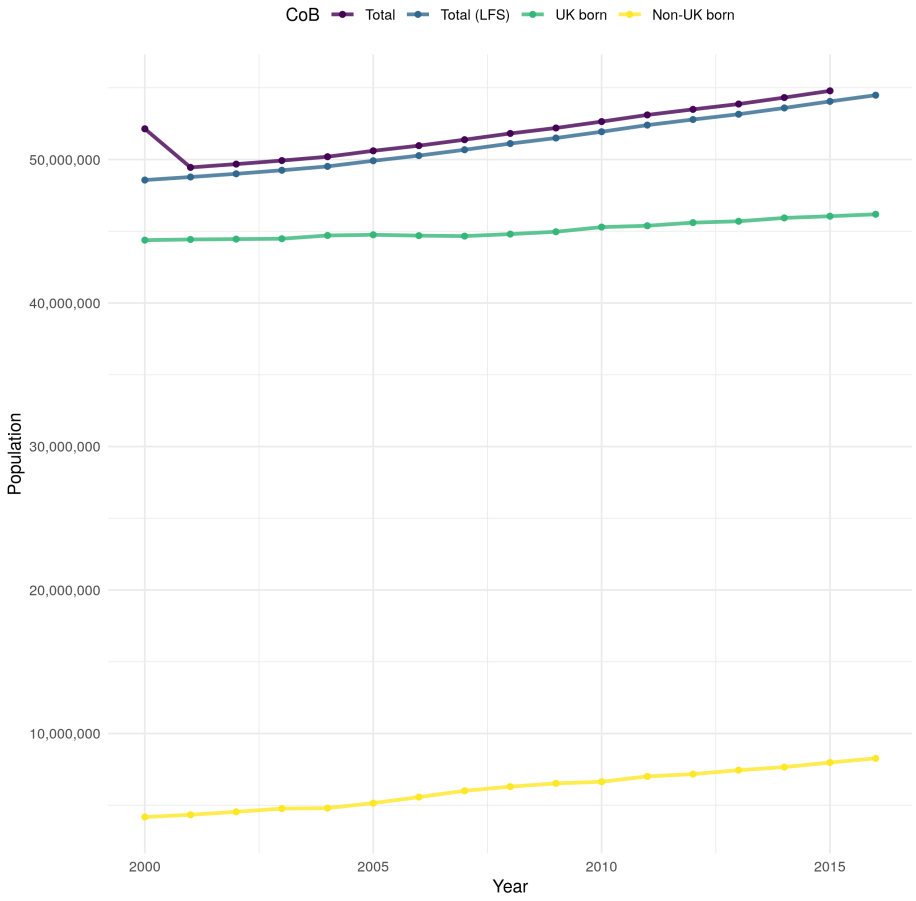


Figure 4.7: Overall population estimates derived using Office for National Statistics (ONS) and Labour Force Survey (LFS) demographic data. The ONS data is likely to be more reliable as the LFS data is derived using a weighted survey. After accounting for missing UK birth status both datasets provide comparable estimates of the population of England, with a clearly increasing trend over time. However the ONS data indicates a reduction in population from 2000 until 2001 that is not seen in the LFS data. The UK born population has also risen slowly from 2000 until 2015, although the biggest increase has been in the non-UK born population.

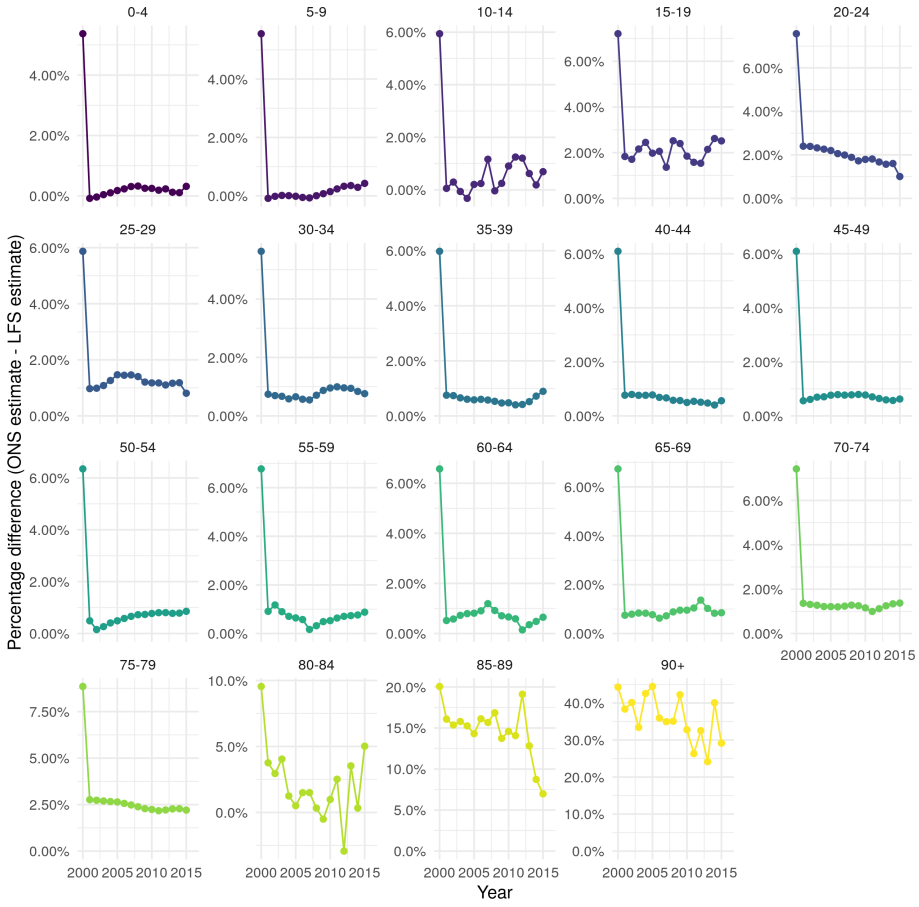


Figure 4.8: Percentage difference between Office for National Statistics (ONS) population estimates and estimates derived from the Labour Force Survey (LFS) by 5 year age group. For most age groups there is less than a 2% difference over time. In older adults (85+) there is a substantially greater difference ranging from 5% to 40%.

4.3 Tuberculosis notifications

4.3.1 Overview

There were 114,820 notifications between 2000-2015 in England of which 67.6% (77669/114820) were non-UK born. Over this period notifications increased in the non-UK born from 2000 until 2011, since when they have decreased year on year (Figure 4.9). In the UK born notifications remained relatively stable from 2000 until 2011, since when there has been some small decrease. Notifications with missing UK birth status have decreased year on year, with only 121 in 2015. The majority of cases were aged between 15-44 years old (60.2% (69106/114820)), with few cases in young children (0-14; 5.1% (5842/114820)) or older adults (65+; 14.4% (16538/114820)). Cases are heterogeneously distributed with the majority of cases in London (42.8% (49142/114820)), with the next highest number of notifications in the West Midlands (12.3% (14100/114820)). Since 2009, 47.2% (24354/51645) of notifications have been BCG vaccinated, with 33.2%

(17133/51645) having a missing BCG status. Of cases with a known BCG status66.8% (34512/51645) were recorded as having been BCG vaccinated. From 2010, when collection of socio-economic status began, 38.6% (16800/43533) of cases have been in the lowest socio-economic quintile. For a more complete breakdown of notifications in the ETS see the yearly PHE TB report.[1] It should be noted that these statistics do not take into account changes in population demographics which may mask underlying changes in TB epidemiology, this is addressed in Section 4.5.

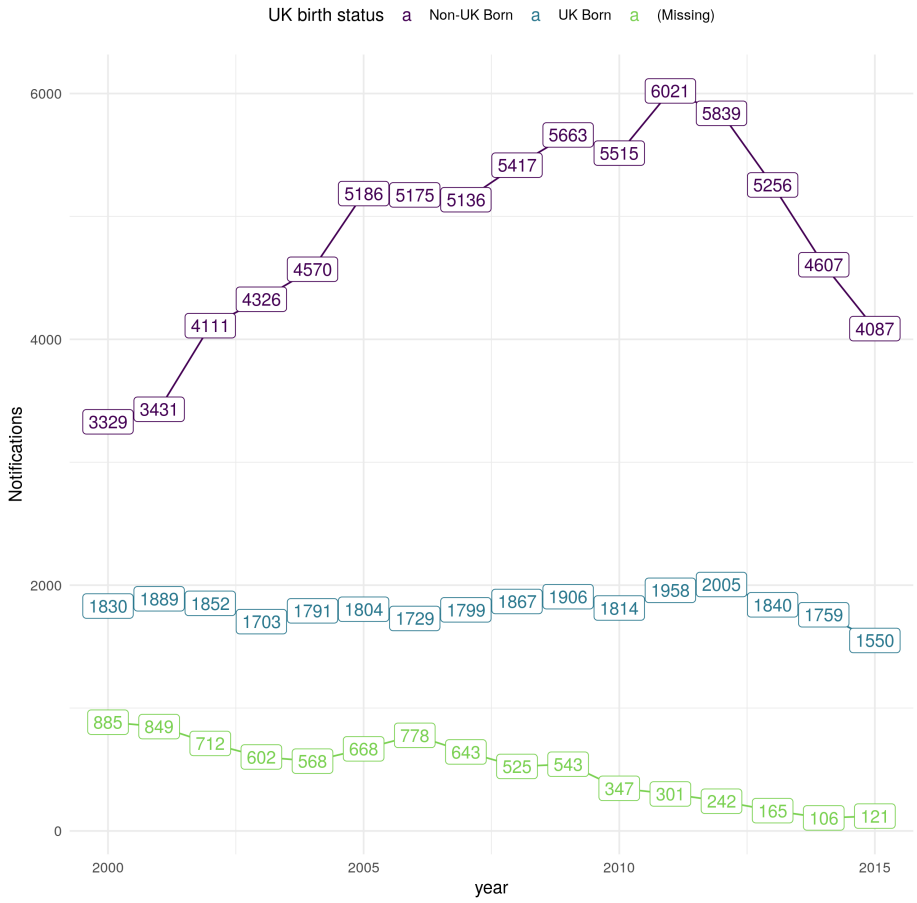


Figure 4.9: Notifications in England from 2000 to 2015 stratified by UK birth status, sourced from the Enhanced Tuberculosis Surveillance (ETS) system. Notifications in the non-UK born doubled from 3329 in 2000 to 6021 in 2011 since when they have decreased year on year. In the UK born notifications have remained comparable over time, with some evidence of a decrease from 2011 until 2015. UK birth status has become increasingly complete over time with notifications without birth status dropping from 885 in 2000 to 121 in 2015.

4.3.2 Age distribution of notifications

Notifications in the ETS are heterogeneous distributed by age as well as by UK birth status.[1] In the non-UK born the majority of cases occur in young adults with few cases in young children or older adults (Figure 4.10). Over time the distribution of cases is

becoming more uniform with a reduction in the proportion of cases in young adults. In the UK born the distribution of cases is more homogeneous, although there is some evidence of a higher proportion of cases in working age adults as opposed to older adults and children. Unlike the non-UK born population there is little evidence of a change in the distribution of cases over time. 0-4 year old UK born children make up a higher proportion of cases than other UK born children. This spike is not observed in the non-UK born population. These conclusions may be biased by changes in underlying population demographics, this is addressed in Section 4.5.

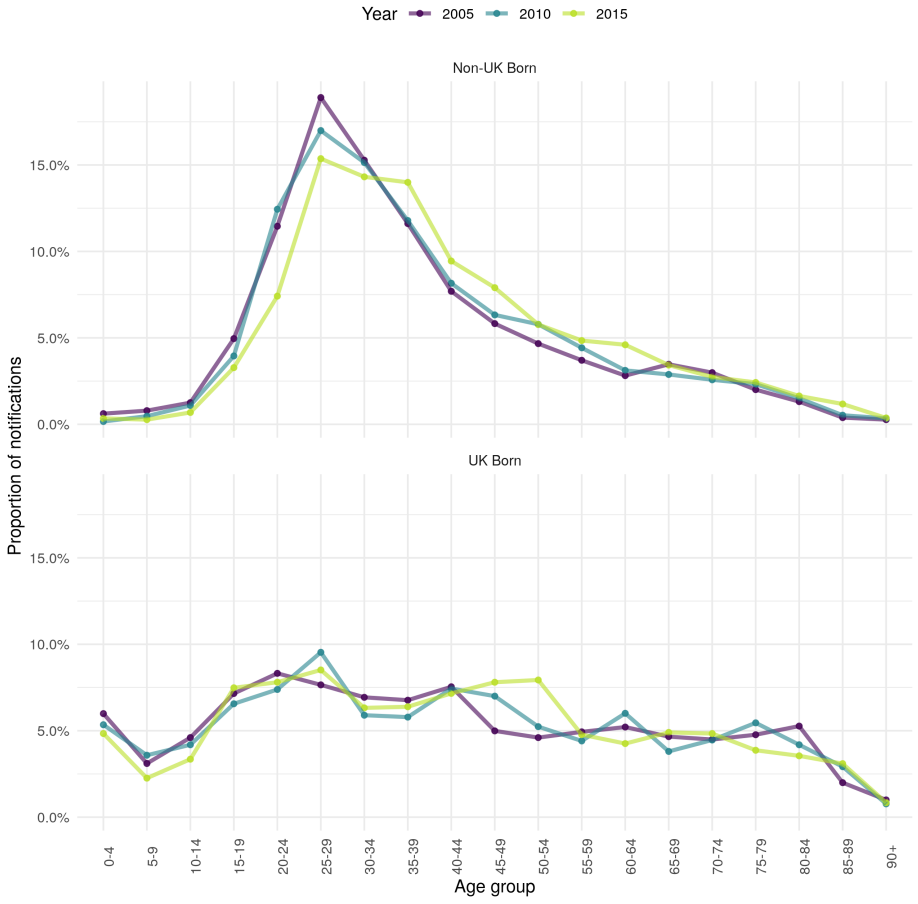


Figure 4.10: Proportion of cases by 5 year age group in the Enhanced Tuberculosis surveillance system in 2005, 2010 and 2015 stratified by UK birth status. Non-UK born cases have a higher proportion of young adult cases with very few cases in children or in older adults. UK born cases have a more uniform distribution of cases with some evidence of a higher proportion of cases in young adults. In the non-UK born the proportion of cases in young adults has decreased over time, with no evidence of a temporal trend in the UK born. These results are not adjusted for population demographics and therefore may be biased.

4.4 Population Demographics in England

Underlying trends in population demographics can be important factors in driving changes in infectious disease dynamic, so it is important to understand these trends before conducting further analysis. England has an increasing population (Figure 4.7), driven by small increases in the UK born population, and larger increases in the non-UK born population. The increase in the non-UK born population is mainly in young adults, with a reduction in the proportion of the non-UK born population that are older (Figure 4.11). In the UK born the proportion of the population that is in late middle age has increased, with the proportion of younger adults decreasing. The proportion of those aged 75+ has remained constant over time in both the UK born and non-UK born populations.

The changes in population demographics, for both the UK and non-UK born, from 2000 to 2015 may have directly impacted the number and age distribution of TB notifications. In the previous section, it appeared that a higher proportion of cases were in young adults in the non-UK born than in other age groups. Figure 4.7 indicates that this maybe due to a higher proportion of the non-UK born population being young adults. Additionally, Figure 4.7 indicates that proportion of the non-UK born population that were young adults has decreased over time, this mirrors the trend in the age distribution of notifications observed in Figure 4.10 and is likely to be driving part of this trend. In the UK born the population has become older in general, this is not clearly reflected in the age distribution of notifications (Figure 4.10). This may indicate changes in the risk of developing TB.

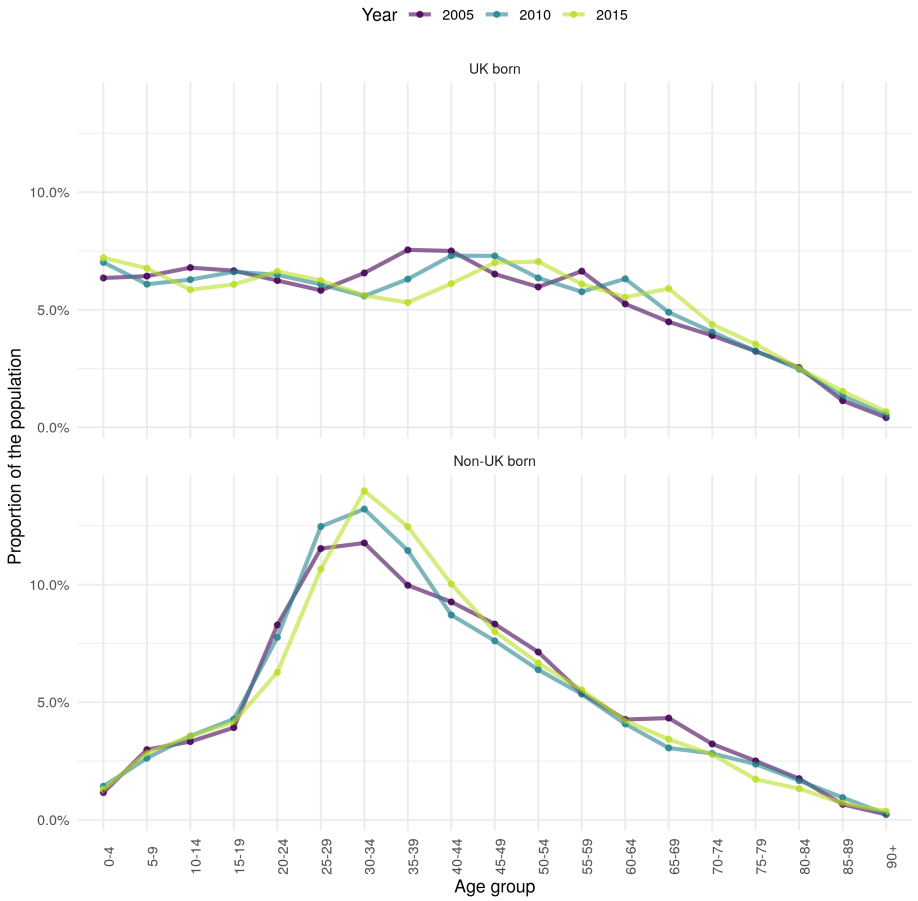


Figure 4.11: The estimate proprotion of the population in each 5 year age group stratified by UK birth status for 2000, 2008, and 2016. The UK born population has become older with a larger segment of the population in late middle age and older in 2016 compared to 2000. In the non-UK born whilst the population as a whole has increased, the majority of the increase is in young adult.

4.5 Tuberculosis incidence rates

4.5.1 Motivation

As discussed in Section 4.3.2 and Section 4.4 changes in underlying population demographics may mask or bias trends in TB notifications. To account for this, incidence rates, which indicate the incidence of TB for a standard population size, may be used. Whilst TB incidence rates are available in the year PHE TB report,[1] they are limited in detail and do not report age, and UK born stratified incidence rates across years. Estimating these incidence rates will allow for novel analyses to be conducted later in this thesis that explore population adjusted trends in TB. The method used to estimate incidence rates is first outlined, then overall trends in incidence rates, stratified by UK birth status are explored. Finally trends in incidence rates, stratified by age and UK birth status, are investigated.

4.5.2 Method

Age-specific incidence rates were calculated as follows:

$$\textit{Incidence rate (over time period, t and age, a)} = \frac{\textit{Number of cases (t,a)}}{\textit{Population}(t,a)} \tag{4.1}$$

Age-standardised rates were calculated using the epiR package for R,[53] using the average age distribution of England, and Wales from 2000-2015 as the standard population to allow comparison between years. Those aged 90+ were excluded as demographic data for this population was unreliable. The code used to calculate incidence rates is available online as an R package (`https://www.samabbott.co.uk/tbinenglanddataclean/reference/clean_munge_ets_2016.html`).

4.5.3 Overall trends in TB incidence rates

Incidence has varied with time, increasing from 11.6 per 100,000 people (95% CI 11.3, 11.9) in 2000 to a maximum of 15.6 per 100,000 people (95% CI 15.3, 15.9) in 2011, since when it was decreased to a low of 10.5 per 100,000 people (95% CI 10.2, 10.8) in 2015 (Figure 4.12). This may indicate that TB control efforts are proving effective in preventing TB outbreaks, or may be driven by changes in the composition of those immigrating to England. It also highlights the lack of progress in reducing TB burden in England over the previous two decades, with little evidence of a decrease in overall incidence rates from 2000 until 2015. In the non-UK born incidence rates increased dramatically from 2000 to 2005, since when they have fallen consistently. In comparison, incidence fell in the UK born from 2000 until 2005 and then increased until 2012, since when they too have decreased year on year. This may indicate that incidence rates in the two populations are linked, with incidence rates in the non-UK born driving incidence rates in the UK born with some time lag. Alternatively it may be that incidence rates in the two populations are only weakly linked, or not linked at all. In this scenario the TB endemic in England would actually be two nearly separate endemics, each with different drivers. These scenarios can be differentiated using trends in age-specific incidence rates, and with statistical (Chapter 7)) and dynamic modelling (Chapter 8).

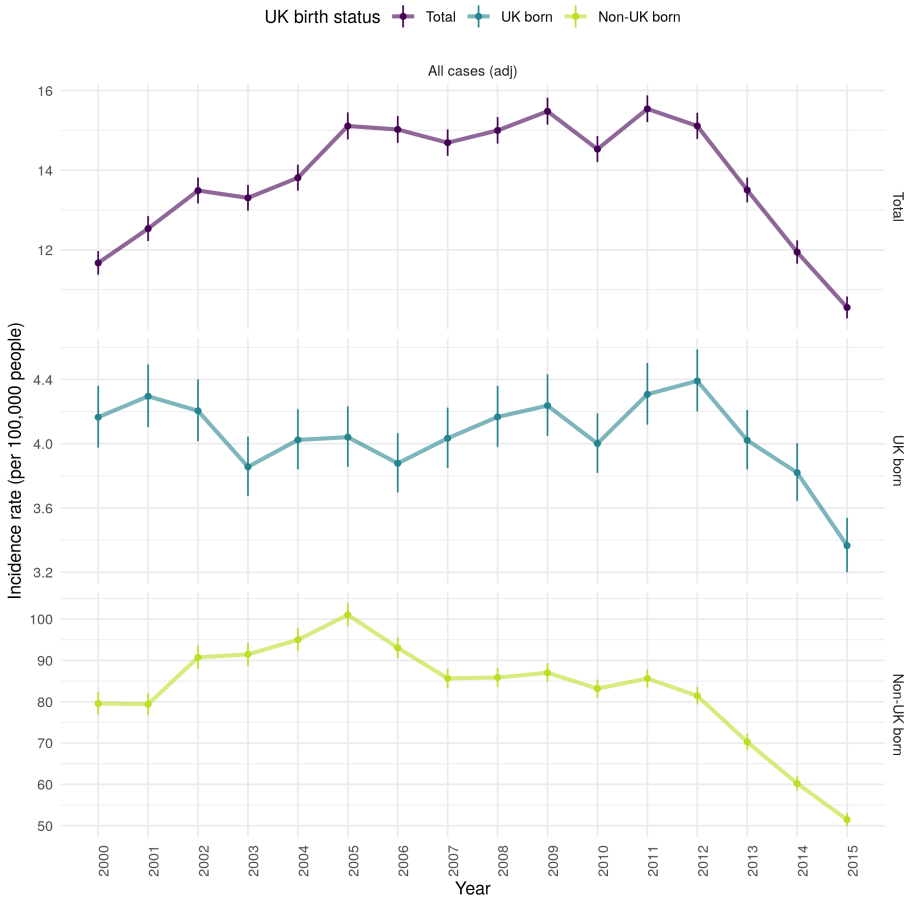


Figure 4.12: Age standardised incidence rates (by 100,000 population) for all notified TB cases from 2000-2015. Overall incidence rates are shown, along with incidence rates in the UK and non-UK born populations. Point estimates are given along with 95% confidence intervals for each incidence rate estimate. Trends over time are highlighted by linking points with a line. Incidence rates increased over time from 2000 until 2011, since when they have falled year on year. This appears to be driven by increasing incidence rates in the non-UK born from 2000 until 2005, since when they have fallen year on year. This trend was not observed in the UK born, in which incidence rates fell from 2000 until 2005 and then increased from 2005 until 2012. As in the non-UK born they have since fallen year on year.

4.5.4 Age stratified incidence rates

Stratifying incidence rates into age groups (children (0-14), adults (15-64) and older adults (65+)) it is clear that the trends observed in the age adjusted overall incidence rates are not seen in all age groups (Figure 4.13). In the 65+ age group there was evidence of a year on year decrease in incidence rates from 14.3 per 100,000 people (95% CI 13.5, 15.1) in 2002, to 8.7 per 100,000 people (95% CI 8.1, 9.3) in 2015. In comparison, in the 15-64 year old age group, which represents the majority of cases, incidence rates rose year on year to a maximum of 19.5 per 100,000 people (95% CI 19.0, 20.0) in 2011 and then fell year on year

to 13.3 per 100,000 people (95% CI 12.9, 13.7) in 2015. In children (0-14) incidence rates peaked earlier, with an incidence rate of 3.5 per 100,000 people (95% CI 3.1, 3.9) in 2000 which increased to 5.0 per 100,000 people (95% CI 4.5, 5.5) in 2007. Since when they have decreased to in 2.2 per 100,000 people (95% CI 2.0, 2.6) in 2015.

Further stratifying incidence rates, by both age group and UK birth status, it is clear that the contribution of the non-UK born dominates that of the UK born in adults (15-64) but that the reverse is true in older adults (65+) and trends appear to be similar in children (0-14), regardless of UK birth status (Figure 4.13). In the non-UK born, incidence rates have fallen year on year in children but increased from 2000 until 2005 in adults, since when they have decreased. In non-UK born older adults there is less clear evidence of a trend over time, although incidence rates have fallen, as in other populations, from 2011 on-wards. In the UK born, incidence rates increased in children from 2000 until 2008, since when they too have consistently fallen. UK born adults had increasing incidence rates year on year until 2012 but incidence rates have since fallen to pre 2000 levels. In older UK born adults incidence rates have consistently fallen, more rapidly from 2000 until 2008 and since 2014.

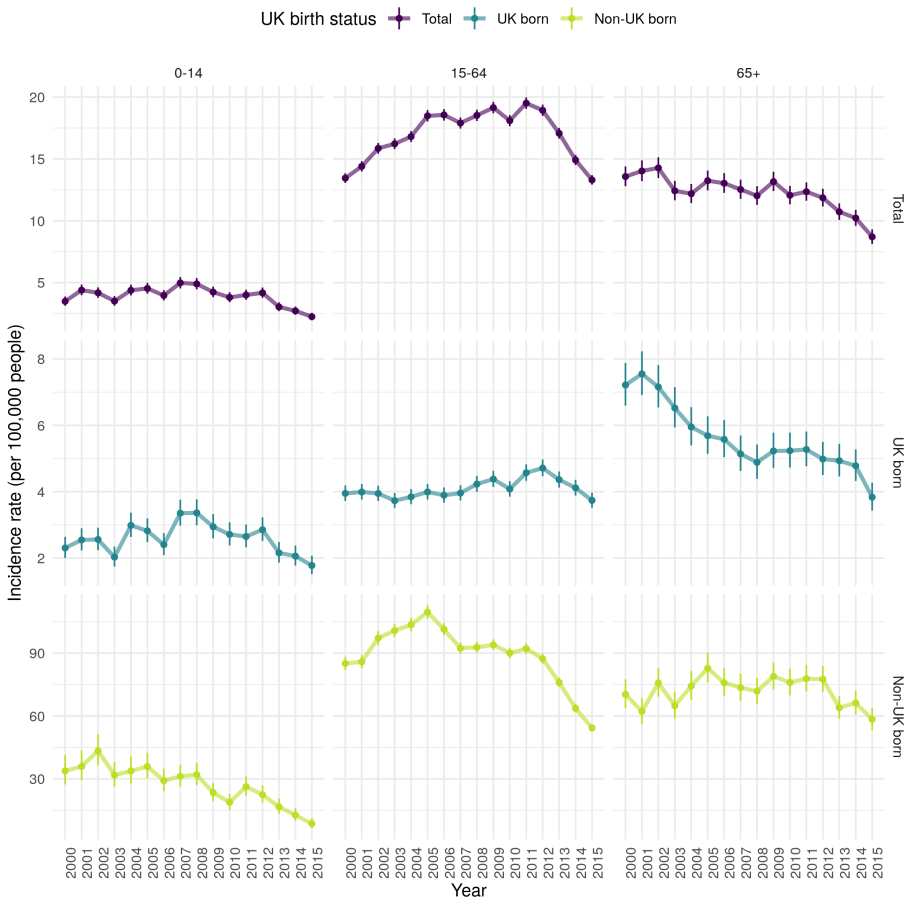


Figure 4.13: Incidence rates (by 100,000 population) for all notified TB cases from 2000-2015, stratified by age group (children (0-14), adults (15-64) and older adults (65+)) and UK birth status. Point estimates are given along with 95% confidence intervals for each incidence rate estimate. Trends over time are highlighed by linking points with a line. Incidence rates declined overall in children over time. In adults incidnce rates incrseased until 2011 and have since fallen. In older adults incidence rates consistently fell. In the non-UK born, incidence rate also fell in childen but peaked earler in adults and showed little evidence of a downwards trends in older adults until 2013. In the UK born, incidence rates increased in children until 2008, since when they havell fallen. Incidence rates also increased over time in UK born adults until 2012 but has consistently fallen in UK born older adults.

Another approach to explore trends in age stratified incidence rates is to visualise them across 5 year age groups, for a selected subset of years. This can be seen in Figure 4.14 stratified by UK birth status. This figure indicates that TB incidence in the non-UK born has been driven by high incidence rates in young adults. Incidence rates in this population increased dramatically between 2000 and 2005 and then fell in all age groups, except 20-24 years old by 2010. In 2015 there was little evidence of this peak in young adults but a secondary spike in much older adults (75+) remained. In the UK born, incidence rates increased with age in 2000, this trend has weakened over time, with a secondary peak

developing in young adults (with a 5 year lag when compared to the peak observed in non-UK born adults). In 2015, incidence rates in the UK born were largely homogeneous except for a gradual increase in much older adults (75+), and lower incidence rates in children. 0-4 year old children have remained at greater risk of TB, compared to other children across the time period for which data is available. There is some evidence that incidence rates fell in this group after the introduction of BCG vaccination in 2005, with incidence rates in older children (5-9) also having fallen by 2015.

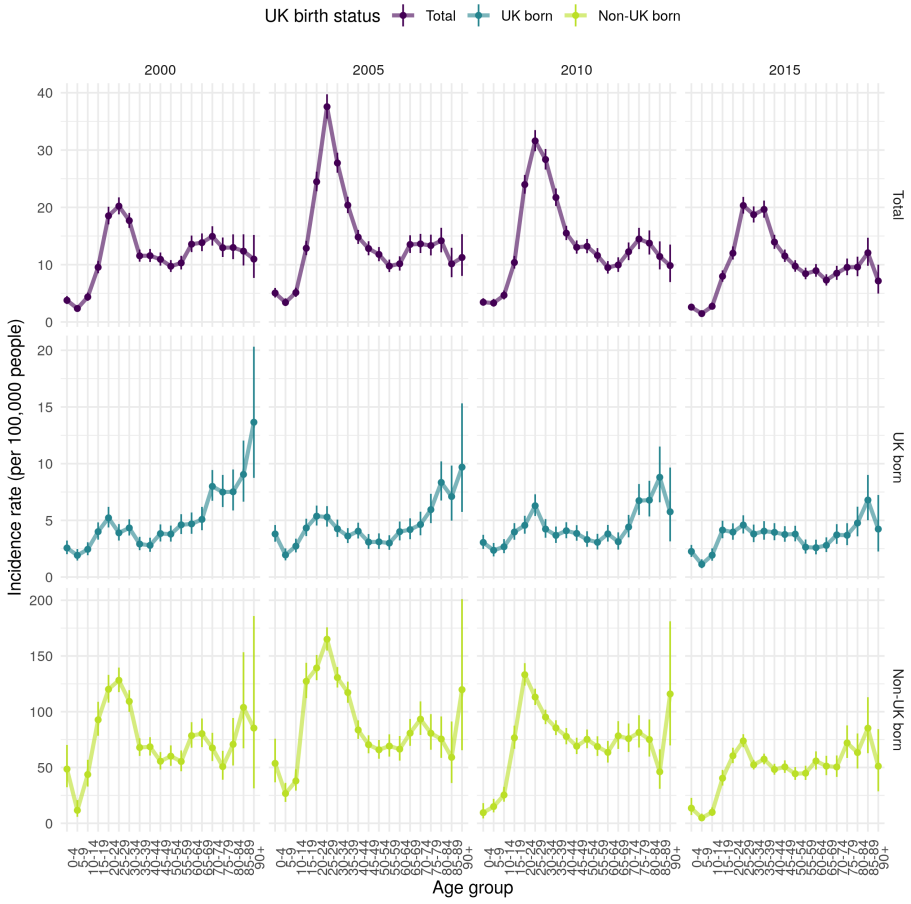


Figure 4.14: Age-specific incidence rates (by 100,000 population) grouped into 5 year age categories for 2000, 2005, 2010 and 2010, stratified by UK birth status. Point estimates are given along with 95% confidence intervals for each incidence rate estimate. Trends across age distributions are highlighed by linking points with a line. This Figure indicates that TB incidence in the non-UK born has been driven by high incidence rates in young adults. Incidence rates in this population increased dramatically between 2000 and 2005 and then fell in all age groups, except 20-24 years old by 2010. In 2015 there was little evidence of this peak in young adults but a secondary spike in much older adults (75+) remained. In the UK born, incidence rates increased with age in 2000, this trend has weakened over time, with a secondary peak developing in young adults (with a 5 year lag when compared to the peak observed in non-UK born adults). In 2015, incidence rates in the UK born were largely homogeneous except for a gradual increase in much older adults (75+), and lower incidence rates in children. 0-4 year old children have remained at greater risk of TB, compared to other children across the time period for which data is avialable.

4.5.5 Incidence rates in children as a proxy for TB transmission

Trends in incidence rates in UK born children are used as a proxy for recent transmission and compared to the overall incidence rate in order to extrapolate the degree of reactivation occurring in older populations.[1] Whilst this proxy approach is limited, in that it assumes that different population groups have an equivalent risk of TB and that TB control measures are the same across age groups, it may be combined with other methods to derive a good understanding of TB transmission. In Figure 4.12 incidence rates in the UK born decreased from 2000 until 2006 and then increased until 2011, since when they have fallen. This trend was not seen in UK born children, in whom incidence rates increased over time until 2008 (Figure 4.13). This can be interpreted as TB transmission increasing and then decreasing from 2000 until 2015. Unfortunately this conclusion is difficult to extrapolate to older populations as it is likely that UK born children (the segment with non-UK born parents) have more interaction with non-UK born adults than UK born adults do. Additionally, BCG vaccination of high risk UK born children was introduced in 2005, which is likely to have depressed incidence rates since then. More complex modelling approaches are required to explore this question in more detail, this is explored in greater detail later in this thesis.

4.6 TB outcomes

4.6.1 Motivation

Whilst TB outcomes are tracked in detail in the yearly PHE TB reports,[1] the role of BCG vaccination has not previously been considered. There is some evidence that BCG vaccination may reduce all-cause mortality,[27,28,54] TB mortality,[22] and improve treatment outcomes.[29] The evidence for this in the ETS will be explored in the following section for; all-cause mortality, TB mortality, successful treatment at 12 months, and lost to follow up. TB outcomes are also likely to vary with age and UK birth status, both of which may mask potential variation due to BCG vaccination if not accounted for. As when identifying trends in TB notifications, relying solely on case counts for TB outcomes gives a biased picture as the underlying number of cases may change. For this reason in this section I explore TB outcomes using case rates.

4.6.2 Method

Case rates were calculated as follows and confidence estimates were estimated using the `prop.test` function from the `stats` package:

$$\textit{Case rate (over time period, t and age, a)} = \frac{\textit{Number of cases with outcome of interest (t,a)}}{\textit{Number of cases with known outcome (t, a)}} \times 100 \tag{4.2}$$

4.6.3 All-cause mortality

From 2000 to 2015 fewer UK born cases died from any cause in the ETS, but the number of non-UK born cases dying remained stable (Figure 4.15). However, the case all-cause fatality rate indicates that the rate of all-cause deaths has increased over time in both the

UK and non-UK born. There is also evidence to suggest that the case all-cause fatality rate is higher in those born in the UK that in the non-born and that it is higher for BCG vaccinated versus unvaccinated cases. The highest case all-cause fatality rate, regardless of UK birth status is observed in those missing UK birth status. In both populations the case all-cause fatality rate increases with age (as might be expected) but also has a secondary peak in early childhood (0-4) (Figure 4.16). The all-cause case fatality rate is higher in BCG unvaccinated cases, compared to vaccinated cases, from early adulthood until 50 years of age in the UK born but there is less evidence of a difference in the non-UK born. Young non-UK born children missing BCG status are particularly at risk of death from any cause.

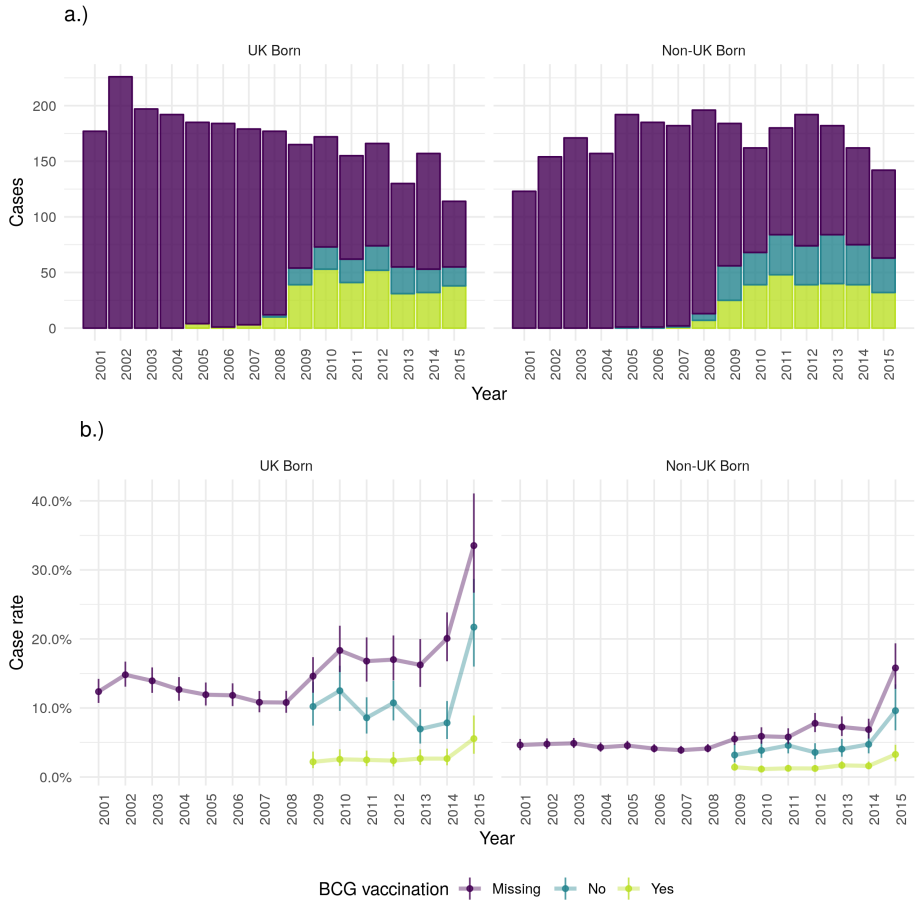


Figure 4.15: a.) Cases that died from any cause by year of notification stratified by UK birth and BCG status, b.) Case all-cause fatality rate stratified by UK birth and BCG status. Point estimates along with 95% confidence are shown for all estimates. All-cause mortality has reduced over time in the UK born but remained stable in the non-UK born. This is also reflected in the case fatality rate with the UK born having a higher rate regardless of BCG status. The recording of BCG status has improved over time but it appears that for years with data BCG unvaccinated cases have a higher all-cause case fatality rate in both the UK and non-UK born. In both populations those missing UK birth status are more likley to die from any cause.

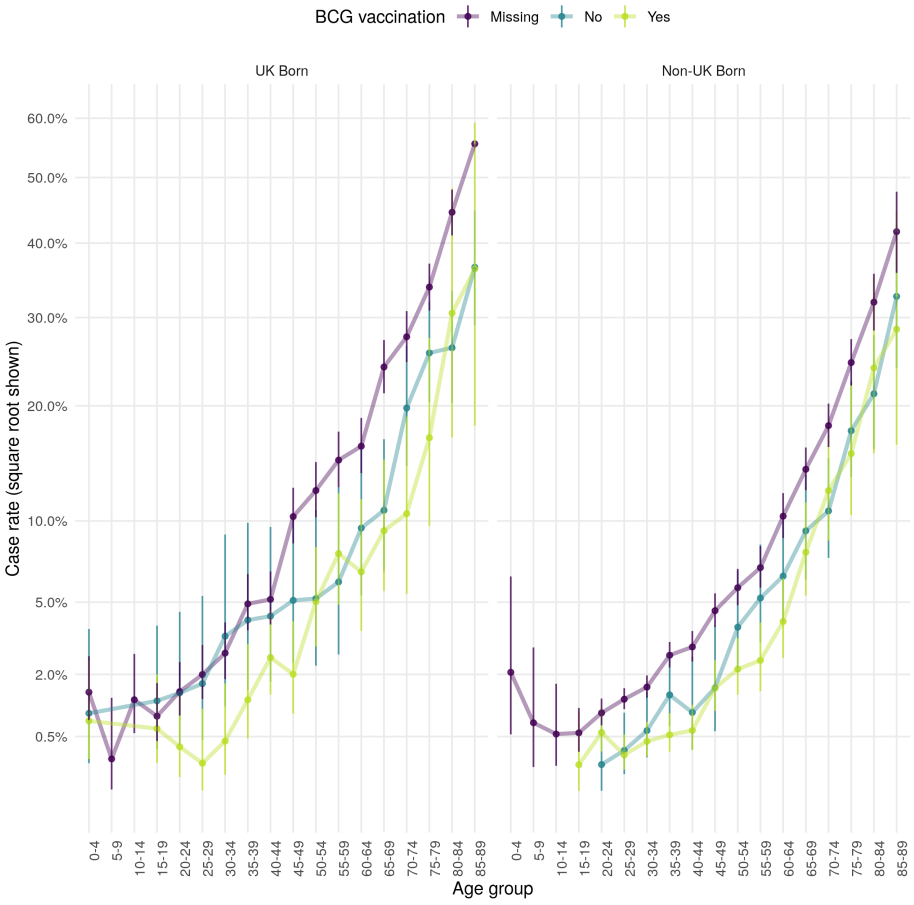


Figure 4.16: Age distribution (in 5 year age groups) of the case all-cause mortality rate presented on a square root scale. Estimates are stratified by BCG and UK birth status. Point estimates and 95% confidence intervals are shown for each case rate estimate. In both populaitons the case all-case fatality rate increases with age, and has a secondary peak in early childhood (0-4). The all-cause case fatality rate is higher in BCG unvaccinated cases, compared to vaccinated cases, from early adulthood until 50 years of age in the UK born. There is less evidence of a difference in case fatality rates in the non-UK born. Case missing BCG status are more likely to die in both populations, with young non-UK born children being particularly at risk.

4.6.4 TB related mortality

Similarly to all-cause deaths, deaths due to TB declined in the UK born over time but remained stable in the non-UK born (Figure 4.17). The case TB fatality rate also increased over time in both populations, with the rate again being higher in the UK born than in the non-UK born. There was still evidence of a higher case rate in those unvaccinated for BCG but the evidence for this was weaker. Comparing case TB fatality rates was difficult due to the large amount of uncertainty (Figure 4.18). However, there is some evidence to suggest that those missing BCG status, who were UK born and those who were older were more likely to die from TB.

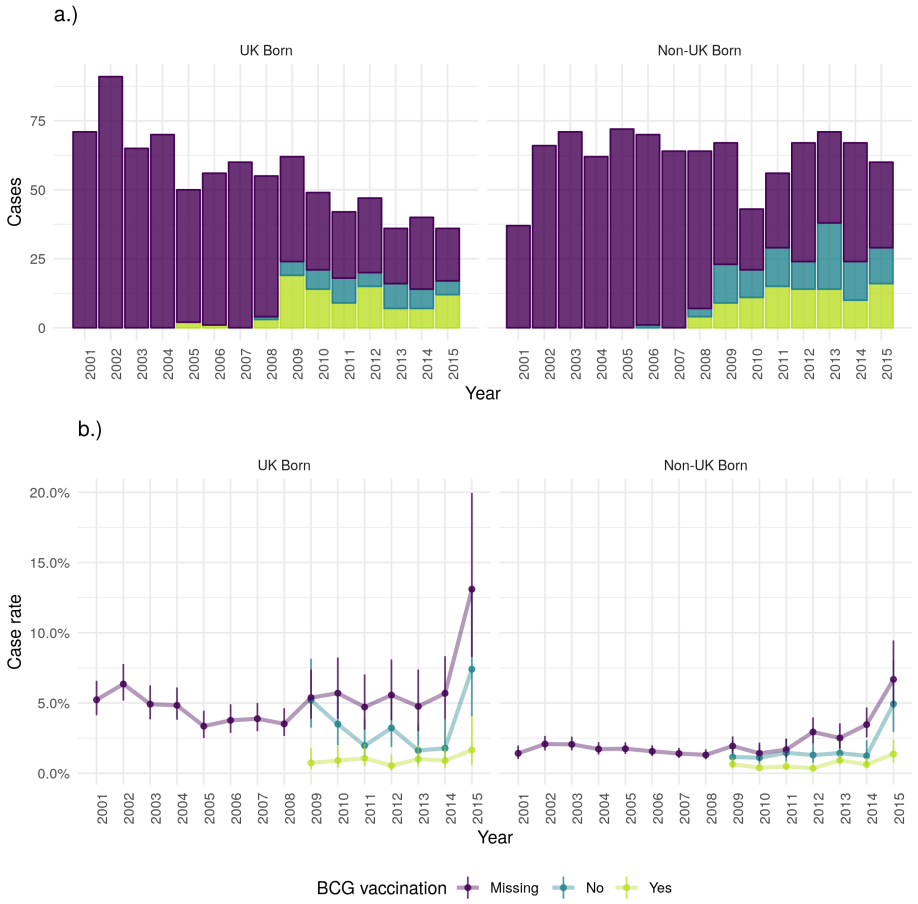


Figure 4.17: a.) Cases that died from TB by year of notification stratified by UK birth and BCG status, b.) Case TB fatality rate stratified by UK birth and BCG status. Point estimates along with 95% confidence are shown for all estimates. TB mortality has reduced over time in the UK born but remained stable in the non-UK born. This is also reflected in the case fatality rate with the UK born having a higher rate regardless of BCG status. The recording of BCG status has improved over time but it appears that for years with data BCG unvaccinated cases have a higher TB case fatality rate in both the UK and non-UK born. In both populations those missing UK birth status are more likley to die from TB. These findings match those found for all-cause mortality.

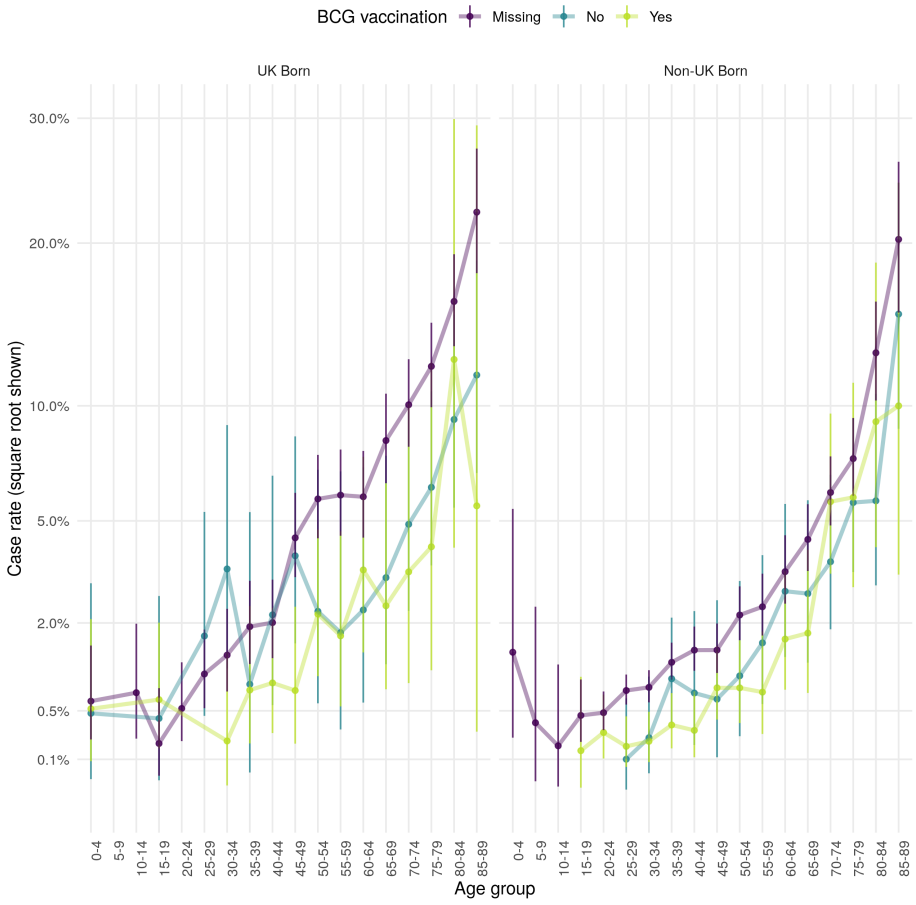


Figure 4.18: Age distribution (in 5 year age groups) of the case TB mortality rate presented on a square root scale. Estimates are stratified by BCG and UK birth status. Point estimates and 95% confidence intervals are shown for each case rate estimate. All estimates have a large degree of uncertainty making drawing conclusions difficult. There is no strong evidence to suggest a difference between those were BCG vaccinated and those that were not. Those that were missing BCG status, were UK born and who were older appeared to be at a greater risk than other cases of death from TB.

4.6.5 Successful treatment

Successful treatment within 12 months has increased in both populations over time in terms of cases (Figure 4.19). The case successful treatment rate initially decreased for both UK and non-UK born populations but since 2012 has improved in the UK born. There is little evidence to suggest that the successful treatment rate varies by BCG status, or by UK birth status. Successful treatment rates appear to be lowest for young adults and highest for young children (Figure 4.20).

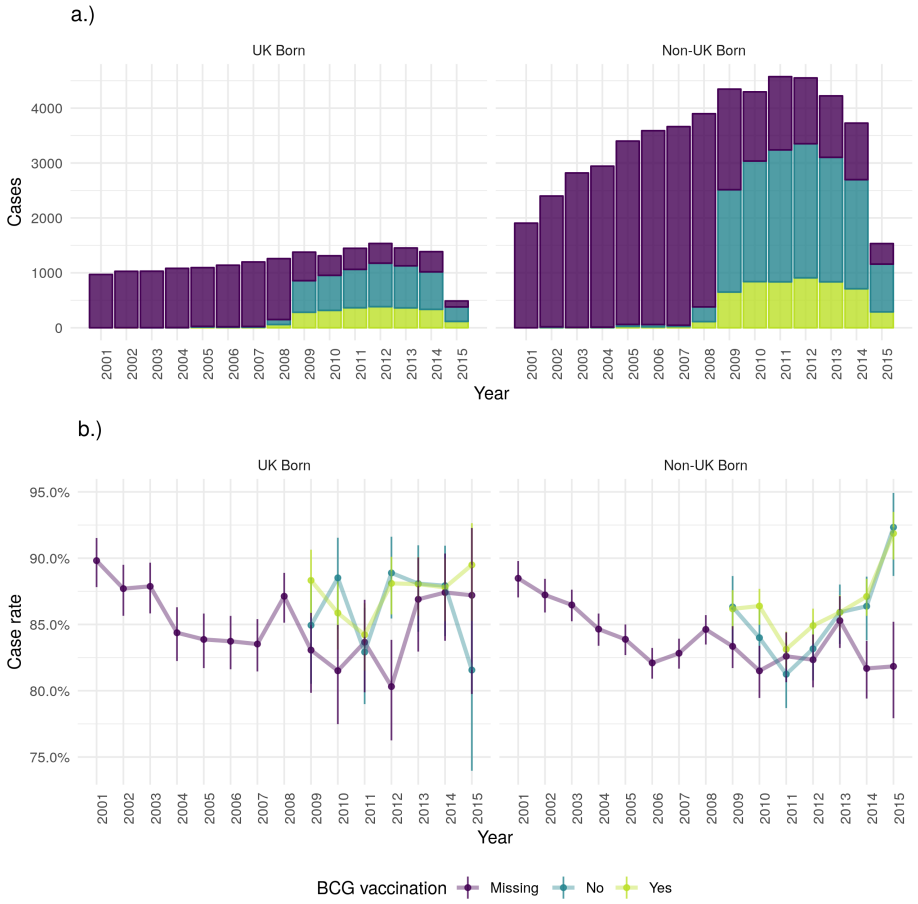


Figure 4.19: a.) Cases that were treated successfully within 12 months by year of notification stratified by UK birth and BCG status, b.) Case successful treatment within 12 months rate stratified by UK birth and BCG status. Point estimates along with 95% confidence are shown for all estimates. Successful treatment within 12 months has increased in both populations over time in terms of cases. The case successful treatment rate initailly decreased for both UK and non-UK born populations but since 2012 has improved in the UK born. There is little evidence to suggest that the case successful treatment rate varies by BCG status.

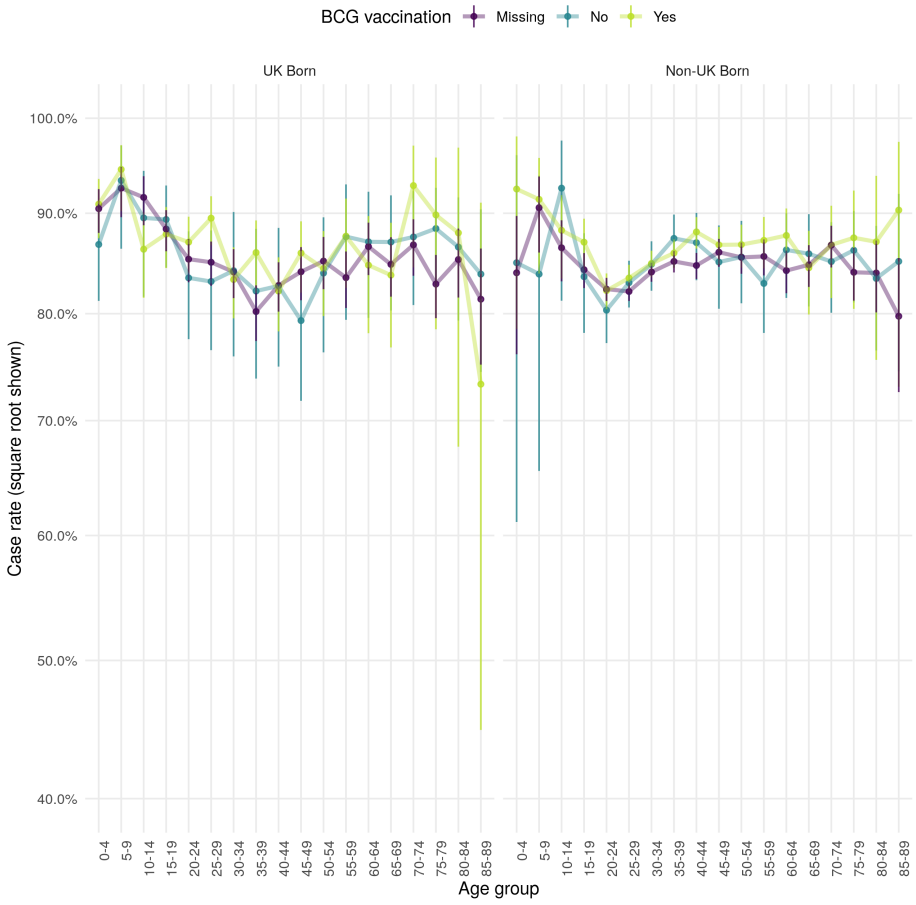


Figure 4.20: Age distribution (in 5 year age groups) of the case successful treatment within 12 months rate presented on a square root scale. Estimates are stratified by BCG and UK birth status. Point estimates and 95% confidence intervals are shown for each case rate estimate. There is little evidence that successful treatment rates differ greatly by BCG or UK birth status when stratified by age. Successful treatment rates appear to be lowest for young adults and highest for young children.

4.6.6 Lost to follow up

As for other outcomes discussed, cases lost to follow up has decreased over time in the UK born, but increased in the non-UK born (with incomplete data for 2015) (Figure 4.21). In all populations the case loss to follow up rate has decreased over time, although this may be biased as cases may not have had sufficient time to be classed as lost to follow up. In both populations there is little evidence to suggest variation by BCG status but the loss to follow up was higher in the non-UK born than in the UK born. This was true across all age groups, and there was again little evidence of variation due to BCG status (Figure 4.22). Young adults were the most likely to be lost follow up in both populations but this is appeared to be a particular issue in the non-UK born.

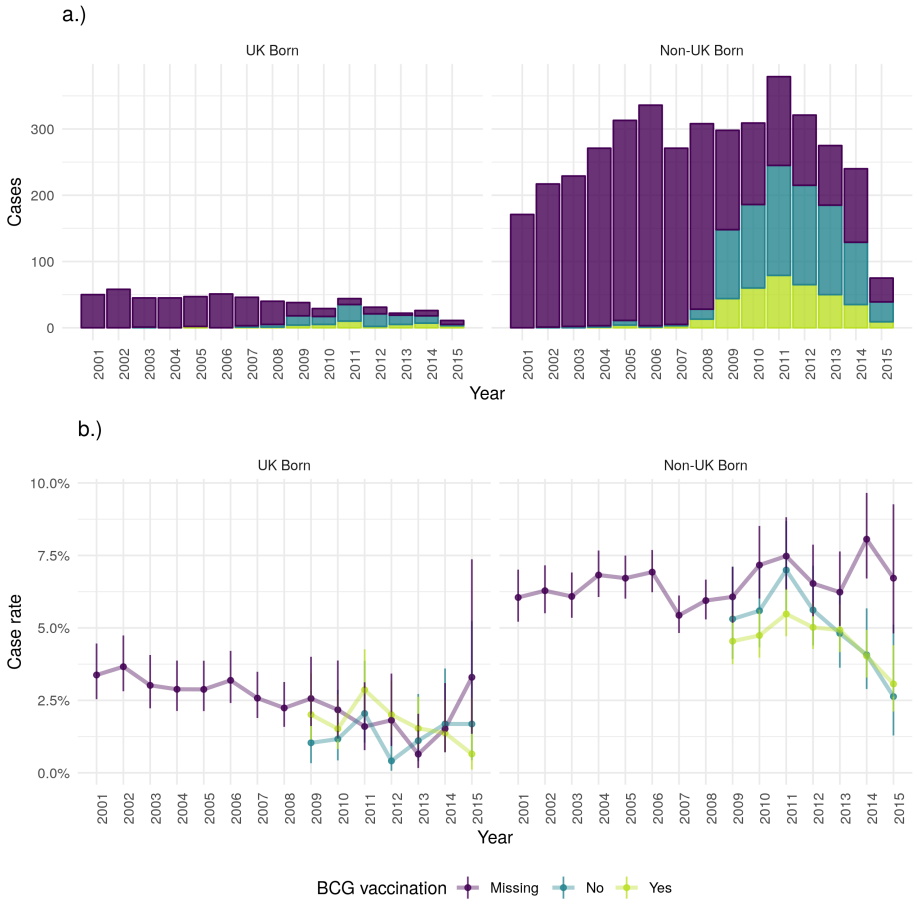


Figure 4.21: a.) Cases that were lost to follow up stratified by UK birth and BCG status, b.) Case lost to follow up rate stratified by UK birth and BCG status. Point estimates along with 95% confidence are shown for all estimates. Loss to follow up has decreased over time in the UK born, but increased in the non-UK born (with incomplete data for 2015). The case loss to follow up rate has decreased over time for the UK born but increased for the non-UK born. In both populations there is little evidence that loss to follow up varies by BCG status.

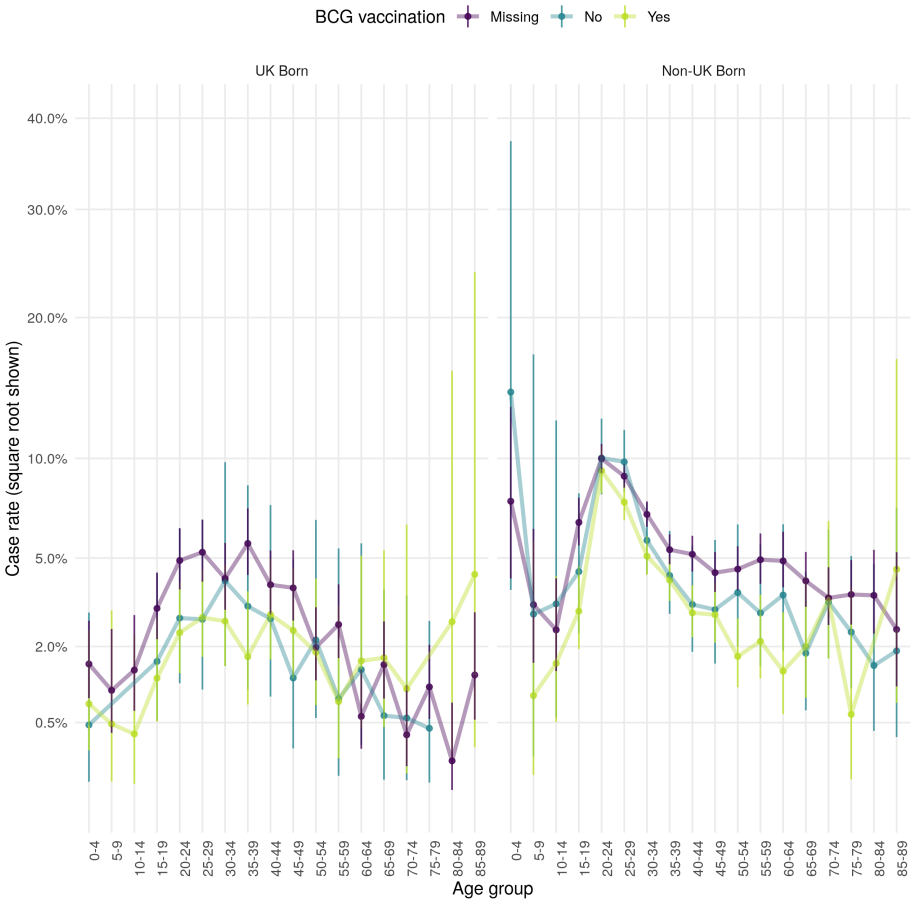


Figure 4.22: Age distribution (in 5 year age groups) of the case loss to follow up rate presented on a square root scale. Estimates are stratified by BCG and UK birth status. Point estimates and 95% confidence intervals are shown for each case rate estimate. There is little evidence of variation by BCG status but loss to follow up is higher in the non-UK born compared to the UK born across all age groups. Young adults are the most likely to be lost follow up in both populations but this is a particular issue in the non-UK born.

4.7 Discussion

In this chapter I have explored the epidemiology of TB in England using routine data-sets, with a particular focus on the impact of missing data, the mechanisms underlying that missing date, seasonal trends, the role of age, UK birth status and BCG status. I have also estimated incidence rates, stratified by UK birth status and age, which I then used to identify trends in TB incidence over time. Finally I explored TB outcomes in England using case rates, again stratified by BCG status and UK birth status.

In the the Enhanced Tuberculosis Surveillance (ETS) system, I found a high degree of missing data for several important variables. I also found that there is likely to be strong missing at random (MAR) mechanism underlying this missing data for multiple variables.

Several factors are strongly associated with data being missing for many variables, including UK birth status, ethnic group, socio-economic status and year. These MAR mechanisms must be adjusted for in future analysis to avoid bias. I found that date variables in particular suffered from changing data completeness over time, which may introduce spurious temporal trends if not fully understood. I also found that for several variables, including the date of symptom onset, that there was a large degree of recall bias both on the scales of days and months. Several variables, including date of notification and date of starting treatment showed a seasonal trend, with a maximum in the summer months. The date of ending treatment showed less evidence of a seasonal trend but there was some evidence of a maximum number of cases completing treatment in the winter months.

As reported elsewhere, I found that TB incidence initially increased from 2000, until 2011, but has since decreased. This was mainly driven by an changing incidence in the non-UK born with only a slight decrease in UK born incidence in recent years. Stratifying by age I found that non-UK born cases were more likely to young adults than any other age group but that the age distribution of the UK born was more nearly uniform. There was some evidence that these trends in TB incidence were driven by changing population demographics, with a large increase in the young adult non-UK born population between 2000 and 2015. In general the population of England is ageing, except for the non-UK born population which is still primarily made up of young adults. This is likely to impact trends in TB over time, with more severe outcomes but potentially less TB transmission.

After estimating incidence rates I found that (again as reported elsewhere) that TB incidence rates increased over time in the UK born from 2000 until 2005, since when they have declined year on year. There appeared to be some linkage between the UK born and the non-UK born, with incidence rates in the UK born initially decreasing until 2005 when they increased year on year until 2012. Since then they have decreased, in line with the decreases seen in the non-UK born. Stratifying incidence rates by age gives insights into what may be driving these mechanisms. In the non-UK born, incidence rates have decreased over time in children (0-14), increased in adults (15-64) through to 2005 before again beginning to decrease year on year, and remained relatively stable in older adults (65+) until 2011 since when they have also fallen. Theses trends are not mirrored in the UK born, with incidence rates initially increasing in children through to 2008 before beginning to decline and also increasing in adults through to 2011 before again beginning to decline. Incidence rates in older adults dramatically decreased between 2000 and 2015, with some evidence of a decline in the rate of this decrease from 2007 on-wards. These findings indicate that current reductions in TB incidence may not be reaching the young UK born adult population, additional control measures may be required to reduce TB incidence in this population further. Finally, I explored the use of incidence rates in UK born children as a proxy of TB transmission in England. There may be issues with this method as UK born children may not be representative of the population as a whole as they may be more likely to mix with higher risk non-UK born adults and because the change of BCG vaccination policy may have depressed incidence rates in children. More work is required, using both dynamic and statistical modelling, to understand whether incidence rates in children may be reliabley used to proxy TB transmission.

Using case rates I found that there was some evidence cases who were not BCG vaccinated may be more likely to suffer from negative TB outcomes. In particular with differences in all-cause mortality and TB mortality. These differences were observable after stratifying by

UK birth status and BCG status, with young adults deriving the greatest apparent benefit from BCG vaccination. TB outcomes were also generally worse in the non-UK born, except for successful treatment.

Findings from this chapter are used throughout the later chapters of this thesis. In particular, Chapter 6 uses statistical modelling to exploring the impact of BCG vaccination on TB outcomes in greater detail, Chapter 7 explores the impact of the change in BCG vaccination policy on TB incidence rates using the incidence rate estimates from this chapter, Chapter 8 uses the understanding of the ETS gained from this chapter to parameterise a dynamic TB transmission model, and Chapter **??** uses the insights gained into the date variables in the ETS to fit a dynamic TB transmission model.

4.8 Summary

- In this chapter the key data sources used in this thesis have been examined in detail, with a particular focus on the role of age, UK birth status and BCG vaccination status. The role of missing data, and potential mechanisms driving it have also been extensively explored. Data completeness was found to increase dramatically over time for many variables, which must be accounted for in any analysis using these variables to identify temporal trends.
- Tuberculosis incidence rates stratified by age and UK birth status have been calculated, along with case rates for TB outcomes. These estimates were then to extensively explore trends in TB in England, identifying possible analysis questions to be addressed later in this thesis.
- The code used in this chapter to import, clean and manipulate the data sources has been made accessible separately as an R package (`https://www.samabbott.co.uk/tbinenglanddataclean/`), along with extensive documentation of the required data sources and package functions. If interested in completely reproducing this work please see this documentation for details.
- Findings from this chapter are used throughout the later chapters of this thesis; to inform analysis questions (Chapter 6 and 7), identify variables for which missing data must be imputed (Chapter 6 and 7), to parameterise a dynamic TB transmission model (Chapter 8), and to fit a dynamic TB transmission model (Chapter **??**).

Chapter 5

Reassessing the evidence for universal school-age Bacillus Calmette Guerin (BCG) vaccination in England and Wales

5.1 Introduction

Prior to the change in BCG vaccination policy in 2005 (see Chapter 2) several studies were carried out to assess the impact of any potential policy change. In this Chapter I aimed to update one of these studies.

I recreated a previous approach for estimating the impact of ending the BCG schools scheme in England and Wales, updating the model with parameter uncertainty, and measurement error. I investigated scenarios considered by the Joint Committee on Vaccination and Immunisation, and explored new approaches using notification data (see Chapter 4). I estimated the number of vaccines needed to prevent a single notification, and the average annual additional notifications caused by ending the BCG schools' scheme. This work was supervised by Hannah Christensen and Ellen Brooks-Pollock.

5.2 Summary

- I found a 1.9% annual decrease in Tuberculosis incidence rates best matched our estimates from notifications.
- I estimate that 1600 (95% Interquartile range (IQR) 1300 to 2100) vaccines would have been required to prevent a single notification in 2004.
- If the scheme had ended in 2001, 302 (95% IQR 238 to 369) additional annual notifications would have occurred compared to if the scheme had continued. If the scheme ended in 2016, 120 (95% IQR 88 to 155) additional annual notifications would have

occurred.

- The estimates of the impact of ending the BCG schools scheme were highly sensitive to the annual decrease in incidence rates.
- The impact of ending the BCG schools scheme was found to be greater than previously thought when parameter values were updated and notifications were used.
- The results highlight the importance of including uncertainty when forecasting the impact of changes in vaccination policy.
- The code for the analysis contained in this Chapter can be found at: doi.org/10.5281/zenodo.2635687[1]

---

[1] Alternatively available from: `https://github.com/seabbs/AssessBCGPolicyChange`

Chapter 6

Exploring the effects of BCG vaccination in patients diagnosed with tuberculosis: observational study using the Enhanced Tuberculosis Surveillance system

6.1 Introduction

Bacillus Calmette–Guérin (BCG) primarily reduces the progression from infection to disease, however there is evidence that BCG may provide additional benefits. In this Chapter I aimed to investigate whether there is evidence in routinely-collected surveillance data (see Chapter 4) that BCG vaccination impacts outcomes for tuberculosis (TB) cases in England. Any impact on TB outcomes could add additional weight to vaccination policies with wider population coverage, as these policies would have benefits beyond reducing TB incidence rates.

To conduct this study, I first obtained all TB notifications for 2009-2015 in England from the Enhanced Tuberculosis surveillance system (see Chapter 4). I then considered five outcomes: All-cause mortality, death due to TB (in those who died), recurrent TB, pulmonary disease, and sputum smear status. I used logistic regression, with complete case analysis, to investigate each outcome with BCG vaccination, years since vaccination and age at vaccination, adjusting for potential confounders. All analyses were repeated using multiply imputed data. This work was supervised by Hannah Christensen and Ellen Brooks-Pollock. Collaborators at Public Health England including, Maeve K Lalor, Dominik Zenner, Colin Campbell, and Mary E Ramsay provided the data and commented on multiple versions of this work.

6.2 Background

Bacillus Calmette–Guérin (BCG) is one of the mostly widely-used vaccines and the only vaccine that protects against tuberculosis (TB) disease. BCG was first used in humans in 1921 and was introduced into the WHO Expanded Program on Immunization in 1974.[30] BCG vaccination has been controversial due to its variable efficacy and possibility of causing a false positive result with the standard skin test for TB.[4] However, the lack of a more effective vaccine and the emergence of drug-resistant TB strains means that BCG vaccination remains an important tool for reducing TB incidence and mortality rates.

BCG's primary mode of action is to directly prevent the development of active, symptomatic disease. Its efficacy in adults is context specific, with estimates ranging between 0% and 78% (see Chapter 2).[19] It has been shown to highly efficacious in England and there is some evidence that efficacy increases with distance from the equator. Efficacy has been shown to be dependent on previous exposure, with unexposed individuals receiving the greatest benefit.[55] Unlike in adults, BCG has consistently been shown to be highly protective against TB and TB meningitis in children.[17,18] For this reason the majority of countries that use BCG, vaccinate at birth.[21] Adult vaccination is no longer common in the UK, where universal BCG vaccination of adolescents was stopped in 2005 in favour of a targeted neonatal programme aimed at high risk children.

Vaccination policy has been primarily based on reducing the incidence of TB disease, and mitigating disease severity, with little attention having been given to any additional effects of BCG vaccination on TB outcomes.[24,25] There is some evidence that BCG vaccination induces innate immune responses which may provide non-specific protection,[26] TB patients with BCG scars were found to respond better to treatment with earlier sputum smear conversion,[29] and there is evidence to suggest that BCG vaccination is associated with reduced all-cause neonatal mortality[27,28] and both reduced TB[22] and all-cause[54] mortality in the general population. Given that the immunology behind TB immunity is not fully understood these findings suggest that BCG may play a more important role in improving TB outcomes than previously thought. I aimed to quantify the effects of BCG vaccination on outcomes for individuals with notified TB in England using routinely collected surveillance data (see Chapter 4) to provide evidence for appropriate public health action and provision. Where I found an association, I additionally explored the role of years since vaccination, and age at vaccination.

6.3 Method

6.3.1 Enhanced Tuberculosis Surveillance (ETS) system

I extracted all notifications from the Enhanced Tuberculosis Surveillance (ETS) system from January 1, 2009 to December 31, 2015 (Chapter 4). BCG vaccination status and year of vaccination have been collected since 2008. The outcomes I considered were: all-cause mortality, death due to TB (in those who died), recurrent TB, pulmonary disease, and sputum smear status. These outcomes were selected based on: their availability in the ETS; evidence from the literature of prior associations with BCG vaccination; associations with increased case infectiousness; or severe outcomes for patients.

All-cause mortality was defined using the overall outcome recorded in ETS, this is based

on up to 36 months of follow up starting from date of starting treatment. Follow up ends when a case is recorded as completing treatment, with treatment status evaluated at 12, 24, and 36 months from starting treatment. Where the treatment start date was not available the notification date was used if appropriate. The date of death was validated against Office for National Statistics (ONS) data. Those that were lost to follow up, or not evaluated were treated as missing. In cases with a known cause of death, death due to TB was defined as those that died from TB, or where TB had contributed to their death. Cause of death was recorded by case managers. TB cases who had recurrent episodes were identified using probabilistic matching. Positive sputum smear status was given to cases that had a sputum sample shown to contain Acid-Fast Bacilli. A positive sputum smear status indicates that cases are more likely to be infectious. Cases were defined as having pulmonary TB if a positive sputum smear sample was recorded, if a positive culture was grown from a pulmonary laboratory specimen, or if they were clinically assessed as having pulmonary TB.

6.3.2 Exposure variables relating to BCG

I included three exposure variables related to BCG: BCG status (vaccinated, yes/no), years since vaccination and age at vaccination.

BCG status was collected and recorded in ETS by case managers. Information on BCG vaccination status may have come from vaccination records, patient recall or the presence of a scar. When cases are uncertain, and there is no evidence of a scar, no BCG status is given. Year of vaccination was collected similarly. Years since BCG vaccination was defined as year of notification minus year of vaccination and categorised into two groups (0 to 10 and 11+ years). This was based on: evidence that the average duration of BCG protection is at least 10-15 years;[22] increasing recall bias with time since vaccination, and any association between years since vaccination and TB outcomes may be non-linear (see Chapter 4).

I calculated age at vaccination as year of vaccination minus year of birth. I categorized age at vaccination into 0 to < 1, 1 to < 12, 12 to < 16 and 16+ years because the distribution was bi-model with modes at 0 and 12 years. This categorization captures the current UK policy of vaccination at birth, historic policy of vaccination at 13-15 years and catch up vaccination for high risk children.

6.3.3 Statistical Analysis

R was used for all statistical analysis.[49] The analysis was conducted in two stages. Firstly, I calculated proportions for all demographic and outcome variables, and compared vaccinated and unvaccinated TB cases using the χ^2 test. Secondly, I used logistic regression, with complete case analysis, to estimate the association between exposures and outcome variables, both with and without adjustment for confounders.

In the multivariable models, I adjusted for sex,[56–58] age,[59] Index of Multiple Deprivation (2010) categorised into five groups for England (IMD rank),[11,60] ethnicity,[56,61] UK birth status,[39,62] and year of notification. As the relationship between age and outcomes was non-linear, I modelled age using a natural cubic spline with knots at the 25%, 50% and 75% quantiles.

I conducted sensitivity analyses to assess the robustness of the results, by dropping each confounding variable in turn and assessing the effect on the adjusted Odds Ratios (aORs) of the exposure variable. I repeated the analysis excluding duplicate recurrent cases, and restricting the study population to those eligible for the BCG schools scheme (defined as UK born cases that were aged 14 or over in 2004) to assess the comparability of the BCG vaccinated and unvaccinated populations. To mitigate the impact of missing data I used multiple imputation, with the MICE package.[63] I imputed 50 data sets (for 20 iterations) using) using all outcome and explanatory variables included in the analysis as predictors along with Public Health England centre. The model results were pooled using the small sample method,[64] and effect sizes compared with those from the main analysis.

6.4 Results

6.4.1 Description of the data

There were 51,645 TB notifications between 2009-2015 in England. Reporting of vaccination status and year of vaccination improved over time: 64.9% (20865/32154) of notifications included vaccination status for 2009 to 2012, increasing to 70% (13647/19491) from 2013 to 2015. The majority of cases that had a known vaccination status were vaccinated (70.6%, 24354/34512), and where age and year of vaccination was known, the majority of cases were vaccinated at birth (60%, 5979/10066).

Vaccinated cases were younger than unvaccinated cases on average (median age 34 years (IQR 26 to 45) compared to 38 years (IQR 26 to 62)). A higher proportion of non-UK borncases were BCG vaccinated, (72.7%, 18297/25171) compared to UK born cases (65.2%, 5787/8871, P: < 0.001) and, of those vaccinated, a higher proportion of non-UK born cases were vaccinated at birth compared to UK born cases (68%, 4691/6896 vs. 40.5%, 1253/3096 respectively, P: < 0.001). See Table 6.1 for the breakdown of outcome variables and Table 6.2 for the breakdown of confounding variables. See Chapter 4 for an extended discussion of the epidemiology of TB in England.

6.4.2 All-cause mortality

In the univariable analysis the odds of death from any cause were lower for BCG vaccinated TB cases compared to unvaccinated cases, with an OR of 0.28 (95% CI 0.24 to 0.32, P: <0.001) (Table 6.3, Table 6.4)); an association remained after adjusting for confounders, but was attenuated with an aOR of 0.76 (95% CI 0.64 to 0.89, P: 0.001). I estimate that if all unvaccinated cases had been vaccinated there would have been on average 19 (95% CI 9 to 29) fewer deaths per year during the study period (out of 81 deaths per year on average in unvaccinated cases). Whilst there was evidence in univariable analyses to suggest all-cause mortality was higher in persons vaccinated more than 10 years prior to notification of TB and that all-cause mortality increased with increasing age group, these disappeared after adjusting for potential confounders (Table 6.5, Table 6.6).

Similar results to the multivariable analysis were found using multiply imputed data for the association between vaccination status and all-cause mortality (aOR: 0.76 (95% CI 0.61 to 0.94), P: 0.013), but not for time since vaccination with a greatly increased risk of all-cause mortality estimated for those vaccinated more than 10 years before case notification, compared to those vaccinated more recently (aOR: 12.19 (95% CI 3.48 to 42.64), (see

Table 6.1: Outcomes for individuals in England notified with tuberculosis between 2009-2015, stratified by BCG vaccination status.

		BCG status		
Outcome	Total	Vaccinated	Unvaccinated	Unknown vaccine status
Total, all cases	51645	24354 {47}	10158 {20}	17133 {33}
All-cause mortality	45588 (88)	21685 (89)	9061 (89)	14842 (87)
No	43024 [94]	21291 [98]	8495 [94]	13238 [89]
Yes	2564 [6]	394 [2]	566 [6]	1604 [11]
Death due to TB (in those who died*)	1373 (3)	276 (1)	320 (3)	777 (5)
No	572 [42]	129 [47]	146 [46]	297 [38]
Yes	801 [58]	147 [53]	174 [54]	480 [62]
Recurrent TB	48497 (94)	23963 (98)	9991 (98)	14543 (85)
No	44869 [93]	22592 [94]	9256 [93]	13021 [90]
Yes	3628 [7]	1371 [6]	735 [7]	1522 [10]
Pulmonary TB	51432 (100)	24289 (100)	10121 (100)	17022 (99)
Extra-pulmonary (EP) only	24280 [47]	12085 [50]	4573 [45]	7622 [45]
Pulmonary, with or without EP	27152 [53]	12204 [50]	5548 [55]	9400 [55]
Sputum smear status - positive	19551 (38)	9768 (40)	3910 (38)	5873 (34)
Negative	11060 [57]	5694 [58]	2231 [57]	3135 [53]
Positive	8491 [43]	4074 [42]	1679 [43]	2738 [47]

{% all cases}(% complete within vaccine status)[% complete within category]
* Death due to TB in those who died and where cause of death was known

Table 6.5, Table 6.7)). For age at vaccination results for the multivariable analysis using multiply imputed data were comparable to those found using complete case analysis, except that there was some evidence that vaccination in adolescence, compared to under 1, was associated with increased, rather than decreased, all-cause mortality (aOR: 1.57 (95% CI 1.13 to 2.19), Table 6.9).

6.4.3 Deaths due to TB (in those who died)

There was little evidence of any association between BCG vaccination and deaths due to TB (in those who died and where cause of death was known) in the univariable analysis (Table 6.4). The adjusted point estimate indicated an association between BCG vaccination and reduced deaths due to TB (in those who died) although the confidence intervals remained wide with a similar result found using multiply imputed data (see Table 6.7). There were insufficient data to robustly estimate an association between deaths due to TB (in those who died) and years since vaccination or age at vaccination (Table 6.5, Table 6.6).

6.4.4 Recurrent TB

In both the univariable and multivariable analysis there was some evidence that BCG vaccination was associated with reduced recurrent TB, although the strength of the evidence

Table 6.2: Confounders for individuals in England notified with tuberculosis between 2009-2015, stratified by BCG vaccination status.

		BCG status		
Confounder	Total	Vaccinated	Unvaccinated	Unknown vaccine status
Total, all cases	51645	24354 {47}	10158 {20}	17133 {33}
Age	51645 (100)	24354 (100)	10158 (100)	17133 (100)
Mean [SD]	40 [19]	36 [16]	44 [22]	45 [20]
Median [25%, 75%]	36 [27, 52]	34 [26, 45]	38 [26, 62]	41 [29, 59]
Sex	51535 (100)	24320 (100)	10136 (100)	17079 (100)
Female	22066 [43]	10791 [44]	4312 [43]	6963 [41]
Male	29469 [57]	13529 [56]	5824 [57]	10116 [59]
IMD rank (with 1 as most deprived and 5 as least deprived)	43525 (84)	21240 (87)	8866 (87)	13419 (78)
1	16800 [39]	7779 [37]	3665 [41]	5356 [40]
2	13057 [30]	6836 [32]	2564 [29]	3657 [27]
3	6838 [16]	3459 [16]	1259 [14]	2120 [16]
4	4045 [9]	1893 [9]	836 [9]	1316 [10]
5	2785 [6]	1273 [6]	542 [6]	970 [7]
UK born	49820 (96)	24084 (99)	9958 (98)	15778 (92)
Non-UK Born	36988 [74]	18297 [76]	6874 [69]	11817 [75]
UK Born	12832 [26]	5787 [24]	3084 [31]	3961 [25]
Ethnic group	50416 (98)	24074 (99)	10024 (99)	16318 (95)
White	10194 [20]	3560 [15]	2695 [27]	3939 [24]
Black-Caribbean	1112 [2]	559 [2]	242 [2]	311 [2]
Black-African	8942 [18]	4620 [19]	1602 [16]	2720 [17]
Black-Other	462 [1]	261 [1]	80 [1]	121 [1]
Indian	12994 [26]	7176 [30]	2061 [21]	3757 [23]
Pakistani	8237 [16]	3512 [15]	1720 [17]	3005 [18]
Bangladeshi	2025 [4]	918 [4]	480 [5]	627 [4]
Chinese	601 [1]	289 [1]	101 [1]	211 [1]
Mixed / Other	5849 [12]	3179 [13]	1043 [10]	1627 [10]
Calendar year	51645 (100)	24354 (100)	10158 (100)	17133 (100)

{% all cases}(% complete within vaccine status)[% complete within category]
* Death due to TB in those who died and where cause of death was known

was weakened after adjusting for confounders (Table 6.4). In the adjusted analysis, the odds of recurrent TB were lower for BCG vaccinated cases compared to unvaccinated cases, with an aOR of 0.90 (95% CI 0.81 to 1.00, P: 0.056). The strength of the evidence for this association was comparable in the analysis using multiply imputed data (see Table 6.7). There was little evidence in the adjusted analysis of any association between recurrent TB and years since vaccination (Table 6.5) or age at vaccination (Table 6.6).

6.4.5 Other Outcomes

After adjusting for confounders there was little evidence for any association between BCG vaccination and pulmonary disease or positive sputum smear status (Table 6.4); similar

Table 6.3: Summary of logistic regression model output with BCG vaccination as the exposure and all-cause mortality as the outcome.

			Univariable		Multivariable	
Variable	Total	All-cause mortality	OR (95% CI)	P-value	aOR (95% CI)	P-value
Total cases	25993	807 (3)				
BCG vaccination				<0.001		0.001
No	7620	473 (6)	1		1	
Yes	18373	334 (2)	0.28 (0.24 to 0.32)		0.76 (0.64 to 0.89)	
Age				<0.001		<0.001
Sex				<0.001		<0.001
Female	11502	296 (3)	1		1	
Male	14491	511 (4)	1.45 (1.34 to 1.58)		1.48 (1.26 to 1.73)	
IMD rank (with 1 as most deprived and 5 as least deprived)				<0.001		0.001
1	9891	298 (3)	1		1	
2	8136	219 (3)	0.85 (0.76 to 0.95)		0.86 (0.70 to 1.04)	
3	4100	120 (3)	1.06 (0.93 to 1.20)		0.66 (0.52 to 0.84)	
4	2341	98 (4)	1.47 (1.28 to 1.70)		0.72 (0.55 to 0.93)	
5	1525	72 (5)	1.70 (1.45 to 1.99)		0.64 (0.47 to 0.85)	
UK born				<0.001		0.136
Non-UK Born	19115	442 (2)	1		1	
UK Born	6878	365 (5)	2.62 (2.40 to 2.85)		1.25 (0.93 to 1.67)	
Ethnic group				<0.001		0.171
White	4699	380 (8)	1		1	
Black-Caribbean	634	25 (4)	0.45 (0.35 to 0.58)		0.95 (0.59 to 1.53)	
Black-African	4681	62 (1)	0.14 (0.12 to 0.17)		0.87 (0.59 to 1.29)	
Black-Other	247	2 (1)	0.13 (0.06 to 0.26)		0.40 (0.10 to 1.69)	
Indian	7041	168 (2)	0.28 (0.25 to 0.31)		0.80 (0.58 to 1.10)	
Pakistani	4067	103 (3)	0.30 (0.27 to 0.34)		0.65 (0.46 to 0.92)	
Bangladeshi	1079	18 (2)	0.21 (0.16 to 0.27)		0.69 (0.40 to 1.22)	
Chinese	286	7 (2)	0.34 (0.23 to 0.51)		0.69 (0.30 to 1.62)	
Mixed / Other	3259	42 (1)	0.16 (0.13 to 0.19)		0.59 (0.39 to 0.91)	
Calendar year			1.06 (1.04 to 1.08)	<0.001	1.10 (1.05 to 1.15)	<0.001

OR (95% CI): unadjusted odds ratio with 95% confidence intervals,
aOR (95% CI): adjusted odds ratios with 95% confidence intervals

results were found using multiply imputed data (see Table 6.7).

Table 6.4: Summary of associations between BCG vaccination and all outcomes

		Univariable				Multivariable			
Outcome	BCG vaccinated	Cases**	Cases with outcome (%)	OR (95% CI)	P-value	Cases***	Cases with outcome (%)	aOR (95% CI)	P-value
All-cause mortality	No	9061	566 (6)	1	<0.001	7620	473 (6)	1	0.001
	Yes	21685	394 (2)	0.28 (0.24 to 0.32)		18373	334 (2)	0.76 (0.64 to 0.89)	
Death due to TB (in those who died*)	No	320	174 (54)	1	0.786	270	143 (53)	1	0.177
	Yes	276	147 (53)	0.96 (0.69 to 1.32)		236	126 (53)	0.76 (0.51 to 1.13)	
Recurrent TB	No	9991	735 (7)	1	<0.001	8502	615 (7)	1	0.056
	Yes	23963	1371 (6)	0.76 (0.70 to 0.84)		20584	1177 (6)	0.90 (0.81 to 1.00)	
Pulmonary TB	No	10121	5548 (55)	1	<0.001	8595	4685 (55)	1	0.769
	Yes	24289	12204 (50)	0.83 (0.79 to 0.87)		20784	10342 (50)	0.99 (0.94 to 1.05)	
Sputum smear status - positive	No	3910	1679 (43)	1	0.187	3367	1435 (43)	1	0.730
	Yes	9768	4074 (42)	0.95 (0.88 to 1.02)		8351	3447 (41)	1.02 (0.93 to 1.11)	

OR (95% CI): unadjusted odds ratio with 95% confidence intervals
aOR (95% CI): adjusted odds ratios with 95% confidence intervals
* Death due to TB in those who died and where cause of death was known
** Univariable sample size for outcomes ordered as in table (% of all cases) = 30746 (60%), 596 (23%), 33954 (66%), 34410 (67%), 13678 (26%)
*** Multivariable sample size with outcomes ordered as in table (% of all cases) = 25993 (50%), 506 (20%), 29086 (56%), 29379 (57%), 11718 (23%)

Table 6.5: Summary of associations between years since vaccination and all outcomes in individuals who were vaccinated. The baseline exposure is vaccination ≤ 10 years before diagnosis compared to vaccination 11+ years before diagnosis. Deaths due to TB (in those who died) had insufficient data for effect sizes to be estimated in both the univariable and multivariable analysis

		Univariable				Multivariable			
Outcome	Years since BCG	Cases**	Cases with outcome (%)	OR (95% CI)	P-value	Cases***	Cases with outcome (%)	aOR (95% CI)	P-value
All-cause mortality	≤ 10	718	5 (1)	1	0.004	554	4 (1)	1	0.897
	11+	8106	166 (2)	2.98 (1.22 to 7.28)		7171	148 (2)	0.91 (0.24 to 3.54)	
Death due to TB (in those who died*)	≤ 10	2	2 (100)	1	-	2	2 (100)	1	-
	11+	108	59 (55)	*Insufficient data*		98	53 (54)	*Insufficient data*	
Recurrent TB	≤ 10	780	22 (3)	1	0.005	613	14 (2)	1	0.515
	11+	9172	451 (5)	1.78 (1.15 to 2.75)		8194	406 (5)	1.24 (0.63 to 2.44)	
Pulmonary TB	≤ 10	770	480 (62)	1	<0.001	601	382 (64)	1	0.309
	11+	9248	4757 (51)	0.64 (0.55 to 0.74)		8254	4232 (51)	0.87 (0.67 to 1.14)	
Sputum smear status - positive	≤ 10	157	81 (52)	1	0.941	122	61 (50)	1	0.920
	11+	3064	1590 (52)	1.01 (0.73 to 1.40)		2734	1405 (51)	1.02 (0.68 to 1.54)	

OR (95% CI): unadjusted odds ratio with 95% confidence intervals
aOR (95% CI): adjusted odds ratios with 95% confidence intervals
* Death due to TB in those who died and where cause of death was known
** Univariable sample size for outcomes ordered as in table (% of vaccinated cases) = 8824 (36%), 110 (28%), 9952 (41%), 10018 (41%), 3221 (13%)
*** Multivariable sample size with outcomes ordered as in table (% of vaccinated cases) = 7725 (32%), 100 (25%), 8807 (36%), 8855 (36%), 2856 (12%)

Table 6.6: Summary of associations between age at vaccination and all outcomes in individuals who were vaccinated - the baseline exposure is vaccination at birth compared to vaccination from 1 to < 12, 12 to < 16, and 16+ years of age.

		Univariable				Multivariable			
Outcome	Age at BCG	Cases**	Cases with outcome (%)	OR (95% CI)	P-value	Cases***	Cases with outcome (%)	aOR (95% CI)	P-value
All-cause mortality	< 1	5234	45 (1)	1	<0.001	4626	43 (1)	1	0.127
	1 to < 12	1915	58 (3)	3.60 (2.43 to 5.34)		1678	52 (3)	1.36 (0.85 to 2.16)	
	12 to < 16	1267	41 (3)	3.86 (2.51 to 5.91)		1094	32 (3)	0.81 (0.45 to 1.46)	
	≥ 16	408	27 (7)	8.17 (5.01 to 13.32)		327	25 (8)	1.41 (0.76 to 2.63)	
Death due to TB (in those who died*)	< 1	27	20 (74)	1	0.118	27	20 (74)	1	0.543
	1 to < 12	43	20 (47)	0.30 (0.11 to 0.87)		39	18 (46)	0.36 (0.08 to 1.51)	
	12 to < 16	23	13 (57)	0.46 (0.14 to 1.50)		17	9 (53)	0.40 (0.06 to 2.52)	
	≥ 16	17	8 (47)	0.31 (0.09 to 1.12)		17	8 (47)	0.35 (0.06 to 2.16)	
Recurrent TB	< 1	5909	284 (5)	1	0.463	5275	258 (5)	1	0.246
	1 to < 12	2174	105 (5)	1.01 (0.80 to 1.26)		1928	92 (5)	0.84 (0.65 to 1.09)	
	12 to < 16	1421	58 (4)	0.84 (0.63 to 1.12)		1242	51 (4)	0.70 (0.48 to 1.02)	
	≥ 16	448	26 (6)	1.22 (0.81 to 1.85)		362	19 (5)	0.82 (0.49 to 1.37)	
Pulmonary TB	< 1	5946	2828 (48)	1	<0.001	5305	2510 (47)	1	0.005
	1 to < 12	2194	1159 (53)	1.23 (1.12 to 1.36)		1941	1033 (53)	1.15 (1.02 to 1.29)	
	12 to < 16	1425	971 (68)	2.36 (2.09 to 2.67)		1245	846 (68)	1.09 (0.92 to 1.29)	
	≥ 16	453	279 (62)	1.77 (1.45 to 2.15)		364	225 (62)	1.47 (1.15 to 1.88)	
Sputum smear status - positive	< 1	1753	836 (48)	1	<0.001	1557	742 (48)	1	0.862
	1 to < 12	755	394 (52)	1.20 (1.01 to 1.42)		682	348 (51)	0.96 (0.79 to 1.17)	
	12 to < 16	556	357 (64)	1.97 (1.62 to 2.40)		486	308 (63)	1.06 (0.81 to 1.39)	
	≥ 16	157	84 (54)	1.26 (0.91 to 1.75)		131	68 (52)	0.93 (0.63 to 1.37)	

OR (95% CI): unadjusted odds ratio with 95% confidence intervals
aOR (95% CI): adjusted odds ratios with 95% confidence intervals
* Death due to TB in those who died and where cause of death was known
** Univariable sample size for outcomes ordered as in table (% of vaccinated cases) = 8824 (36%), 110 (28%), 9952 (41%), 10018 (41%), 3221 (13%)
*** Multivariable sample size with outcomes ordered as in table (% of vaccinated cases) = 7725 (32%), 100 (25%), 8807 (36%), 8855 (36%), 2856 (12%)

6.4.6 Sensitivity analysis of the missing data using multiple imputation

As discussed in the previous sections, I found that repeating the analysis with an imputed data set had some effect on the results from the complete case analysis. There was a decrease in the accuracy of effect size estimates for BCG vaccination, some increase in p-values (Table 6.7). However, none of the estimated effects changed their direction, and there were no detectable systematic changes in the results. For the secondary exposure variables

Table 6.7: Summary of associations between BCG vaccination and all outcomes, using pooled imputed data.

	Univariable			Multivariable		
Outcome	OR (95% CI)	P-value	fmi	aOR (95% CI)	P-value	fmi
All-cause mortality	0.44 (0.35 to 0.56)	<0.001	90	0.76 (0.61 to 0.94)	0.013	85
Death due to TB (in those who died*)	0.94 (0.57 to 1.56)	0.810	85	0.89 (0.52 to 1.51)	0.651	85
Recurrent TB	0.83 (0.75 to 0.92)	<0.001	56	0.90 (0.81 to 1.00)	0.058	54
Pulmonary TB	0.84 (0.79 to 0.90)	<0.001	70	0.99 (0.93 to 1.06)	0.814	62
Sputum smear status - positive	0.88 (0.82 to 0.94)	<0.001	65	1.01 (0.94 to 1.08)	0.886	60

OR: odds ratio with 95% confidence intervals
aOR: adjusted odds ratio with 95% confidence intervals
fmi: fraction of missing information
* Death due to TB in those who died and where cause of death was known

(years since vaccination and age at vaccination, (Table 6.8 and Table 6.9), I found a change in direction of the point estimate between years since vaccination and all-cause mortality and recurrent TB, but similar results for age at vaccination and outcomes.

6.4.7 Sensitivity analysis

Dropping duplicate recurrent TB notifications increased the magnitude, and precision, of the effect sizes for recurrent TB, all-cause mortality, and deaths due to TB (in those who died) (see Table 6.10). Restricting the analysis to only cases that were eligible for the BCG schools scheme reduced the sample size of the analysis (from an initial study size of 51645, of which 12832 were UK born, to 9943 cases that would have been eligible for the BCG schools scheme). With this reduced sample size, there was strong evidence in adjusted analyses of an association between BCG vaccination and reduced recurrent TB, and evidence of an association with decreased all-cause mortality (see Table 6.10).

6.5 Discussion

Using TB surveillance data collected in England I found that BCG vaccination, prior to the development of active TB, was associated with reduced all-cause mortality and fewer

Table 6.8: Summary of associations between years since vaccination and all outcomes, using pooled imputed data. There was insufficient data to estimate an effect for deaths due to TB (in those who died)

	Univariable			Multivariable		
Outcome	OR (95% CI)	P-value	fmi	aOR (95% CI)	P-value	fmi
All-cause mortality	3.28 (1.85 to 5.79)	<0.001	50	12.19 (3.48 to 42.64)	<0.001	70
Death due to TB (in those who died*)	0.00 (0.00 to Inf)	0.974	0	0.00 (0.00 to Inf)	0.972	0
Recurrent TB	1.29 (1.00 to 1.66)	0.050	39	0.81 (0.59 to 1.11)	0.187	44
Pulmonary TB	0.58 (0.52 to 0.66)	<0.001	33	0.99 (0.84 to 1.17)	0.913	40
Sputum smear status - positive	0.99 (0.82 to 1.19)	0.891	70	0.95 (0.77 to 1.18)	0.648	60

OR: odds ratio with 95% confidence intervals
aOR: adjusted odds ratio with 95% confidence intervals
fmi: fraction of missing information
* Death due to TB in those who died and where cause of death was known

Table 6.9: Summary of associations between age at vaccination and all outcomes, using pooled imputed data (reference is vaccination at <1 year).

		Univariable			Multivariable		
Outcome	Age group	OR (95% CI)	P-value	fmi	aOR (95% CI)	P-value	fmi
All-cause mortality	1 to < 12	6.48 (4.71 to 8.91)	<0.001	70	1.69 (1.18 to 2.40)	0.004	68
	12 to < 16	3.33 (2.50 to 4.43)	<0.001	78	1.57 (1.13 to 2.19)	0.008	79
	≥ 16	3.36 (2.56 to 4.41)	<0.001	69	1.01 (0.70 to 1.46)	0.948	71
Death due to TB (in those who died*)	1 to < 12	0.45 (0.22 to 0.92)	0.028	62	0.47 (0.21 to 1.04)	0.063	62
	12 to < 16	0.41 (0.22 to 0.75)	0.004	67	0.40 (0.20 to 0.78)	0.008	67
	≥ 16	0.53 (0.28 to 1.00)	0.051	54	0.47 (0.20 to 1.12)	0.088	62
Recurrent TB	1 to < 12	1.39 (1.11 to 1.73)	0.004	41	1.04 (0.82 to 1.32)	0.736	41
	12 to < 16	1.01 (0.88 to 1.16)	0.892	45	0.86 (0.75 to 1.00)	0.052	44
	≥ 16	0.95 (0.79 to 1.15)	0.598	53	0.77 (0.61 to 0.98)	0.034	55
Pulmonary TB	1 to < 12	1.83 (1.59 to 2.10)	<0.001	46	1.36 (1.17 to 1.58)	<0.001	44
	12 to < 16	1.28 (1.19 to 1.36)	<0.001	35	1.12 (1.04 to 1.21)	0.002	36
	≥ 16	2.28 (2.10 to 2.48)	<0.001	34	1.10 (0.98 to 1.23)	0.107	40
Sputum smear status - positive	1 to < 12	1.49 (1.21 to 1.84)	<0.001	74	1.08 (0.85 to 1.37)	0.549	76
	12 to < 16	1.29 (1.17 to 1.43)	<0.001	65	1.09 (0.97 to 1.22)	0.158	67
	≥ 16	2.40 (2.16 to 2.66)	<0.001	58	1.20 (1.04 to 1.37)	0.011	59

OR: odds ratio with 95% confidence intervals
aOR: adjusted odds ratio with 95% confidence intervals
fmi: fraction of missing information
* Death due to TB in those who died and where cause of death was known

recurrent TB cases, although the evidence for this association was weaker. There was some suggestion that the association with all-cause mortality was due to reduced deaths due to TB (in those who died), though the study was underpowered to definitively assess this. I

Table 6.10: Summary of associations between BCG vaccination and all outcomes; cases that have no recurrent flag in the ETS (n=50407), and cases that would have been eligible for the BCG schools scheme (n=9943). Those defined to be eligible for the schools scheme are the UK born, that were aged 14 or over in 2004

			Univariable		Multivariable	
Study population	Outcome	BCG	OR (95% CI)	P-value	aOR (95% CI)	P-value
Recurrent cases dropped						
	All-cause mortality	No	1	<0.001	1	<0.001
		Yes	0.27 (0.23 to 0.31)		0.73 (0.61 to 0.86)	
	Death due to TB (in those who died*)	No	1	0.709	1	0.147
		Yes	0.94 (0.68 to 1.31)		0.74 (0.49 to 1.11)	
	Recurrent TB	No	1	<0.001	1	<0.001
		Yes	0.61 (0.55 to 0.69)		0.76 (0.66 to 0.87)	
	Pulmonary TB	No	1	<0.001	1	0.672
		Yes	0.83 (0.79 to 0.87)		0.99 (0.93 to 1.04)	
	Sputum smear status - positive	No	1	0.141	1	0.871
		Yes	0.94 (0.88 to 1.02)		1.01 (0.92 to 1.10)	
Cases eligible for the schools scheme						
	All-cause mortality	No	1	<0.001	1	0.018
		Yes	0.24 (0.19 to 0.29)		0.72 (0.55 to 0.95)	
	Death due to TB (in those who died*)	No	1	0.893	1	0.987
		Yes	0.96 (0.57 to 1.63)		0.99 (0.49 to 2.03)	
	Recurrent TB	No	1	<0.001	1	<0.001
		Yes	0.51 (0.42 to 0.61)		0.66 (0.52 to 0.84)	
	Pulmonary TB	No	1	0.017	1	0.417
		Yes	0.87 (0.78 to 0.98)		0.94 (0.82 to 1.08)	
	Sputum smear status - positive	No	1	0.613	1	0.588
		Yes	1.04 (0.89 to 1.22)		1.05 (0.87 to 1.27)	

OR: odds ratio with 95% confidence intervals
aOR: adjusted odds ratio with 95% confidence intervals
fmi: fraction of missing information
* Death due to TB in those who died and where cause of death was known

did not find evidence of an association between BCG status and positive smear status or pulmonary TB. Analysis with multiply imputed data indicated that notification 10+ years after vaccination was associated with increased all-cause mortality. In separate analyses, there was some evidence that vaccination at birth, compared to at any other age, was associated with reduced all-cause mortality, and increased deaths due to TB (in those who died).

This study used a large detailed dataset, with coverage across demographic groups, and standardized data collection from notifications and laboratories. The use of routine surveillance data means that this study would be readily repeatable with new data. The surveillance data contained multiple known risk factors, this allowed us to adjust for these confounders in the multivariable analysis, which attenuated the evidence for an association with BCG vaccination for all outcomes. However, there are important limitations to consider. The

study was conducted within a population of active TB cases, therefore the association with all-cause mortality cannot be extrapolated to the general population. Additionally, vaccinated and unvaccinated populations may not be directly comparable because vaccination has been targeted at high-risk neonates in the UK since 2005. I mitigated this potential source for bias by conducting a sensitivity analysis including only those eligible for the universal school age scheme, and whilst the strength of associations were attenuated there remained some evidence of improved outcomes. Sensitivity analysis excluding recurrent cases indicated their inclusion may have biased our results towards the null.

Variable data completeness changed with time, with both BCG vaccination status and year of vaccination having a high percentage of missing data, which may not be missing completely at random. I therefore checked the robustness of our results with multiple imputation including regional variability, however an unknown missing not at random mechanism, or unmeasured confounding may still have introduced bias. I found a greatly increased risk of all-cause mortality for those vaccinated more than 10 years ago in the analysis with multiply imputed data, compared to the complete case analysis. This is likely to be driven by a missing not at random mechanism for years since vaccination, with older cases being both more likely to have been vaccinated more than 10 years previously and to also have an unknown year of vaccination. The high percentage of missing data also means that I was likely to be underpowered to detect an effect of BCG vaccination on sputum smear status and deaths due to TB (in those who died), with years since vaccination, and age at vaccination likely to be underpowered for all outcomes. I was not able to adjust for either tuberculin skin test (TST) stringency, or the latitude effect, although I was able to adjust for UK birth status.[65] However, the bias induced by these confounders is likely to be towards the null, meaning that our effect estimates are likely to be conservative. Finally, BCG vaccination status, and year of vaccination, may be subject to misclassification due to recall bias; validation studies of the recording of BCG status in the ETS would be required to assess this.

Little work has been done to assess the overall effect of BCG on outcomes for active TB cases although the possible non-specific effects of BCG are an area of active research.[28,66,67] Whilst multiple studies have investigated BCG's association with all-cause mortality, it has been difficult to assess whether the association continues beyond the first year of life.[67] The effect size of the association I identified between BCG and all-cause mortality in active TB cases was comparable to that found in a Danish case-cohort study in the general population (aHR: 0.58 (95% CI 0.39 to 0.85).[54] A recent systematic review also found that BCG vaccination was associated with reduced all-cause mortality in neonates, with an average relative risk of 0.70 (95% CI 0.49 to 1.01) from five clinical trials and 0.47 (95% CI 0.32 to 0.69) from nine observational studies at high risk of bias.[28] I found some weak evidence that BCG vaccination was associated with reduced deaths due to TB (in those who died), although our point estimate had large confidence intervals. Several meta-analyses have found evidence supporting this association,[18,22] with one meta-analysis estimating a 71% (RR: 0.29 95% CI 0.16 to 0.53) reduction in deaths due to TB in individuals vaccinated with BCG.[18] The meta-analysis performed by Abubakar et al. also found consistent evidence for this association, with a Rate Ratio of 0.22 (95% CI 0.15 to 0.33).[22] In contrast to our study, both of these meta-analyses estimated the protection from TB mortality in BCG vaccinated individuals rather than in BCG vaccinated cases who had died from any cause. Additionally, neither study explored the association between BCG vaccination and all-cause mortality or recurrent TB. This study could not determine the possible causal pathway for

the association between BCG vaccination all-cause mortality, and recurrent TB. These are important to establish in order to understand the effect of BCG vaccination on TB outcomes.

I found that BCG vaccination was associated with reduced all-cause mortality, with some weaker evidence of an association with reduced recurrent TB. A plausible mechanism for this association is that BCG vaccination improves treatment outcomes,[29] which then results in decreased mortality, and reduced recurrent TB. However, these effects may also be independent and for all-cause mortality may not be directly related to active TB. In this case, a possible mechanism for the association between BCG vaccination and all-cause mortality is that BCG vaccination modulates the innate immune response, resulting in non-specific protection.[26] For low incidence countries, where the reduction in TB cases has been used as evidence to scale back vaccination programs,[21] these results suggest that BCG vaccination may be more beneficial than previously thought. In countries that target vaccination at those considered to be at high risk of TB the results from this study could be used to help drive uptake by providing additional incentives for vaccination. The evidence I have presented should be considered in future cost-effectiveness studies of BCG vaccination programs.

Several other Chapters (Chapter 5, Chapter 7, and Chapter 10) in this thesis assess the impact of moving from universal school age vaccination to selective high risk neonatal vaccination. The reduction in BCG coverage that this implies means that on top of any potential increase in TB incidence rates there may also have been a reduction in the benefical effects from the BCG vaccine discussed in this Chapter. However, as outlined in the previous paragraph, the evidence of reductions in both all-cause, and TB specific mortality, is strongest in the early years of life. This means that the move to neonatal vaccination may have actually led to an increase in the non-specific benefits.

Further work is required to determine whether years since vaccination and age at vaccination are associated with TB outcomes as this study was limited by low sample size, missing data for year of vaccination, and the relative rarity of some TB outcomes. However, due to the continuous collection of the surveillance data used in this analysis, this study could be repeated once additional data have been collected. If this study were to be repeated with a larger sample size particular attention should be given to the functional form of any decay in protection from negative TB outcomes. Additionally, a larger sample size would allow investigation of the associations identified between TB outcomes and BCG vaccination stratified by pulmonary, extrapulmonary, and disseminated TB disease. The results from this study require validation in independent datasets and the analysis should be reproducible in other low incidence countries that have similarly developed surveillance systems. If validated in low incidence countries, similar studies in medium to high incidence countries should be conducted because any effect would have a greater impact in these settings.

6.6 Summary

- I found evidence of an association between BCG vaccination and reduced all-cause mortality (aOR:0.76 (95%CI 0.64 to 0.89), P:0.001) and weak evidence of an association with reduced recurrent TB (aOR:0.90 (95%CI 0.81 to 1.00), P:0.056). Analyses using multiple imputation suggested that the benefits of vaccination for all-cause mor-

tality were reduced after 10 years.

- There was some suggestion that the association with all-cause mortality was due to reduced deaths due to TB (in those who died), though the study was underpowered to definitively assess this.
- There was little evidence for other associations.
- The code for the analysis contained in this Chapter can be found at: `doi.org/10.5281/zenodo.1213799`[1]

[1] Alternatively available from: `https://github.com/seabbs/ExploreBCGOnOutcomes`

Chapter 7

Estimating the effect of the 2005 change in BCG policy in England: A retrospective cohort study

7.1 Introduction

In 2005, England changed from universal Bacillus Calmette–Guérin (BCG) vaccination of school-age children to targeted BCG vaccination of high-risk children at birth. In this Chapter I aimed to assess the effects of this change in vaccination policy on the populations targeted by each vaccination scheme.

I combined notification data from the Enhanced Tuberculosis Surveillance system, with demographic data from the Labour Force Survey to construct retrospective cohorts of individuals in England relevant to both the universal, and targeted vaccination programmes between Jan 1, 2000 and Dec 31, 2010. For each cohort, I estimated incidence rates over a 5 year follow-up period and used Poisson and Negative Binomial regression models in order to estimate the impact of the change in policy on TB. This work was supervised by Hannah Christensen and Ellen Brooks-Pollock. Nicky Weltton provided guidance on the statistical methods used.

7.2 Background

In 2005 England changed its Bacillus Calmette–Guérin (BCG) vaccination policy against tuberculosis (TB) from a universal programme aimed at 13 and 14 year olds to a targeted programme aimed at high-risk neonates (see Chapter 2). High risk babies are identified by local TB incidence and by the parents' and grandparents' country of origin. The change in policy was motivated by evidence of reduced TB transmission,[15,24,25] and high effectiveness of the BCG vaccine in children,[3,17,18] and variable effectiveness in adults.[21] Little work has been done to evaluate the impact of this change in vaccination policy.

Globally, several countries with low TB incidence have moved from universal vaccination, either of those at school-age or neonates, to targeted vaccination of neonates considered at high-risk of TB (see Chapter 2).[4] In Sweden, which discontinued universal vaccina-

tion of neonates in favour of targeted vaccination of those at high risk, incidence rates in Swedish-born children increased slightly after the change in policy. [68] In France, which also switched from universal vaccination of neonates to targeted vaccination of those at high-risk, a study found that targeted vaccination of neonates may have reduced coverage in those most at risk.[69]

The number of TB notifications in England increased from 6929 in 2004 to 8280 in 2011 but has since declined to 5137 in 2017 (see Chapter 4).[15] A recent study found that this reduction may be linked to improved TB interventions.[70] Directly linking trends in TB incidence to transmission is complex because after an initial infection an individual may either develop active disease, or enter a latent stage which then may later develop into active disease. Incidence in children is a proxy of TB transmission, because any active TB disease in this population is attributable to recent transmission. Using this approach it is thought that TB transmission has been falling in England for the last 5 years, a notion supported by strain typing.[15] However, this does not take into account the change in BCG policy, which is likely to have reduced incidence rates in children.

Although the long term effects of BCG vaccination such as reducing the reactivation of latent cases and decreasing on-wards transmission are not readily detectable over short time scales the direct effects of vaccination on incidence rates can be estimated in vaccinated populations, when compared to comparable unvaccinated populations.[71] Here, I aimed to estimate the impact of the 2005 change in BCG policy on incidence rates, in both the UK and non-UK born populations, directly affected by it.

7.3 Methods

7.3.1 Data source

Data on all notifications from the Enhanced Tuberculosis Surveillance (ETS) system from Jan 1, 2000 to Dec 31, 2015 were obtained from Public Health England (PHE). The ETS is maintained by PHE, and contains demographic, clinical, and microbiological data on all notified cases in England (see Chapter 4). A descriptive analysis of TB epidemiology in England is published each year, which fully details data collection and cleaning.[15]

I obtained yearly population estimates from the April to June Labour Force Survey (LFS) for 2000-2015. The LFS is a study of the employment circumstances of the UK population, and provides the official measures of employment and unemployment in the UK (see Chapter 4). Reporting practices have changed with time so the appropriate variables for age, country of origin, country of birth, and survey weight were extracted from each yearly extract, standardised, and combined into a single data-set.

7.3.2 Constructing Retrospective cohorts

I constructed retrospective cohorts of TB cases and individuals using the ETS and the LFS. TB cases were extracted from the ETS based on date of birth and date of TB notification.

Cohort 1: individuals aged 14 between 2000 and 2004, who were notified with TB whilst aged between 14 and 19 years old.

Comparison cohort 1: individuals aged 14 between 2005 and 2010, who were notified with TB whilst aged between 14 and 19 years old.

Cohort 2: individuals born between 2005 and 2010, who were notified with TB whilst aged 0-5 years old

Comparison cohort 2: individuals born between 2000 and 2004, who were notified with TB whilst aged 0-5 years old

Each cohort was stratified by UK birth status, with both non-UK born and UK born cases assumed to have been exposed to England's vaccination policy. Corresponding population cohorts were calculated using the LFS population estimates, resulting in 8 population level cohorts, each with 5 years of follow up (Table 7.1).

Table 7.1: Summary of relevance and eligibility criteria for each cohort.

Cohort	Vaccination programme	Eligible for the programme*	Birth status	Age at study entry	Year of study entry
Cohort 1	Universal	Yes	UK born	14	2000-2004
Comparison cohort 1	Universal	No	UK born	14	2005-2010
Cohort 1	Universal	Yes	Non-UK born	14	2000-2004
Comparison cohort 1	Universal	No	Non-UK born	14	2005-2010
Comparison cohort 2	Targeted	No	UK born	Birth	2000-2004
Cohort 2	Targeted	Yes	UK born	Birth	2005-2010
Comparison cohort 2	Targeted	No	Non-UK born	Birth	2000-2004
Cohort 2	Targeted	Yes	Non-UK born	Birth	2005-2010

* Eligible signifies that the cohort fit the criteria for the programme and entered the study during the time period it was in operation not that the cohort was vaccinated by the programme.

7.4 Statistical methods overview

I estimated incidence rates (with 95% confidence intervals) by year, age and place of birth as (number of cases) divided by (number of individuals of corresponding age) (see Chapter 4). UK birth status was incomplete, with some evidence of a missing not at random mechanism (MNAR). I imputed the missing data using a gradient boosting method (see Section 7.4.2). I then used descriptive analysis to describe the observed trends in age-specific incidence rates over the study period, comparing incidence rates in the study populations relevant to both vaccination programmes before and after the change in BCG policy.

I calculated Incidence Rate Ratios (IRRs) for the change in incidence rates associated with

the change in BCG vaccination policy (modelled as a binary breakpoint at the start of 2005) for both the UK born and non-UK born populations that were relevant to the universal programme, and for the targeted programme using a range of models. I considered the following covariates: age,[15,21] incidence rates in both the UK born and non-UK born who were not in the age group of interest,[15] and year of study entry (as a random intercept). I first investigated a univariable Poisson model, followed by combinations of covariates (Table 7.2). I also investigated a Negative Binomial model adjusting for the same covariates as in the best fitting Poisson model. The models were estimated with a Bayesian approach using Markov Chain Monte Carlo (MCMC), with default weakly informative priors (see Section 7.4.3). Model fit, penalised by model complexity, was assessed using the leave one out cross validation information criterion (LOOIC) and its standard error.[72] Models were ranked by goodness of fit, using their LOOIC, with a smaller LOOIC indicating a better fit to the data after adjusting for the complexity of the model. No formal threshold for a change in the LOOIC was used, with changes in the LOOIC being evaluated in the context of their standard error. The inclusion of the change in policy in the best fitting model was tested by refitting the model excluding the change in policy and estimating the improvment in the LOOIC. Once the best fitting model had been identified I estimated the number of cases prevented, from 2005 until 2015, for each vaccination programme in the study population relevant to that programme (see Section 7.4.4).

7.4.1 Implementation overview

R `3.5.0` was used for all analysis.[49] Reproducibility was ensured by using R package infrastructure (`https://github.com/seabbs/DirectEffBCGPolicyChange`). Missing data imputation using a GBM was implemented using the `h2o` package (see Section 7.4.2).[73] Incidence rates, with 95% confidence intervals, were calculated using the `epiR` package (see Chapter 4)[1].[74] The `brms` package,[75] and STAN,[76] was used to perform MCMC. Models were run until convergence (4 chains with a burn in of 10,000, and 10,000 sampled iterations each), with convergence being assessed using trace plots and the R hat diagnostic.[76] All numeric confounders were centered and scaled by their standard deviation, and age was adjusted for using single year of age categories.

7.4.2 Imputation of UK birth status

As I was imputing a single variable I reformulated the imputation as a categorical prediction problem. This allowed the use of techniques from machine learning to improve the quality of our imputation, whilst also validating it using metrics supported by theory. I included year of notification, sex, age, Public Health England Centre (PHEC), occupation, ethnic group, Index of Multiple Deprivation (2010) categorised into five groups for England (IMD rank), and risk factor count (risk factors considered; drug use, homelessness, alcohol misuse/abuse and prison). However, I could not account for a possible missing not at random mechanism not captured by these covariates. To train the model I first split the data with complete UK birth status into a training set (80%), a calibration set (5%), and a test set (15%). I then fit a gradient boosted machine with 10,000 trees, early stopping (at a precision of $1e-5$, with 10 stopping rounds), a learning rate of 0.1, and a learn rate annealing of 0.99. Gradient boosted machines are a tree based method that can incorporate complex non-linear relationships and interactions. Much like a random forest model they work by

[1] An R package is available containing the required code DOI: 10.5281/zenodo.2551554

Table 7.2: Complete definition of each model, ordered by increasing complexity.

Model	Description
Model 1	Poisson model adjusting for no fixed effects.
Model 2	Poisson model adjusting with fixed effects for the change in policy.
Model 3	Poisson model adjusting with fixed effects for the change in policy and incidence rates in the UK born.
Model 4	Poisson model adjusting with fixed effects for the change in policy and incidence rates in the non-UK born.
Model 5	Poisson model adjusting with fixed effects for the change in policy and incidence rates in the UK born and non-UK born populations.
Model 6	Poisson model adjusting with fixed effects for the change in policy and age.
Model 7	Poisson model adjusting with fixed effects for the change in policy, age, and incidence rates in the UK born.
Model 7 (Negative Binomial)	Negative binomial model adjusting with fixed effects for the change in policy, age, and incidence rates in the UK born.
Model 8	Poisson model adjusting with fixed effects for the change in policy, age, and incidence rates in the non-UK born.
Model 8 (Negative Binomial)	Negative binomial model adjusting with fixed effects for the change in policy, age, and incidence rates in the non-UK born.
Model 9	Poisson model adjusting with fixed effects for the change in policy, age, and incidence rates in the UK born and non-UK born populations.
Model 10	Poisson model with a random intercept for year of study entry, adjusting for no fixed effects.
Model 11	Poisson model with a random intercept for year of study entry, adjusting with fixed effects for the change in policy.
Model 12	Poisson model with a random intercept for year of study entry, adjusting with fixed effects for the change in policy and incidence rates in the UK born.
Model 13	Poisson model with a random intercept for year of study entry, adjusting with fixed effects for the change in policy and incidence rates in the non-UK born.
Model 14	Poisson model with a random intercept for year of study entry, adjusting with fixed effects for the change in policy and incidence rates in the UK born and non-UK born populations.
Model 15	Poisson model with a random intercept for year of study entry, adjusting with fixed effects for the change in policy and age.
Model 16	Poisson model with a random intercept for year of study entry, adjusting with fixed effects for the change in policy, age, and incidence rates in the UK born.
Model 16 (Negative Binomial)	Negative binomial model with a random intercept for year of study entry, adjusting with fixed effects for the change in policy, age, and incidence rates in the UK born.
Model 17	Poisson model with a random intercept for year of study entry, adjusting with fixed effects for the change in policy, age, and incidence rates in the non-UK born.
Model 17 (Negative Binomial)	Negative binomial model with a random intercept for year of study entry, adjusting with fixed effects for the change in policy, age, and incidence rates in the non-UK born.
Model 18	Poisson model with a random intercept for year of study entry, adjusting with fixed effects for the change in policy, age, and incidence rates in the UK born and non-UK born populations.

ensembling a group of trees, but unlike a random forest model each tree is additive aiming to reduce the residual loss from previous trees. Once the model had been fit to the training set I performed platt scaling using the calibration data set. Our fitted imputation model

had a Logloss of 0.28 on the test set, with an AUC of 0.93, both of which indicate a robust out of bag performance. I found that ethnic group was the most important variable for predicting UK birth status, followed by age and PHEC.

Using the fitted model I predicted the birth status for notifications where this was missing, using the F1 optimal threshold as the probability cut-off. It is common to impute missing values multiple times, to account for within- and between imputation variability. However, I considered this unnecessary for this analysis as the amount of missing data was small, this analysis considered only aggregate counts, my model metrics indicated a robust level of performance out of bag and any unaccounted for uncertainty would be outweighed by the uncertainty in the population denominator.[70] I found that cases with imputed birth status had a similar proportion of UK born to non-UK born cases as in the complete data (Table 7.3). Inclusion of imputed values for UK birth status should reduce bias caused

Table 7.3: Comparison of UK birth status in cases with complete or imputed records.

Status	Birth Status	Proportion of Cases (%)	Cases
Complete			106765
	UK Born	27.3	29096
	Non-UK Born	72.7	77669
Imputed			8055
	UK Born	32.7	2634
	Non-UK Born	67.3	5421

by any missing not at random mechanism captured by predictors included in the model. Graphical evaluation of UK birth status indicated that missingness has reduced over time, indicating a missing not at random mechanism (see Chapter 4). If only the complete case data had been included in the analysis then incidence rates would have reduced over the study period due to this mechanism, this may have biased the estimate of the impact of the change in policy.

7.4.3 Prior choice

Default weakly informative priors were used based on those provided by the `brms` package.[75] For the population-level effects this was an improper flat prior over the reals. For both the standard deviations of group level effects and the group level intercepts this was a half student-t prior with 3 degrees of freedom and a scale parameter that depended on the standard deviation of the response after applying the link function.

7.4.4 Estimating the magnitude of the estimated impact of the change in BCG policy

I estimated the magnitude of the estimated impact from the change in BCG policy by applying the IRR estimates from the best fitting model for each cohort to the observed number of notifications from 2005 until 2015 in the study population. For the cohorts relevant to the universal school-age vaccination scheme I estimated the number of prevented cases by first aggregating cases (C_O) and then using the following equation,

$$C_P^i = C_O * (1 - I^i), \text{ Where } i = e,\ l,\ u.$$

Where C_P^i is the predicted number of cases prevented using the mean (e), 2.5% bound (l) and 97.5% bound (u) of the IRR estimate I^i. For the cohorts relevant to the targeted high-risk neonatal scheme I used a related equation, adjusting for the fact that the populations were exposed to the scheme and I therefore had to first estimate the number of cases that would have been observed had the scheme not been implemented. After simplification this results in the following equation,

$$C_P^i = C_0 \left(\frac{1}{I^i} - 1 \right), \text{ Where } i = e,\ l,\ u.$$

7.5 Results

7.5.1 Descriptive analysis

During the study period there were 114,820 notifications of TB in England, of which 93% (106765/114820) had their birth status recorded. Of notifications with a known birth status 27% (29096/106765) were UK born, in comparison to 33% (2634/8055) in cases with an imputed birth status (see Chapter 4 for details). There were 1729 UK born cases and 2797 non-UK born cases in individuals relevant to the universal schools scheme, and 1431 UK born cases and 238 non-UK born cases relevant to the targeted neonatal scheme, who fit the age criteria during the study period. Univariable evidence for differences between mean incidence rates before and after the change in BCG policy in the UK born was weak. In the non-UK born incidence rates were lower after the change in BCG policy in both the cohort relevant to the universal school-age scheme and the cohort relevant to the targeted neonatal scheme (Figure 7.1).

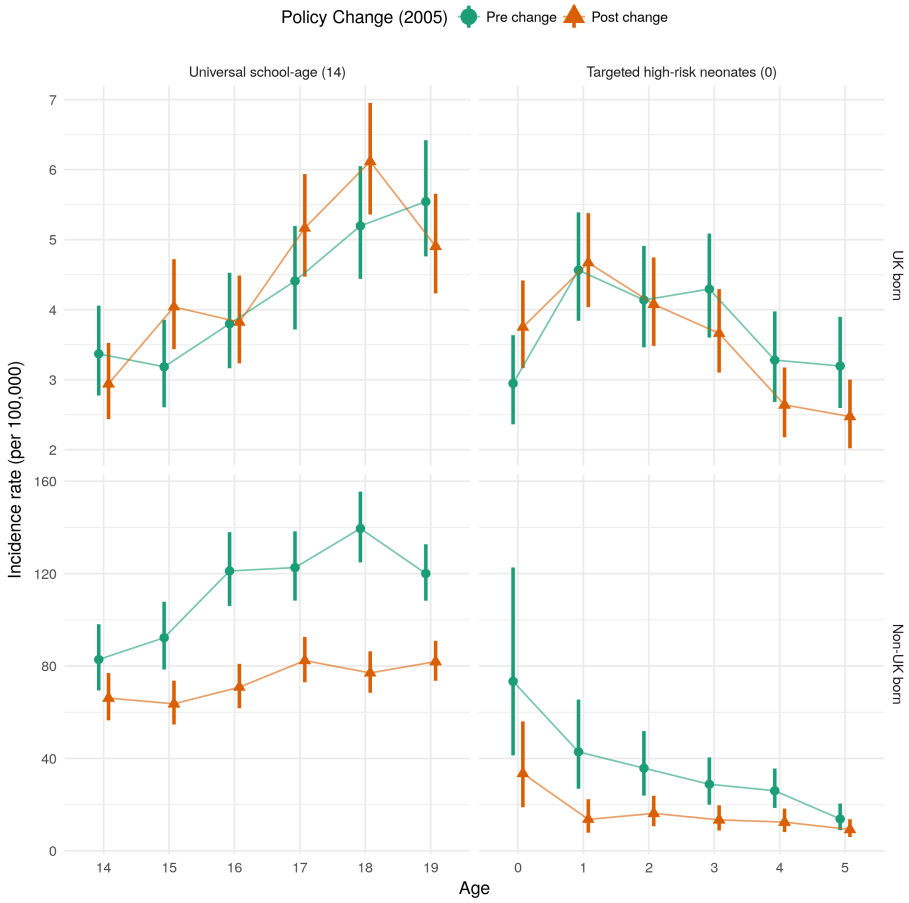


Figure 7.1: Mean incidence rates per 100,000, with 95% confidence intervals for each retrospective cohort (see Table 7.1 for cohort definitions), stratified by the vaccination policy and UK birth status. The top and bottom panels are on different scales in order to highlight trends in incidence rates over time.

Trends in incidence rates varied by age group and UK birth status. From 2000 until 2012 incidence rates in the UK born remained relatively stable but have since fallen year on year. In comparison incidence rates in the non-UK born increased from 2000 until 2005, since when they have also decreased year on year (see Chapter 4). In 14-19 year old's, who were UK born, incidence rates remained relatively stable throughout the study period, except for the period between 2006 to 2009 in which they increased year on year. This trend was not observed in the non-UK born population aged 14-19, where incidence rates reached a peak in 2003, since when they have consistently declined. In those aged 0-5, who were UK born, incidence rates also increased year on year after the change in BCG policy, until 2008 since when they have declined. This does not match with the observed trend in incidence rates in the non-UK born population, aged 0-5, in which incidence rates declined steeply between 2005 and 2006, since when they have remained relatively stable (Figure 7.2).

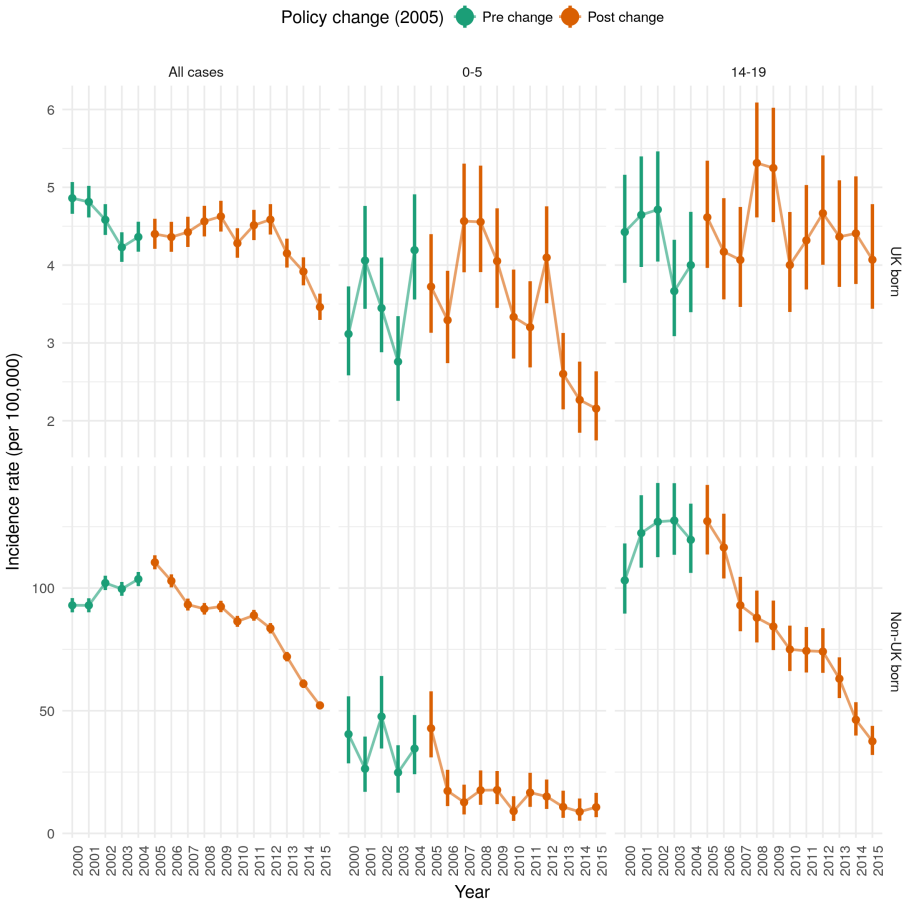


Figure 7.2: Incidence rates per 100,000 for UK born population and non-UK born population, aged 0-5 and therefore directly affected by the targeted neonatal vaccination programme, and aged 14-19 and therefore directly affected by the universal school-age scheme.

7.5.2 Adjusted estimates of the effects of the change in policy on school-age children

In the UK born cohort relevant to universal vaccination there was some evidence, across all models that adjusted for age, that ending the scheme was associated with a modest increase in TB rates (Table 7.4). Using the LOOIC goodness of fit criteria the best fitting model was found to be a Negative Binomial model that adjusted for the change in policy, age, and incidence rates in the UK born (Table 7.5). In this model there was some evidence of an assocation between the change in policy and an increase in incidence rates in those at school-age who were UK born, with an IRR of 1.08 (95%CI 0.97, 1.19). Dropping the change in policy from the model resulted in a small decrease in the LOOIC (0.52 (SE 2.63)) but the change was too small, with too large a standard error, to conclusively state that the excluding the change in policy from the model improved the quality of model fit. I found that it was important to adjust for UK born incidence rates, otherwise the impact from the change in BCG vaccination policy was over-estimated.

For the comparable non-UK born cohort who were relevant to the universal vaccination

there was evidence, in the best fitting model, that ending the scheme was associated with a decrease in incidence rates (IRR: 0.74 (95%CI 0.61, 0.88)). The best fitting model was a Negative Binomial model which adjusted for the change in policy, age, incidence rates in the non-UK born, and year of eligibility as a random effect (Table 7.5). I found that omitting the change in policy from the model resulted in poorer model fit (LOOIC increase of 3.02 (SE 3.52)), suggesting that the policy change was an important factor explaining changes in incidence rates, after adjusting for other covariates. All models that adjusted for incidence rates in the UK born or non-UK born estimated similar IRRs (Table 7.6).

Table 7.4: Comparison of models fitted to incidence rates for the UK born population that were relevant to the universal vaccination programme of those at school-age (14). Models are ordered by the goodness of fit as assessed by LOOIC, the degrees of freedom are used as a tiebreaker.

		Variable							
Model	IRR (CI 95%)*	Policy Change	Age	UK born rates	Non-UK born rates	Year of study entry	DoF**	LPD***	LOOIC (se)****
Model 7 (Negative Binomial)	1.08 (0.97, 1.19)	Yes	Yes	Yes	No	No	9	-211	439 (10)
Model 7	1.08 (1.00, 1.17)	Yes	Yes	Yes	No	No	8	-211	443 (14)
Model 9	1.12 (1.01, 1.25)	Yes	Yes	Yes	Yes	No	9	-210	445 (14)
Model 16	1.08 (0.97, 1.21)	Yes	Yes	Yes	No	Yes	20	-207	445 (14)
Model 18	1.12 (0.97, 1.28)	Yes	Yes	Yes	Yes	Yes	21	-207	447 (15)
Model 8	1.16 (1.04, 1.29)	Yes	Yes	No	Yes	No	8	-213	449 (17)
Model 6	1.06 (0.98, 1.15)	Yes	Yes	No	No	No	7	-215	452 (17)
Model 17	1.15 (1.00, 1.32)	Yes	Yes	No	Yes	Yes	20	-209	452 (17)
Model 15	1.06 (0.94, 1.20)	Yes	Yes	No	No	Yes	19	-209	453 (17)
Model 1	0.00 (0.00, 0.00)	No	No	No	No	No	1	-254	513 (26)
Model 2	1.06 (0.98, 1.14)	Yes	No	No	No	No	2	-252	515 (25)
Model 4	1.00 (0.90, 1.10)	Yes	No	No	Yes	No	3	-251	516 (25)
Model 3	1.06 (0.98, 1.15)	Yes	No	Yes	No	No	3	-252	518 (26)
Model 5	0.98 (0.89, 1.09)	Yes	No	Yes	Yes	No	4	-249	518 (24)
Model 13	0.94 (0.78, 1.12)	Yes	No	No	Yes	Yes	15	-237	518 (27)
Model 10	0.00 (0.00, 0.00)	No	No	No	No	Yes	13	-244	521 (28)
Model 11	1.06 (0.94, 1.20)	Yes	No	No	No	Yes	14	-244	522 (28)
Model 14	0.93 (0.78, 1.11)	Yes	No	Yes	Yes	Yes	16	-236	522 (27)
Model 12	1.06 (0.93, 1.20)	Yes	No	Yes	No	Yes	15	-243	526 (28)

* Incidence Rate Ratio, with 95% credible intervals,
** Degrees of Freedom,
*** Computed log pointwise predictive density,
**** Leave one out information criterion, with standard error,

Table 7.5: Summary table of incidence rate ratios, in the UK born and non-UK born cohorts relevant to the targeted neonatal scheme, using the best fitting models as determined by comparison of the LOOIC (UK born: Negative binomial model adjusting with fixed effects for the change in policy, age, and incidence rates in the UK born (Model 7 (Negative Binomial)), Non-UK born: Negative binomial model with a random intercept for year of study entry, adjusting with fixed effects for the change in policy, age, and incidence rates in the non-UK born (Model 17 (Negative Binomial))). Model terms which were not included in a given cohort are indicated using a hyphen (-).

	IRR (95% CrI)*	
Variable	UK born	Non-UK born
Policy change**		
Pre-change	*Reference*	*Reference*
Post-change	1.08 (0.97, 1.19)	0.74 (0.61, 0.88)
Age		
14	*Reference*	*Reference*
15	1.18 (0.98, 1.42)	1.03 (0.87, 1.22)
16	1.24 (1.03, 1.50)	1.25 (1.07, 1.47)
17	1.59 (1.33, 1.91)	1.40 (1.19, 1.63)
18	1.92 (1.60, 2.30)	1.47 (1.26, 1.73)
19	1.80 (1.49, 2.17)	1.47 (1.24, 1.73)
UK born incidence rate (per standard deviation)	1.08 (1.03, 1.14)	-
Non-UK born incidence rate (per standard deviation)	-	1.11 (1.03, 1.19)
Year of study elibility, group level	-	
Intercept (standard deviation)	-	1.13 (1.05, 1.26)
Year of study elibility, individual level	-	
2000	-	1.10 (0.96, 1.29)
2001	-	1.06 (0.93, 1.24)
2002	-	1.07 (0.94, 1.25)
2003	-	0.90 (0.76, 1.03)
2004	-	0.89 (0.75, 1.02)
2005	-	0.98 (0.85, 1.12)
2006	-	1.13 (0.99, 1.33)
2007	-	1.04 (0.91, 1.20)
2008	-	0.96 (0.83, 1.09)
2009	-	0.95 (0.81, 1.08)
2010	-	0.96 (0.82, 1.11)

* Incidence Rate Ratio (95% Credible Interval),

**There was an improvement in the LOOIC score of 0.52 (SE 2.63) from dropping the change in policy from the model in the UK born cohort and a -3.02 (SE 3.52) improvement in the non-UK born cohort.

Table 7.6: Comparison of models fitted to incidence rates for the non-UK born population that were eligible to the universal vaccination programme of those at school-age (14). Models are ordered by the goodness of fit as assessed by LOOIC, the degrees of freedom are used as a tiebreaker.

		Variable							
Model	IRR (CI 95%)*	Policy Change	Age	UK born rates	Non-UK born rates	Year of study entry	DoF**	LPD***	LOOIC (se)****
Model 17 (Negative Binomial)	0.74 (0.61, 0.88)	Yes	Yes	No	Yes	Yes	21	-228	483 (10)
Model 17	0.74 (0.62, 0.87)	Yes	Yes	No	Yes	Yes	20	-223	492 (16)
Model 18	0.73 (0.61, 0.87)	Yes	Yes	Yes	Yes	Yes	21	-222	493 (16)
Model 15	0.64 (0.53, 0.78)	Yes	Yes	No	No	Yes	19	-224	496 (18)
Model 16	0.65 (0.54, 0.78)	Yes	Yes	Yes	No	Yes	20	-223	496 (17)
Model 8	0.79 (0.73, 0.86)	Yes	Yes	No	Yes	No	8	-239	507 (20)
Model 9	0.79 (0.72, 0.86)	Yes	Yes	Yes	Yes	No	9	-238	511 (20)
Model 11	0.64 (0.52, 0.78)	Yes	No	No	No	Yes	14	-241	522 (22)
Model 10	0.00 (0.00, 0.00)	No	No	No	No	Yes	13	-241	523 (22)
Model 12	0.64 (0.53, 0.79)	Yes	No	Yes	No	Yes	15	-241	525 (22)
Model 13	0.64 (0.52, 0.79)	Yes	No	No	Yes	Yes	15	-241	526 (23)
Model 14	0.64 (0.52, 0.79)	Yes	No	Yes	Yes	Yes	16	-241	530 (23)
Model 7	0.66 (0.62, 0.70)	Yes	Yes	Yes	No	No	8	-248	532 (23)
Model 6	0.65 (0.61, 0.69)	Yes	Yes	No	No	No	7	-253	539 (27)
Model 4	0.70 (0.65, 0.76)	Yes	No	No	Yes	No	3	-270	556 (31)
Model 5	0.70 (0.64, 0.76)	Yes	No	Yes	Yes	No	4	-270	559 (31)
Model 2	0.65 (0.61, 0.69)	Yes	No	No	No	No	2	-275	561 (33)
Model 3	0.65 (0.61, 0.69)	Yes	No	Yes	No	No	3	-273	561 (32)
Model 1	0.00 (0.00, 0.00)	No	No	No	No	No	1	-341	692 (51)

* Incidence Rate Ratio, with 95% credible intervals,
** Degrees of Freedom,
*** Computed log pointwise predictive density,
**** Leave one out information criterion, with standard error,

7.5.3 Adjusted estimates of the effect of the change in policy in those relevant to the targeted neonatal programme

For the UK born cohort relevant to the targeted neonatal vaccination programme the evidence of an association, across all models, was mixed and credible intervals were wide compared to models for the UK born cohort relevant to the universal school-age vaccination programme (Table 7.7). The best fitting model was a Poisson model which adjusted for the change in policy, age, UK born incidence rates, and year of study entry with a random effect (Table 7.8). In this model, there was weak evidence of an association between the change in BCG policy and an decrease in incidence rates in UK born neonates, with an IRR of 0.96 (95%CI 0.82, 1.14). There was weak evidence to suggest that dropping the change in policy from this model improved the quality of the fit, with an improvement in the LOOIC score of 0.92 (SE 1.07). This suggests that the change in policy was not an important factor for explaining incidence rates, after adjusting for covariates. Models which also adjusted for non-UK born incidence rates estimated that the change in policy was associated with no change in incidence rates in the relevant cohort of neonates.

For the comparable non-UK born cohort who were relevant to the targeted neonatal vaccination programme there was evidence, across all models, that change in policy was associated with a large decrease in incidence rates (IRR: 0.62 (95%CI 0.44, 0.88)) (Table 7.8 in the best fitting model). The best fitting model was a Negative Binomial model that adjusted for the change in policy, age, and non-UK born incidence rates (Table 7.8). All models which at least adjusted for age estimated comparable effects of the change in policy (Table 7.9).

Table 7.7: Comparison of models fitted to incidence rates for the UK born population that were elgible to the targeted vaccination programme of neonates. Models are ordered by the goodness of fit as assessed by LOOIC, the degrees of freedom are used as a tiebreaker.

		Variable							
Model	IRR (CI 95%)*	Policy Change	Age	UK born rates	Non-UK born rates	Year of study entry	DoF**	LPD***	LOOIC (se)****
Model 16	0.96 (0.82, 1.14)	Yes	Yes	Yes	No	Yes	20	-192	415 (12)
Model 16 (Negative Binomial)	0.96 (0.82, 1.13)	Yes	Yes	Yes	No	Yes	21	-196	415 (10)
Model 18	0.99 (0.82, 1.18)	Yes	Yes	Yes	Yes	Yes	21	-192	417 (13)
Model 7	0.96 (0.88, 1.05)	Yes	Yes	Yes	No	No	8	-200	420 (15)
Model 9	1.00 (0.89, 1.12)	Yes	Yes	Yes	Yes	No	9	-200	422 (15)
Model 8	1.02 (0.91, 1.15)	Yes	Yes	No	Yes	No	8	-203	427 (16)
Model 6	0.95 (0.87, 1.03)	Yes	Yes	No	No	No	7	-204	428 (16)
Model 15	0.95 (0.83, 1.09)	Yes	Yes	No	No	Yes	19	-198	428 (14)
Model 17	1.02 (0.87, 1.20)	Yes	Yes	No	Yes	Yes	20	-198	429 (14)
Model 14	1.10 (0.92, 1.33)	Yes	No	Yes	Yes	Yes	16	-206	442 (16)
Model 5	1.08 (0.97, 1.21)	Yes	No	Yes	Yes	No	4	-216	445 (18)
Model 12	0.98 (0.83, 1.15)	Yes	No	Yes	No	Yes	15	-209	448 (17)
Model 4	1.12 (1.00, 1.24)	Yes	No	No	Yes	No	3	-219	449 (18)
Model 3	0.97 (0.89, 1.06)	Yes	No	Yes	No	No	3	-219	450 (19)
Model 13	1.14 (0.97, 1.35)	Yes	No	No	Yes	Yes	15	-211	452 (16)
Model 1	0.00 (0.00, 0.00)	No	No	No	No	No	1	-229	462 (21)
Model 2	0.95 (0.87, 1.03)	Yes	No	No	No	No	2	-228	463 (20)
Model 10	0.00 (0.00, 0.00)	No	No	No	No	Yes	13	-220	466 (19)
Model 11	0.95 (0.83, 1.09)	Yes	No	No	No	Yes	14	-219	467 (19)

* Incidence Rate Ratio, with 95% credible intervals,
** Degrees of Freedom,
*** Computed log pointwise predictive density,
**** Leave one out information criterion, with standard error,

Table 7.8: Summary table of incidence rate ratios, in the UK born and non-UK born cohorts relevant to the targeted neonatal scheme, using the best fitting models as determined by comparison of the LOOIC (UK born: Poisson model with a random intercept for year of study entry, adjusting with fixed effects for the change in policy, age, and incidence rates in the UK born (Model 16), Non-UK born: Negative binomial model adjusting with fixed effects for the change in policy, age, and incidence rates in the non-UK born (Model 8 (Negative Binomial))). Model terms which were not included in a given cohort are indicated using a hyphen (-).

	IRR (95% CrI)*	
Variable	UK born	Non-UK born
Policy change**		
Pre-change	*Reference*	*Reference*
Post-change	0.96 (0.82, 1.14)	0.62 (0.44, 0.88)
Age		
0	*Reference*	*Reference*
1	1.39 (1.20, 1.61)	0.49 (0.30, 0.83)
2	1.24 (1.06, 1.44)	0.49 (0.30, 0.80)
3	1.21 (1.03, 1.41)	0.42 (0.26, 0.68)
4	0.90 (0.76, 1.06)	0.41 (0.25, 0.66)
5	0.89 (0.75, 1.06)	0.27 (0.16, 0.45)
UK born incidence rate (per standard deviation)	1.12 (1.06, 1.18)	-
Non-UK born incidence rate (per standard deviation)	-	1.25 (1.04, 1.51)
Year of study elibility, group level		-
Intercept (standard deviation)	1.13 (1.04, 1.26)	-
Year of study elibility, individual level		-
2000	0.83 (0.68, 0.99)	-
2001	0.93 (0.79, 1.07)	-
2002	1.08 (0.95, 1.28)	-
2003	1.07 (0.93, 1.26)	-
2004	1.12 (0.97, 1.32)	-
2005	1.02 (0.89, 1.17)	-
2006	1.02 (0.89, 1.17)	-
2007	0.97 (0.83, 1.11)	-
2008	1.01 (0.88, 1.15)	-
2009	1.01 (0.88, 1.16)	-
2010	0.98 (0.85, 1.13)	-

* Incidence Rate Ratio (95% Credible Interval),

**There was an improvement in the LOOIC score of 0.92 (SE 1.07) from dropping the change in policy from the model in the UK born cohort and a -3.45 (SE 4.63) improvement in the non-UK born cohort.

Table 7.9: Comparison of models fitted to incidence rates for the non-UK born population that were revelant to the targeted vaccination programme of neonates. Models are ordered by the goodness of fit as assessed by LOOIC, the degrees of freedom are used as a tiebreaker.

		Variable							
Model	IRR (CI 95%)*	Policy Change	Age	UK born rates	Non-UK born rates	Year of study entry	DoF**	LPD***	LOOIC (se)****
Model 8 (Negative Binomial)	0.62 (0.44, 0.88)	Yes	Yes	No	Yes	No	9	-138	293 (15)
Model 8	0.64 (0.47, 0.86)	Yes	Yes	No	Yes	No	8	-137	295 (18)
Model 9	0.62 (0.45, 0.85)	Yes	Yes	Yes	Yes	No	9	-137	297 (18)
Model 6	0.47 (0.38, 0.58)	Yes	Yes	No	No	No	7	-139	298 (19)
Model 7	0.48 (0.39, 0.60)	Yes	Yes	Yes	No	No	8	-139	298 (19)
Model 17	0.63 (0.44, 0.89)	Yes	Yes	No	Yes	Yes	20	-135	298 (18)
Model 18	0.61 (0.42, 0.87)	Yes	Yes	Yes	Yes	Yes	21	-135	300 (18)
Model 15	0.47 (0.35, 0.62)	Yes	Yes	No	No	Yes	19	-136	301 (20)
Model 16	0.48 (0.36, 0.63)	Yes	Yes	Yes	No	Yes	20	-136	301 (19)
Model 4	0.82 (0.61, 1.10)	Yes	No	No	Yes	No	3	-147	304 (17)
Model 5	0.78 (0.58, 1.06)	Yes	No	Yes	Yes	No	4	-147	306 (18)
Model 13	0.83 (0.59, 1.16)	Yes	No	No	Yes	Yes	15	-145	308 (18)
Model 14	0.78 (0.55, 1.12)	Yes	No	Yes	Yes	Yes	16	-144	310 (19)
Model 3	0.52 (0.42, 0.64)	Yes	No	Yes	No	No	3	-152	314 (22)
Model 12	0.51 (0.38, 0.69)	Yes	No	Yes	No	Yes	15	-148	317 (23)
Model 2	0.49 (0.40, 0.61)	Yes	No	No	No	No	2	-156	319 (22)
Model 11	0.49 (0.37, 0.65)	Yes	No	No	No	Yes	14	-152	322 (23)
Model 10	0.00 (0.00, 0.00)	No	No	No	No	Yes	13	-150	330 (25)
Model 1	0.00 (0.00, 0.00)	No	No	No	No	No	1	-171	346 (27)

* Incidence Rate Ratio, with 95% credible intervals,
** Degrees of Freedom,
*** Computed log pointwise predictive density,
**** Leave one out information criterion, with standard error,

7.5.4 Magnitude of the estimated impact of the change in BCG policy

I estimate that the change in vaccination policy was associated with preventing 385 (95%CI -105, 881) cases from 2005 until the end of the study period in the directly impacted populations after 5 years of follow up (Table 7.10). The majority of the cases prevented were in the non-UK born, with cases increasing slightly overall in the UK born. This was due to cases increasing in the UK born at school-age, and decreasing in UK born neonates, although both these estimates had large credible intervals.

Table 7.10: Estimated number of cases prevented, from 2005 until 2015, for each vaccination programme in the study population relevant to that programme, using the best fitting model for each cohort.

Vaccination Programme	Birth Status	Cases Prevented (95% CI*)	Notified Cases
Universal school-age (14)		-291 (24, -571)	2364
	UK born	76 (188, -26)	969
	Non-UK born	-367 (-165, -546)	1395
Targeted high-risk neonates (0)		94 (-81, 310)	906
	UK born	30 (-95, 173)	800
	Non-UK born	65 (14, 137)	106
Change in Policy**		385 (-105, 881)	3270
	UK born	-46 (-284, 199)	1769
	Non-UK born	431 (179, 682)	1501

*95% CI: 95% Credible Interval,
** Estimated total number of cases prevented due to the change in vaccination policy in 2005

7.6 Discussion

In the non-UK born I found evidence of an association between the change in BCG policy and a decrease in TB incidence rates in both those at school-age and neonates, after 5 years of follow up. I found some evidence that the change in BCG policy was associated with a modest increase in incidence rates in the UK born population who were relevant to the universal school-age scheme and weaker evidence of a small decrease in incidence rates in the UK born population relevant to the targeted neonatal scheme. Overall, I found that the change in policy was associated with preventing 385 (95%CI -105, 881) cases in the study population, from 2005 until the end of the study period, with the majority of the cases prevented in the non-UK born.

I was unable to estimate the impact of the change in BCG policy after 5 years post vaccination, so both the estimates of the positive and negative consequences are likely to be underestimates of the ongoing impact. Tuberculosis is a complex disease and the BCG vaccine is known to offer imperfect protection, which has been shown to vary both spatially and with time since vaccination (see Chapter 2).[19,22] By focusing on the impact of the change in policy on the directly affected populations within a short period of time, and by employing a multi-model approach I have limited the potential impact of these issues.

This study was based on a routine observational data set (ETS), and a repeated survey (LFS) both of which may have introduced bias. Whilst the LFS is a robust data source, widely used in academic studies,[39,77,78] it is susceptible to sampling errors particularly in the young, and in the old, which may have biased the estimated incidence rates. As the ETS is routine surveillance system some level of missing data is inevitable (see Chapter 4). However, UK birth status is relatively complete (93% (106765/114820)) and I imputed missing values using an approach which accounted for MNAR mechanisms for the variables included in the imputation model. I was unable to adjust for known demographic risk factors for TB, notably socio-economic status,[11,56] and ethnicity.[11,56,61] However, this confounding is likely to be mitigated by the use of multiple cohorts and the adjustment for incidence rates in the UK born and non-UK born. Finally, I have assumed that the effect I have estimated for the change in BCG policy is due to the changes in BCG vaccination policy as well as other associated changes in TB control policy, after adjusting for hypothesised confounders. However, there may have been additional policy changes which I have not accounted for.

Whilst little work has been done to assess the impact of the 2005 change in BCG vaccination several other studies have estimated the impact of changing BCG vaccination policy, although typically only from universal vaccination of neonates to targeted vaccination of high-risk neonates. A previous study in Sweden found that incidence rates in Swedish-born children increased after high-risk neonatal vaccination was implemented in place of a universal neonatal program, this corresponds with our finding that introducing neonatal vaccination had little impact on incidence rates in UK born neonates. Theoretical approaches have indicated that targeted vaccination of those at high-risk may be optimal in low incidence settings.[79] Our study extends this work by also considering the age of those given BCG vaccination, although I was unable to estimate the impact of a universal neonatal scheme as this has never been implemented nationally in England. It has previously been shown that targeted vaccination programmes may not reach those considered most at risk,[80] our findings may support this view as I observed only a small decrease in incidence rates in UK born neonates after the introduction of the targeted neonatal vaccination programme. Alternatively, the effectiveness of the BCG in neonates, in England, may be lower than previously thought as I only observed a small decrease in incidence rates, whilst a previous study estimated BCG coverage at 68% (95%CI 65%, 71%) amongst those eligible for the targeted neonatal vaccination programme.[81] Chapter 5 also found evidence that incidence rates would increase in UK born population relevant to school-age BCG programme.

This study indicates that the change in England's BCG vaccination policy was associated with a modest increase in incidence in the UK born that were relevant to the school-age vaccination programme, and with a small reduction in incidence in the UK born that were relevant to the high-risk neonatal vaccination programme, although both these estimates had wide credible intervals. I found stronger evidence of an association between the change in policy and a decrease in incidence rates in the non-UK born populations relevant to both programmes. This suggests that the change of vaccination policy to target high-risk neonates may have resulted in an increased focus on high-risk non-UK born individuals who may not have been the direct targets of the vaccination programme. Further validation is required using alternative study designs, but this result should be considered when vaccination policy changes are being considered.

It is well established that interventions against infectious diseases, such as TB, should

be evaluated not only for their direct effects but also for future indirect effects via ongoing transmission. Statistical approaches such as those used in this paper are not appropriate for capturing these future indirect effects, and instead dynamic disease models should be used. In Chapter 8 I develop such a dynamic disease model, Chapter 9 then fits this model to the available data, and Chapter 10 compares the impact of continuing with the BCG school's scheme post 2005 compared to universal neonatal vaccination. In addition, this study could not evaluate the impact of the neonatal programme on the high-risk population it targets, due to a lack of reliable data. This is a limitation of all of the analyses presented in this thesis. Improved coverage data for the BCG programme is required to more fully evaluate its ongoing impact.

7.7 Summary

- In the non-UK born, I found evidence for an association between a reduction in incidence rates and the change in BCG policy (school-age IRR: 0.74 (95%CI 0.61, 0.88), neonatal IRR: 0.62 (95%CI 0.44, 0.88)).
- I found some evidence that the change in BCG policy was associated with a increase in incidence rates in the UK born school-age population (IRR: 1.08 (95%CI 0.97, 1.19)) and weaker evidence of an association with a reduction in incidence rates in UK born neonates (IRR: 0.96 (95%CI 0.82, 1.14)).
- Overall, we found that the change in BCG policy was associated with directly preventing 385 (95% CI -105, 881) TB cases.
- Withdrawing universal vaccination at school-age and targeting BCG vaccination towards high-risk neonates was associated with reduced incidence of TB in England. This was largely driven by reductions in the non-UK born. There was a slight increase in UK born school-age cases.
- The code for the analysis contained in this Chapter can be found at: `doi.org/10.5281/zenodo.2583056`[2]

[2] Alternatively available from: `https://github.com/seabbs/DirectEffBCGPolicyChange`

Chapter 8

Devoloping a dynamic transmission model of Tuberculosis

8.1 Introduction

In the previous chapter (Chapter 7) I estimated the direct impact of the change in BCG policy on the subset of the population who were directly impacted by the change. The time horizon of any estimated impact was limited by the available data. If there is a non-negligible amount of Tuberculosis (TB) transmission amongst the UK born then any change in BCG vaccination policy will also have in-direct impacts via on-wards transmission. In order to capture these impacts a dynamic transmission model (see Chapter 1) may be used, this explicitly models the rate that individuals are infected using the mass action assumption. A dynamic transmission model also allows estimates to be made of the long term impact of BCG policy changes via model simulation, although this does require some assumptions to be made about future population demographics and TB incidence rates in the non-UK born population.

This chapter presents the development and parameterisation of a, semi-stochastic, dynamic model of TB transmission, incorporating BCG vaccination, in the UK born population of England. The key features of TB transmission, and BCG vaccination are discussed, with details of pertinent TB models given. An appropriate model structure for answering the study question is then outlined, along with a justification of the choices made and details of required sensitivity analyses. The model structure is then defined mathematically and parameterised using literature sources as well as data from the Enhanced Tuberculosis Surveillance System (ETS), Labour Force Survey (LFS) and Office for National Statistics (ONS) (see Chapter 4). The assumptions made during model building and parameterisation are highlighted in preparation for evaluation during model fitting.

8.2 Choice of model structure

When developing an infectious disease dynamic model there is a trade-off between reproducing reality and interpretability.[5] A model that includes all known features of a disease may not be able to answer questions of interest as it is too complex to interpret or because data does not exist to calibrate many of its parameters. A highly complex model, or indeed

an overly simplistic one, may also be at risk of bias. The optimal model is therefore as parsimonious as possible, whilst still capturing the key features of a disease and making best use of all available data.[5] In this section the key features of TB and BCG vaccination that must be captured in order to produce meaningful output are discussed, as well as the features that can be excluded for this study question. Data from the Enhanced Tuberculosis Surveillance (ETS) (Chapter 4) is used to support evidence from the literature. Further background information can be found in Chapter 2 and Chapter 4.

8.3 Tuberculosis disease

The key features of TB transmission in England which must be captured in order to develop a methodologically sound model, are as follows:

1. **Latency** - after an initial infection 5-10% of individuals develop symptomatic TB within 1-2 years. The majority of individuals enter a latent state in which they passively carry TB mycobacterium but are not symptomatic. Reactivation of the bacilli can then occur many years later due to a lose of immune control.[8] Simplistically latent TB may be modeled with a single latent compartment[82], more commonly an additional transition rate between the susceptible and active disease states is added.[83] This represents rapid progression to active disease, and slower progression via a low risk latent stage. Both of these model structures have been shown to not fit activation data well.[83,84] More complex structures that are commonly used incorporate either parallel or serial latency (Figure **??**). Both of these structures incorporate both slow and fast latent periods and have been shown to produce identical activation dynamics.[84] This is unfortunate as they represent two disparate biological mechanisms, with the serial assumption representing decreasing risk over time for individuals and the parallel assumption suggesting that a subset of individuals are at a greater risk of developing active TB disease. For models that seek to investigate interventions targeted at latent cases this structural uncertainty is problematic. However, as BCG vaccination occurs prior to infection both structures will produce comparable results for study questions evaluating this intervention. The model presented here uses a serial latent structure. This is commonly used in the literature; simplifies modelling other aspects of TB; and has a plausible biological underpinning.[84]

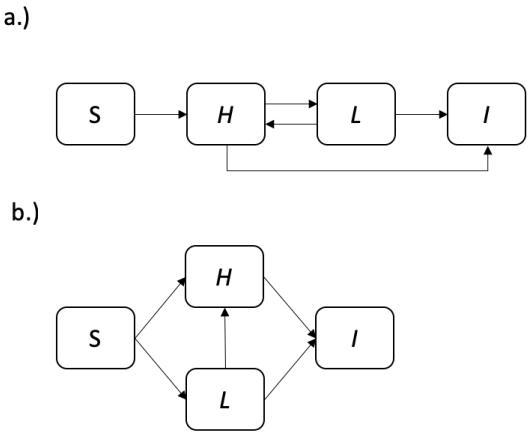


Figure 8.1: Flow diagrams of a.) the serial latency assumption and b.) the parallel latency assumption. The flow diagrams contain the following compartments; Susceptible (S), high risk latent (H), low risk latent (L), and infected (I). Solid arrows represent transition rates. Note that in both models repeated transmission to low risk latents is possible. This allows low risk latent cases to become high risk latent cases. For some varients of the parallel latency assumption, where it is assumed being high risk is inherent to individuals, this may not be appropriate.

1. **Pulmonary/Extra-Pulmonary TB** - active TB disease can be defined as any symptomatic TB infection but it may present with a range of diverse individual states. Commonly, TB cases are stratified into pulmonary and extra-pulmonary TB cases, with pulmonary cases being individuals who present with TB present in the lungs, and extra-pulmonary TB cases being cases that present with TB symptoms that do not involve the lungs (Chapter 2). Often pulmonary cases also present with extra-pulmonary symptoms. It is thought that pulmonary TB cases make up for the vast majority of TB transmission,[9,85] as TB is primarily spread by aerosol transmission but that extra-pulmonary cases have worse outcomes. The proportion of pulmonary to extra-pulmonary cases has increased over time from 26.2% (1944/7410) in 1982 to 45.8% (2634/5748) in 2016. This may be attributed to the age distribution of TB cases changing, as different age-groups are more likely to progress to pulmonary vs extra-pulmonary TB.[86] It may also be related to the increase of non-UK born cases, as a higher proportion of non-UK born cases have extra-pulmonary disease only (51.4%, 2,103/4,089, in 2016), compared to UK born cases (31.9%, 467/1,465, in 2016) [1]. The model presented here includes both pulmonary and extra-pulmonary cases, with only pulmonary cases contributing to onwards transmission. Extra-pulmonary cases are included so that the full impacts of any intervention can be correctly estimated.

2. **Smear status** - microscopic analysis of sputum smear samples for acid-fast bacilli is widely used as a means of diagnosis for TB. There is evidence that smear positive cases are responsible for the majority of transmission,[87] with smear negative cases contributing approximately 76% (95% CI 70%, 80%) less to transmission than smear positive cases.[88] The proportion of smear positive cases varies with age,[89] with 30.2% (95% CI 26.3%, 33.7%) in 0-14 year-olds, 65.2% (95% CI 64.2%, 66.2%) in 15-59 year-olds and 53.6% (95% CI 51.9%, 55.3%) in 60-89 year-olds in the ETS between

2000 and 2015. the model presented here includes sputum status via the force of infection.

3. **Re-infection** - individuals with latent TB, or who have recovered from active TB, may be at risk of re-infection. It is thought that latent individuals gain some partial protection from prior infection but estimates for the magnitude of this protection vary widely.[90] A review of prospective cohort studies of persons exposed to individuals with infectious TB that was published prior to the widespread treatment of latent TB found that prior TB infection provided partial protection of 79% (70%, 86%).[91] This is included in the model presented below via a the force of infection.

4. **Re-activation/Re-infection of recovered cases** - individuals who have recovered from active TB disease are at risk of both re-infection and re-activation. As in many dynamic transmission models, this has been modelled here by treating recovered cases as having low risk latent TB.[92,93] This provides recovered cases with the same protection against re-infection as low risk latent cases. However, this means that vaccinated cases receive the benefits of BCG protection even after they have recovered from active TB disease.This may not be realistic but due to the low burden of TB in England is unlikely to lead to significant bias.

5. **TB treatment** - standard treatment consists of a 6 month course of multiple antibiotics, usually consisting of isoniazid, rifampicin, pyrazinamide and ethambutol. If treatment is unsuccessful using these first line drugs, second line drugs are then proscribed which have more severe side effects and a longer treatment regime (12-24 months).[1,14] Individuals on treatment may be considered non-infectious but are still at risk of negative outcomes including death. 4.9% (4847/98124) of cases in the ETS were lost to follow up within the first year of starting treatment between 2000 and 2014. A treatment term has been included in the model presented here along with potential treatment failure. Multi-stage treatment has not been modelled as this would add complexity but would not improve the models performance in other areas.

6. **TB related mortality** - within the first 12 months of starting treatment 6% (5884/98124) of cases, with complete data and who were evaluated, died in the ETS between 2000 and 2014. Of these 60.5% (1984/3290) had TB as a cause of death or had a cause of death that was related to active TB. The rate of TB mortality varies with age, with the very old and the very young at the greatest risk. Age-stratified TB mortality is important to include in any policy relevant model of TB transmission as reducing mortality is a major public health goal. There is little data on the rate of TB mortality in those untreated for TB, so all TB mortality will be modelled using a single, age stratified, term.

7. **Age related presentation of TB** - there is evidence to suggest that the risk of TB activation varies by age,[84] as does the proportion of cases that develop pulmonary TB,[86], the proportion of cases that are smear positive, and the risk of TB mortality. It may also be the case that the transmission probability varies by age, after accounting for the proportion of cases that are pulmonary and the proportion of cases that are smear positive. In the model presented here age has been included by stratifying the population into age-groups.

8. **Demographic changes** - TB dynamics develop over a long timespan, because of the potential for cases to develop active TB disease many years after infection. Over these

long timespans population demographics can play an important role. An approach to include demographics is to link birth and death rates so that the modeled population is static over time. This has the advantage of making it easier to identify changes that are linked to the disease dynamics. In the model presented here birth and death processes have been incorporated based on available, age-specific, data. For years with available data this has the advantage of producing demographics which match those observed in the study population, allowing for policy relevant forecasts to be made. However, for years with limited data assumptions must be made about the likely birth and death rates.

9. **Non-UK born TB Cases** - TB incidence in England is highly heterogeneous with over 70% of cases occurring in the non-UK born population.[1] The age distribution of cases in the UK born and non-UK born populations differ, with the UK born population having a relatively uniform distribution. Meanwhile, the non-UK born have higher incidence rates in those aged 80 years and older (69.3 per 100,000 in 2016), those aged 75 to 79 years (62.9 per 100,000 in 2016) and those aged 25-29 years old (61.6 per 100,000 in 2016).[1] Exposure to England's BCG vaccination policy is difficult to assess for the non-UK born as is the degree of transmission occurring in the UK as opposed to cases being imported from abroad, or acquired from trips to cases countries of origin. For this reason the model presented here does not explicitly include non-UK born cases. Instead it imports non-UK born cases into the force of infection, with the addition of a mixing parameter that controls the degree of contact between non-UK born cases and those born in the UK.

8.3.1 BCG vaccination

The key features of the BCG vaccine that must be considered in order to forecast the impacts of vaccine policy are:

1. **Protection from active disease** - the BCG vaccine has been shown to primarily protect against the progression from latent to active TB disease (Chapter 2). It has been shown to be highly protective in children,[3,17,18] but to have variable protection in adults ranging from 0-80%.[21] This variation in protection is thought to be linked to the equator, with the vaccine becoming increasing effective at higher, and lower latitudes. In England an MRC trial in the 1950's found that the BCG vaccine was highly effective, there is little evidence to suggest that this has changed in the UK born population.[20]

2. **Duration of protection** - BCG protection wanes with time, with the greatest protection shortly after vaccination. There is good evidence to suggest that the effectiveness of BCG vaccination lasts up to 15 years,[22] and a recent study suggests that this protection may last later into adulthood in the UK born.[23] However, the it's effectiveness reduces with time.

3. **Protection from initial infection** - there is evidence that the BCG vaccine provides partial protection against initial infection.[3] This may impact transmission dynamics. Not including it would lead to a higher proportion of latent cases in those vaccinated with BCG. One complicating factor is that the majority of the estimates of the protection offered by BCG vaccination from active TB disease include the protection from initial infection.

4. **Age structure** - BCG vaccination has previously been targeted at those at school-age and is currently targeted at neonates. There is also evidence that the effectiveness of BCG vaccination varies with age,[17,19] although there is little evidence of this in England. In order to answer questions relevant to BCG vaccination, TB disease must be modeled in young children and young adults. To capture the waning of BCG protection age structure must be modeled beyond these age groups.[22]

5. **Non-UK born TB Cases** - the majority of cases that occur in the non-UK born would not have been exposed to England's BCG vaccination. In the majority of high incidence countries BCG vaccination is common, with most countries vaccinating young children as early in life as possible.[4] Based on this it could be assumed that all non-UK born cases were vaccinated at birth. However, this high level of coverage is unlikely. As the BCG vaccine has not been shown to decrease the likelihood of transmission from vaccinated TB cases assuming that all non-UK born cases are unvaccinated does not impact the dynamics in the modeled UK born population.

6. **Additional benefits of BCG vaccination** - there is some evidence that the BCG vaccine may reduce all-cause mortality both in the general population and specifically for TB cases (Chapter 6). There is weaker evidence that this reduction in all-cause mortality for TB cases may be associated with a reduction in TB specific mortality. This was not included in the model presented here as the evidence was not conclusive. This means the benefits of the BCG vaccine may have been underestimated.

8.4 A dynamic model of TB transmission

8.4.1 Model outline

The dynamic model of TB implemented here may be considered as 3 nested models; these are a TB transmission model; a demographic processes model; and a BCG vaccination model. For an overview of the model structure see the flow diagram (Figure 8.2) and for full details see the model equations (Section 8.4.2). Model parameters are discussed in detail in Section 8.5.2.

Disease model

The model includes the following compartments: Susceptible (S), high risk latent (H), low risk latent (L), active TB cases with pulmonary TB (P), active TB cases with extra-pulmonary TB disease only (E), pulmonary cases on treatment (T_P), and extra-pulmonary case on treatment (T_E). Cases that were previously infected and considered at low risk of developing active disease may be reinfected, although their latent infection provides partial protection. Treatment is assumed to be the only pathway for recovery for active TB disease, with a single rate used to model the heterogeneity of treatment times. A fraction of those on treatment are assumed to be lost to follow up, with these cases returned to active pulmonary or extra-pulmonary disease. Cases that start treatment immediately stop being infectious and upon treatment completion are treated as if they have low risk latent TB disease. TB mortality is included for both active TB cases on, and off, treatment. TB mortality is stratified by disease type and age. TB transmission is assumed to act under the mass action assumption. Non-UK born cases are included into the force of infection.

Demographic model

The model is stratified into 5 year age groups from 0 to 49, with a single age group from 50-69, and from 70 to 89. 5 year age groups are most commonly used in the literature when age stratifying TB and BCG parameters. Using 5 year age groups in the model presented here allows for this data to be utilised to its maximum potential. Older adults were grouped into larger age groups as they are thought to be responsible for a small amount of TB transmission and because fine scale BCG mechanisms do not need to be modelled in these age groups. Adults aged 90+ were not modeled due to large amounts of uncertainty in the demographic data and because cases in these population represent a small fraction of total TB cases (see Chapter 4). The number of births in a given year is incorporated as a time varying parameter. The natural mortality rate is also allowed to vary with time and is stratified by age. Immigration and emigration were not included in the demographic model as reliable age stratified data was unavailable and it is unlikely that either immigration or emigration of the UK born population is a significant driver of overall population size, or structure.

Vaccination model

The vaccination model is nested into the demographic process model and therefore vaccination is possible upon entry to each modeled age group. The target age group can be varied to represent changing BCG vaccination policy. The vaccinated population is then modeled explicitly throughout all disease compartments. The primary action of the BCG vaccine is to prevent the transition from latent to active disease, this is included for both high and low risk latent cases. Waning vaccination effectiveness has been included by stratifying vaccine effectiveness by age group. The partial protection offered by BCG vaccination against initial infection has been included as a negative modifier on the protection from latent to active disease and as a modifier on the proportion of cases that are initially infected. This allows estimates of the effectiveness of BCG vaccination at preventing active TB disease in the susceptible population to be used, as these estimates have the most robust data sources. It is assumed that latently infected individuals do not gain additional protection from re-infection from the BCG vaccine. The BCG vaccine has been modeled as being partially protective for all individuals rather than as a "take" vaccine (i.e all or nothing protection). This assumption simplifies the model and will not impact the dynamics of TB transmission, assuming that protected and unprotected BCG vaccinated individuals obey the mass action assumption (See Chapter 1 and [5]).

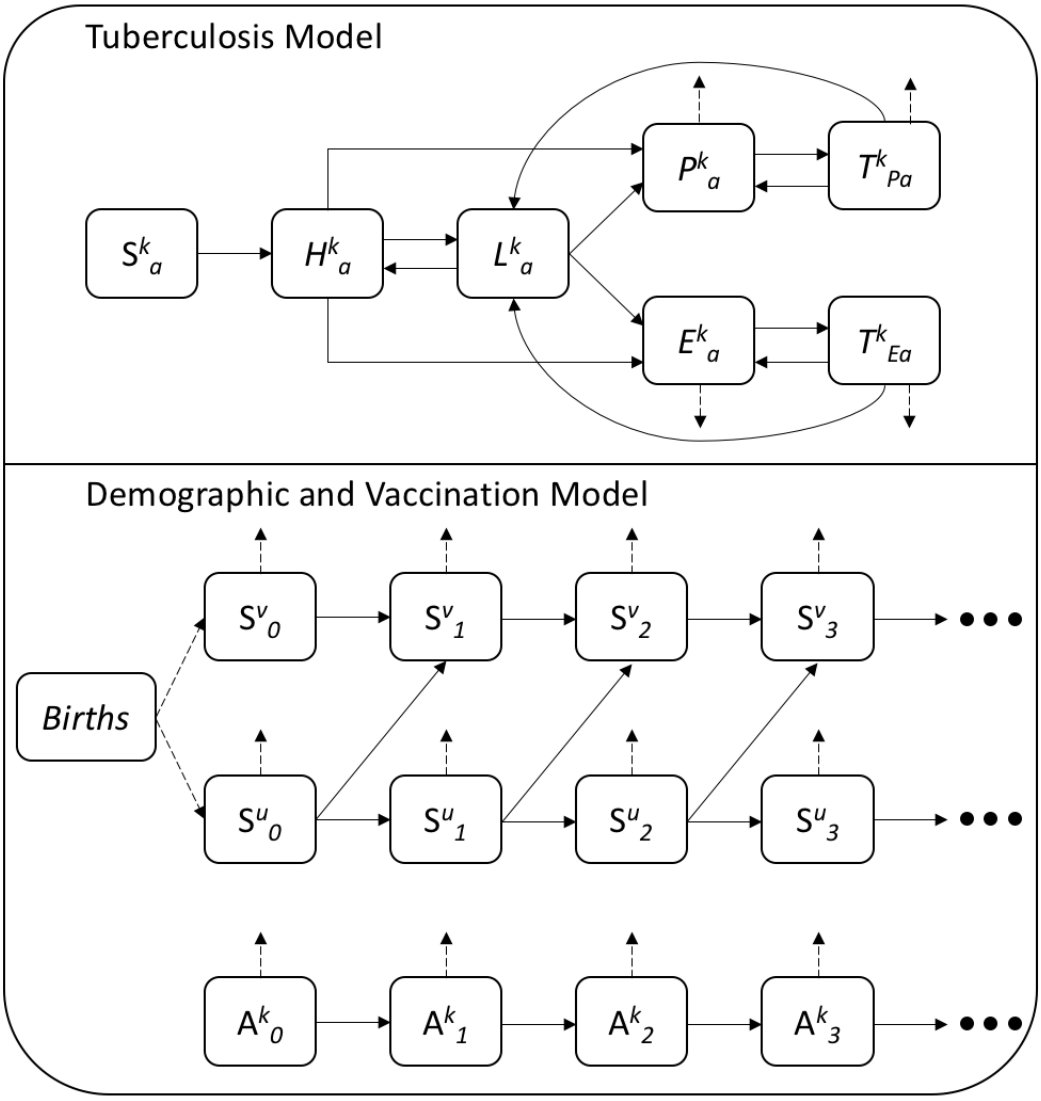


Figure 8.2: Flow diagram for the dynamic Tuberculosis (TB) disease model with demographics and vaccination described in the text. The TB model contains the following compartments; Susceptible (S), high risk latent (H), low risk latent (L), active cases with pulmonary TB (P), active TB cases with extra-pulmonary TB only (E), pulmonary cases on treatment (T_P), and extra-pulmonary cases on treatment (T_E). The vaccinated (v) and unvaccinated (u) populations are represented by k, such that $k = u, v$. Age stratification is represented by a (where $a = 1, 2, ..., 15$) in the disease model and the $0, 1, 2, 3$ subscripts in the demographic model. Five year age groups are modelled (i.e $0-4$, $5-9$, $10-14$, ...) up to 49 years old, with a single age group for those aged 50-69 years old and those aged 70-89 years old. Individuals aged 90+ are not explicitly modelled. In the demographic and vaccination model the A compartment represents the demographic processes modelled in all population compartments except for the vacinated and unvaccinated susceptible populations. Solid arrows represent transition rates within the modelled populations and dashed arrows represent transition rates into, or out of the modelled populations (i.e birth and death processes).

8.4.2 Model equations

In order to simplify the model equations the disease (d) and demographic and vaccination models (p) have been separated such that (where $C = S,\ H,\ L,\ P,\ E,\ T_P,\ T_E$),

$$\frac{dC}{dt} = \frac{dC^d}{dt} + \frac{dC^p}{dt} \tag{8.1}$$

The disease model (C^d) is then defined as,

$$\frac{dS_a^{kd}}{dt} = -(1-\chi_a^k)\lambda_a S_a^k \tag{8.2}$$

$$\frac{dH_a^{kd}}{dt} = (1-\chi_a^k)\lambda_a S_a^k + (1-\delta)\lambda_a L_a^k - (1-\alpha_a^k)\epsilon_H^a H_a^k - \kappa_a H_a^k \tag{8.3}$$

$$\frac{dL_a^{kd}}{dt} = \kappa_a H_a^k - (1-\delta)\lambda_a L_a^k - (1-\alpha_a^k)\epsilon_L^a L_a^k + \phi_a^P T_{Pa}^k + \phi_a^E T_{Ea}^k \tag{8.4}$$

$$\frac{dP_a^{kd}}{dt} = \Upsilon_a(1-\alpha_a^k)(\epsilon_H^a H_a^k + \epsilon_L^a L_a^k) + \zeta_a^P T_{Pa}^k - \nu_a^P P_a^k - \mu_a^P P_a^k \tag{8.5}$$

$$\frac{dE_a^{kd}}{dt} = (1-\Upsilon_a)(1-\alpha_a^k)(\epsilon_H^a H_a^k + \epsilon_L^a L_a^k) + \zeta_a^E T_a^k - \nu_a^E E_a^k - \mu_a^E E_a^k \tag{8.6}$$

$$\frac{dT_{Pa}^{kd}}{dt} = \nu_a^P P_a^k - \zeta_a^P T_{Pa}^k - \mu_a^P T_{Pa}^k - \phi_a^P T_{Pa}^k \tag{8.7}$$

$$\frac{dT_{Ea}^{kd}}{dt} = \nu_a^E E_a^k - \zeta_a^E T_{Ea}^k - \mu_a^E T_{Ea}^k - \phi_a^E T_{Ea}^k \tag{8.8}$$

Where the unvaccinated (u) and vaccinated (v) populations are represented by $k = u, v$ and age groups are represented by $a = 0, 1, 2, 3, ...10$ for 0-4 year olds ($a = 0$), 5-9 year olds ($a = 1$), up to 45-49 year olds ($a = 10$). 50-69 year olds are modelled as a single group ($a = 11$), as are 70-89 year olds ($a = 12$). Individuals 90+ are excluded from the model. The disease models parameters are defined as follows: ϵ_a is the age-specific rate of activation from each latent population, κ_a is the age-specific rate of transition into the low risk latent population, ν_a is the rate of starting treatment, δ is the protection from re-infection conferred by prior latent infection, Υ_a is the age-specific proportion of cases that develop pulmonary TB, with or without extra-pulmonary TB, $\mu_a^{P,E}$ is the age-specific mortality from active pulmonary (P) and extra-pulmonary (E) TB, and α_a is the age-specific effectiveness of the BCG vaccine at preventing active TB disease. In the unvaccinated population $\alpha = 0$. χ_a^k is the age-specific protection inferred due to vaccination from initial infection, in the unvaccinated population it is defined to be 0.

The demographic and vaccination model (C^p) is then defined as ($A = H,\ L,\ P,\ E,\ T_P,\ T_E$),

$$\frac{dS_a^{up}}{dt} = (1-sgn(a))(1-\gamma_a)\omega(t) + sgn(a)(1-\gamma_a)\theta_{a-1}S_{a-1}^u - \theta_a S_a^u - \mu_a(t)S_a^u \tag{8.9}$$

$$\frac{dS_a^{vp}}{dt} = (1-sgn(a))\gamma_a\omega(t) + sgn(a)\gamma_a\theta_{a-1}S_{a-1}^u - \theta_a S_a^v - \mu_a(t)S_a^v \tag{8.10}$$

$$\frac{dA_a^{kp}}{dt} = sgn(a)\theta_{a-1}A_{a-1}^k - \theta_a A_a^k - \mu_a(t)A_a^k \tag{8.11}$$

Where $\omega(t)$ is the time varying number of births, γ_a is the age-specific proportion that are vaccinated, θ_a is the rate of ageing, and $\mu_a(t)$ is the time varying natural mortality rate.

The signum function used above is defined as follows;

$$sgn(x) := \begin{cases} -1 & \text{if } x < 0, \\ 0 & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases} \tag{8.12}$$

8.4.3 Force of infection

The age-specific (a), and vaccination dependent (k), force of infection (λ_a^k) is defined as follows,

$$\lambda_a^k = \frac{\beta_a}{N} \sum_{i=1}^{A} \rho_i C_{ai} (\frac{M\iota_i}{\nu_a^P} + \sum_{j=u,v} P_i^j) \tag{8.13}$$

Where ι_a is the age stratified number of non-UK born pulmonary cases notified in a given year, ρ_a is the age-specific proportion of cases that are smear positive, ν is the rate of starting treatment for active TB, C_{ai} is the age stratified contact matrix, and M is the mixing rate between the UK born and non-UK born population. The force of infection is age stratified by contact rates and by the proportion of cases that are smear positive in a given age group.

8.5 Parameterisation and data synthesis

Parameters distributions were either estimated from the available data or assumed based on common values found in the literature. Where no comparable estimates were found in the literature a range of distributions have been considered, with the most plausible adopted as the baseline. The discussion of model parameters has been split into disease model parameters, vaccination model parameters, and demographic model parameters. The data sources used to estimate model parameters have been detailed although for the Enhanced TB Surveillance system (ETS) and the Labour Force Survey more detail is provided elsewhere (Chapter 4).

8.5.1 Data sources

Enhanced TB Surveillance System

Model parameters were estimated using the ETS system where possible, with data on all notified cases in England from Jan 1, 2000 to Dec 31, 2015. The ETS is a robust national surveillance network that collects collects demographic, clinical, and microbiological data; a yearly report is published detailing data collection, cleaning, and trends in TB incidence (Chapter 4).[1]

Labour Force Survey

Yearly population estimates, stratified by age and UK birth status, were extracted from the April to June LFS from 2000 to 2015. As detailed previously (Chapter 4) the LFS

is a study of the employment circumstances of the UK population, providing the official measures of employment and unemployment in the UK. As the LFS is based on a sample the population estimates are subject to sampling errors.

8.5.2 Model Parameters

Disease model parameters

Details of the prior distributions used for each disease model parameter are given in Table 8.1. Table 8.2 contains details of the sources used to parameterise the model. More detail is given in the following sub-sections.

Table 8.1: Dynamic disease model parameters, descriptions, prior distributions, units, method used to derive the prior distribution and the type (i.e data derived, literature, assumption). All data based parameters are included. P = pulmonary TB, E = extra-pulmonary TB, v = vaccinated, i = age at vaccination, $\mathcal{U}$ = Uniform, $\mathcal{N}$ = Normal

Parameter	Description	Distribution	Units	Method	Type
C_{eff}	The assumed effective number of contacts per infectious TB case.	$\mathcal{U}(0,5)$	-	Estimated using a dynamic model of TB transmission in England which found an effective contact rate of 1 in 1990. A conservative interval has been chosen around this value to represent the large amount of uncertainty as to its current value.	Literature
C_{eff}^{hist}	The assumed historic effective number of contacts per infectious TB case. Additive with C_{eff}	$\mathcal{U}(10,15)$	-	Estimated using a dynamic model of TB transmission in England which found an effective contact rate of 1 in 1990 and 20 in 1901. A conservative interval has been chosen to represent the parameter uncertainty. Assumed to decrease linearly over time from 1931 until 1990.	Literature
C_{half}^{hist}	Half life of the initial historic effective contact rate. The historic effective contact rate is assumed to decay with an exponential decay. C_{half}^{hist} is half life of this decay process.	$\mathcal{U}(0,20)$	Years	The prior distribution is informed by historic TB notification and the timeline being simulated.	Assumption
Υ	The age-specific proportion of cases that have pulmonary TB	$\Upsilon_{0-14,15-59,60-89} = \mathcal{N}(6.29e^{-1}, 1.01e^{-2})$, $\mathcal{N}(7.06e^{-1}, 4.11e^{-3})$, $\mathcal{N}(7.50e^{-1}, 5.69e^{-3})$	Proportion	Estimated using the age-specific proportion of cases that had pulmonary TB in the ETS.	Derived from data
ρ	The age-specific proportion of pulmonary TB cases that are smear positive	$\rho_{0-14,15-59,60-89} = \mathcal{N}(3.02e^{-1}, 1.89e^{-2})$, $\mathcal{N}(6.52e^{-1}, 5.18e^{-3})$, $\mathcal{N}(5.36e^{-1}, 8.45e^{-3})$	Proportion	Estimated using the age-specific proportion of pulmonary TB cases that were smear postive in the ETS.	Derived from data
C	Matrix of contact rates between each age group	-	Non-unique yearly contacts.	For each parameter sample a contact matrix was bootstrapped from the POLYMOD survey data, standardised using the UK born population in 2005, and then averaged to provided a symmetric contact matrix.	Literature
$\iota(t)$	The age-specific number of non-UK born pulmonary TB cases in England each year	-	Cases	The number of pulmonary non-UK born cases for each year were extracted from the ETS and grouped by age.	Derived from data

Table 8.1: Dynamic disease model parameters, descriptions, prior distributions, units, method used to derive the prior distribution and the type (i.e data derived, literature, assumption). All data based parameters are included. P = pulmonary TB, E = extra-pulmonary TB, v = vaccinated, i = age at vaccination, $\mathcal{U}$ = Uniform, $\mathcal{N}$ = Normal *(continued)*

Parameter	Description	Distribution	Units	Method	Type
M	The proportion of mixing between the UK born and non-UK born population.	$\mathcal{U}(0, 0.5)$	Proportion	Hetergeneous mixing is assumed as the non-UK born population is spatially clustered.	Assumption
χ	Age-specific protection from infection with TB due to BCG vaccination	$\chi_i^v = \mathcal{N}(0.185,\ 5.36e-02)$, where i is the age group vaccinated.	Proportion	A meta-analysis of the protection from infection due to BCG vaccinatiion in children. It has been assumed that there is no reduction in protection in UK born adults.	Literature
ϵ_H	The age-specific rate of transition to active disease during high risk latent period.	$\epsilon_H^{0-4,5-14,15-89} = \mathcal{N}(6.95e-3, 1.30e-3)$, $\mathcal{N}(2.8e-3, 5.61e-4)$, $\mathcal{N}(3.35e-4, 8.93e-5)$	$days^{-1}$	From fitting a similar model to contact data in Australia, and Holland. Distribution derived by the assumption of a normal distribution based on 95% credible intervals.	Literature
κ	1 over the age-specific average high risk latent period.	$\kappa^{0-4,5-14,15-89} = \mathcal{N}(1.33e^{-2}, 2.42e^{-3})$, $\mathcal{N}(1.20e^{-2}, 2.07e^{-3})$, $\mathcal{N}(7.25e^{-3}, 1.91e^{-3})$	$days^{-1}$	From fitting a similar model to contact data in Australia, and Holland. Distribution derived by the assumption of a normal distribution based on 95% credible intervals.	Literature
ϵ_L	1 over the age-specific average low risk latent period.	$\epsilon_L^{0-4,5-14,15-89} = \mathcal{N}(8.00e^{-6}, 4.08e^{-6})$, $\mathcal{N}(9.84e^{-6}, 4.67e^{-6})$, $\mathcal{N}(5.95e^{-6}, 2.07e^{-6})$	$days^{-1}$	From fitting a similar model to contact data in Australia, and Holland. Distribution derived by the assumption of a normal distribution based on 95% credible intervals.	Literature
α_i^T	The BCG vaccine effectiveness at preventing the development of active TB disease in a TB free population	$\alpha^T_{i,i+5,i+10,i+15,i+20,i+25} = 1 - e^{\alpha^{log(T)}_{i,i+5,i+10,i+15,i+20,i+25}}$, where $\alpha^{log(T)}_{i,i+5,i+10,i+15,i+20,i+25} = \mathcal{N}(-1.86, 0.22)$, $\mathcal{N}(-1.19, 0.24)$, $\mathcal{N}(-0.84, 0.22)$, $\mathcal{N}(-0.84, 0.2)$, $\mathcal{N}(-0.28, 0.19)$, $\mathcal{N}(-0.23, 0.29))$ and i is the age group vaccinated	Proportion	Poisson regression used to calculate Risk Ratios from raw study values. A distribution is then found using the log normal approximation. Effectiveness estimates are caculated using 1 minus the exponentiated log normal distribution.	Literature
δ	Reduction in susceptibilty to infection for low risk latent cases.	$\mathcal{N}(0.78,\ 4.08e-02)$	Proportion	A review of prospective cohort studies of persons exposed to individuals with infectious tuberculosis that was published prior to the widespread treatment of latent tuberculosis.	Literature

Table 8.1: Dynamic disease model parameters, descriptions, prior distributions, units, method used to derive the prior distribution and the type (i.e data derived, literature, assumption). All data based parameters are included. P = pulmonary TB, E = extra-pulmonary TB, v = vaccinated, i = age at vaccination, $\mathcal{U}$ = Uniform, $\mathcal{N}$ = Normal *(continued)*

Parameter	Description	Distribution	Units	Method	Type
$\nu^{P,E}$	1 over the average infectious period	$\nu^{P}_{(0-14,15-89)} = \mathcal{N}(0.181, 0.310)^{-1}$, $\mathcal{N}(0.328, 0.447)^{-1}$, $\nu^{E}_{(0-14,15-89)} = \mathcal{N}(0.306, 0.602)^{-1}$, $\mathcal{N}(0.480, 0.866)^{-1}$	$years^{-1}$	Estimated based on the time from initial symptoms to starting treatment	Derived from data
ν_{scale}	Rate of treatment scale up. Time to treatment is assumed to scale using logit function from 1953 to 2000. This parameter controls the rate of this scale up.	$\mathcal{U}(0, 5)$	-	The prior distribution is based on values that map the scale up from an approximately linear relationship to a binary switch half between 1953 and 2000.	Assumption
$\phi^{P,E}$	1 over the time to succesful treatment completion	$\phi^{P,E}_{0-14,15-69,70-89} = \mathcal{N}(0.606, 0.237)^{-1}$, $\mathcal{N}(0.645, 0.290)^{-1}$, $\mathcal{N}(0.616, 0.265)^{-1}$ and truncated to be greater than 4 months	$years^{-1}$	Estimated based on the time from starting treatment to treatment completion.	Derived from data
$\mu^{P,E}$	Rate of age-specific pulmonary/extra-pulmonary TB TB mortality	$\mu^{P,E}_{0-14,15-59,60-89} = \mathcal{N}(3.9e-3, 1.8e-2)$, $\mathcal{N}(2.26e-2, 7.87e-3)$, $\mathcal{N}(1.17e-1, 1.65e-2)$ truncated to be greater than 0.	$years^{-1}$	Estimated based on outcomes at 12 months where cause of death was known, including all-cause deaths in the denominator.	Derived from data
$\zeta^{P,E}$	Rate of loss to follow up	$\zeta^{P,E}_{0-14,15-59,60-89} = \mathcal{N}(9.76e-3, 1.79e-2)$, $\mathcal{N}(3.04e-2, 7.64e-3)$, $\mathcal{N}(6.14e-3, 1.59e-2)$, truncated to be greater than 0.	$years^{-1}$	Estimated based on outcomes at 12 months for TB cases	Derived from data

Table 8.2: Sources used to parameterise the disease and demographic models. Parameters that use the source are given, as well as the study type, setting, year/years studied and a description of the study/data source.

Parameters	Study Type	Setting	Year	Description	Source
$\iota(t)$, $\mu^{P,E}$, $\nu^{P,E}$, $\phi^{P,E}$, ρ, Υ, $\zeta^{P,E}$	-	England	2000-2015	The Enhanced Tuberculosis Surveillance System (ETS) is a robust national data collection system that collects demographic and microbiological data on all notified cases in England.	ETS
$\mu^{all-cause}(t)$, $\omega(t)$	-	England	-	The Office for National Statistics (ONS) compiles demographic, health, enconomic, and social data for the United Kingdom	ONS
C_{eff}, C_{eff}^{hist}	Dynamic modelling study	England	Up to 1990	Used a dynamic model of tuberculosis, robustly parameterised to the available evidence and including realistic population demographis to estimate the effective contact rate of TB over time until the 1990's in the UK born white male population.	[94]
C	Contact survey	Europe - including the United Kingdom	2005	Conducted contact surveys, based on a contact diary, in multiple European countries. Contacts were stratified by age and type of contact. In the United Kingdom over a thousand people were surveyed.	[95]

Parameters	Study Type	Setting	Year	Description	Source
χ	Systematic review and meta-analysis	Global	Up to 2014	A meta-analysis; conducted with the aim of determining whether BCG vaccination protects against Mycobacterium tuberculosis infection as assessed by interferon γ release assays (IGRA) in children. Estimated both protection from initial latent infection and active TB disease.	[3]
ϵ_H, ϵ_L, κ	Systematic review	Global	Up to 2017	Aimed to determine which dynamic TB model structure best captured the observed activation dynamics of TB. Identified 6 different commonly used model structures and compared them by fitting to activation data from the Netherlands and Australia.	[84]
α_i^T	Clinical trial	England	1950-1965	Investigated the effectiveness of the BCG vaccine at preventing TB disease when given at what as then school-leaving age. Followed the cohort over 15 years and estimated the effectiveness of the BCG vaccien in 2.5 year intervals from vaccination.	[20]

Parameters	Study Type	Setting	Year	Description	Source
α_i^T	Population based case-control study	England	2002-2014	Recruited UK-born White subjects with TB and randomly smapled White community controls. Cox regression was used to adjusted for known confounders and the effectiveness of the BCG vaccine was estimated from 10 years after vaccination until 30 years after vaccinaton.	[23]
δ	Systematic review and meta-analysis	Global	Up to 2012	Reviewed prospective cohort studies of persons exposed to individuals with infectious TB. Only included studies that were published before the widespread treatment of latent TB. Aimed to estimate the reduction in re-infection for latent TB cases.	[91]
γ, ν_{scale}, M	-	England	-	Where data, or literature, sources were not available assumed values were used based on expert opinion	Assumption

Non-UK born pulmonary cases Non-UK born pulmonary cases were estimated using the ETS (see Chapter 4) for each age-group included in the model from 2000 until 2015. Prior to 2000 a linear scale up of non-UK born cases, starting in 1960 and reaching the average number of cases observed in 2000 by December 1999 was assumed. A log link scale (i.e $log(t - t_{start})/log(t_{max} - t_{start})$) is also explored as a scenario during model fitting. To incorperate the uncertainty in the number of non-UK born cases a normal distribution was used, with the standard deviation determined using parameters from the observation model (Chapter 9).

Probability of transmission The probability of transmission can be defined as the probability that a single contact between an infectious active TB case and a susceptible individual will lead to TB infection. The probability of transmission (β_a) can be redefined in terms of effective contacts (C_{eff}), historic effective contacts (C_{eff}^{hist}), actual average yearly total contacts (C_{actual}), the average period of time infectious ($\frac{1}{\nu_{avg}^P}$), and the average

mortality rate (μ_{age}) as follows,

$$\beta_a = \frac{(\nu^P_{avg} + \mu_{avg})\left(C_{eff} + C^{hist}_{eff} * t_{hist}\left(1 - \frac{log(t-1931)}{log(1990-1931)}\right)\right)}{C_{actual}} \tag{8.14}$$

Where,

$$t_{hist} := \begin{cases} 1 & \text{if } t < 1990, \\ 0 & \text{if } t >= 1990. \end{cases} \tag{8.15}$$

Vynnycky et al. found that the effective contact rate (i.e the transmission probability times the contact rate) for TB was approximately 22 in 1900 and fell to approximately 1 in 1990.[94] Incidence rates have increased since 1990 (Chapter 4) and it is unclear what impact this has had on the effective contact rate. I have assumed that the effective contact rate is uniformly distributed between 0 and 5 contacts. Similarly, for the historic effective contact rate I have assumed a uniform distribution between 10 and 15. This gives a maximum effective contact rate of 20, inline with Vynnycky et al.'s estimate for the effective contact rate in 1900. I have also assumed that the historic contact rate declines over time using an exponential model, reducing to 0 in 1990. The prior for the half life of this decline (c^{hist}_{half}) has been assumed to be uniformly distributed between 0 and 20 years. Additional age stratification of β is explored by including modifiers for certain age-groups. The baseline scenario is that no modification is required, with variation between children and adults (β_{child}) and between children, adults and older adults (β_{child} and $\beta_{olderadults}$) also being considered. The priors for these modifiers are normally distributed around 1 with a standard deviation of 0.5 (truncated to be greater than 0). The contact rate is estimated by averaging the total age-specific contact rates estimated from POLYMOD data (Section 8.5.2) on an annual basis.

Rate of recovery from active disease The rate of recovery from active TB disease was estimated as 1 over the time with active disease using the ETS with UK born cases from 2000 until 2012. Cases with a period of time symptomatic that was less than 0 days were removed as these are likely to be spurious. Figure 8.3 indicates that the distribution of time to treatment differs between children and adults and by pulmonary/extra-pulmonary TB status. There was little evidence that time to treatment differed between adults and older adults. A normal distribution was used for each age group, truncated to be greater than 0 months. Prior to 1952, and the introduction of isoniazid, I have assumed that the time to recovery from active TB disease is 20 years, representing natural recovery.

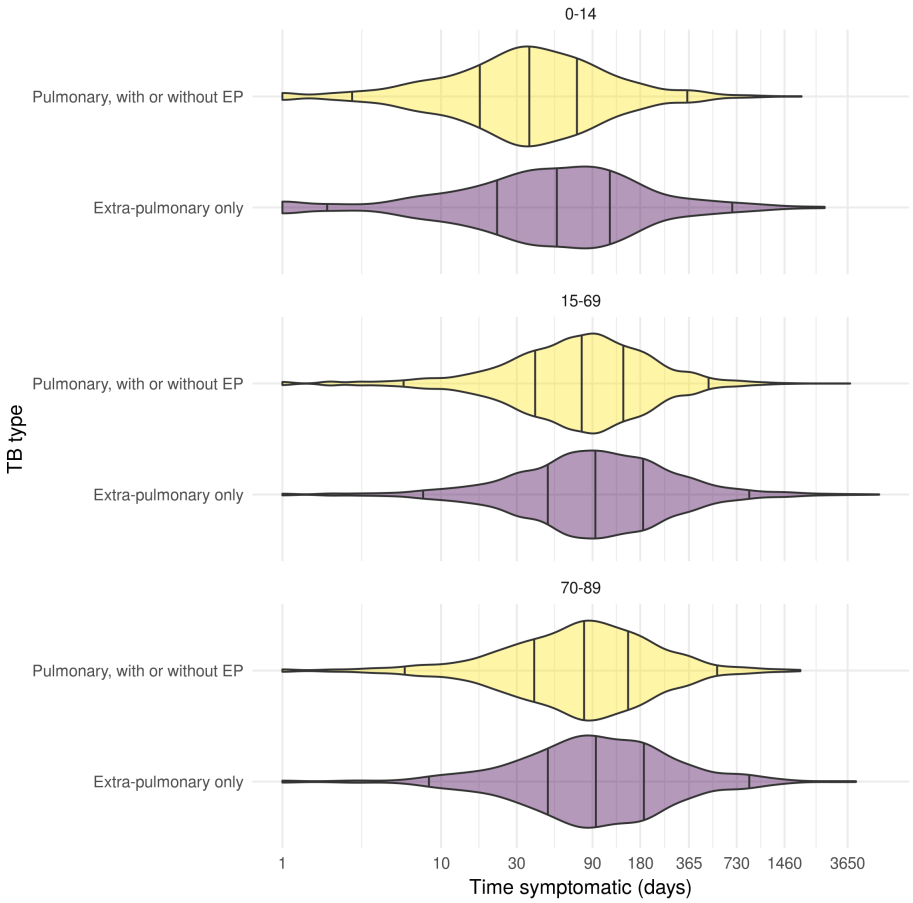


Figure 8.3: Distribution of time to treatment (days) from the date of reported symptom onset until the date started treatment for the UK born, stratified by age group and pulmonary/extra-pulmonary TB status in the Enhanced Tuberculosis Surveillance system for notifications between 2000 and 2012. Age is stratified into three groups; children (0-14), adults (15-59) older adults (15-59). The time from symptom onset to starting treatment is shorter for cases with pulmonary TB cases across age groups, with younger cases starting treatment more rapidly than older cases.

Rate of successful treatment The rate successful treatment was estimated as 1 over the period of time on treatment using the ETS, with UK born cases between 2000 and 2012. Cases with a treatment time less than 1 month were removed as TB treatment is standardised and should take at least several months. There was little evidence that time to treatment completion differed between pulmonary and extra-pulmonary TB cases but there was some evidence that older TB cases were more likely to be on treatment for longer than younger cases (Figure 8.4). A normal distribution was used for children, adults and older adults, with each truncated to be greater than 4 months. This truncation was introduced as a faster treatment time than this was considered implausible.

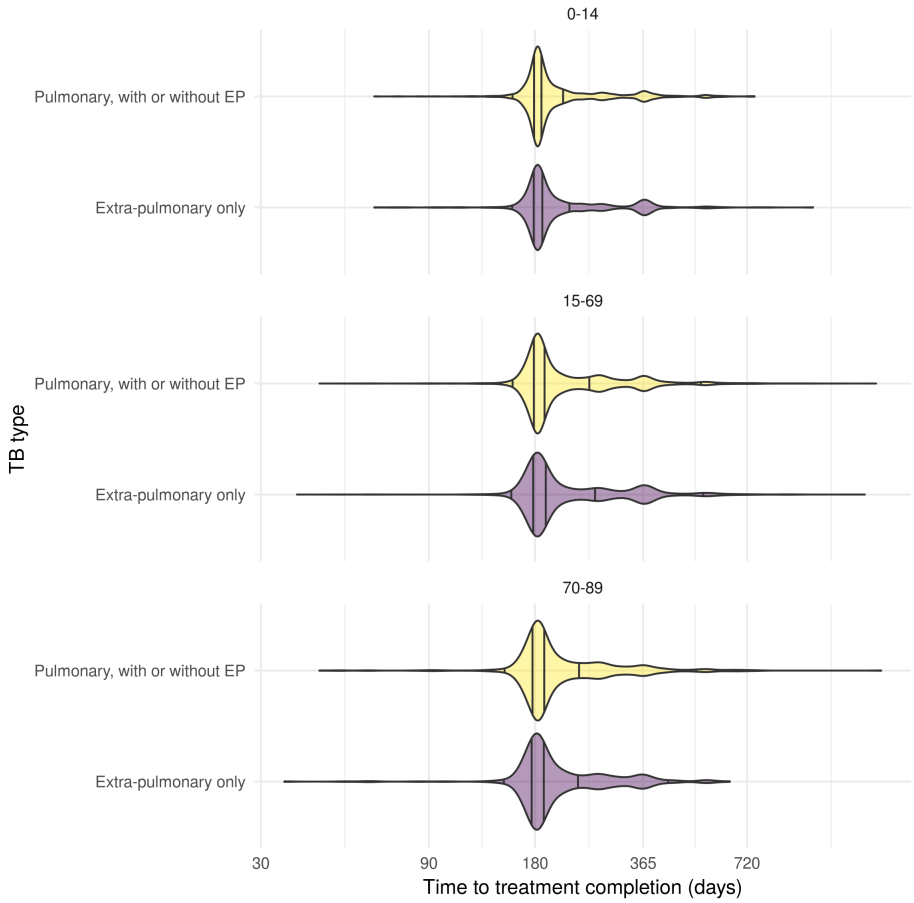


Figure 8.4: Distribution of time to treatment completion in the UK born successfully treated (days), stratified by age group and pulmonary/extra-pulmonary TB status in the Enhanced Tuberculosis Surveillance system for notifications between 2000 and 2012. Age is stratified into three groups; children (0-14), adults (15-59) older adults (15-59). There is little evidence that the time to successful treatment differs between pulmonary and extra-pulmonary cases only but older cases appear to have a high likelihood of longer treatment times.

Age-stratified contact matrix The previously defined age-stratified contact matrix has 72 free parameters, assuming that the contact matrix is symmetric. Whilst these parameters could conceivably be fitted to the available age-stratified incidence data it is likely that doing so would result in over-fitting and potentially obscure other age related differences. An alternative is to specify the contact matrix using available data sources. This is commonly achieved using survey data on the number of self reported contacts between individuals.[95]

The POLYMOD contact survey The POLYMOD survey, which was conducted between May 2005 and September 2006, asked 7,290 participants across eight European countries (Belgium, Germany, Finland, Great Britain, Italy, Luxembourg, the Netherlands, and Poland) about the number of unique contacts on a randomly assigned day of the week. Sur-

vey participants were recruited to be broadly representative of the population in terms of geographical spread, age, and sex. Children and adolescents were deliberately over-sampled due to the important role they typically play in the transmission of infectious diseases. Contacts were defined as either physical (skin-to-skin contact) or as nonphysical (two-way conversation of 3 or more words in the presence of an individual but without physical contact). The age and gender of contacts was recorded as was the duration and location of the contact event. The locations were stratified into: home; school; work; transport; leisure; and other. In total 97,904 contacts were recorded, with both physical and nonphysical contacts showing large amounts of assortativity by age.

In the model presented here unstratified social (nonphysical) contacts are used to generate an age-stratified contact matrix. There are several reasons for this. Firstly, stratifying by home, school, work, transport or leisure contacts whilst initially appealing as doing so may lead to insights as to the nature of the type of contacts required for TB transmission may also lead to over-fitting without a strong apriori hypothesis. In high and medium burden countries it has been shown that within household transmission is not a major driver of overall transmission.[85] Until recently it has been thought that household transmission plays a more dominated role in low burden settings, such as England, which would indicate that home contacts should be considered. However, it has recently been found that only 7.7% (1849/24,060) of cases in England between 2010 and 2012 lived in a household with another case.[96] The same study estimated that overall only 3.9% of cases were due to recent household transmission, and there was no evidence that cases within households were more likely to transmit within the household than outside of it. There is little evidence to suggest that school, work, transport or leisure contacts are more likely to transmit TB in England than any other contact. The choice of contact type is disease dependent; for TB it is likely that closer contacts result in a greater likelihood of transmission.[85] Unfortunately the physical contacts recorded in the POLYMOD survey represent a poor proxy to closeness of contacts as physical contact can be a little as a handshake and because TB is a respiratory disease physical contact is not required. For this reason physical contacts have not been further evaluated. Instead, the uncertainty in age-dependent transmission rates has been explored by allowing for scenarios in which the transmission probability varies between children and adults, between children, adults and older adults and in which it does not vary.

Generation of the symmetric contact matrix As the POLYMOD contact data was collected using a survey there is likely to be measurement error and missing data for the number of contacts reported and the age that contacts were reported to be. Some participants also recorded contacts with an estimated age range rather than with a point estimate. In addition, as the survey had a relatively low sample size (1,011) in the UK the estimated contact matrices contain considerable uncertainty. These considerations are often not considered in modelling studies but may introduce significant bias. Here the `socialmixr` R package is used to generate 1000 bootstrapped contact matrix samples using the following steps,

1. Missing or estimated ages are sampled from the appropriate ranges.
2. Using data on the participants of the POLYMOD study, and the contacts that they recorded, participants are randomly sampled (with replacement) and the mean number of contacts is then calculated from each age group (using 5 year age groups from

0-5 to 49, 50-69, and then 70+).

3. Each sampled contact matrix is then averaged to be symmetric, as logically contacts should be mutual. This can be represented mathematically as follows,

$$C_{ij}N_i = C_{ji}N_j \tag{8.16}$$

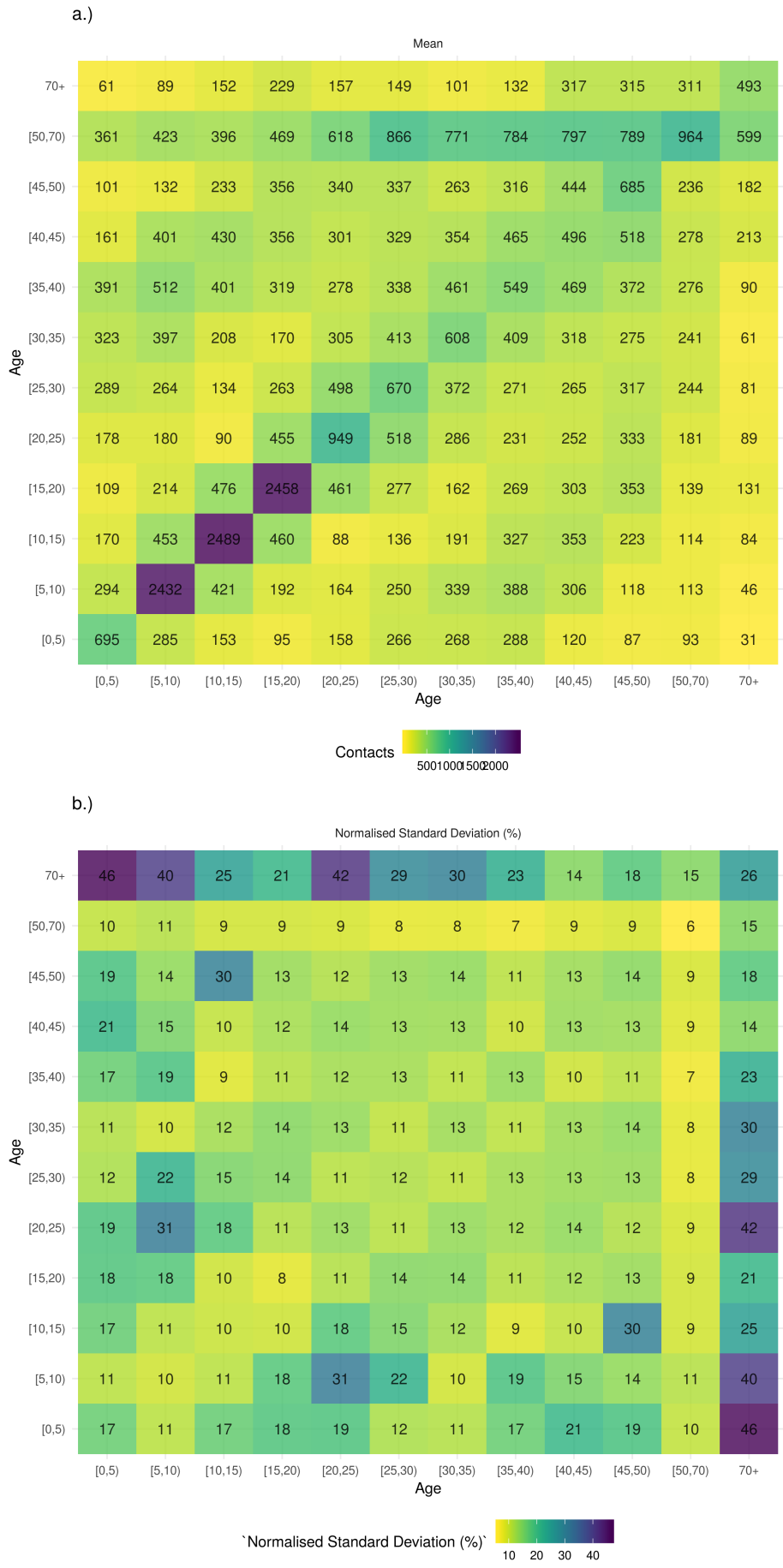
Where N_i is the number of people in age group i, N_j is the number of people in age group j, c_{ij} is the number of contacts between members of group i with group j and c_{ji} is the number of contacts between members of group j with group i. In the POLYMOD survey this relationship does not hold exactly due to random variation. A symmetric contact matrix (C'_{ij}) can be derived by averaging the contacts between the i and j groups and the j and i groups for all age groups using the following equation,

$$C'_{ij} = \frac{C_{ij}N_i + C_{ji}N_j}{2N_i} \tag{8.17}$$

The above equation requires data on the population in which the survey was undertaken in order to create a symmetric contact matrix. Here we use the 2005 population of the UK as it is most representative of the POLYMOD study population.

This results in 1000 bootstrapped symmetric contact matrices based on the reported social contacts in the POLYMOD survey for the UK. In order to be used in the model the mean and standard deviation are calculated for the number of contacts between each age group, the data is also scaled to represent non-unique yearly contacts by multiplying by 365.25. Contacts are then modelled noisily using a normal distribution around the mean number of contacts with the standard deviation as calculated above.

The final mean contact matrix is visualised in Figure 8.5, along with the normalised standard deviation. It is clear that the POLYMOD mixing is highly assortative with the majority of contacts occurring between those close to the same age. The highest number of contacts were between children and young adults (between 5 and 20), with the number of within age groups contacts reducing as age increased. There was some outside age group mixing for all age groups with a large amount of mixing between children and middle aged adults (i.e parents and children). There was some uncertainty for all contact rates with the minimum normalised standard deviation being 10% of mean contact rates. Contact rates between older adults and children were highly uncertain and contact rates for older adults were also generally more uncertain.


a.)
Mean
70+ 61 89 152 229 157 149 101 132 317 315 311 493
[50,70) 361 423 396 469 618 866 771 784 797 789 964 599
[45,50) 101 132 233 356 340 337 263 316 444 685 236 182
[40,45) 161 401 430 356 301 329 354 465 496 518 278 213
[35,40) 391 512 401 319 278 338 461 549 469 372 276 90
[30,35) 323 397 208 170 305 413 608 409 318 275 241 61
[25,30) 289 264 134 263 498 670 372 271 265 317 244 81
[20,25) 178 180 90 455 949 518 286 231 252 333 181 89
[15,20) 109 214 476 2458 461 277 162 269 303 353 139 131
[10,15) 170 453 2489 460 88 136 191 327 353 223 114 84
[5,10) 294 2432 421 192 164 250 339 388 306 118 113 46
[0,5) 695 285 153 95 158 266 268 288 120 87 93 31
Age
[0,5) [5,10) [10,15) [15,20) [20,25) [25,30) [30,35) [35,40) [40,45) [45,50) [50,70) 70+
Age
Contacts
500 1000 1500 2000
b.)
Normalised Standard Deviation (%)
70+ 46 40 25 21 42 29 30 23 14 18 15 26
[50,70) 10 11 9 9 9 8 8 7 9 9 6 15
[45,50) 19 14 30 13 12 13 14 11 13 14 9 18
[40,45) 21 15 10 12 14 13 13 10 13 13 9 14
[35,40) 17 19 9 11 12 13 11 13 10 11 7 23
[30,35) 11 10 12 14 13 11 13 11 13 14 8 30
[25,30) 12 22 15 14 11 12 11 13 13 13 8 29
[20,25) 19 31 18 11 13 11 13 12 14 12 9 42
[15,20) 18 18 10 8 11 14 14 11 12 13 9 21
[10,15) 17 11 10 10 18 15 12 9 10 30 9 25
[5,10) 11 10 11 18 31 22 10 19 15 14 11 40
[0,5) 17 11 17 18 19 12 11 17 21 19 10 46
Age
[0,5) [5,10) [10,15) [15,20) [20,25) [25,30) [30,35) [35,40) [40,45) [45,50) [50,70) 70+
Age
`Normalised Standard Deviation (%)`
10 20 30 40

Vaccination model parameters

An overview of the vacination model parameters can be found in: Table 8.1 for parameters that impact the natural history of TB; Table 8.4 for parameters that impact the population level distribution of BCG vaccination; and Table 8.2 for an overview of the sources used to generate prior distributions. More detail is given, where required, in the following section.

Effectiveness of the BCG vaccine at preventing active TB The effectiveness of the BCG vaccine is usually estimated using it's effectiveness at reducing the incidence of active TB cases in a susceptible population. In the model outlined in this chapter the action of the BCG vaccine has been split into its main effect of reducing the rate of latent TB cases developing active disease and its secondary effect of reducing the likelihood of initial infection. There are few estimates of the effectiveness of the BCG vaccine at preventing active TB in cases that are already latently infected and where these estimates do exist they are not stratified by time since vaccination, or age at vaccination.[3] As the overall effectiveness (α_a^T) of the BCG vaccine can be estimated from the combined effectiveness at preventing initial infection (χ_a) and then preventing activation in latently infected individuals (α_a) using the following equation,

$$\alpha_a^T = \chi_a + (1 - \chi_a)\alpha_a \tag{8.18}$$

The effectiveness of the BCG vaccine at preventing active TB in those latently infected can then be found via rearrangement as follows,

$$\alpha_a = \frac{\alpha_a^T - \chi_a}{1 - \chi_a} \tag{8.19}$$

There is strong evidence that the overall effectiveness of the BCG vaccine reduces over time.[22,23] For this reason the effectiveness of the BCG vaccination overall (α_a^T) has been stratified by the time since vaccination (by 5 year age groups). This step-wise approach has been chosen as the majority of studies report estimates for these groups and the precise functional form of the reduction in protection is unknown. For 0-5, and 5-10, years since vaccination estimates of the effectiveness of the BCG vaccine were extracted from the MRC trial.[20] Using the published data, Poisson regression was used to estimate Rate Ratios and 95% confidence intervals. For 10-29 years after vaccination Rate Ratio estimates from a more recent case control cohort study in the UK born vaccinated at school-age have been used.[23] Table 8.3 details the estimated effectiveness for each five yearly band after initial vaccination. I have assumed that the BCG vaccine is equally effective regardless of when the vaccine is given as there is no evidence that protection reduces when given to older age groups in England. Using the literature derived estimates for the Risk Ratio (RR) of the BCG vaccine at different periods after vaccination, the log normal approximation for the distribution of Risk Ratios, and the relationship between vaccination effectiveness and the Risk Ratio (Effectiveness = 1 - Risk Ratio) I derived a prior distribution for the overall effectiveness of the BCG vaccine (α_a^T). This can be summarised by the following equation,

$$\alpha_a^T = 1 - e^{\mathcal{N}(log(RR),SE)} \tag{8.20}$$

Where SE is the standard error of the logged Risk Ratio. The transformed values used as the prior distribution are detailed in Table 8.1.

Table 8.3: Estimates of the effectiveness of the BCG vaccine at preventing active TB disease stratified by years since vaccination. For 0-9 years since vaccination estimates were derived using Poisson regression from the MRC BCG trial and for 10-29 years since vaccination estimates were extracted from a more recent case control cohort study in the UK born vaccinated at school-age.[@Hart1972; @Mangtani2017] BCG effectiveness at preventing active TB disase used as prior distributions in the model.

Time since vaccination (years)	Effectiveness (%)
0-4	84 (76, 90)
5-9	69 (51, 81)
10-14	56 (33, 72)
15-19	57 (36, 71)
20-24	25 (-10, 48)
25-29	21 (-39, 55)

Effectiveness of the BCG vaccine at preventing initial infection Roy et al. published a meta-analysis that estimated the effectiveness of the BCG vaccine at preventing initial infection in children.[3] This has been used as the primary source for this parameter, with the assumption being made that the effectiveness is the same in adults as it is in children. This is reasonable to assume as there is little evidence the the overall effectiveness of the BCG vaccines reduces with the age it is given in England. Unfortunately the meta-analysis by Roy et al. did not include an estimate of the effectiveness of the BCG vaccine at preventing initial TB infection stratified by time since vaccination. This is problematic as there is a large amount of evidence that the overall effectiveness of the BCG vaccine wanes with time,[22,23] and if the protection from initial infection does not also reduce over time then as overall effectiveness decreases the contribution from the prevention of initial infection will increase. For this reason I have assumed that the protection from initial infection reduces over time with the same functional form as for the overall effectiveness of BCG vaccination. This relation can be formalised using the following equation,

$$\chi_j = \frac{\alpha_j^T \chi_i}{\alpha_i^T} \tag{8.21}$$

Where i is the age at vaccination and j is any subsequent age group.

Demographic model parameters

The demographic model paramters are outlined in Table 8.4, additional details are given in the following section. Table 8.2 contains details of the sources used to parameterise the demographic model, again more detail is given in the following section for complex parameters.

Table 8.4: Demographic model parameters, descriptions, prior distributions, units, method used to derive the prior distribution and the type (i.e data derived, literature, assumption). $\mathcal{U}$ = Uniform and i = age at vaccination.

Parameter	Description	Distribution	Units	Method	Type
$\omega(t)$	Time varying births	-	-	The dataset contains the estimated number of births from 1929-2015 in England. From 2016 onwards the numbers of births are projections as published by ONS.	Derived from data
γ	BCG vaccination coverage	$\gamma_i = \mathcal{N}(0.8, 0.05)$ Where i is the age group vaccinated.	Proportion	England has a robust national health service and an established system for providing BCG vaccination.	Assumption
θ	Rate of ageing	-	$years^{-1}$	Defined as the 1 over the width of the modelled age groups.	Model defined
$\mu^{all-cause}(t)$	Time varying all-cause age-specific mortality rate	-	$years^{-1}$	Age specific mortality averaged across age group from 1981-2015. From 2016 onwards, and prior to 1981, mortality rates are modelled using a exponential model fit to data from 1981 until 2015.	Derived from data

Age-stratified population estimates Age-stratified and UK birth stratified population estimates for England were estimated using the Labour Force Survey (Chapter 4). Figure 8.6 indicates that the age distribution of the UK born population changed over the study period (2000 to 2015), with an increase in those in late middle age and older and a decrease in those in early middle age. The proportion of young adults and young children also increased. This may have impacted TB incidence as young adults are thought to be responsible for the majority of transmission. Data from the 1931 census was also used to estimate the population of England in 1931 stratified into the modeled age groups.

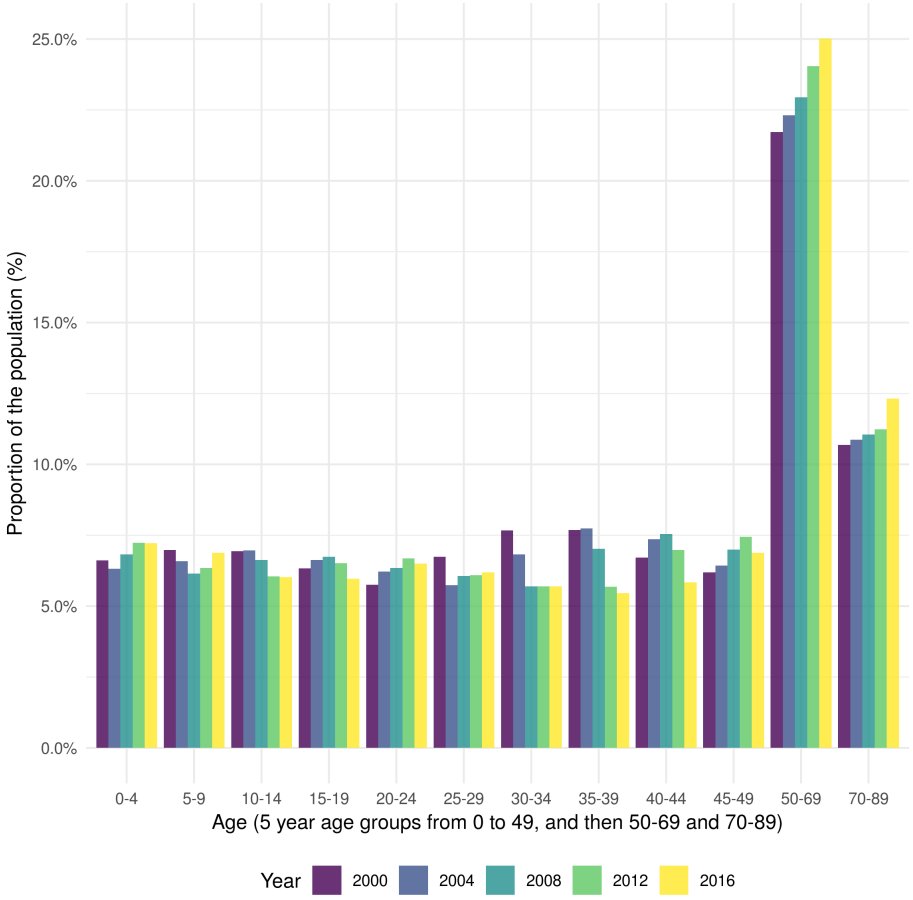


Figure 8.6: Distribution of the UK born population of England in 2000, 2004, 2008, and 2012. Age is grouped into 5 year age groups from 0 to 49, from 50-69, and from 70 to 89. Those aged 90+ are excluded due to low quality data. The age groups used here represent those used in the model. The figure indicates that the population has skewed older overall over the last two decades, although the proportion of young children has increased in the last 10 years.

Observed and projected births The number of births is incorporated into the demographic model as a time varying, noisy, parameter ($\omega(t)$). It is parameterised from the data published by the Office for National Statistics (ONS), with the available data covering all years modeled. The ONS publishes the recorded number of births in England each

year starting from 1929 through to 2015, with projections avaiable through to 2101 (Figure @ref(fig:births_england)). As there is some uncertainty as to the number of births in each year I included normally distributed noise with a standard deviation of 5% of annual births.

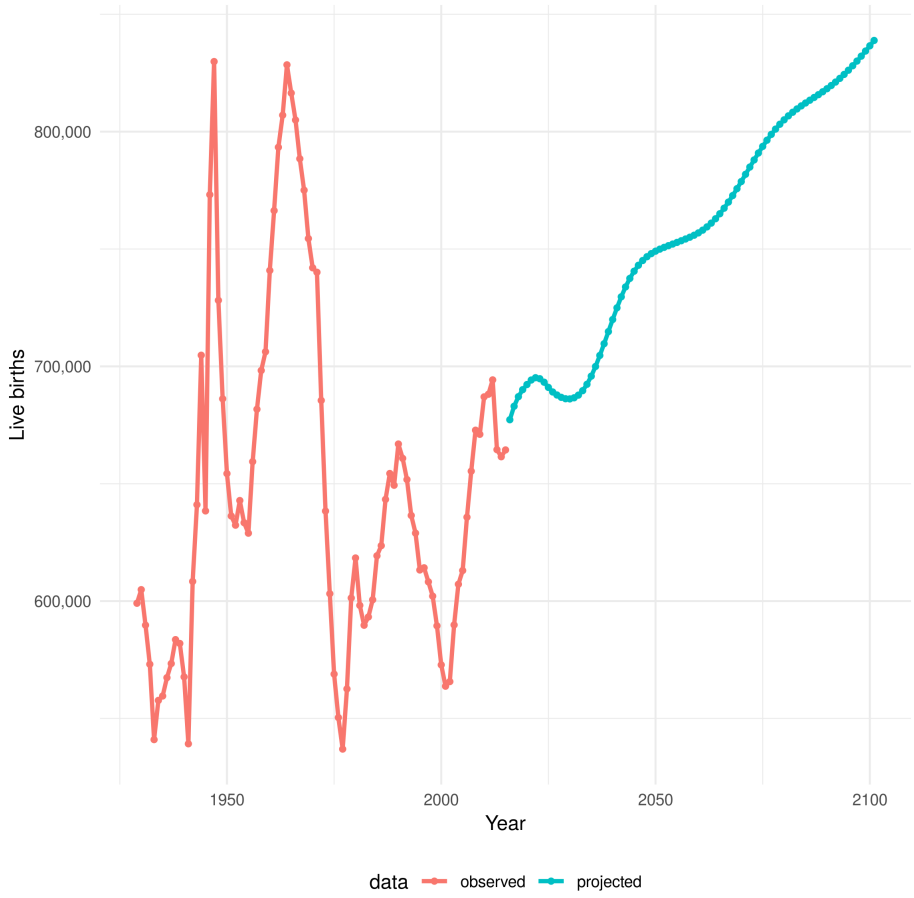


Figure 8.7: Estimated and projected live births in England from 1929 until 2101. The red line indicates estimated data and the blue line indicates projected data. Data is sourced from the Office for National Statistics (ONS).

Age-specific mortality rates The time varying, age-specific, noisy, all-cause mortality rates ($\mu_a^{all-cause}(t)$) included in the demographic model are sourced from Office for National Statistics (ONS) estimates from 1981 until 2015. For years outside of the available data mortality rates are forecast using an age-stratified exponential model (8.8. This model was used as it constrains mortality rates above zero and decreases yearly changes in mortality rates over time. To model the uncertainty in the estimate of the annual number of deaths a normally distributed noise term was introduced with a standard deviation of 5%. In order to calculate the all-cause dynamic mortality rate ($\mu_a^{all-cause}(t)$), excluding deaths from, or related to, TB the following equation is used,

$$\mu_a(t) = \mu_a^{all-cause}(t) - \left(\frac{\mu_a^P(P_a + T_{Pa}) + \mu_a^E(E_a + T_{Ea})}{N_a} \right) \tag{8.22}$$

Where $\mu_a(t)$ is constrained to be greater than or equal to zero.

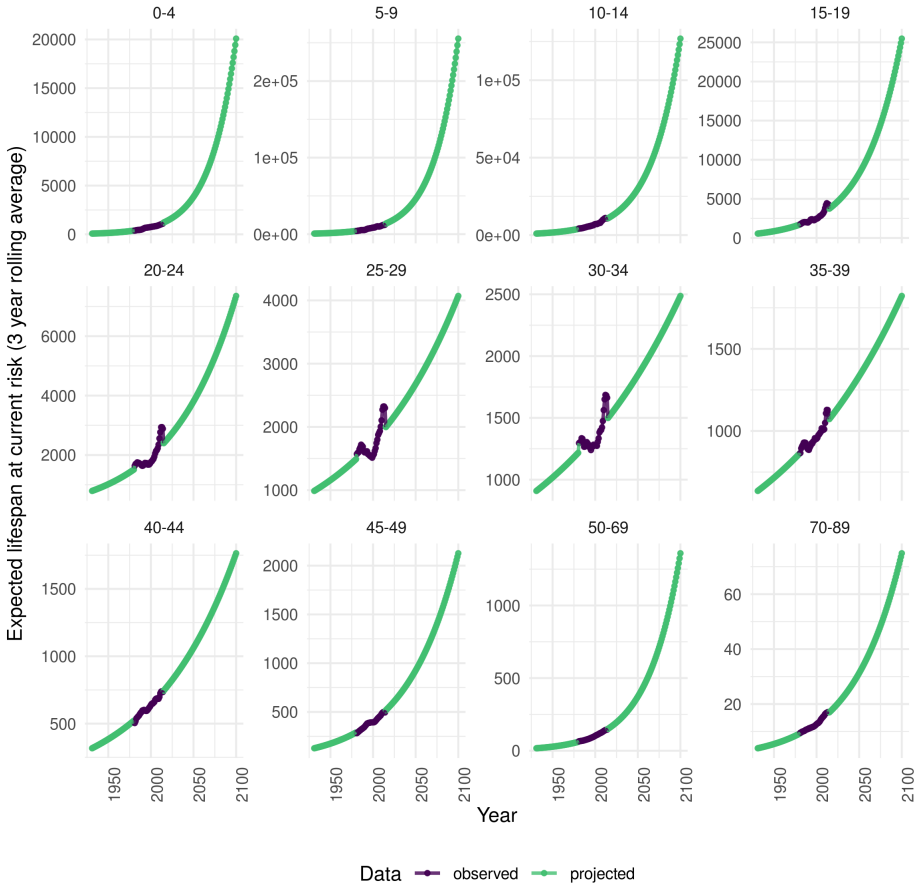


Figure 8.8: 3 year rolling average expected remaining lifespan stratified by age group in England from 2000 to 2014. Age is grouped into 5 year age groups from 0 to 49, from 50-69, and from 70 to 89. Those aged 90+ are excluded due to low quality data. The age groups used here represent those used in the model. Data from this figure was sourced from the Office for National Statistics age-specific mortality rate estimates.

8.6 Initialisation

Dynamics transmission models are susceptible to the conditions under which they are initialised.[5] For models of endemic disease this can be problematic as the full disease outbreak can often not be modeled, due to a lack of data and the changing nature of the endemic over time. A common approach to minimise this issue is to initialise the model with an uninformative set of initial conditions and then run the model for a period of time, known as the burn-in period, until steady state dynamics have developed.[5] Models that include demographic processes are more complex to burn-in as demographic data is typically required to initialise the model so that it has the demographics observed during the period of time modeled.

8.6.1 Starting simulation date, initial population and changes over time.

Model simulations were begun in 1931 due to the availability of population data from the 1931 census and because data on live births was only available until 1929. The demographic model is initialised using the age grouped 1931 census data, with the assumption that the entirety of the population is UK born. Initially it is assumed that there is no BCG vaccination and no TB treatment. This means that the coverage of BCG vaccination is zero. The rate of recovery from active disease is to be a minimum of 1/20, this reflects natural clearance of TB. TB treatment is assumed to begin in 1952 with the discovery of isoniazid and BCG vaccination begins at school-age in 1953. BCG vaccination coverage is assumed to vary randomly over the time horizon of the model but to have the same distribution. The assumed distribution is normal with a mean of 80% and a standard deviation of 5%. The duration with active TB is assumed to decrease from the introduction of treatment in 1952 through to 1990 when it is assumed that detection rates were equivalent to those seen today. I have assumed the following logistic model for the increase in the rate of starting treatment,

$$\nu_a(t) = \nu_a \frac{1}{1 + exp(\nu_{scale} * (t - 1952) - (1990 - 1952)/2} \frac{log(t - 1952)}{log(1990 - 1952)} \tag{8.23}$$

This leaves ν_{scale} as a free scaling parameter.

8.6.2 Initial disease distribution

The model is initialised with the number of pulmonary and extra-pulmonary cases reported in 1931. To account for possible measurement error a normal distribution is sampled around the number of cases reported with a standard deviation of 5% of the reported cases. Cases are distributed across age groups based on the age distribution of the population.

8.7 Discussion

In this chapter, I have outlined the requirements for a dynamic transmission model of TB in order for it to be able to answer policy relevant questions relating to BCG vaccination in England. I then outlined, and gave the equations for, a semi-stochastic model that met these requirements and made best use of the data available. Finally, I outlined prior distributions for each model parameter and initialisation conditions. I detailed the data sources used for parameterisation, approximations required to make best use of the available data, and the scenario analyses needed to explore model, parameterisation and initialisation assumptions.

All dynamic transmission models require a series of assumptions to be made. These assumptions fall into two categories: structural assumptions and parameter assumptions.[97] Structural assumptions, such as the choice of serial latency in the model presented here, maybe difficult to test as they require the development of a parallel model structure. In the model presented here I have chosen to base the model structure on the known epidemiology of TB in England and the effects of the BCG vaccine. Structural assumptions have been discussed as have their potential impacts but a full scenario analysis of all potential model structures is beyond the scope of this work. Instead, I have focused on parameter assumptions which are more likely to directly impact the evaluation of BCG vaccination.

Table 8.5: Summary of planned scenario analyses to be carried out in the next chapter as part of model fitting by comparision of the goodness of fit to the data.

Parameter	Scenario
β_a - transmission probability	Constant across age groups
	Variable between children (0-14) and adults (15-89)
	Variable between children (0-14), adults (15-59), and older adults (60-80)
Historic non-UK born cases	Increasing from 1960 to 1999 based on notifications in 2000 using a log link function.
	Linear increase from 1960 to 1999 based on notifications in 2000

In the next chapter, I will consider an age modified transmission probablity, and the change in historic non-UK born cases over time as distinct scenarios (Table 8.5). Identifying if the transmission probablity varies with age is important as it may alter the distribution and number of TB cases. This would impact the observed effects of the BCG vaccine and is therefore of primary importance. The historic number of non-UK born cases is likely to be less translatable to other models, or scenarios, but is important to fully investigate here as it determines the proportion of the population with different stages of latent TB in 2000 when data on TB notifications is first available. This is important as a higher proportion of latent TB cases in the population leads to a greater number of re-activations. This in turn results in a reduction in TB transmission when fitting to observed incidence. This chapter has outlined a realistic dynamic transmission model of TB that includes key features required to investigate BCG vaccination policy and is robustly parameterised from an extensive, and previously unused in a TB model, routine surveillance data-set. Transformations, and approximations, of parameters have been used to make the best use of available data. However, there are several key limitations. Firstly, the model presented here does not explicitly model TB transmission in the non-UK born. This means that in order to initialise the model assumptions must be made about the historic number of non-UK born cases and the future incidence in the non-UK born must also be assumed in order to produce projections of future TB incidence. However, this simplification allows many complexities of TB in the non-UK born to be discounted, such as the rate of case importation, heterogeneity amongst the non-UK born from different countries, and mixing within the non-UK born. It also means that, as stated above, explicit assumptions must be made about the number of historic, and future, TB cases in the non-UK born. This makes the impact of these assumptions clearer as a scenario analysis can be conducted, whereas if non-UK born cases were explicitly modeled incidence in these populations would depend on more abstract assumptions. Finally, the majority of the non-UK born population would not be exposed to England's BCG vaccination policy and so are unlikely to be impacted by changes in this

policy. Secondly, the model presented here does not include high and low risk stratification within the UK born. Individuals that are from countries with incidence above 40 per 100,000, or that have parents/grandparents from countries with incidence above 40 per 100,000, are considered at higher risk of TB.[41] In addition, individuals living in areas of the UK with incidence above this threshold are also considered at higher risk. Current BCG vaccination policy targets high risk neonates for vaccination, with low risk neonates not being vaccinated. Ideally this high/low risk stratification would be included as it would allow the evaluation of the current BCG vaccination policy. Unfortunately this has not been possible as there is little data from which to extrapolate either the number of high risk notifications in the ETS, or the size of the high risk population. There is also little evidence to suggest the degree of mixing between the non-UK born, the high risk UK born population and the low risk non-UK born population. It is likely that introducing this structure into the model, without the data outlined above, would lead to the model being poorly specified and therefore failing to fit to the observed data. Instead, in the final chapter in this thesis, the high risk neonatal programme will be proxied by a universal neonatal programme. This will allow for a comparison to be made between school-age and neonatal vaccination but does not allow for the impact of targeting high risk individuals to be evaluated. Finally, the model presented here does not include the full complexity of TB epidemiology. Drug resistant TB may have more severe outcomes, standard TB treatment may fail resulting in a longer period on treatment, and TB outcomes may vary by risk group.[1] However, drug resistant TB cases are known to make up a small fraction of TB cases in England in the UK born, variable treatment times have been included in the prior distribution of treatment times and TB outcomes. Model parameters have also been stratified by pulmonary status and age group where appropriate. Additionally, complexity has been included for the action of the BCG vaccine, with realistic waning in effectiveness. Observed age-specific mortality rates and the number of live births has also been included, allowing for realistic population demographics. This means that estimates of the impact of the change in vaccination policy are likely to be more accurate, whereas a more complex model of the epidemiology of TB would likely have little impact on these results.

There are several key differences between the model presented here and others that have been previously been published that modeled TB transmission in low incidence settings or evaluated BCG vaccination policy. These are: the inclusion of dynamic TB transmission; robust parameterisation from an extensive surveillance data-set; realistic population demographics; and detailed modelling of the action of the BCG vaccine. Several previous studies have evaluated the role of BCG vaccination at a population level and estimated the impact of targeting different age groups and populations. Manusseri et al. estimated the impact of various BCG vaccination strategies in low-intermediate incidence settings using an annual risk of infection model based on an approach previously published by Trunz et al. [**???**,98] This approach estimates the number of new cases generated by a single smear positive case per year in a birth cohort. Only a single year of data was used to parameterise the model and age structure, the duration of protection from BCG vaccination and the different types of protection conferred by BCG vaccination were not considered. Rahman et al. compared the cost effectiveness of universal BCG vaccination to no vaccination using a cohort model of Japanese infants.[99] Their model did not include TB transmission and used an estimated duration of protection from BCG of 10 years. Similarly Usher et al. used a decision analytical model to follow a birth cohort to compare universal, selective or no BCG vaccination.[100] As in the previous study, TB transmission was not included. The

model I have presented here includes TB transmission and uses more recent estimates of the effectiveness of the BCG vaccine to capture the full benefits of vaccination.

Several studies have made use of dynamic TB transmission models to evaluate BCG vaccination or future vaccines.[101–103] In general these studies used less detailed models than the one presented here, typically because they were modelling TB in a more generic setting or because more information about TB epidemiology, TB natural history, and the BCG vaccine has become available over time. In addition, no dynamic model of TB, including BCG vaccination, has currently been published that includes both protection from initial infection and protection from active TB due to BCG vaccination. There have also been no studies that use the current best estimates for the duration of BCG protection in developed countries away from the equator. Harris et al. reviewed mathematical models that explored the epidemiological impacts of future TB vaccines. They found that vaccines targeted at all-ages or at adolescents/adults were more effective at eradicating TB than neonatal programmes when vaccine effectiveness was not assumed to degrade with age. The majority of studies included in their review used deterministic, compartmental, dynamic models. Model structures were found to have evolved over time as TB natural history and epidemiology is better understood, with the majority of models having at least susceptible, latent, active disease, and recovered states. Treatment status, variable infectiousness of active disease, vaccine waning, and age stratification were included in some of the models evaluated.[101] Recently it has been shown that only models that include at least two latent compartments are able to reproduce the observed activation dynamics of TB.[84] The model presented here is based on the serial latency archetype identified in this study. It has also been shown that realistic age structure and population demographics, included in the model presented in this chapter, are critical for reproducing TB epidemiology.[104] Egbetade et al. presented a dynamic model of TB that included BCG vaccination but did not include age structure. They found that universal vaccination increased the stability of the disease free equilibrium in countries with high TB burden. However the model presented was not rigorously parameterised with data and only a single latent TB compartment was used.[102] Bhunu et al. developed a dynamic transmission model of TB that in order to investigate the effects of pre- and post-exposure vaccines for tuberculosis control. Again their model did not include multiple latent compartments or age structure unlike the model presented here.[103]

Vynnycky et al. modeled the long term dynamics of pulmonary TB, in England and Wales, in the white male population using a deterministic TB transmission model that included; high and low risk latent periods, reinfection, BCG vaccination, TB specific and all-cause mortality.[105] Whilst this is a highly detailed and well parameterised modelling study more recent developments such as survey derived age stratified contact matrices, evidence that BCG provides protection against initial infection as well as active TB disease and parameter estimates for TB activation stratified by age are included in the model presented here. In addition, their study only modeled TB transmission until 1990, allowing them to ignore the contribution of non-UK born cases. The model presented in this chapter includes non-UK born cases, via the force of infection, as they are now thought to be a key driver of TB transmission in England.

Dowdy et al. presented a data wish list for evidence base decision making using TB models, which maybe used to assess the usefulness of a TB model for policy makers. The data requirements included: the rate of TB transmission; probability of developing active disease after an initial infection; the rate of activation amongst cases with risk factors; protection

afforded by latent TB infection; the duration of infectiousness; treatment success; and the rate of spontaneous recovery. The model presented here fulfills the majority of these criteria. The rate of TB transmission is parameterised using previously published estimates of the effective contact rate in England,[94] this parameterisation will be refined in the following chapter using incidence data from the ETS. The probability of developing active TB has been sourced from recently published modelling work that fit a model of TB transmission to contact data in low incidence countries,[84] and is stratified by age as considered important by Dowdy et al. The rate of activation amongst cases with risk factors has not been included as it has been assumed that the proportion of UK born cases in the ETS with risk factors such as HIV is low. The duration of infectiousness, and treatment success have been parameterised using the ETS, although this approach is limited by possible reporting biases in the data available. The rate of spontaneous recovery has not been modeled as it is assumed that individuals are likely to be notified before clearing TB and are also likely to rapidly be started on TB treatment. This assumptions is likely to be valid as England has a robust national health service and a strong notification framework for TB. The protection afforded by latent TB infection has been included using the most recent literature sources available. All other parameters have been parameterised using the ETS were possible and otherwise from the most robust literature sources available. In particular the effectiveness of the BCG vaccine has been parameterised using data from studies that took place in England, where available, and both the protection from initial infection and the protection from developing active disease in those latently infected has been included along with estimates of the reduction in protection over time.

The semi-stochastic transmission dynamic model of TB transmission and BCG vaccination presented in this chapter provides a detailed overview of the features required to reproduce the observed epidemiology of TB in England. The model was robustly parameterised using routine surveillance data where available and otherwise using literature sources. The assumptions required by the model can be explored by fitting the model to observed data and assessing the goodness of fit. This is the focus of the next chapter. In addition the model may also be used to explore the impact of current and historic BCG vaccination policy, both in the observed data and projected into the future. This is explored in the final chapter of this thesis.

8.8 Summary

- This chapter presents a semi-stochastic transmission dynamic model of TB transmission and BCG vaccination. The model includes; age structure, pulmonary and extra-pulmonary TB, re-infection and re-activation, serial latency, TB treatment, treatment failure, TB mortality, non-UK born cases and details of the historic TB endemic.

- The model has been robustly parameterised to a rich routine surveillance data set, which has allowed more complex features to be modeled than in previously published models. Parameter transformation and approximations, that make the best use of the available data, have been detailed.

- The assumptions required by the model have been explored in detail, with the required sensitivity analyses listed. These sensitivity analyses will be explored in the following chapter by comparing the goodness of fit of the model to the available data.

- The strengths and weaknesses of the model have been discussed as well as its context within the literature. It appears that few models are parameterised to a comparably rich surveillance data source, that few models capture the full complexity of BCG vaccination and that few models include realistic population demographics to the same extent as included in the model presented in this chapter.
- Chapter 5 used a simple simulation model to estimate the impact of the 2005 change in BCG vaccination policy and Chapter 7 used Poisson and Negative Binomial multilevel models to estimate the observed impact of the change in policy on incidence rates in the directly effected populations. Whilst these approaches are valid they cannot estimate the indirect effects of policy changes, nor can they predict the future impacts of BCG vaccination policy. For this a transmission dynamic model, as presented here, is required. In the following chapter this model will be fit to available TB data and the impact of various BCG vaccination policies will be explored.

Chapter 9

Fitting a dynamic transmission model of Tuberculosis

9.1 Introduction

In the previous chapter I outlined a mechanistic model of Tuberculosis (TB) transmission. Whilst this model made use of the best available evidence there remains a large degree of uncertainty regarding its structure and parameterisation. The majority of this uncertainty relates to the amount of TB transmission occuring in England. An approach to deal with this uncertainty is to fit the model to available observed data. Model fitting involves optimising over the available parameter space to return parameter sets that fit the data in some quantative way "better" than other parameter sets.

This chapter details an approach to challenging a infectious disease model to data using the state-space model formulation and bayesian model fitting techniques. It first outlines the infectious disease model discussed in the previous chapter as a state-space model, as well as detailing the data used for fitting the model, and the parameters that are fitted. It then outlines the theoretical, and practical, justification for the model fitting pipeline used to calibrate and fit this state space model. Finally it discusses the quality of the model fit, strengths and limitations of the approach, and areas for further work.

9.2 Formulation as a state-space models

State space models (SSMs) may be used to model dynamic systems that are partially observed (such as all but the most contained infectious disease outbreak or endemic). They consist of a set of parameters, a latent continuous/discrete time state process and an observed continuous/discrete time process.[106] The model developed in the previous section represents the state process of the SSM, with the parameters estimated for the model representing the model initial conditions and parameter set. To complete the mapping to an SSM an observational model is required. This observational model takes the latent estimates from the dynamic model and forecasts the observed data. We specifiy such an observational model in the Section 9.2.2.

9.2.1 Observed data

Before the observation model can be detailed the data that will be used to fit the state-space model must be detailed.

The primary data source used was the reported UK born TB notifications from 2000 to 2004 as recorded by the Enhanced Tuberculosis Surveillance (ETS) system (see Chapter 4). These are the only years for which notifications are available stratified by UK birth status and for which universal school-age BCG vaccination was in place. As the model of TB transmission outlined in the previous chapter had multiple age-dependent parameters UK born TB notifications stratified by age group are also fitted to. The age groups considered are children (aged 0-14 years old), adults (15-69 years old), and older adults (70-89 years old). These age groups were used, rather than the more detailed 5 year age groups used in the TB transmission model (see Chapter 8), for three reasons. Firstly, by using condensed age groups the number of notifications in each group increased. Having a larger number of notifications in each group reduces the impact of stochastics noise, which makes fitting the model easier. Secondly, reducing the number of data points, whilst still capturing the important age dynamics, reduces the compute requirements of the model (see Section 9.3.2). As will be discussed in Section 9.3 this was a major consideration as the model fitting approached used was highly compute intensive. Finally, the condensed age groups used were chosen to be meaningful. This allows model results to be more easily reported and interpreted.

In addition to the notification data available via the ETS pulmonary TB notifications (including both UK born and non-UK born cases) from 1990 were also fitted too. These TB notifications made to Statutory Notifications of Infectious Diseases (NOIDS) from 1913 to 1999. These data were sourced from Public Health England,[1] and made available in R using `tbinenglanddataclean`[1]. Data from 1990 was used as using data from a decade prior to the time period of interest allows the long term trends to be fitted to. Only a single data point was used as this limited the impact on the compute time of the fitting pipeline and would have brought only marginal improvements to the model fit (see Section 9.3.2).

9.2.2 Observational model

There are three major considerations to account for when developing an observed disease notification model (i.e a reporting model). These are: systematic reporting error over time; systematic changes in reporting error over time; and reporting noise. For the notification data outlined in the previous section there is little evidence to suggest the form that this measurement model should take. For this reason I have assumed that all reporting errors are gaussian and that their are no time variable reporting errors. The reporting model can be then defined as follows,

$$O = \mathcal{N}\left(E_{syst}A, E_{noise}A\right)$$

Where O are the observed notifications, A are the incident cases of disease as forecast by the disease transmission model, E_{syst} is the systematic reporting error, E_{noise} is the reporting noise, and $\mathcal{N}$ represents the gaussian (normal) distribution. The priors for the model are

[1]Historic TB notification data via `tbinenglanddataclean`: `https://www.samabbott.co.uk/tbinenglanddataclean/`

defined in Table 9.1. The prior for systematic reporting error is based on the assumption that cases are less likely to be reported than they are to be over reported. The prior for the reporting noise is an assumption. This observation model was also used for the non-UK born pulmonary cases.

A potential limitation of this model is that reporting of TB cases is likely to have improved over time. This is especially true of notifications reported prior to the introduction of the ETS in 2000 (see Chapter 4). A potential improvement to this model would be to introduce seperate systematic reproting errors for notifications pre and post the introduction of the ETS. However, given that I have only included a single point prior to the introduction to the ETS doing so in this instance would lead to overfitting of these parameters.

Table 9.1: Measurement model parameters, descriptions, prior distributions, units, method used to derive the prior distribution and the type (i.e data derived, literature, assumption). $\mathcal{U}$ = Uniform

Parameter	Description	Distribution	Units	Method	Type
E_{syst}	Systematic reporting error of incident TB cases	$\mathcal{N}(0.9, 0.1)$ truncated to be greater tha 0.	Proportion	Assumption is that underreporting of TB cases is more likely that exact, or over reporting.	Assumption
E_{noise}	Magnitude of reporting noise for incidence TB cases.	$\mathcal{N}(0, 0.2)$ truncated to be greater than 0.	Proportion	Reporting noise should allow for over reporting in some years when in general underreporting is the norm.	Assumption

9.2.3 Fitted parameters

The model outlined in the Chapter 8 has a large number of free parameters for which prior distributions have been specified based on the observed data, the literature, and expert knowledge. In theory the model fitting pipeline outlined below could be used to produce posterior distributions for all these parameters. However, in practise this is not feasible as the data discussed in Section 9.2.1 only covers notifications and therefore does not contain sufficient information. If every parameter was allowed to update based on the data then it is likely that the resulting posterior distributions would not match with alternative data sources and the literature. Another potential issue is that by allowing all parameters to be fitted the meaningful information in the observed data may be lost.

For this reason in the model fitting pipeline outlined here I have only allowed parameters relating to TB transmission, and measurement model parameters, to have their posterior distributions updated by the model fitting pipeline. All other parameters have posterior distributions that exactly match their prior distributions. Parameters that have updated posterior distributions based on the data are,

- Mixing rate between UK born and non-UK born (M).
- Effective contact rate (c_{eff}).
- Historic effective contact rate (c_{eff}^{hist}).
- Half life of the effective contact rate (c_{half}^{hist}).
- Systematic reporting error (E_{syst}).
- Reporting noise (E_{noise}).

In addition for scenarios with age variable transmission probabilities the following paramters may also be fitted to,

- Transmission probablity modifier for children (β_{child}).
- Transmission probablity modifier for older adults ($\beta_{olderadults}$).

9.3 Model fitting pipeline

9.3.1 Introduction

Fitting dynamic transmisson models is complex and requires the use of specialist statistical techniques. There are a variety of these tools available. Ranging for tried and tested to cutting edge. Historically many modellers have used maximum likelihood methods to fit deterministic models. Unfortunately, whilst computationally efficient, these methods only provide point estimates and are limited to relatively simple models. More recently Bayesian methods have become popular. These have numerous benefits including: explicit inclusion of prior knowledge via prior distributions for all parameters; ability to handle complex stochastic models; and provide parameter distributions (posterior distribution) of best fitting parameters rather than single point estimates. Unfortunately these methods also require calibration prior to use. This section outlines the theoretical justification, and implementation details, of an automated model fitting pipeline used to calibrate, and fit, the previously detailed state space model of TB tranmission in England.

Libbi was used for all model fitting.[106] LibBi is a software package for state-space modelling and Bayesian inference. It uses a domain specifc language for model specification, which is then optimised and compiled to provide highly efficient model code. It focusses

on full information model fitting approaches including: particle Markov chain Monte Carlo (PMCMC), and SMC^2 methods for parameter estimation. All fitting algorithms are highly efficient and scalable across mutliple CPUs or GPUs. `RBI` and `RBI.helpers` were used to interface with LibBi from R.[???,??] `RBI.helpers` was also used to optimise the model fitting pipeline as detailed in the calibration section. As model fitting using `LibBi` is compute intensive a workstation was built, and overclocked, with these compute requirements in mind[2]. All model fitting code is available on GitHub as an R package[3].

9.3.2 The particle filter

In order to fit a model to data it is neccessary to estimate, or calculate, the marginal likelihood. Mathematically, the marginal likelihood is the plausibility that a parameter set, given the specified statistical model and the initial conditions, describes the observed data. For complex state space models, such as that discussed in the previous chapter, calculating the marginal likelihood is not possible.[106] The particle filter provides a model-agnostic approach, based on importance sampling, to estimate the marginal likelihood. The variant used in this thesis, the bootstrap particle filter, is described below. See [106] for a more technical discussion of the bootstrap particle filter.

1. For a given parameter set the particle filter is initialised by drawing a number of random samples (state particles) from the initial conditions of the model under consideration. These samples are then given a uniform weighting.

2. Sequentially for the each observed data point the particle filter is then advanced through a series of *propagation*, *weighting*, and *resampling* steps.

- *Propagation*: For each particle the model is simulated, producing a forecast of the observed data data point.

- *Weighting:* The likelihood of the new observation, given the predicted state, is then computed for each state particle. State particles are then weighted based on this likelihood.

3. The marginal likelihood (likelihood of the observed data given the parameter set, marginalised across the initial conditions) can then be estimated by taking the product of the mean likelihood at each observed data point. A sample trajectory can also be calculated using the estimated weights from each time point.

The particle filter has been shown to provide an unbiased estimate of the likelihood for arbitary state-space models.[106] As a full information technique the particle filter provides a more accurate estimate of the likelihood than other approximate techniques (such as approximate bayesian computation), with relatively little tuning or user interaction.[106] The downside of the particle filter, compared to these approaches, is the high compute requirements, with each particle requiring a full model simulation. For highly complex models, the particle filter approach may not be tractable or a reduced level of accuracy of the marginal likelihood estimate must be accepted.

[2]See these blog posts for details: `https://www.samabbott.co.uk/post/building-an-rstats-workstation/`, `https://www.samabbott.co.uk/post/benchmarking-workstation-xgboost/`, `https://www.samabbott.co.uk/post/benchmarking-workstation-benchmarkme/`

[3]See here for details: `https://github.com/seabbs/ModelTBBCGEngland`

9.3.3 Sequential Monte Carlo

The particle filter approach outlined above, is a member of a family of sequential monte carlo (SMC) methods. These methods all initialise particles and then follow the same *propagation*, *weighting*, and *resampling* steps as previously detailed. SMC may also be used to sample from the posterior distribution of a given set of priors and a specified model. This works as follows,

1. Initially a number of samples (parameter particles) is taken from the prior distribution of the parameters and assigned a uniform weighting.

2. These parameter particles are then iterated sequentially over each observed data point, undergoing the same *propagation*, *weighting*, and *resampling* steps as in the particle filter, as well as an additional *rejuvenation* step.

    - *Propagation:* The model is simulated to the next observed data point.

    - *Weighting:* Parameter particles are weighted using the marginal likelihood. In principle this could be computed exactly, but is most commonly estimated using a nested particle filter for each state particle (i.e as outlined in the previous section). For a subset of models a Kalman filter maybe used instead.[106] The marginal likelihood may also be estimated using other partial information techniques such as approximate bayesian computation. In the case where a particle filter is used the full algorithm is known as Sequential Monte Carloˆ2 (SMCˆ2). This algorithm is used for all dynamic model fitting in this thesis.

- *Resampling:* The parameter particle stock is restored to equal weights by resampling particles, with replacement, with the probability of each sample being drawn being proportional to its weight.

- *Rejuvenation: Resampling* of the parameter particles at each time point leads to a reduction in the number of unique values present. For state particles (when estimating the marginal likelihood using a particle filter) particles are diversifed with each propagation but as parameters do not change in time parameter particles cannot diversify in this way. To account for this the *rejuvenation* step is inserted after the resampling of parameter particles at each time point. The *rejuvenation* step is a single, or multiple depending on the acceptance rate, Metropolis-Hastings step for each parameter particle. This step aims to preserve the distribution of the parameter particles, whilst increasing their diversity. To minimise unnessary rejuvenation an effective sample size thresold can be used. This only triggers rejuveation when particle diversity has decreased below the target effective sample size thresold.

Marginal Metropolis-Hastings

The Metropolis-Hasting step may be used as a model fitting approach in it's own right (MCMC) when repeated sequentially. It works by proposing a new value from the proposal distribution, estimating the marginal likelihood using the attached particle filter (or using any other exact or inexact method), and then accepting or rejecting the move based on the acceptance probability.[106] Where the acceptance probability is given by,

$$\min\left(1, \frac{p(y(t_{1:T})|\theta')p(\theta')q(\theta|\theta')}{p(y(t_{1:T})|\theta)p(\theta)q(\theta'|\theta)}\right)$$

Where y is the observed data, θ is the current parameter set, θ' is some proposed parameter set sampled from some proposal distribution $q(\theta'|\theta)$.[106] By construction samples drawn using this rule are ergodic to the posterior distribution. This means that after convergance samples drawn using this rule may be considered as samples from the posterior distribution.

Strengths and limitations.

SMC has numerous advantages over MCMC approaches. The first of these is that MCMC aproaches are sensitive to their initial conditions. If a model has multiple local best fits MCMC may only converge to a single minima rather than fully exploring the posterior distribution. Multiple MCMC chains may be used to try and account for this but as each chain must be independently run to convergance only a few concurrent chains are likely to be practical. SMC on the other hand is initialised with a large sample from the prior distribution, meaning that local minimas are more likely to be explored. Parameter particle weighting and resampling then balances the contribution to the posterior distribution of these local minimas based on their fit to the observed data.

Additionally, MCMC approaches are by definition sequential,[106] although if they make use of particle filters these can be run in parallel. Increasing the number of particles in a filter may lead to an increase in the chain mixing rate of the MCMC chain but as particles numbers are increased any returns will decrease. To account for this multiple chains are often used, but as outlined above the burn-in required for each chain limits the potential speed-up. In comparision, each SMC parameter particle can have it's marginal likelihood computed seperately. Although the resampling step remains a bottleneck as it can only be completed once all marginal likelihoods have been computed.

On the other hand SMC is less interative than MCMC meaning that model fitting is harder to inspect when it is in progress. This is because SMC is not sequential, unlike an MCMC run for which each draw can be inspected as it is computed. Similarly, as SMC is not a sequential technique multiple runs cannot be combined. This means that model fitting must be done in a single run using a priori knowledge to judge the number of MCMC rejuvernation steps required, and the expected total run time. SMC will also have a variable run time based on the effective sample size as rejuvernation only happens when parameter particles have been depleted beyond a certain point.

9.3.4 Calibration

Particle calibration

The accuracy of the marginal likelihood estimate returned by the particle filter is dependent on the number of particles used, the number of observed data points, the parameter sample, and the complexity of the model. As the number of particles tends towards infinity the likelihood estimate provided by the particle filter should tend towards the exact solution. This suggests that choosing a very high number of particles maybe the optimal solution in terms of accuracy. Unfortunately, each particle requires a full model simulation, which for

complex models can be computationally costly. This means that using very large numbers of particles is not tractable. For this reason it is neccassary to determine an optimal number of particles that both provides an adequately accurate estimate of the likelihood whilst being computationally tractable.

The `rbi.helpers` R package attempts to solve this issue by adopting the following strategy.[**???**] First, the approximate mean of the posterior distribution is obtained, as accurate likelihood estimates near the posterior mean are of the most interest. Repeated model simulations are then run using the same set of parameters, with the marginal likelihood being estimated each time using a given number of particles. The variance of these log-likelihood estimates is then calculated. This process is then repeated for increasing numbers of particles until the log-likelihood variance is below some target thresold, commonly 1.[**???**]

I have implemented this as a two step process for each fitted scenario. Firstly, I used the Nelder-Mead simplex method, via LibBi,[106] to find a parameter set that optimised the maximum likelihood. I then initialised a 1000 step PMCMC chain with this parameter set, using 256 particles in the particle filter. I then used `rbi.helpers`,[**???**] as outlined above, to estimate the number of particles required to produce a log-likelihood variance of less than 1 for this sample of the posterior distribiution, using 250 samples per step and starting with 64 particles. I initially planned to repeat this process for multiple draws from the posterior distribution but this proved to be infeasible given the compute available. A target of 5 for the log-likelihood variance was chosen as a smaller target could not be feasibly achieved given the compute resources available. Additionaly 256 was specified as the maximum number of feasible particles to use in the particle filter.

Proposal calibration

When using an MCMC alogorithm a proposal distribution is required to provide new parameter samples to evaluate. For SMC^2 a proposal distribution is required to inform the MCMC sampler that is run during the rejuvernation step. By default if no proposal distribution is provided Libbi uses the prior distribution.[106] The prior distribution can be an inefficient proposal distribution as it is likely to have a low acceptance rate (from the MCMC sampler).[106] Having a low acceptance rates means that many more MCMC steps are required to generate a successful parameter sample. This results in slow mixing and computationally expensive MCMC steps may make model fitting intractable.

A more efficient approach is to specify a proposal distribution that draws parameter samples that are closer to the current state of the MCMC chain than the overall prior distribution. There is an extensive literature examing how to optimise the proposal distribution to achieve an good acceptance rate. In practise it has been shown that a rate of between 10% and 20% is optimal for upwards of 5 parameters.[106] This strikes a balance between allowing the chain to fully explore the posterior distribution whilst still being as efficient as possible.

A simple approach to setting the proposal is to run a series of MCMC steps and then calculate the acceptence rate. Based on the acceptence rate the width of the proposal distributions can then be adapted. By repeating these steps multiple times a proposal distribution which gives an acceptance rate within the desired bounds can be arrived at. This adaption can either be independent for each parameter or dependent (taking into account empirical correlations). The `adapt_proposal` function, from the `rbi.helpers` R package,[**???**] implements this approach and is used in this model fitting pipeline. In many

models parameters are likely to have strong correlations (i.e between UK and Non-UK born mixing rate and effective contact rate) and in these scenarios it is likely that a dependent strategy for adapting the proposal distribution will more efficiently explore the posterior distribution. However, the downside of adapting the proposal distribution using dependent methods is that the resulting proposal is highly complex, is computationally expensive to compute and may breakdown in some areas of the posterior distribution.

In this model fitting pipeline I have used a maximum of 5 iterations of, manual, independent proposal adaption, drawing 250 samples in each iteration, starting with the prior distributions and halfing the scale each time. The proposal distribution is also truncated so that parameter samples cannot be drawn from areas without support from the prior distributions. As for the particle calibration, I initially used a maximum likelihood method to provide a point estimate of the best fitting parameter set, followed by 1000 PMCMC steps, using a 256 particle filter. This means that the proposal distribution is adapted near to the posterior mean rather than in the tails of the posterior distribution.

I chose to use, manual, independent proposal adaption method for several reasons. Firstly, when developing this pipeline the approaches implemented in `rbi.helpers` produced multiple transient errors in other `rbi` and `rbi.helpers` code. Secondly, the resulting dependent proposal distribution was highly complex, slow to compute, and diffult to debug. Finally, for SMC^2 efficient exploration of the proposal distribution is less important than when using MCMC alone as SMC^2 is initialised with multiple samples from the prior distribution. This means that multiple local maximas can be efficiently explored regardless of the proposal distribution used. The MCMC rejuvernation step then serves to provide additional samples from these local maximas. Proposal adaption was only carried out for the main model scenario with all other scenarios using this proposal distribution.

9.3.5 Model comparision

In the previous chapter multiple potential model structures were outlined, each of which could be valid based on theoretical considerations. The observed data can be used to identify which of these structures best reflects reality. This can be done using the deviance information criterion (DIC). The DIC is a hierarchical modeling generalization of the Akaike information criterion (AIC) and can be used to compare nested models.[**???**]

Smaller DIC values should indicate a better fit to data than larger DIC values. The DIC is composed of the deviance, which favours a good fit, and the effective number of parameters, which penalises overfitting.[**???**] Unlike the AIC the DIC can be estimated using samples from the posterior distribution and so is more readily calculated for models estimated using bayesian methods. It can be defined as,

$$DIC = D(\bar{\theta}) + 2p_D$$

Where $\bar{\theta}$ is the expectation of θ, with θ being defined as the unknown parameters of the model. p_D is the effective number of parameters in the model and is used to penalise more comple complex models. It can be estimated as follows,[**???**]

$$p_D = p_V = \frac{1}{2}\widehat{\text{var}}\left(D(\theta)\right).$$

Finally the deviance is defined as,

$$D(\theta) = -2\log(p(y|\theta)) + C$$

Where y are the data, $p(y|\theta)$ is the likelihood function and C is a constant. C cancels out when comparing different models and therefore does not need to be calculated.

The DIC has two limitations. The first of these is that in its derivation it is assumed that the model that generates future observations encompasses the true model. This assumption may not hold in all circumstances. The second limitations is that the observed data is used to construct both the posterior distribution and to estimate the DIC. This means that the DIC tends to select for over-fitted model.[**???**]

In this chapter I have used the DIC, as estimated by the `DIC` function from `rbi.helpers`,[**???**] to evaluate the various model structures outlined in the previous chapter.

9.3.6 Parameter sensitivity

Understanding the impact of parameter variation can help when interpreting findings from a model, targeting interventions, and identifying parameters for which improved estimates are needed. Often parameter sensitivity is assessed using single-parameter or local sensitivity analyses. Unfortunately, these techniques do not accurately capture uncertainty or sensitivity in the system as they hold all other parameters fixed.[107] Various techniques exist that can globally study a multi-dimensional parameter space. In this section, I will outline the partial rank correlation coefficient method (PRCC) and discuss it's strengths and weaknesses. The implementation of PRCC to estimate the parameter sensitivity of parameters fitted using the model fitting pipeline presented above is then discussed.

PRCC is a sampling based approach which can be computed with minimal computational cost from a sample of the prior or posterior distributions of a model. It estimates the degree of correlation between a given parameter input and an output after adjusting (using a linear model) for variation in other inputs. It is an extension of more simplistic sampling techniques, the most basic of which, is simply examining scatter plots of a sampled parameter set against the outcome of interest. PRCC is required as these more simplistic techniques become intractable with higher dimensionality as they do not account for between paremeter correlation or are just difficult to interpret with multiple dimensions.[107] PRCC can be understood by first outlining the individual steps. These are:

1. **Correlation:** Provides a measure of the strength of a linear association between an input and and output (scaled from -1 to 1). It is calculated as follows,

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Where cov is the covariance, σ_X is the standard deviation of X, and σ_Y is the standard deviation of Y. Where X is the input and Y is the output.

2. **Rank Correlation:** This is defined as for correlation but with the data being rank transformed. Rank transformation reorders inputs and outputs in magnitude order.

Unlike non-rank transformed correlation it can handle non-linear relationships but still requires monotonicity.

3. **Partial Rank Correlation:** Inputs and outputs are first rank transformed as above. Linear models are then built which adjust for the effects of the other inputs on Y, and on the current input X_i. Correlation is the calculated as above using the residuals from these models.

A limitation of PRCC is that it whilst it can capture non-lnear relationships between outputs and inputs these relationships must be monotonic.[107] For relationships that are non-monotic methods that rely on the decompositon of model output variance, such as the extended Fourier amplitude sensitivity test,[107] are more appropriate. However, these approaches are computationally demanding as they typically require mutliple iterations of model simulation. Additionally, they cannot be used on a previous parameter samples, instead needing to sample and simulate the model within the parameter sensitivity algorithmn. This means that they cannot be used for "free" (i.e with neglible additional compute cost), unlike PRCC which can be estimated using a sample from the posterior distribution. For this reason these approaches have not been further explored in this thesis.

I have implemented PRCC using the `epiR` R package[4],[74] using the samples from the posterior distribution of the model calculated during the SMC^2 step. Parameter sensitivity measures such as PRCC must be calculated seperately for each output time point. To account for this I calculated the PRCC for each fitted paramter, at each time point fitted too, for both overall TB incidence rates and TB incidence rates in children. These results are then summarised by plotting the absolute PRCC values for each output measure, indicating the direction of correlation using colour[5].

9.3.7 Pipeline overview

The full model fitting pipeline can be summarised as follows[6]:

1. Model initialisation using minimal error checking and single precision computation. Implemented using the `disable-assert` adn `enable-single` flags in LibBi.[106] Outputs are only given for times with oberserved data and a subset of parameters are recorded for final reporting[7].

2. 2500 parameter sets were taken from the prior distribution and the model was then simulated for each one.

3. Maximum likelihood optimisation with 100 steps, using the Nelder-Mead simplex method, via LibBi.[106] This approximates the mean of the posterior distribution.

4. 1000 PMCMC steps, with 256 particles used in the particle filter. This provides a better estimate of the mean of the posterior distribution.

5. Particle adaption via `rbi.helpers` at the approximate mean of the posterior distribution.[**???**] A minimum of 64 particles and a maximum of 256 particles are assessed

[4]Code: `https://github.com/seabbs/ModelTBBCGEngland/blob/master/R/test_sensitivity.R`
[5]Code: `https://github.com/seabbs/ModelTBBCGEngland/blob/master/R/plot_sensitivity.R`
[6]Code: `https://github.com/seabbs/ModelTBBCGEngland/blob/master/R/fit_model.R`
[7]Model code: `https://github.com/seabbs/ModelTBBCGEngland/blob/master/inst/bi/BaseLineModel.bi`

with the target of a log-likelihood variance of less than 5. 250 PMCMC steps were used at each stage to estimate the log-likelihood variance.

6. Manual independent proposal adaption at the approximate mean of the posterior distribution. The size of the proposal distribution is halved each iteration, with at most 5 iterations of adaption. The minimum target acceptance rate specified was 10% and the maximum was 20%. 250 PMCMC samples were used each time to estimate the acceptance rate. Proposal adaption was only carried out for the main model scenario, with other scenarios using this proposal.

7. SMC^2 model fitting with 3000 initial parameter particles. Particle rejuvernation was set to trigger when the effective sample size decreased below 1000, with 4 MCMC steps used each time.

8. For each sample from the posterior distribution the model was then simulated for all time points.

9. The model DIC was computed using `rbi.helpers`.[**???**] This gives a model agnostic approach to evaluate the fit to the observed data.

10. Parameter sensitivity was estimated by calculating the partial rank correlation coffecient (PRCC) for each model parameter, at each time point fitted too, for both overall TB incidence rates and TB incidence rates in children. Results were then plotted in order of the absolute magnitude of the correlation, with the direction of the correlation determined using colour. The `epiR` package was used to compute the PRCC.[74]

9.4 Results

9.4.1 Calibration

- Report on proposal acceptance rate and number of particles selected in main scenario
- Report on run time and feasibilty

9.4.2 Model comparision

- Compare Scenarios using DIC and discuss results.
- Table of DIC results with scenarios

9.4.3 Model fit to historic TB incidence rates

- Plot of historic TB data + fit from model
- Table of data + model fit + uncertainty
- Comment on model fit

9.4.4 Model Fit to TB incidence rates from the ETS

- Plot of TB data and fit
- Table of data and fit
- Comment on fit

9.4.5 Model fit to TB age distribution

- Age group plot over time + fit
- Age distribution over time + fit
- Comment on fit

9.4.6 Posterior parameter distributions

- Plot of Posterior Parameter distributions
- Table of Posterior Parameter distributions
- Comment on posteriors

9.4.7 Parameter Sensitivity

- Plot parameter sensitivty
- Comment on parameter sensitivity in light of strength of data

9.5 Discussion

In this chapter I have formulated the disease transmission model developed in the previous chapter as a state-space model, developed a model fitting pipeline to fit this model to observed data, and presented the results of this fitted model.

- Summary of work from chapter
- Method strengths and weaknesses
- Software strengths and weaknesses
- Strengths and weaknesses
- Next chapter
- Conclusions

9.6 Summary

- Defined the disease transmission model from the previous chapter as a state-space model.
- Outlined the observed data and defined an appropriate measurement model for this data.
- Developed a model fitting pipeline based on SMC^2 to fit the previously defined state-space model to TB notification data.
- Evaluated the various scenarios discussed in the previous chapter using DIC and discussed the implications of these findings.
- Presented the model fit to data from the best fitting scenario as established using the DIC.
- Discussed the posterior distributions of the fitted parameters and compared these distributions to the prior distributions.
- Estimated the model sensitivity to parameter changes and discussed the impact that this sensitivity may have.

Chapter 10

Investigating the impact of the 2005 change in BCG vaccination policy using a fitted dynamic transmission model of TB

10.1 Introduction

- Why
- What did I do in the last chapter and how does this extend it
- What I did summary

10.2 Method

10.2.1 Scenarios considered

- Outline scenarios considered

10.2.2 Future Non-UK born cases assumption

- Exponential decay
- Remain constant
- Compare and constrast what these mean at different timelines

10.3 Results

10.4 Impact between 2005 and 2010

- Comparision of cases at multiple time horizons.
- Comparision of age distribution of cases.
- Comparision of deaths at multiple time horizons.
- Comparision of age distribution of deaths.

- All in reflection of reality.

10.5 Non-UK born cases decline exponentially over time

- Comparision of cases at multiple time horizons (10 and 25 years).
- Comparision of age distribution of cases.
- Comparision of deaths at multiple time horizons (10 and 25 years).
- Comparision of age distribution of deaths

10.6 Non-UK born cases stay at todays levels

- Comparision of cases at multiple time horizons (10 and 25 years).
- Comparision of age distribution of cases.
- Comparision of deaths at multiple time horizons (10 and 25 years).
- Comparision of age distribution of deaths

10.7 Discussion

- Summary of work from chapter
- Strengths and weaknesses
- Literature comparisions
- Concliusions

10.8 Summary

- Best vaccination scenario according to model
- Implications for actual policy
-

Chapter 11

Discussion

This thesis has assessed the impact of the 2005 change in BCG vaccination policy in some detail. The aim of this chapter is to provide an overview of the principle findings of this thesis; interpret these findings; discuss the overall strengths and weaknesses of this thesis; outline the potential implications; explore the opportunities for public engagement that this work allowed; and describe potential future research. Each results chapter contains a detailed discussion of the approach used, and the results, for that chapter. Consequently, the aim of this chapter is to summarise and discuss the findings from this thesis as a whole.

11.0.1 Principal findings

In Chapter 4 I found that there was some evidence that negative outcomes were more frequent in TB cases not BCG vaccinated than in those that were. I also found that missingness in routine surveillance sources of Tuberculosis data was associated with multiple risk factors. In Chapter **??** I found supporting evidence that BCG vaccination was associated with reduced all-cause morality with some evidence that this may have been due to reduced TB mortality. I found little evidence for any other association with TB outcomes, after adjusting for confounding. In Chapter 5 I explored some of the evidence that was used by policy makers to assess the impact of the change in policy. I found that the previous approach had drastically underestimated the amount of uncertainty surrounding the effect estimates. Using newly available data, I also found that ending the scheme was projected to result in greater number of notifications in the UK born than previously thought. These findings were confirmed in Chapter 7 where I evaluated the evidence in the surveillance data that the change in policy had impacted Tuberculosis incidence rates in the target populations. However, in this chapter I found that any increase in TB notifications in the UK born was likely far outweighed by reductions in the number of notifications in the non-UK born. Although I was unable to rule out an unrelated policy change. Finally in Chapter 10 I found that the BCG schools' scheme was projected to *placeholder* TB incidence rates compared to both neonatal and no vaccination over a range of time horizons.

11.1 Strengths and limitations

This thesis has used multiple methods, and data sources, to explore the impact of the 2005 change in BCG policy. This triangulation allows for more certainty in the findings than if only a single approach had been used. A limitation of the work in this thesis is that all results were based on surveillance data. Surveillance data is subject to multiple bias issues (see Chapter 4 for details). Ideally multiple different data types would have been used to more effectively triangulate the impact of the change in policy. However, the data used in this thesis represents the best available. No similarly thorough use of an equivalent data source exists for Tuberculosis. A major strength of the work in this thesis is the attention that has been paid to make it both open and reproducible. Hopefully, this will allow these findings to be more easily validated, and built upon, by others. Finally the work in this thesis generated several peer reviewed tools as a by-product of the main research question. These tools have benefited other research projects.

11.2 Implications for policy makers

This thesis has highlighted the trade-off between vaccinating those at school-age and neonates in settings where the waning of BCG effectiveness when given later in life is minimal. Whilst policy makers were previously aware of this trade-off, their was little quantitative evidence exploring it explicitly. Globally, BCG policy does not account for areas where the BCG vaccine may be equally effective regardless of when it is given.[2] The evidence from this thesis should, therefore, be considered when outlining future BCG policy. In addition, new Tuberculosis vaccines are in development that may be less susceptible to waning effectiveness when given later in life over a greater geographic area.[16] The findings from this thesis may be applicable to these vaccines in areas where the BCG vaccine is currently known to be ineffective. This may mean that these newly developed vaccines may be better targeted at those at school-age, rather than neonates, depending on the duration of protection that they provide. The work from Chapter 8 and Chapter 9 may be particularly suitable to adaption for this use case. This thesis has also explored the potential benefits of BCG vaccination on TB outcomes. The evidence of a reduction in all-cause TB mortality in TB cases may add additional weight to the argument that wider vaccination maybe more cost effective than previously thought in low incidence countries. These findings may also be used to drive vaccine uptake as they provide additional incentives for vaccination. Finally, this thesis has shown the that the impact of the BCG policy has varied depending on UK birth status. This may further strengthen the case for varying vaccination policies depending on the country of origin of the target of vaccination policy, and their immediate families country of origin.

11.3 Open reproducible research

Open reproducible research has been a primary focus of this thesis. A version controlled archive of this thesis is available from GitHub[1], with a formatted version available on my personal site[2]. This thesis relies on data from the Enhanced Tuberculosis Surveillance system and the Labour Force Survey. The cleaning and munging of this data has been

[1] `https://github.com/seabbs/thesis`
[2] `https://www.samabbott.co.uk/thesis`

standardised as an R package, `tbinenglanddataclean`[3], and is available for download. All chapters that contain analysis are linked to their own GitHub repositories, each of which is fully reproducible (discounting the raw data which cannot be released due confidentiality reasons). Literate coding was used to link analysis code with documentation using the R tool chain. An R package, `prettypublisher`[4], was developed to augment these tools. Where possible open source tooling has been used to provide a working analytical environment for each chapter. All chapters that have been peer reviewed, or are undergoing peer review, have been preprinted. The model developed in Chapter 8 has been released as an R package along with the fitting pipeline developed in Chapter 9. Tools used to develop the figures in Chapter 2, using World Health Organization data, were expanded into an R package (see Chapter 3). Tooling developed alongside this thesis follows open source best practices. See Chapter 1 for full details on the open source projects developed as part of this thesis.

11.4 Public engagement

Public engagement has also been a primary focus of this thesis. This is closely linked with the previous aim of open and reproducible research. Effort has been taken so that all peer reviewed content is available for the wider public, with Twitter used disseminate findings. Where appropriate, interactive applications have been developed that seek to explore some of the key findings of this thesis, as well as teaching more theoretical concepts used throughout (see Chapter 1). Numerous case studies have also been produced that outline these theoretical concepts using some of the open source tools developed alongside this thesis[5]. These tools were themselves developed to lower the barrier of entry to infectious disease research. One of these tools, `idmodelr`, has been released to CRAN and peer reviewed by JOSS. Finally, in 2017 I spent a week at the Green Man Festival exploring the mathematics of vaccination with the general public. This made use of several simple games, as well as an interactive online tool[6].

11.5 Future research

The finding that BCG vaccination may reduce mortality in TB cases from Chapter **??** require validation in other data sources and settings. A larger sample size may be required in order to unpick the association between BCG vaccination and the cause of mortality. These findings could also be included in a cost effectiveness study of the BCG vaccine. The dynamic model developed in Chapter 8 and fitted in 9 does not currently include targeted vaccination of neonates and is only fitted to data up to 2004. In order to be able to more accurately explore current, and future, vaccination policy the extension of the model to the present data would be required. This would potentially be of great use for policy makers. An alternative would be to develop a comparable models in different settings. This would allow the generalisability of the findings to be explored. The dynamic model could also be further generalised to include hypothetical future vaccines with differing characteristics. This would allow vaccine characteristics and optimal deployment strategies to be explored, via simulations, ahead of further development. Both `getTBinR` and `idmodelr` have ac-

[3] `https://www.samabbott.co.uk/tbinenglanddataclean/`

[4] `https://github.com/seabbs/prettypublisher`

[5] See my personal site: `https://www.samabbott.co.uk`

[6] Available here: `https://github.com/seabbs/prettypublisher`

tive user bases and further developments are planned. This includes: additional tooling, documentation, and case studies. Further development of several of the interactive tools discussed in Chapter 1 is also planned.

11.6 Conclusions

This thesis has provided new evidence concerning the 2005 change in BCG policy in England. A simulation study that was used as part of the quantitative evidence for the change in policy was recreated and updated. The results from this updated model suggested that the change in policy was likely to have a greater impact on the UK born, at school-age, than previously thought. This finding was supported by a regression modelling study on the impact on TB incidence rates from the policy change. However, this study also found that the change in policy was associated with some benefits in UK born neonates and a much larger reduction in TB incidence rates in both non-UK born neonates and those at school-age. An additional regression study looking at the possible link between BCG vaccination and improved TB outcomes found some evidence that BCG vaccination was associated with reduced all-cause mortality, with little evidence of any other benefits. This result strengthens the case for wider vaccination. Additionally, a dynamic model of TB transmission was developed to provide a more detailed tool for evaluating the impact of the change in policy. Using this model it was found that *placeholder*. These findings suggest a stronger case for the use of the BCG vaccine in school-age populations; in areas with an equivalent level of TB transmission to England; and where the effectiveness of the BCG vaccine has been shown to not reduce with age. They also indicate that a future vaccine, without the reduced effectiveness observed in some geographic areas, may be more effectively targeted at those at school-age that at neonates. However, this depends on the potential duration of protection inferred by vaccination. The findings from this thesis may be of use to policy-makers to inform vaccine usage both in the UK and globally. As a by-product of the work conducted in this thesis several open source tools have been developed. These tools maybe used as learning resources, for public engangement, and as part of other research projects.

References

1 Public Health England. Tuberculosis in England 2017 report (presenting data to end of 2016) About Public Health England. 2017.

2 The World Health Organization. BCG vaccne:WHO position paper. *Weekly epidemiological record* 2018;1–24.

3 Roy a, Eisenhut M, Harris RJ *et al.* Effect of BCG vaccination against Mycobacterium tuberculosis infection in children: systematic review and meta-analysis. *BMJ (Clinical research ed)* 2014;**349**:g4643–3.

4 Zwerling A, Behr MA, Verma A *et al.* The BCG world atlas: A database of global BCG vaccination policies and practices. *PLoS medicine* 2011;**8**:e1001012.

5 Anderson RM, May RM. Infectious Diseases of Humans: Dynamics and Control (Oxford Univ. *Press, Oxford* 1991.

6 Keeling MJ, Rohani P. *Modeling Infectious Diseases in Humans and Animals.* Epidemiology Department, Ben-Gurion University of the Negev, Beer-Sheva, Israel. rbalicer@netvision.net.il: 2007.

7 King A a, Nguyen D, Ionides EL. Statistical Inference for Partially Observed Markov Processes via the R Package pomp. *The American Statistician* 2015.

8 Gideon HP, Flynn JL. Latent tuberculosis: What the host "sees"? *Immunologic Research* 2011;**50**:202–12.

9 Sepkowitz K a. How contagious is tuberculosis? *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 1996;**23**:954–62.

10 Rottenberg ME, Pawlowski A, Jansson M *et al.* Tuberculosis and HIV Co-Infection. *PLoS Pathogens* 2012;**8**:e1002464.

11 Bhatti N, Law MR, Morris JK *et al.* Increasing incidence of tuberculosis in England and Wales: a study of the likely causes. *BMJ (Clinical research ed)* 1995;**310**:967–9.

12 Narasimhan P, Wood J, Macintyre CR *et al.* Risk Factors for Tuberculosis. 2013;**2013**.

13 Story A, Murad S, Roberts W *et al.* Tuberculosis in London: the importance of homelessness, problem drug use and prison. *Thorax* 2007;**62**:667–72.

14 World Health Organization. Global Tuberculosis Report. 2016.

15 PHE. Tuberculosis in England 2016 Report (presenting data to end of 2015). *Public Health England* 2016;**Version 1.**:173.

16 Medicine C. History of BCG Vaccine. 2013;**8**:53–8.

17 Rodrigues LC, Diwan VK, Wheeler JG. Protective effect of BCG against tuberculous meningitis and miliary tuberculosis: a meta-analysis. *International journal of epidemiology* 1993;**22**:1154–8.

18 Colditz GA, Brewer TF, Berkey CS *et al.* Efficacy of BCG Vaccine in the Prevention of Tuberculosis. *JAMA* 1994;**271**:698.

19 Mangtani P, Abubakar I, Ariti C *et al.* Protection by BCG Vaccine Against Tuberculosis: A Systematic Review of Randomized Controlled Trials. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 2014;**58**:470–80.

20 Hart PDA, Sutherland IAN. BCG and vole bacillus vaccines in the prevention of tuberculosis in adolescence and early adult life. *The American Statistician* 1972;**46**:371–85.

21 Zwerling A, Behr MA, Verma A *et al.* The BCG World Atlas: a database of global BCG vaccination policies and practices. *PLoS medicine* 2011;**8**:e1001012.

22 Abubakar I, Pimpin L, Ariti C *et al.* Systematic review and meta-analysis of the current evidence on the duration of protection by bacillus Calmette-Guérin vaccination against tuberculosis. *Health technology assessment* 2013;**17**:1–372, v–vi.

23 Mangtani P, Nguipdop-Djomo P, Keogh RH *et al.* Original article The duration of protection of school-aged BCG vaccination in England : a population -based case control study. *International journal of epidemiology* 2017;**0**:1–9.

24 Fine P. Stopping routine vaccination for tuberculosis in schools. *BMJ (Clinical research ed)* 2005;**331**:647–8.

25 Teo SSS, Shingadia DV. Does BCG have a role in tuberculosis control and prevention in the United Kingdom? *Archives of Disease in Childhood* 2006;**91**:529–31.

26 Kleinnijenhuis J, Quintin J, Preijers F *et al.* Bacille Calmette-Guerin induces NOD2-dependent nonspecific protection from reinfection via epigenetic reprogramming of monocytes. *Proceedings of the National Academy of Sciences of the United States of America* 2012;**109**:17537–42.

27 Garly ML, Martins CL, Balé C *et al.* BCG scar and positive tuberculin reaction associated with reduced child mortality in West Africa: A non-specific beneficial effect of BCG? *Vaccine* 2003;**21**:2782–90.

28 Higgins JPT, Soares-weiser K, López-lópez JA *et al.* Association of BCG , DTP , and measles containing vaccines with childhood mortality : systematic review. *BMJ (Clinical research ed)* 2016;i5170.

29 Jeremiah K, Praygod G, Faurholt-Jepsen D *et al.* BCG vaccination status may predict sputum conversion in patients with pulmonary tuberculosis: a new consideration for an

old vaccine? *Thorax* 2010;**65**:1072–6.

30 The World Health Organization. BCG Vaccine. *Weekly epidemiological record* 2004;**79**:27–48.

31 World Health Organization. *Global Tuberculosis Report*. 2017.

32 Mangtani P, Abubakar I, Ariti C *et al.* Protection by BCG vaccine against tuberculosis: A systematic review of randomized controlled trials. *Clinical Infectious Diseases* 2014;**58**:470–80.

33 Sutherland I, Springett VH. 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**:15–24.

34 Schrager LK, Harris RC, Vekemans J. Research and development of new tuberculosis vaccines: a review. *F1000Research* 2018;**7**:1732–32.

35 R Core Team. R: a language and environment for statistical computing. 2019.

36 World Health Organisation. Global Tuberculosis Report. 2018.

37 Abbott S. seabbs/getTBinR 0.5.7. 2019.

38 Abbott S. Exploring Estimates of the Tuberculosis Case Fatality Ratio - with getTBinR. 2018.

39 French CE, Antoine D, Gelb D *et al.* Tuberculosis in non-UK-born persons, England and Wales, 2001-2003. *Int J Tuberc Lung Dis* 2007;**11**:577–84.

40 Kruijshaar M, French C, Anderson C *et al.* Tuberculosis in the UK, Annual report on tuberculosis surveillance and control in the UK 2007. *Thorax* 2007;**50**:703–3.

41 Public Health England. The Green Book. 2013;391–409.

42 Pillaye J, Clarke A. An evaluation of completeness of tuberculosis notification in the United Kingdom. *BMC Public Health* 2003;**3**:31.

43 PHE. Tuberculosis in England 2016 Report (presenting data to end of 2015). 2016.

44 Groothuis-oudshoorn K. Journal of Statistical Software MICE : Multivariate Imputation by Chained.;**VV**.

45 Sterne JA, Davey Smith G. Sifting the evidence-what's wrong with significance tests? *Bmj* 2001;**322**:226–31.

46 Benchimol EI, Smeeth L, Guttmann A *et al.* The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *The American Statistician* 2016;**115-116**:1–22.

47 Fewell Z, Davey Smith G, Sterne JAC. The impact of residual and unmeasured confounding in epidemiologic studies: A simulation study. *American Journal of Epidemiology* 2007;**166**:646–55.

48 Wickham Rstudio H. Tidy Data. *JSS Journal of Statistical Software MMMMMM*

YYYY;**VV**.

49 R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: 2016.

50 Storey A. Accuracy of official high-age population estimates , in England and Wales : an evaluation. 2016;1–67.

51 NHS Trafford CCG Governing Body. Performance and Quality Report. 2015;1–47.

52 Date W, Area G, Authority L *et al.* A comparison of the 2011 Census and the Labour Force Survey (LFS) labour market indicators. 2012;1–23.

53 Kazuki Yoshida GJAMS with contributions from TNCHJMJSRTJRJR-CPSPS. epiR: Tools for the Analysis of Epidemiological Data. 2016.

54 Rieckmann A, Villumsen M, Sørup S *et al.* Vaccinations against smallpox and tuberculosis are associated with better long-term survival: a Danish case-cohort study 19712010. *International journal of epidemiology* 2016;**0**:1–11.

55 Barreto ML, Pilger D, Pereira SM *et al.* Causes of variation in BCG vaccine efficacy: Examining evidence from the BCG REVAC cluster randomized trial to explore the masking and the blocking hypotheses. *Vaccine* 2014;**32**:3759–64.

56 Parslow R, El-Shimy NA, Cundall DB *et al.* Tuberculosis, deprivation, and ethnicity in Leeds, UK, 1982-1997. *Archives of disease in childhood* 2001;**84**:109–13.

57 Roth A, Sodemann M, Jensen H *et al.* Tuberculin reaction, BCG scar, and lower female mortality. *Epidemiology (Cambridge, Mass)* 2006;**17**:562–8.

58 Aaby P, Nielsen J, Benn CS *et al.* Sex-differential and non-specific effects of routine vaccinations in a rural area with low vaccination coverage: An observational study from Senegal. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 2014;**109**:77–84.

59 Teale C, Goldman JM, Pearson SB. The association of age with the presentation and outcome of tuberculosis: a five-year survey. *Age and ageing* 1993;**22**:289–93.

60 Thomas P, Delincée H, Diehl JF. Thin-layer isoelectric focusing of polyphenoloxidase of Sephadex and its detection by the print technique. *Analytical biochemistry* 1978;**88**:138–48.

61 Abubakar I, Laundy MT, French CE *et al.* Epidemiology and treatment outcome of childhood tuberculosis in England and Wales: 1999-2006. *Archives of Disease in Childhood* 2008;**93**:1017–21.

62 Djuretic T, Herbert J, Drobniewski F *et al.* Antibiotic resistant tuberculosis in the United Kingdom : 2002;477–82.

63 Van Buuren S, Groothuis-Oudshoorn K. Multivariate Imputation by Chained Equations. *Journal Of Statistical Software* 2011;**45**:1–67.

64 Barnard J, Rubin DB. Miscellanea. Small-sample degrees of freedom with multiple imputation. *Biometrika* 1999;**86**:948–55.

65 Roy a, Eisenhut M, Harris RJ *et al.* Effect of BCG vaccination against Mycobacterium tuberculosis infection in children: systematic review and meta-analysis. *BMJ (Clinical research ed)* 2014;**349**:g4643–3.

66 Kandasamy R, Voysey M, McQuaid F *et al.* Non-specific immunological effects of selected routine childhood immunisations: systematic review. *BMJ (Clinical research ed)* 2016;**355**:i5225.

67 Pollard AJ, Finn A, Curtis N. Non-specific effects of vaccines: plausible and potentially important, but implications uncertain. *Archives of Disease in Childhood* 2017;**102**:archdischild–2015–310282.

68 Romanus V, Svensson Å, Hallander HO. The impact of changing BCG coverage on tuberculosis incidence in Swedish-born children between 1969 and 1989. *Tubercle and Lung Disease* 1992;**73**:150–61.

69 Guthmann JP, Antoine D, Fonteneau L *et al.* Assessing BCG vaccination coverage and incidence of paediatric tuberculosis following two major changes in BCG vaccination policy in France. 2011;1–6.

70 Thomas HL, Harris RJ, Muzyamba MC *et al.* Reduction in tuberculosis incidence in the UK from 2011 to 2015: a population-based study. *Thorax* 2018;thoraxjnl–2017–211074.

71 Parikh SR, Andrews NJ, Beebeejaun K *et al.* Eff ectiveness and impact of a reduced infant schedule of 4CMenB vaccine against group B meningococcal disease in England : a national observational cohort study. *The Lancet* 2013;**388**:2775–82.

72 Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *The American Statistician* 2016;**27**:1–20.

73 ai H. R Interface for H2O. 2018.

74 Stevenson M, Nunes T, Heuer C *et al. epiR: Tools for the Analysis of Epidemiological Data*. 2017.

75 Bürkner P-C. brms: An R Package for Bayesian Multilevel Models using Stan. *Journal Of Statistical Software*

76 Stan Development Team. RStan: the R interface to Stan. 2016.

77 Davies R, Jones M, Lloyd-Williams H. Age and Work-Related Health: Insights from the UK Labour Force Survey. *British Journal of Industrial Relations* 2016;**54**:136–59.

78 Lindley J. The over-education of UK immigrants and minority ethnic groups: Evidence from the Labour Force Survey. *Economics of Education Review* 2009;**28**:80–9.

79 Manissero D, Lopalco PL, Levy-Bruhl D *et al.* Assessing the impact of different BCG vaccination strategies on severe childhood TB in low-intermediate prevalence settings. *Vaccine* 2008;**26**:2253–9.

80 Feiring B, Laake I, Molden T *et al.* Do selective immunisation against tuberculosis and hepatitis B reach the targeted populations ? A nationwide register-based study evaluating the recommendations in the Norwegian Childhood Immunisation Programme.

Vaccine 2016;**34**:2015–20.

81 Nguipdop-Djomo P, Mangtani P, Pedrazzoli D *et al.* Uptake of neonatal BCG vaccination in England: Performance of the current policy recommendations. *Thorax* 2014;**69**:87–9.

82 Brooks-Pollock E, Cohen T, Murray M. The impact of realistic age structure in simple models of tuberculosis transmission. *PLoS ONE* 2010;**5**:3–8.

83 Menzies NA, Wolf E, Connors D *et al.* Review Progression from latent infection to active disease in dynamic tuberculosis transmission models : a systematic review of the validity of modelling assumptions. *Lancet Infect Dis* 2018;**3099**.

84 Ragonnet R, Trauer JM, Scott N *et al.* Optimally capturing latency dynamics in models of tuberculosis transmission. *Epidemics* 2017;**21**:39–47.

85 Mathema B, Andrews JR, Cohen T *et al.* Drivers of Tuberculosis Transmission. *J Infect Dis* 2018;**216**:S644–53.

86 Lefebvre N, Sotgiu G, Falzon D *et al.* Determinants of site of tuberculosis disease : An analysis of European surveillance data from 2003 to 2014. 2017;1–14.

87 SHAW JB, WYNN-WILLIAMS N. Infectivity of pulmonary tuberculosis in relation to sputum status. *American review of tuberculosis* 1954;**69**:724–32.

88 Tostmann A, Kik SV, Kalisvaart NA *et al.* Tuberculosis Transmission by Patients with Smear- Negative Pulmonary Tuberculosis in a Large Cohort in The Netherlands. *Clinical Infectious Diseases* 2008;**47**:1135–42.

89 Piccini P, Chiappini E, Tortoli E *et al.* Clinical peculiarities of tuberculosis. 2014;**14**:1–12.

90 Houben RMGJ, Lalli M, Sumner T *et al.* TIME Impact - a new user-friendly tuberculosis (TB) model to inform TB policy decisions. *BMC Medicine* 2016;**14**:1–10.

91 Andrews JR, Lawn SD, Rusu C *et al.* The cost-effectiveness of routine tuberculosis screening with Xpert MTB/RIF prior to initiation of antiretroviral therapy: a model-based analysis. *Aids* 2012;**26**:987–95 10.1097/QAD.0b013e3283522d47.

92 Vynnycky E, Fine PEM. The natural history of tuberculosis : the implications of age-dependent risks of disease and the role of reinfection. 2018;183–201.

93 Houben RM, Lalli M, Sumner T *et al.* TIME Impact - a new user-friendly tuberculosis (TB) model to inform TB policy decisions. *BMC Med* 2016;**14**:56.

94 Vynnycky E, Fine PE. Interpreting the decline in tuberculosis: the role of secular trends in effective contact. *Int J Epidemiol* 1999;**28**:327–34.

95 Hens N, Jit M, Beutels P *et al.* Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases. *PLoS medicine* 2008;**5**:e74.

96 Lalor MK, Anderson LF, Hamblion EL *et al.* Recent household transmission of tuberculosis in England, 2010-2012: Retrospective national cohort study combining epidemiological and molecular strain typing data. *BMC Medicine* 2017;**15**:1–10.

97 Dowdy DW, Dye C, Cohen T. Data needs for evidence-based decisions : a tuberculosis modeler ' s ' wish list '. *International Journal of Tuberculosis and Lung Disease* 2013;**17**:866–77.

98 Trunz BB, Fine P, Dye C. Effect of BCG vaccination on childhood tuberculous meningitis and miliary tuberculosis worldwide: a meta-analysis and assessment of cost-effectiveness. *Lancet* 2006;**367**:1173–80.

99 Rahman M, Sekimoto M, Takamatsu I *et al.* Economic evaluation of universal BCG vaccination of Japanese infants. *International journal of epidemiology* 2001;**30**:380–5.

100 Usher C, Adams R, Schmitz S *et al.* Evaluating the neonatal BCG vaccination programme in Ireland. *Archives of Public Health* 2016;**74**:1–12.

101 Harris RC, Dodd PJ, White RG. The potential impact of BCG vaccine supply shortages on global paediatric tuberculosis mortality. *BMC Med* 2016;**14**:138.

102 Egbetade S a, Polytechnic T, Ibrahim MO. Modelling The Impact of BCG Vaccines on Tuberculosis Epidemics. 2011;**1**:49–55.

103 Bhunu CP, Garira W, Mukandavire Z *et al.* Modelling the effects of pre-exposure and post-exposure vaccines in tuberculosis control. *Journal of Theoretical Biology* 2008;**254**:633–49.

104 Brooks-Pollock E, Cohen T, Murray M. The impact of realistic age structure in simple models of tuberculosis transmission. *PLoS ONE* 2010;**5**:3–8.

105 Vynnycky E, Fine PE. The natural history of tuberculosis: the implications of age-dependent risks of disease and the role of reinfection. *Epidemiology and infection* 1997;**119**:183–201.

106 Murray LM. Bayesian State-Space Modelling on High-Performance Hardware Using LibBi. *Journal of Statistical Software* 2015;**67**:1–36.

107 Marino S, Hogue IB, Ray CJ *et al.* A methodology for performing global uncertainty and sensitivity analysis in systems biology. *J Theor Biol* 2008;**254**:178–96.