

---

---

# 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.

AUGUST 2019

Word count: 55,287



# Abstract

Bacillus Calmette–Guérin (BCG) remains the only licensed vaccine against Tuberculosis (TB). 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 TB 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. I recreated one such study, correcting a methodological flaw, 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 TB 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 affected by the policy change. I found that 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. Therefore, I developed a dynamic model of TB, fit to available data, to forecast the impact of the policy change. Although the fit to the data was poor, the forecasts suggested that continuing school-age vaccination reduced TB incidence in the UK born compared with neonatal vaccination. Neonatal vaccination reduced incidence in children but had little impact on other age groups.



# 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 thank Matthew Hickman for providing additional supervisory support.

This thesis could not have been completed without 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.



# Contributorship 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., Brooks-Pollock E. *Reassessing the evidence for universal school-age Bacillus Calmette Guerin (BCG) vaccination in England and Wales*. doi: <https://doi.org/10.1101/567511>

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., Lalor M. K., Zenner D., Campbell C., Ramsay M. E., & Brooks-Pollock, E. (2019). *Exploring the effects of BCG vaccination in patients diagnosed with tuberculosis: Observational study using the Enhanced Tuberculosis Surveillance system*. Vaccine, 1–6. doi: <http://doi.org/10.1016/j.vaccine.2019.06.056> (preprint: <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 N.J., Brooks-Pollock E. (2019) *Estimating the effect of the 2005 change in BCG policy in England: a retrospective cohort study, 2000 to 2015*. Eurosurveillance, 49(24). doi: <https://doi.org/10.2807/1560-7917.ES.2019.24.49.1900220>

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:

# **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 others, is indicated as such. Any views expressed in this thesis are those of the author.

Signed:

Date:



# Table of Contents

<b>Chapter 1: Introduction . . . . .</b>	<b>1</b>
1.1 Theoretical framework . . . . .	2
1.1.1 Data exploration and visualisation . . . . .	2
1.1.2 Statistical modelling . . . . .	2
1.1.3 Mechanistic modelling . . . . .	3
1.2 Aims and objectives of the thesis . . . . .	3
1.2.1 Aim . . . . .	3
1.2.2 Objectives . . . . .	3
1.3 Chapter overview . . . . .	4
1.4 Thesis output . . . . .	5
1.4.1 Peer reviewed papers . . . . .	5
1.4.2 Papers under review . . . . .	5
1.4.3 Software . . . . .	5
1.4.4 Talks . . . . .	7
1.5 Summary . . . . .	7
<b>Chapter 2: Background . . . . .</b>	<b>9</b>
2.1 Tuberculosis . . . . .	9
2.1.1 Natural history of TB . . . . .	9
2.1.2 Risk factors . . . . .	9
2.1.3 Treatments . . . . .	10
2.1.4 Global TB . . . . .	13
2.1.5 TB in the England and Wales . . . . .	17
2.2 The Bacillus Calmette–Guérin Vaccine . . . . .	20
2.2.1 Vaccine action . . . . .	20
2.2.2 Vaccine effectiveness . . . . .	21
2.2.3 Duration of protection . . . . .	21
2.2.4 Additional effects of BCG vaccination . . . . .	21
2.2.5 Usage globally . . . . .	22
2.2.6 Usage in England . . . . .	23
2.2.7 Replacement vaccines . . . . .	23
2.3 Summary . . . . .	23
<b>Chapter 3: getTBinR: an R package for accessing and summarising World Health Organization Tuberculosis data . . . . .</b>	<b>25</b>
3.1 Introduction . . . . .	25
3.2 Installation . . . . .	26

3.3	Data extraction and variable look-up . . . . .	26
3.4	Data visualisation . . . . .	27
3.4.1	Mapping TB burden metrics . . . . .	27
3.4.2	Plotting an overview for a given TB metric . . . . .	28
3.4.3	Plotting a comparison between country, regional and global metric values . . . . .	29
3.5	Plotting country level trends for a given metric . . . . .	30
3.6	Data summarisation . . . . .	31
3.7	Dashboard . . . . .	32
3.8	Country report . . . . .	32
3.9	Package infrastructure . . . . .	33
3.10	Discussion . . . . .	33
3.11	Summary . . . . .	34
<b>Chapter 4: The epidemiology of tuberculosis and the role of BCG vaccination in England . . . . .</b>		<b>35</b>
4.1	Introduction . . . . .	35
4.2	Data sources . . . . .	35
4.2.1	Enhanced tuberculosis surveillance (ETS) system . . . . .	35
4.2.2	Demographic data . . . . .	58
4.3	TB notifications . . . . .	61
4.3.1	Overview . . . . .	61
4.3.2	Age distribution of notifications . . . . .	62
4.4	Population Demographics in England . . . . .	63
4.5	TB incidence rates . . . . .	65
4.5.1	Motivation . . . . .	65
4.5.2	Method . . . . .	66
4.5.3	Overall trends in TB incidence rates . . . . .	66
4.5.4	Age stratified incidence rates . . . . .	67
4.5.5	Incidence rates in children (0-14 years old) as a proxy for TB transmission . . . . .	71
4.6	TB outcomes . . . . .	71
4.6.1	Motivation . . . . .	71
4.6.2	Method . . . . .	71
4.6.3	All-cause mortality . . . . .	71
4.6.4	TB related mortality . . . . .	74
4.6.5	Successful treatment . . . . .	76
4.6.6	Lost to follow up . . . . .	78
4.7	Discussion . . . . .	80
4.8	Summary . . . . .	82
<b>Chapter 5: Reassessing the evidence for universal school-age BCG vaccination in England and Wales . . . . .</b>		<b>83</b>
5.1	Introduction . . . . .	83
5.2	Background . . . . .	83
5.3	Methods . . . . .	84
5.3.1	Modelling the impact of ending the BCG schools scheme . . . . .	84

5.3.2	Updating model parameter estimates and incorporating parameter uncertainty . . . . .	87
5.3.3	Statistical analysis . . . . .	88
5.4	Results . . . . .	88
5.4.1	Annual change in TB incidence rates . . . . .	91
5.5	Vaccines required to prevent a single notification . . . . .	92
5.5.1	Average annual additional cases from ending the BCG schools scheme at various dates . . . . .	95
5.6	Discussion . . . . .	98
5.7	Summary . . . . .	100
<b>Chapter 6: Exploring the effects of BCG vaccination in patients diagnosed with tuberculosis: observational study using the Enhanced Tuberculosis Surveillance system . . . . .</b>		<b>101</b>
6.1	Introduction . . . . .	101
6.2	Background . . . . .	102
6.3	Method . . . . .	102
6.3.1	Enhanced TB Surveillance (ETS) system . . . . .	102
6.3.2	Exposure variables relating to BCG . . . . .	103
6.3.3	Statistical Analysis . . . . .	103
6.4	Results . . . . .	104
6.4.1	Description of the data . . . . .	104
6.4.2	All-cause mortality . . . . .	106
6.4.3	Deaths due to TB (in those who died) . . . . .	107
6.4.4	Recurrent TB . . . . .	108
6.4.5	Other Outcomes . . . . .	108
6.4.6	Sensitivity analysis of the missing data using multiple imputation . . . . .	112
6.4.7	Sensitivity analysis . . . . .	114
6.5	Discussion . . . . .	115
6.6	Summary . . . . .	118
<b>Chapter 7: Estimating the effect of the 2005 change in BCG policy in England: A retrospective cohort study . . . . .</b>		<b>119</b>
7.1	Introduction . . . . .	119
7.2	Background . . . . .	119
7.3	Methods . . . . .	120
7.3.1	Data source . . . . .	120
7.3.2	Constructing Retrospective cohorts . . . . .	120
7.4	Statistical methods overview . . . . .	121
7.4.1	Implementation overview . . . . .	122
7.4.2	Imputation of UK birth status . . . . .	122
7.4.3	Prior choice . . . . .	124
7.4.4	Estimating the magnitude of the estimated impact of the change in BCG policy . . . . .	125
7.5	Results . . . . .	126
7.5.1	Descriptive analysis . . . . .	126
7.5.2	Adjusted estimates of the effects of the change in policy on school-age children . . . . .	128

7.5.3	Adjusted estimates of the effect of the change in policy in those relevant to the targeted neonatal programme . . . . .	132
7.5.4	Magnitude of the estimated impact of the change in BCG policy . . . . .	136
7.6	Discussion . . . . .	136
7.7	Summary . . . . .	138
<b>Chapter 8: Developing a dynamic transmission model of Tuberculosis . . . . .</b>	<b>139</b>	
8.1	Introduction . . . . .	139
8.2	Previously published TB models . . . . .	139
8.3	Choice of model structure . . . . .	140
8.3.1	TB disease . . . . .	140
8.3.2	BCG vaccination . . . . .	144
8.4	A dynamic model of TB transmission . . . . .	145
8.4.1	Model outline . . . . .	145
8.4.2	Model equations . . . . .	148
8.4.3	Force of infection . . . . .	149
8.5	Parameterisation and data synthesis . . . . .	149
8.5.1	Data sources . . . . .	150
8.5.2	Model Parameters . . . . .	150
8.6	Initialisation . . . . .	170
8.6.1	Starting simulation date, initial population and changes over time . . . . .	171
8.6.2	Initial disease distribution . . . . .	171
8.7	Scenarios . . . . .	171
8.8	Discussion . . . . .	172
8.9	Summary . . . . .	175
<b>Chapter 9: Fitting a dynamic transmission model of Tuberculosis . . . . .</b>	<b>177</b>	
9.1	Introduction . . . . .	177
9.2	Formulation as a state-space models . . . . .	177
9.2.1	Observed data . . . . .	178
9.2.2	Observational model . . . . .	178
9.2.3	Fitted parameters . . . . .	181
9.2.4	Stochastic parameters . . . . .	181
9.3	Model fitting pipeline . . . . .	182
9.3.1	The particle filter . . . . .	182
9.3.2	Sequential Monte Carlo . . . . .	183
9.3.3	Calibration . . . . .	184
9.3.4	Model comparison . . . . .	186
9.3.5	Parameter sensitivity . . . . .	187
9.3.6	Pipeline overview . . . . .	188
9.4	Results . . . . .	189
9.4.1	Ad hoc calibration . . . . .	189
9.4.2	Particle and proposal calibration . . . . .	191
9.4.3	Model comparison . . . . .	191
9.4.4	Model Fit to TB incidence from the ETS . . . . .	192
9.4.5	Posterior parameter distributions . . . . .	194
9.4.6	Parameter Sensitivity . . . . .	195
9.5	Discussion . . . . .	197

9.6 Summary . . . . .	200
<b>Chapter 10: Investigating the impact of the 2005 change in BCG vaccination policy using a fitted dynamic transmission model of TB . . . . .</b>	<b>201</b>
10.1 Introduction . . . . .	201
10.2 Methods . . . . .	201
10.2.1 Scenarios considered . . . . .	201
10.2.2 Forecasting assumptions . . . . .	202
10.2.3 Analytical methods . . . . .	202
10.3 Results . . . . .	202
10.3.1 Forecasting the long-term impact of each vaccination scenario. . . . .	203
10.4 Discussion . . . . .	207
10.5 Summary . . . . .	209
<b>Chapter 11: Discussion . . . . .</b>	<b>211</b>
11.1 Principal findings . . . . .	211
11.2 Strengths and limitations . . . . .	212
11.3 Implications for policy makers . . . . .	212
11.4 Open reproducible research . . . . .	214
11.5 Public engagement . . . . .	214
11.6 Future research . . . . .	214
11.7 Conclusions . . . . .	215
<b>References . . . . .</b>	<b>217</b>



# List of Tables

2.1	A timeline of interventions against TB . . . . .	11
2.2	Percentage (%) of rifampicin resistant TB cases that have multi-drug resistant TB in Russia and regional medians, with interquartile ranges. The estimate for Europe includes data from the Russian Federation. . . . .	17
4.1	Variables derived or modified from the ETS system for use in the analyses throughout this thesis. . . . .	37
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. . . . .	40
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 grouped as 0-14 year olds, 15-44 year olds, 45-64 year olds, and 65+), ethnic group, UK birth status and socio-economic status (national quintiles). . . . .	43
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 grouped as 0-14 year olds, 15-44 year olds, 45-64 year olds, and 65+), ethnic group, UK birth status and socio-economic status (national quintiles). . . . .	44
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 grouped as 0-14 year olds, 15-44 year olds, 45-64 year olds, and 65+), ethnic group, UK birth status and socio-economic status (national quintiles). . . . .	45
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 grouped as 0-14 year olds, 15-44 year olds, 45-64 year olds, and 65+), ethnic group, UK birth status and socio-economic status (national quintiles). . . . .	47
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 grouped as 0-14 year olds, 15-44 year olds, 45-64 year olds, and 65+), ethnic group, UK birth status and socio-economic status (national quintiles). . . . .	48

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 grouped as 0-14 year olds, 15-44 year olds, 45-64 year olds, and 65+), ethnic group, UK birth status and socio-economic status (national quintiles). . . . .	49
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 grouped as 0-14 year olds, 15-44 year olds, 45-64 year olds, and 65+), ethnic group, UK birth status and socio-economic status (national quintiles). . . . .	50
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 grouped as 0-14 year olds, 15-44 year olds, 45-64 year olds, and 65+), ethnic group, UK birth status and socio-economic status (national quintiles). . . . .	51
5.1	Comparison of results published by Sutherland et al. vs. the recreated model.	89
5.2	Continued: Comparison of results published by Sutherland et al. vs. the recreated model. . . . .	90
5.3	The median number (with the 2.5% and 97.5% quantiles) of vaccines required to prevent a single case of TB within 15 years in a ethnic White UK born adult vaccinated at 13 years old. The percentage annual decrease scenarios considered were based on those considered by the JCVI BCG subgroup, with the addition of a scenario using aggregate notification data and a scenario using estimates of age-specific incidence rates in the UK born. . . . .	94
5.4	The median number (with the 2.5% and 97.5% quantiles) of additional annual notifications due to ending the BCG schools scheme in selected years. The percentage annual decrease scenarios considered were based on those considered by the JCVI BCG subgroup. Data based scenarios and the JCVI 0% decrease scenario were not presented here as the updated transmission model could not support these scenarios. . . . .	97
6.1	Outcomes for individuals in England notified with TB between 2009-2015, stratified by BCG vaccination status. . . . .	105
6.2	Potential confounders for individuals in England notified with TB between 2009-2015, stratified by BCG vaccination status. . . . .	106
6.3	Summary of logistic regression model output with BCG vaccination as the exposure and all-cause mortality as the outcome. . . . .	107
6.4	Summary of associations between BCG vaccination and all outcomes . . . .	109
6.5	Summary of associations between years since vaccination and all outcomes in individuals who were vaccinated. . . . .	110
6.6	Summary of associations between age at vaccination and all outcomes in individuals who were vaccinated. . . . .	111
6.7	Summary of associations between BCG vaccination and all outcomes, using pooled imputed data. . . . .	113
6.8	Summary of associations between years since vaccination and all outcomes, using pooled imputed data. . . . .	113

6.9	Summary of associations between age at vaccination and all outcomes, using pooled imputed data (reference is vaccination at <1 year) . . . . .	114
6.10	Summary of associations between BCG vaccination and all outcomes. . . . .	115
7.1	Summary of relevance and eligibility criteria for each cohort. . . . .	121
7.2	Complete definition of each model, ordered by increasing complexity. . . . .	123
7.3	Comparison of UK birth status in cases with complete or imputed records.	124
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). . . . .	129
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. . . . .	130
7.6	Comparison of models fitted to incidence rates for the non-UK born population that were eligible for the universal vaccination programme of those at school-age (14). . . . .	131
7.7	Comparison of models fitted to incidence rates for the UK born population that were eligible for the targeted vaccination programme of neonates. . . . .	133
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. . . . .	134
7.9	Comparison of models fitted to incidence rates for the non-UK born population that were relevant to the targeted vaccination programme of neonates.	135
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. . . . .	136
8.1	Dynamic disease model parameters, descriptions, prior distributions, units, method used to derive the prior distribution and the type. . . . .	151
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. . . . .	156
8.3	Estimates of the effectiveness of the BCG vaccine at preventing active TB disease stratified by years since vaccination. . . . .	166
8.4	Demographic model parameters, descriptions, prior distributions, units, method used to derive the prior distribution and the type. . . . .	167
8.5	Summary of planned scenario analyses to be carried out in the next chapter as part of model fitting by comparison of the goodness of fit to the data. .	172
9.1	Measurement model parameters, descriptions, prior distributions, units, method used to derive the prior distribution and the type. . . . .	180
9.2	DIC values for each scenario evaluated during model fitting - arranged from best to worst model fit. Using 1000 samples from the posterior distribution of the fitted model for each scenario. . . . .	192
9.3	Observed versus predicted overall TB cases for years that the model was fitted to. . . . .	192
9.4	Prior and posterior medians for fitted model parameters . . . . .	194
10.1	Forecast of overall TB incidence for each scenario evaluated from 2005 to 2040.	203

10.2 Forecast of overall TB mortality for each scenario evaluated from 2005 to 2040.205

# List of Figures

2.1	TB incidence rates (per 100,000) by region and globally from 2000 until 2017.	13
2.2	Global map of country level TB incidence rates (per 100,000 population) in 201 . . . . .	14
2.3	TB mortality rates (per 100,000 cases) by region and globally from 2000 until 2017. . . . .	15
2.4	Global map of estimated HIV in incidence TB (percent) in 2018. . . . .	16
2.5	Global map of the estimated percentage of new TB cases with rifampicin resistance (percent) in 2018. . . . .	17
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. . . . .	18
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. . . . .	20
3.1	Map of global TB incidence rates in 2017 as generated by getTBinR. . . . .	28
3.2	Dot plot showing trends over time in TB incidence rates in Europe ordered by TB incidence rates in 2017. . . . .	29
3.3	TB incidence by region and globally as computed and visualised by getTBinR. .	30
3.4	TB incidence rates over time, with confidence intervals, in the UK. As produced by getTBinR. . . . .	31
3.5	Snapshot of the built in package dashboard. . . . .	32
3.6	Screenshot of the start of the built-in package summary report, for the United Kingdom. . . . .	33
4.1	Summary plot of missing data in the extract of the ETS data used in this thesis. . . . .	39
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. . . . .	53
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. . . . .	54
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. . . . .	55
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. . . . .	57
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. . . . .	58

4.7	Overall population estimates in England derived using ONS and LFS demographic data. . . . .	60
4.8	Percentage difference between ONS population estimates and estimates derived from the LFS by 5 year age group. . . . .	61
4.9	Notifications in England from 2000 to 2015 stratified by UK birth status, sourced from the ETS system. . . . .	62
4.10	Proportion of total yearly notification by 5 year age group in the ETS system in 2005, 2010 and 2015 stratified by UK birth status. . . . .	63
4.11	The estimated proportion of the population in each 5 year age group stratified by UK birth status for 2000, 2008, and 2016. . . . .	65
4.12	Age standardised (adjusted) incidence rates (by 100,000 population) for all notified TB cases from 2000-2015. . . . .	67
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. . . . .	68
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. . . .	70
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. . . . .	73
4.16	Age distribution (in 5 year age groups) of the case all-cause mortality rate presented on a square root scale. . . . .	74
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. . . . .	75
4.18	Age distribution (in 5 year age groups) of the case TB mortality rate presented on a square root scale. . . . .	76
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. . . . .	77
4.20	Age distribution (in 5 year age groups) of the case successful treatment within 12 months rate presented on a square root scale. . . . .	78
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. . . .	79
4.22	Age distribution (in 5 year age groups) of the case loss to follow up rate presented on a square root scale. . . . .	80
5.1	Annual percentage change in ethnic White UK born incidence rates for those aged 15-19, 20-24, and 25-29 years old under different scenarios. . . . .	91
5.2	Vaccines required in a cohort of those vaccinated at school-age to prevent a single case of Tuberculosis within 15 years of vaccination in 2004, 2009, 2014, or 2019. . . . .	93
5.3	Annual additional (stratified into primary and secondary) notifications in 15-29 year olds from stopping the BCG schools scheme in 2006, and 2011 until 2028. . . . .	96
7.1	Mean incidence rates per 100,000, with 95% confidence intervals for each retrospective cohort, stratified by the vaccination policy and UK birth status.	126

7.2	Incidence rates (with 95% confidence intervals) per 100,000 per year 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. . . . .	127
8.1	Flow diagrams of a.) the serial latency assumption and b.) the parallel latency assumption. . . . .	142
8.2	Flow diagram for the dynamic TB disease model with demographics and vaccination described. . . . .	147
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 ETS system for notifications between 2000 and 2012. . . . .	160
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 ETS for notifications between 2000 and 2012. . . . .	161
8.5	a.) Mean contacts (non-unique social contacts per year) and the b.) normalised standard deviation (%) of 1000 bootstrapped samples of social contacts from the POLYMOD social contact survey using 5-year age groups up to 49 years old and then a single group for 50-69 year olds. . . . .	164
8.6	Distribution of the UK born population of England in 2000, 2004, 2008, and 2012. . . . .	168
8.7	Estimated and projected live births in England from 1929 until 2101. . . . .	169
8.8	Three year rolling average expected remaining lifespan stratified by age group in England from 2000 to 2014. . . . .	170
9.1	Observed and predicted TB incidence stratified by model age group . . . . .	193
9.2	Prior and posterior distributions for fitted model parameters. . . . .	195
9.3	Partial rank correlation coefficients for each parameter fitted too. . . . .	196
10.1	Forecast of TB incidence for each scenario evaluated from 2005 to 2040, stratified by age group. . . . .	204
10.2	Forecast of TB mortality for each scenario evaluated from 2005 to 2040, stratified by age group, stratified by age group. . . . .	206



# Acronyms

- ABC** - Approximate Bayesian Computation
- aHR** - adjusted Hazard Ratio
- AIC** - Akaike Information Criterion
- aOR** - adjusted Odds Ratio
- AUC** - Area Under the Curve
- BCG** - Bacillus Calmette–Guérin
- CI** - Confidence Interval
- COVER** - Cover of Vaccination Evaluated Rapidly
- CRAN** - Comprehensive R Archive Network
- CrI** - Credible Interval
- DHSS** - Department of Health and Social Security
- DIC** - Deviance Information Criterion
- DOTS** - Directly Observed Treatment Short-course
- ETS** - Enhanced Tuberculosis Surveillance System
- HIV** - Human Immunodeficiency Viruse
- IMD** - Index of Multiple Deprivation
- IQR** - Interquartile Range
- IRR** - Incidence Rate Ratio
- JCVI** - Joint Committee on Vaccination and Immunisation
- LFS** - Labour Force Survey
- LOESS** - Locally Estimated Scatterplot Smoothing
- looIC** - Leave One Out Information Criterion
- LRT** - Likelihood Ratio Test
- MAR** - MissingaAt Random

**MCAR** - Missing Completely at Random

**MCMC** - Markov Chain Monte Carlo

**MDR** - Multi-drug Resistant

**MNAR** - Missing Not at Random

**MRC** - Medical Research Council

**MRSA** - Methicillin-resistant *Staphylococcus aureus*

**NTM** - Non-Tuberculosis Mycobacteria

**ONS** - Office for National Statistics

**OR** - Odds Ratio

**PHE** - Public Health England

**PMCMC** - Particle Markov Chain Monte Carlo

**POLYMOD** - Improving Public Health Policy in Europe through the Modelling and Economic Evaluation of Interventions for the Control of Infectious Diseases

**PRCC** - Partial Rank Correlation Coefficient

**QALY** - Quality-adjusted Life-Year

**RR** - Risk Ratio

**SD** - Standard Deviation

**SE** - Standard Error

**SMC** - Sequential Monte Carlo

**SMC-SMC** - Sequential Monte Carlo - Sequential Monte Carlo

**TB** - Tuberculosis

**TST** - Tuberculin Skin Testing

**UK** - United Kingdom

**WHO** - World Health Organization

# 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.[1] 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 populations; where incidence rates can be as much as 15 times higher than in the general population.[2] Globally, TB remains the leading cause of death from infectious disease.[3]

The Bacillus Calmette–Guérin (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.[3] However, there is strong evidence that BCG offers high levels of protection for children, and more generally within the UK born population.[4] It remains the only licensed TB vaccine with over 100 million doses given globally each year. Serious side effects are 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.[5] 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.[5] Since this change in policy, declining incidence appears to support this decision.[2] 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.[2] This dataset 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 datasets 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 in-

forming global efforts to control TB. As a first step I outline some of the key background information and discuss the tooling that I developed in order to explore this information more fully. I then make use of the detailed PHE routine TB surveillance data to explore the current epidemiology of TB in England. Next, I 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, semi-stochastic, 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. This thesis is also available as a website<sup>1</sup>, a pdf<sup>2</sup>, and a reproducible Rmarkdown document<sup>3</sup>

The remainder of this chapter outlines: the theoretical framework used in this thesis; the aims and objectives 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 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 data and results.

### **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.[6] 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. Use cases include: adjusting for confounding variables when estimating the

---

<sup>1</sup>Website: <https://www.samabbott.co.uk/thesis/>

<sup>2</sup>PDF: <https://www.samabbott.co.uk/thesis/thesis.pdf>

<sup>3</sup>Rmarkdown: <https://github.com/seabbs/thesis>

relationship between BCG vaccination and TB outcomes (Chapter 6); 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.[7] Mechanistic models can be used to simplify complex real-world systems, whilst retaining a linkage to real-world processes.[8] They are 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.[7,8] Both of these approaches can be represented as deterministic, semi-stochastic (deterministic with some stochastic elements),[9] or fully stochastic - both of the latter include 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.[10] 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 semi-stochastic compartmental model of TB 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.[7,8] Movements between these groups are then modeled using a series of differential equations. Transmission is modeled using mass action,[7,8] where infected cases are assumed to randomly interact with susceptible individuals at a rate dictated by the concentration of susceptibles in the population. 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. Transition between compartments is assumed to be exponential. See [7,8] for a theoretical introduction to infectious disease models and [10] 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 TB 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 TB in England, in the context of global TB 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 TB that 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 TB and the BCG vaccine. This information helps motivate future chapters and may be useful for non-subject area experts.
- **Chapter 3:** `getTBinR`, an R package that facilitates downloading TB 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 TB in England, using routine surveillance datasets. Focusing on: the impact of missing data; the mechanisms underlying that missing data; seasonal trends; the role of age; UK birth status; BCG status; trends in TB incidence rates 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 capturing 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 a series of 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 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.

- **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 ongoing impact of each policy is then discussed through to 2040.
- **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 outlined.

## 1.4 Thesis output

This thesis has produced: peer reviewed papers; preprints; talks at academic conferences; open source research software; open source 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>
- Abbott, S., Christensen, H., Lalor, M. K., Zenner, D., Campbell, C., Ramsay, M. E., & Brooks-Pollock, E. (2019). *Exploring the effects of BCG vaccination in patients diagnosed with tuberculosis: Observational study using the Enhanced Tuberculosis Surveillance system*. Vaccine, 1–6. doi: <http://doi.org/10.1016/j.vaccine.2019.06.056> (preprint: <https://doi.org/10.1101/366476>)
- Abbott S., Christensen H., Welton N.J., Brooks-Pollock E. (2019) *Estimating the effect of the 2005 change in BCG policy in England: a retrospective cohort study, 2000 to 2015*. Eurosurveillance, 49(24). doi: <https://doi.org/10.2807/1560-7917.ES.2019.24.49.1900220> (preprint: <https://doi.org/10.1101/567511>)

### 1.4.2 Papers under review

- Abbott S., Christensen H., Brooks-Pollock E. *Reassessing the evidence for universal school-age BCG vaccination in England and Wales*, 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 TB 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 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 analyses in this thesis and to expedite the work of others using data from the Enhanced Surveillance System (ETS) (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 of 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; and a scenario analysis function. Used by the explore infectious disease model dashboard (<http://seabbs.co.uk/shiny/exploreidmodels/>) for all functionality. Available from CRAN using `install.packages("idmodelr")` or install the development version from GitHub with `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: <https://github.com/seabbs/getTBinR/>
- **Explore Tuberculosis in England and Wales:** Developed to allow public Public Health England TB Notification data to be explored interactively. Key interventions are highlighted and link to trends in TB notifications. This app is used in its static form in Chapter 2. Link: [https://github.com/seabbs/TB\\_England\\_Wales/](https://github.com/seabbs/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 viewable format to allow for users to develop their own models. Link: <https://github.com/seabbs/exploreidmodels/>
- **Introduction to Tuberculosis models:** Developed to allow simple TB models to be explored in an interactive session. Inspired by practicals from the Introduction to

## 1.5. Summary

---

TB, run by TB MAC (<http://tb-mac.org/>) at the 2017 Union conference. Link: [https://github.com/seabbs/intro\\_to\\_tb\\_models/](https://github.com/seabbs/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: <https://github.com/seabbs/thepebblegame/>

### 1.4.4 Talks

- **Assessing the Evidence for Universal 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/>
- **What do we really know about BCG?** - UK Clinical Vaccine Network Conference 2019, Oxford, United Kingdom. Link: <https://www.samabbott.co.uk/talk/uk-vac-network-2019/>

## 1.5 Summary

- This chapter provides an introduction to TB 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 (symptomatic) TB in their lifetime.[11] Of those with active TB around 10% are likely to die from TB or TB related causes.[3] TB is preventable and curable, but the majority of active cases occur in less economically developed countries and may not be diagnosed.[3] In the following section, the natural history of TB, TB risk factors, TB treatment, global impact, and the impact TB in England and Wales are explored.

#### 2.1.1 Natural history of TB

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 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.[3] The remaining 90-95% 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.[12]

Both active and latent TB represent a range of diverse individual states. Pulmonary cases are typically responsible for the vast majority of transmission.[13] Latent cases may completely clear the bacterium or be asymptotically carrying reproducing active TB bacterium.[12] Adolescents have the highest risk of developing active TB, usually in the form of pulmonary TB.[3] The risk of developing pulmonary TB, versus extra-pulmonary TB, varies with age. For instance, younger children are more likely to develop pulmonary TB.[3]

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 Risk factors

TB has been associated with several risk factors, the most common of which is human immunodeficiency virus (HIV). HIV increases the rate of activation 20-fold and TB is the

most common cause of AIDS-related death.[14] 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.[15–17]

### **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 Organization (WHO) now recommends the use of the Directly Observed Treatment short-course (DOTS), which focuses on 5 action points.[18] 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 extit-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 and there is only strong evidence of protection for 10-15 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 antibiotic 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 prophylactic 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 TB.
1952	Pyrazinamide	Antibiotic	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 its 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 pregnancy, 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 introduced by the World Health Organization 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 extit-Staphylococcus aureus; DOTS - Directly Observed Treatment Short-course  
*(continued)*

Year	Intervention	Type	Line	Detail
2012	Bedaquiline	Antibiotic	Second	The first new antibiotic 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,[3] of which 4.3 million were estimated to have been missed by health systems.[19] Global incidence rates have decreased year-on-year since the early 2000s, 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: TB incidence rates (per 100,000) by region and globally from 2000 until 2017. Globally incidence rates have been declining since the early 2000s but this decline varies with region.

Regional incidence rates only tell part of the story, as TB incidence rates vary significantly within regions. India, Indonesia, China, Nigeria, Pakistan, and South Africa account for 60% of new cases. India, Indonesia, and Nigeria are thought to be responsible for nearly half of all undiagnosed cases.[18] 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 TB 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. Incidence rates in the legend refer to the lower bound for each colour.

TB remains one of the top 10 causes of death worldwide, leading to 1.7 million deaths in 2016 alone.[3] The absolute number of deaths due to TB has fallen since 2000, with an average annual global rate of decline in TB mortality rates of 2.9% between 2000-2016. 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.

## 2.1. Tuberculosis

---



Source: World Health Organisation

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

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% confidence interval (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 2018. Note the high percentage of TB cases with HIV in sub-saharan Africa. The percentages of HIV in incident TB in the legend refer to the lower bound for each colour.

Multi-drug-resistant TB (MDR-TB), which is defined as being resistant to at least isoniazid and rifampin, made up 4.6% of all incident TB cases globally 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.[18] 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. 92% of rifampicin cases in Russia in 2018 also had MDR-TB, which is comparable to the global median of 81% (Table 2.2). Across all regions, Europe had the highest median percentage of rifampicin cases with MDR-TB (87%), with the Western Pacific having the lowest (78%).

## 2.1. Tuberculosis

---



Figure 2.5: Global map of the estimated percentage of new TB cases with rifampicin resistance (percent) in 2018. Note that a far higher percentage of TB cases have rifampicin resistance in the former Soviet Union than in the rest of the world. The percentages of rifampicin resistances in incident TB in the legend refer to the lower bound for each colour.

Table 2.2: Percentage (%) of rifampicin resistant TB cases that have multi-drug resistant TB in Russia and regional medians, with interquartile ranges. The estimate for Europe includes data from the Russian Federation.

WHO region	Median Proportion of Rifampicin resistant cases with MDR* (2.5% to 97.5% quantiles)
Global	81.0 (6.8 to 100.0)
Western Pacific	78.0 (0.0 to 100.0)
South-East Asia	100.0 (77.0 to 100.0)
Europe	87.0 (17.2 to 100.0)
Eastern Mediterranean	83.0 (29.5 to 100.0)
Americas	73.0 (34.2 to 93.9)
Africa	81.0 (50.5 to 100.0)
Russian Federation	92.00 ***

\*\* Multi-Drug Resistant TB

\*\*\* Only point estimates available

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<sup>1</sup>). However, in the past several decades, incidence rates first stabilised and have since increased since their lowest point in the 1990s. 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.[20] Figure 2.6 includes the interventions discussed above

<sup>1</sup> or see [http://www.seabbs.co.uk/shiny/TB\\_England\\_Wales](http://www.seabbs.co.uk/shiny/TB_England_Wales) for an interactive dashboard

(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.

### 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 per 100,000, in 2016).[2] 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) [2]. 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 (when it was 49%).[2] 23.3% (420/1,800) of non-UK born cases had traveled outside the UK in 2016 with the majority returning to their country of origin. Incidence rates in

## **2.1. Tuberculosis**

---

the UK born are between 3 and 14 times higher in non-White ethnic groups compared to the White ethnic groups.[2]

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, in 2016).[2] 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.[2] 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.[2] 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.[2]

### **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 often 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.[2] 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%).[2] 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 [2]. 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).[2]

### **Pulmonary Vs. Extra-Pulmonary TB**

Figure 2.7 shows that since the 1980s the proportion of extra-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).[2] For more details on TB in England, see Chapter 4 and the Public Health England 2017 and 2018 TB reports from which the summary data discussed above was extracted.[2,21]

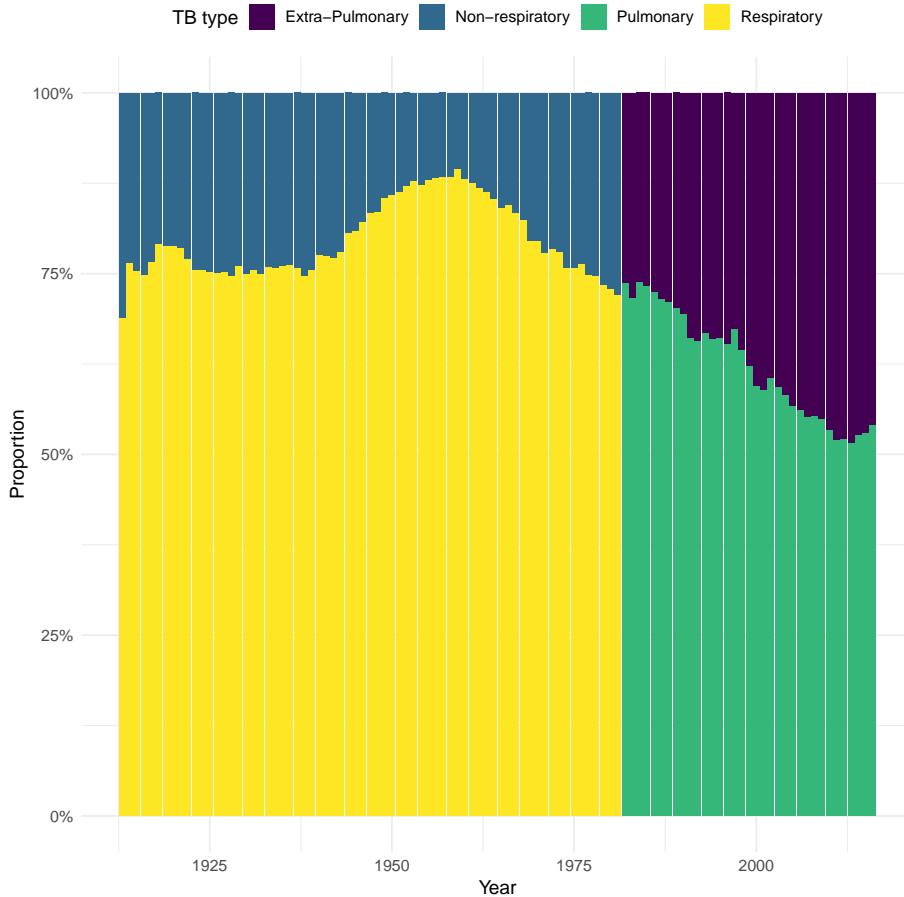


Figure 2.7: From 1913 until 1981 the figure shows the proportion respiratory vs. non-respiratory cases and from 1982 it shows the proportion of pulmonary vs. non-pulmonary TB.

## 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.[22] 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.[3] Serious side effects are rare, although a small scar at the injection site is common. Initially, public acceptance was slow, with low take up until after the Second World War. However, controversy remains and several countries have scaled back, or retargeted, their usage in recent years. 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.[4] There is no evidence that BCG vaccination of latent TB cases provides protection from developing active TB disease.[3]

### **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 for both pulmonary TB and TB meningitis.[23] Efficacy in adults ranges from 0% to 78%,[25] with an MRC trial in England finding that BCG was 78% effective in the White UK-born.[26] Effectiveness at preventing initial infection has been estimated at 19% (95% CI: 8% to 29%) globally.[4]

A meta-analysis of RCTs indicated that increased protection is associated with distance from the equator.[25] One hypothesis for this is that there is a greater density of mycobacteria near the equator that may mask, or block, the protection offered by the BCG vaccine.[27] Recently it has been found that much of this latitude effect may be due to the robustness of tuberculin skin testing (TST), with lower stringency near the equator.[3] TST screening tests for the presence of TB 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 latent TB 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. Regardless of the mechanism this effect means that vaccination early in life maximises the protection conferred by the BCG vaccine in countries with evidence of decreasing protection with age.

### **2.2.3 Duration of protection**

The effectiveness of the BCG vaccine has been shown to reduce over time.[28] However, there is good evidence that protection can last up to 10 years, with limited evidence of protection beyond 15 years.[28] Although, a recent study found that protection from active TB may extend later into life in England.[29] There is little evidence to suggest that re-vaccination boosts the protection offered by initial vaccination.[3]

The limited duration of protection has informed vaccination policy globally.[5] 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.[30,31] However, there is now some evidence that BCG vaccination induces innate immune responses that may provide non-specific protection[32] and reduce all-cause neonatal mortality.[33,34] There is some evidence that this reduction in all-cause mortality extends later in life.[35] 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 (detection of TB bacterium in a sputum sample).[36] There is also evidence to support an association between BCG vaccination and reduced TB mortality.[28] The evidence for additional effects of BCG vaccination on outcomes in individuals with notified TB, in England, is explored further in Chapter 6.

In addition to its effect on TB outcomes, the BCG vaccine has also been found to be effective at preventing leprosy (with a risk ratio (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.[3] Additionally, there is some evidence that the BCG vaccine can provide protection against Non-TB mycobacteria (NTM) infections, with an estimated effectiveness of 50%. [3] This protection has been shown to greatly reduce the incidence of NTM lymph node inflammation and protect against Buruli ulcers for the first 12 months following vaccination (RR, 0.50 [95% CI, 0.37–0.69]). A study also found that individuals with Buruli ulcers are less likely to develop osteomyelitis (bone infection) if they have a BCG scar (RR, 0.36 (95% CI: 0.22 to 0.58)).[37]

### **2.2.5 Usage globally**

The BCG vaccine is one of the mostly widely-used vaccines worldwide, with approximately 100 million doses given annually.[38] However, due to the variable estimates of BCG efficacy, vaccination has been controversial since its development. The 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.[3,39] This recommendation is based on the strong evidence that the BCG is highly protective in children (particularly against disseminated TB and TB meningitis),[23,24] whilst its effectiveness has been shown to vary with latitude when given later in life.[40] Historically, different strategies have been utilized worldwide. These include universal vaccination of those at most risk of onwards transmission and high-risk group vaccination targeting either neonates or children.[28]

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.[5] 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 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.[5] Most countries began universal programmes between the 1940s and 1980s due to high levels of TB incidence and strong evidence of the effectiveness of the BCG vaccine.[26] 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.[5] Globally, in countries that have BCG vaccination policies in place, coverage is thought to range from 70% to 100%.

### 2.2.6 Usage in England

In England, universal school-aged vaccination was introduced after an MRC trial in the 1950s estimated BCG's effectiveness at 78% in the white UK born population.[26] In 2005, the UK shifted from this strategy to targeted vaccination in neonates deemed at high risk.[28] This change was a reflection of current WHO vaccination policy,[39] falling TB incidence rates, an increasing proportion of TB cases occurring in the non-UK born,[2] and modelling evidence that suggested stopping the BCG schools scheme would have minimal long term effects on incidence rates.[41] 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 and 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%. [2] These estimates may not be reliable as COVER has only relatively recently begun to include returns for BCG, meaning that data quality may be poor. Prior to the switch to targeted neonatal vaccination, coverage in those at school leaving age was thought to be approximately 75%. [41]

### 2.2.7 Replacement vaccines

Multiple replacement vaccines are currently in clinical trials.[5,42] Vaccine candidates include both live and sub-unit vaccines. Many of these candidate vaccines serve as a boost to the BCG vaccine, with the BCG vaccine being administered prior to the candidate vaccine.[5] Several BCG replacements are also being trialed, both based on alternative methods of attenuating TB mycobacteria and using other approaches.[42] However, in the short-to-midterm it is unlikely that a new vaccine will replace the BCG vaccine. This means that its 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 TB.
- This chapter also details the action, effectiveness, duration of protection, effects and usage of the BCG vaccine. The only licensed vaccine for TB.
- Motivation is given for the remaining chapters in this thesis.



# Chapter 3

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

### 3.1 Introduction

Developing tools for rapidly accessing and exploring datasets benefits the public health research community by enabling multiple datasets to be combined in a consistent manner, increasing the visibility of key datasets, 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 datasets 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[43] that I developed to facilitate working with the data[19] collated by the World Health Organization (WHO) on the country level epidemiology of Tuberculosis (TB). All data is freely available from the WHO<sup>1</sup> and the package code is archived on Zenodo<sup>2</sup> and Github<sup>3</sup>. The aim of getTBinR is to allow researchers, and other interested individuals, to quickly and easily gain access to a detailed TB dataset 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 dataset before they are explored further using more complex methods or more detailed data. Prior to the development of getTBinR access to the WHO data was ad-hoc and there were no standardised visualisation or summary tools.

The remainder of this chapter outlines the structure, and key functionality, of getTBinR 0.6.1. The use of getTBinR in this thesis is explored as well as the use of the package in work external to this thesis and by others. Much of the work done for this chapter was code, documentation, and case study development and so is not fully captured here. The

---

<sup>1</sup>WHO TB data: <https://www.who.int/tb/country/data/download/en/>

<sup>2</sup>Zenodo: <https://zenodo.org/badge/latestdoi/112591837>

<sup>3</sup>GitHub: <https://github.com/seabbs/getTBinR/>

GitHub repository contains a full development history of the package, as well as providing links to the documentation and accompanying case studies.

## 3.2 Installation

`getTBinR` has been released to the Comprehensive R Archive Network (CRAN) and can therefore 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,

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

## 3.3 Data extraction and variable look-up

The data sourced by `getTBinR` is collated by the WHO, via member governments, and used to compile the yearly global TB report.[19] Data collation and estimation 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 [19]. 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.

`getTBinR` provides a single user facing function for data extraction, `get_tb_burden`. This function downloads multiple datasets from the WHO, cleans variables names where required, and finally joins all datasets together. On top of the core datasets provided by default, `getTBinR` also supports importing multiple other datasets. These include data on latent TB, HIV surveillance, intervention budgets, and outcomes (see `?additional_datasets` for a full list of available datasets). 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, key phrases, and datasets. 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
```

### 3.4. Data visualisation

---

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

`getTBIInR` 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) +
  theme(text = element_text(size = 12))
```



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) +
  theme(text = element_text(size = 12))
```

### 3.4. Data visualisation

---

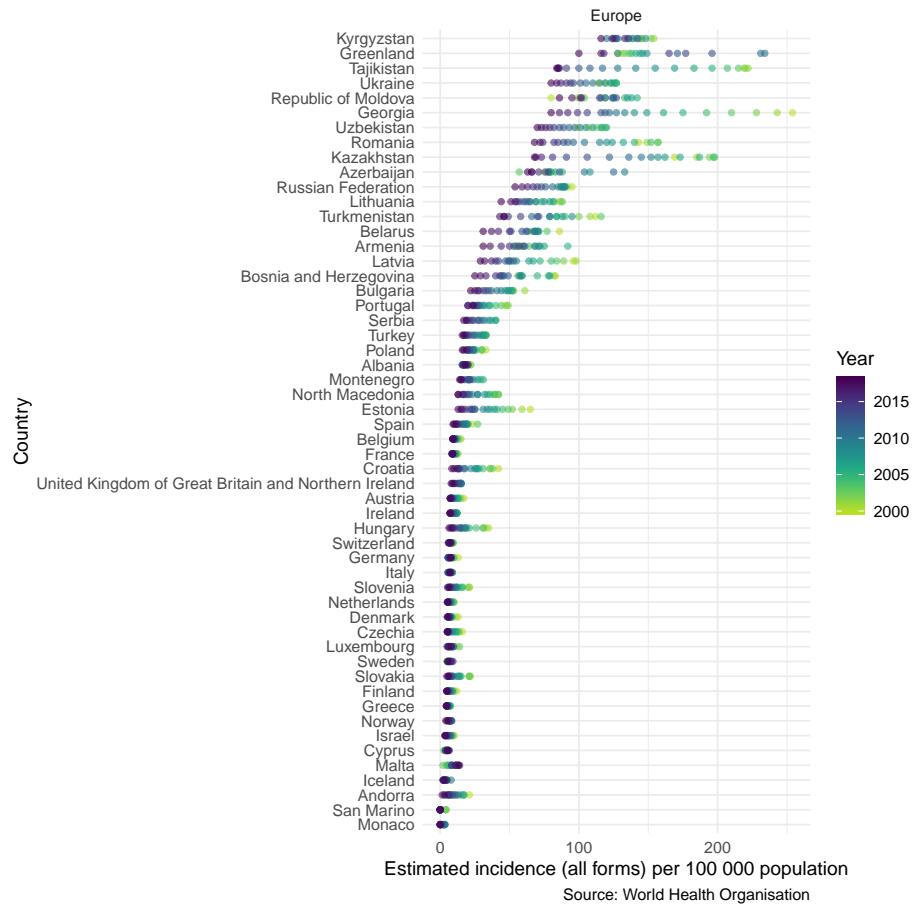


Figure 3.2: Dot plot showing trends over time in TB incidence rates in Europe ordered by TB incidence rates in 2017.

#### 3.4.3 Plotting a comparison 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) +
  theme(text = element_text(size = 12))
```



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) +
  theme(text = element_text(size = 12))
```



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 datasets 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: 152 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 Britain and Northern Ireland 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      14.6      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 142 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()
```

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.

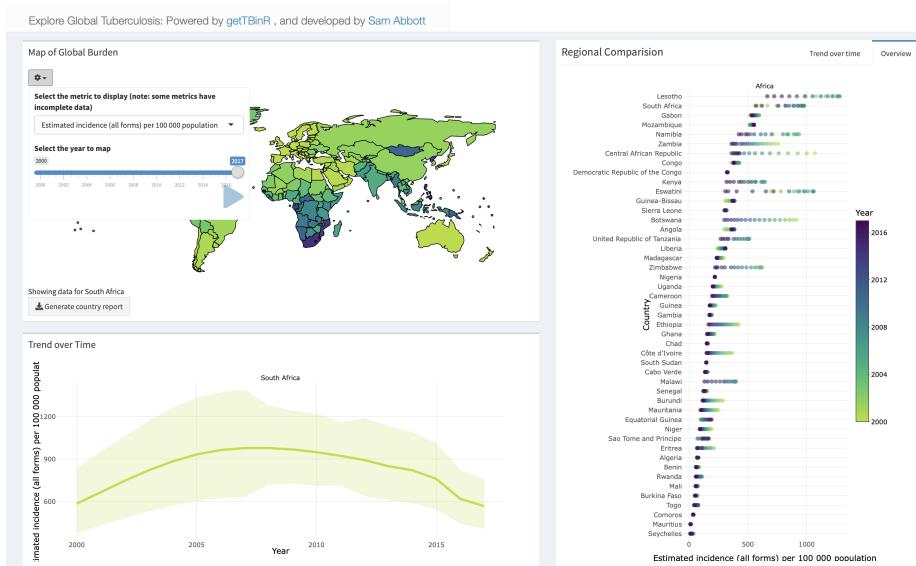


Figure 3.5: Snapshot of the built in package dashboard.

## 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

### 3.9. Package infrastructure

---

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.

## Tuberculosis Report

### TB incidence rates

In 2017 United Kingdom had an estimated Tuberculosis incidence rate of 8.9 (8.1 - 9.8) per 100,000 people making it number 165 in the world and number 32 regionally. In the last 10 years this has changed by -4.9% on average each year.

### Regional and Global Trends Comparison

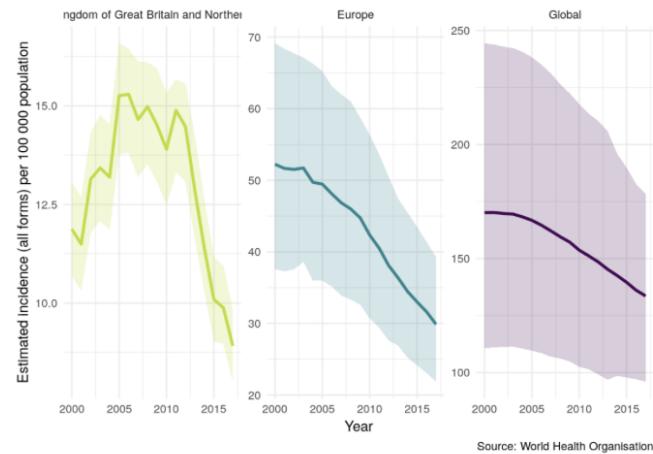


Figure 3.6: Screenshot of the start of the built-in package summary report, for the United Kingdom.

## 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<sup>4</sup> and a development environment can be launched with a single button press<sup>5</sup>. Use cases for the package have been outlined using multiple case studies, see the package documentation for details.

## 3.10 Discussion

In this chapter I have introduced the `getTBinR` R package. `getTBinR` facilitates downloading the most up-to-date version of multiple TB relevant data sources from the WHO, along with

<sup>4</sup>Website: <https://www.samabbott.co.uk/getTBinR/>

<sup>5</sup>Binder: <https://mybinder.org/v2/gh/seabbs/getTBinR/master?urlpath=rstudio>

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. It was developed using R package development best practices and has been peer-reviewed.[44]

As of the 8th of August 2019, `getTBinR` has been released on CRAN for over a year. It has been downloaded over 10,000 times, has a growing user base, no outstanding bug-related issues, and has received multiple updates greatly expanding the functionality available. The standalone dashboard hosted online<sup>6</sup> has had over 3000 unique users. It has been used as a teaching aid, as an example of open science, to facilitate exploratory data analysis and to provide context for other research. In this thesis, it has been used extensively in Chapter 2 to provide context and was also used as a hypothesis generating tool in all other chapters. Outside of the work presented in this thesis I have used `getTBinR` extensively as a data analysis tool<sup>7</sup>, to widen the appreciation of TB as a global health problem<sup>8</sup>, and to provide contextual information for other research<sup>9</sup>.

Whilst `getTBinR` is feature complete, and stable, development work continues. Future projects include: using the `shinymeta` R package to provide downloadable R code to users of the interactive application; iterating on the current automated report to improve formatting and to increase the amount of information displayed; and expanding the range of visualisation functions available. As additional WHO TB data are released they will be added to `getTBinR`.

### 3.11 Summary

- In this chapter I have introduced `getTBinR` an R package for accessing, summarising and visualising the WHO TB surveillance dataset 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.
- The current impact and future direction of `getTBinR` has also been detailed.

---

<sup>6</sup>Dashboard: <http://www.seabbs.co.uk/shiny/ExploreGlobalTB/>

<sup>7</sup>Blog: <https://www.samabbott.co.uk/tags/who/>

<sup>8</sup>Twitter: <https://twitter.com/search?q=getTBinR&src=typd>; Reddit: <https://www.reddit.com/user/seabbs/posts/>

<sup>9</sup>Presentation: <https://www.samabbott.co.uk/what-do-we-really-know-about-bcg/presentation.html#1>

# Chapter 4

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

### 4.1 Introduction

Although the characteristics of tuberculosis (TB) in England have been reported elsewhere,[2,21] and key risk factors such as non-UK birth status have been identified,[45] 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 datasets with a particular focus on the impact of missing data, the mechanisms underlying that missing data, seasonal trends, the role of age, UK birth status and BCG status. I also estimate incidence rates, stratified by UK birth status and age, which I use 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 are then 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 9).

### 4.2 Data sources

#### 4.2.1 Enhanced tuberculosis surveillance (ETS) system

##### Background

The ETS is a database that collects demographic, clinical, and microbiological data on all notified TB cases in England and is maintained by Public Health England (PHE). Notification is required by law, with health service providers having to inform PHE of

all confirmed TB cases.[2] Data collection began in 2000 and was expanded, with additional variables, with the launch of a web based system in 2008.[46] It is updated annually with de-notifications, late notifications and other updates. A descriptive analysis of TB 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.[2]

### **Data extraction and management**

Data on all notifications (114,820 notifications) from the ETS system from 2000 to 2015 were obtained from PHE via an application to the TB monitoring team. Data fields included: notification date, age, PHE 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. Notifications were assessed for identifiability and the data release was conditional on the raw data not being shared further. 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 dataset. 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<sup>1</sup>.

---

<sup>1</sup>Data cleaning code: [https://www.samabbott.co.uk/tbinenglanddataclean/reference/clean\\_munge\\_ets\\_2016.html](https://www.samabbott.co.uk/tbinenglanddataclean/reference/clean_munge_ets_2016.html)

## 4.2. Data sources

---

Table 4.1: Variables derived or modified from the ETS 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 $\leq 10$ and $11+$ due to the evidence of waning protection for the BCG vaccine.[28]
Age at BCG	Derived using year of vaccination and age at vaccination. Categorised into $< 1$ , $1$ to $11$ , $12$ to $16$ and $16+$ to capture historic vaccination policy.[47]
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 where TB contributed to their death, or 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 death, and survival is defined as those that completed treatment, were still on treatment, or stopped treatment. Those that were lost to follow up, or not evaluated were treated as missing
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 TB 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 are missing completely at random (MCAR), data that are missing at random (MAR) and data that are missing not at random (MNAR).[48] Data that are MAR are missing with a mechanism that is conditional on observed variables, whilst MNAR are 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 scenarios, and multiple imputation of missing data using additional variables in the dataset to inform the imputation model.[48]

As the ETS is aggregated across England, from a variety of sources, some level of missing data are inevitable. This takes two forms: under-reporting of notified cases, of which there is some evidence in the literature,[49] and data missing for a notified case. The former is particularly problematic as apart from using comparative studies the characteristics of those that are not notified is unknown. For variables that are missing data within the dataset 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. After accounting for nested variables, there was 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,[2,50] 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 ETS 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), PHE Centre (phec), Occupation (occat), Ethnic group (ethgrp), UK birth status (ukborn), Time since entry (timesinceent), date of symptom onset (symponset), 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 (tomdeathrelat), socio-economic status quintiles (natquintile), and date of death (dateofdeath). Nested variables have been accounted for (i.e date 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), PHE Centre (phec), Occupation (occatt), 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 (tomdeathrelat), 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.

Variable	Pre 2009		Post 2008	
	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
occatt	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

## 4.2. Data sources

---

For nested variables with rare outcomes an alternative approach for estimating the proportion of missing data is to first 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 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 where 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 potential MAR mechanisms from observed variables that would not necessarily be included in a model used for analysis, although any associations may themselves be caused by an external factor. 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 estimated. Unlike classic approaches to missing data, such as multiple imputation by chained regression (MICE),[51] this is not an imputation. The details of the approach are discussed below.

**Method** In order to reformulate missing data as a logistic regression I took the following steps:

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 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.[52] 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 were both highly complete and also plausibly linked to missingness for all outcomes considered. These were: year, sex, age (grouped as grouped as 0-14 year olds, 15-44 year olds, 45-64 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. 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 missingness 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.

## 4.2. Data sources

---

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, 15-45 year olds and 65+), ethnic group, UK birth status and socio-economic status (national quintiles). 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 was 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
Year	2010	31.3% (2235)	7143		
	2011	29.8% (2319)	7781	0.93 (0.87, 1.00)	
	2012	27.9% (2164)	7755	0.85 (0.79, 0.91)	
	2013	27.1% (1907)	7034	0.79 (0.74, 0.86)	
	2014	30.1% (1907)	6327	0.91 (0.85, 0.98)	
Sex	2015	29.7% (1668)	5619	0.89 (0.82, 0.96)	
	Female	27.4% (4847)	17664		8.74e-11
Age	Male	30.6% (7353)	23995	1.16 (1.11, 1.21)	
	0-14	13.1% (235)	1793		1.67e-157
	15-44	26.0% (6557)	25235	2.10 (1.82, 2.43)	
	45-64	32.8% (2964)	9026	2.84 (2.45, 3.30)	
	65+	43.6% (2444)	5605	4.42 (3.80, 5.15)	
Ethnic group	White	35.4% (2959)	8359		2.15e-41
	Black-Caribbean	24.6% (228)	928	0.62 (0.52, 0.72)	
	Black-African	27.3% (1966)	7204	0.73 (0.67, 0.80)	
	Black-Other	24.1% (89)	369	0.65 (0.51, 0.83)	
	Indian	25.9% (2805)	10848	0.62 (0.58, 0.68)	
	Pakistani	33.2% (2258)	6806	0.89 (0.82, 0.97)	
	Bangladeshi	27.9% (469)	1680	0.71 (0.62, 0.80)	
	Chinese	33.6% (166)	494	0.88 (0.72, 1.07)	
	Mixed / Other	25.3% (1260)	4971	0.65 (0.59, 0.71)	
	Non-UK Born	29.5% (9104)	30880		7.2e-18
UK birth status	UK Born	28.7% (3096)	10779	0.75 (0.70, 0.80)	
	Socio-economic status	1	16131		4.88e-08
	2	26.8% (3383)	12621	0.84 (0.80, 0.89)	
	3	29.2% (1905)	6530	0.92 (0.86, 0.98)	
	4	30.1% (1142)	3796	0.91 (0.84, 0.99)	
	5	31.8% (822)	2581	0.94 (0.85, 1.03)	

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

**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

## 4.2. Data sources

---

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

**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.

## 4.2. Data sources

---

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

**Date of symptom onset** For date of symptom onset there was strong evidence of an MAR 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 grouped as 0-14 year olds, 15-44 year olds, 45-64 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 was 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
Year	2010	34.0% (2426)	7143		0
	2011	30.1% (2339)	7781	0.83 (0.78, 0.89)	
	2012	24.2% (1878)	7755	0.61 (0.57, 0.66)	
	2013	17.5% (1233)	7034	0.41 (0.38, 0.44)	
	2014	11.8% (744)	6327	0.25 (0.23, 0.28)	
Sex	2015	6.9% (390)	5619	0.14 (0.13, 0.16)	
	Female	22.0% (3894)	17664		0.93
Age	Male	21.3% (5116)	23995	1.00 (0.95, 1.05)	
	0-14	38.1% (684)	1793		3.59e-73
	15-44	20.5% (5182)	25235	0.35 (0.31, 0.39)	
Ethnic group	45-64	20.7% (1870)	9026	0.37 (0.33, 0.42)	
	65+	22.7% (1274)	5605	0.43 (0.38, 0.49)	
	White	20.9% (1749)	8359		3.98e-09
	Black-Caribbean	23.1% (214)	928	1.04 (0.88, 1.23)	
	Black-African	23.0% (1654)	7204	0.89 (0.80, 0.98)	
UK birth status	Black-Other	18.7% (69)	369	0.79 (0.60, 1.04)	
	Indian	22.2% (2404)	10848	0.86 (0.79, 0.94)	
	Pakistani	19.2% (1305)	6806	0.75 (0.68, 0.83)	
	Bangladeshi	23.9% (401)	1680	1.05 (0.91, 1.20)	
	Chinese	18.8% (93)	494	0.74 (0.58, 0.94)	
	Mixed / Other	22.6% (1121)	4971	0.93 (0.83, 1.03)	
	Non-UK Born	21.9% (6774)	30880		5.44e-12
	UK Born	20.7% (2236)	10779	0.77 (0.71, 0.83)	
	1	19.9% (3218)	16131		5e-17
	2	22.9% (2888)	12621	1.22 (1.15, 1.29)	
Socio-economic status	3	24.2% (1578)	6530	1.33 (1.24, 1.43)	
	4	22.0% (837)	3796	1.20 (1.09, 1.31)	
	5	18.9% (489)	2581	1.00 (0.89, 1.12)	

**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 date 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.

## 4.2. Data sources

---

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 grouped as 0-14 year olds, 15-44 year olds, 45-64 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 was 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
Year	2010	26.9% (1918)	7143		1.65e-283
	2011	22.3% (1736)	7781	0.78 (0.72, 0.84)	
	2012	18.8% (1458)	7755	0.63 (0.58, 0.68)	
	2013	12.9% (909)	7034	0.41 (0.37, 0.44)	
	2014	10.4% (659)	6327	0.32 (0.29, 0.35)	
	2015	7.4% (415)	5619	0.22 (0.19, 0.24)	
Sex	Female	16.9% (2984)	17664		0.0296
	Male	17.1% (4111)	23995	1.06 (1.01, 1.12)	
Age	0-14	19.4% (348)	1793		0.000164
	15-44	17.8% (4504)	25235	0.76 (0.67, 0.87)	
	45-64	15.9% (1434)	9026	0.73 (0.64, 0.84)	
	65+	14.4% (809)	5605	0.72 (0.62, 0.84)	
Ethnic group	White	12.5% (1043)	8359		2.91e-67
	Black-Caribbean	25.2% (234)	928	2.21 (1.87, 2.61)	
	Black-African	21.9% (1577)	7204	1.49 (1.34, 1.66)	
	Black-Other	17.9% (66)	369	1.32 (0.98, 1.74)	
	Indian	18.0% (1957)	10848	1.09 (0.99, 1.21)	
	Pakistani	11.8% (805)	6806	0.75 (0.67, 0.84)	
UK birth status	Bangladeshi	21.5% (361)	1680	1.57 (1.35, 1.82)	
	Chinese	13.4% (66)	494	0.82 (0.61, 1.07)	
	Mixed / Other	19.8% (986)	4971	1.32 (1.18, 1.48)	
	Non-UK Born	18.4% (5696)	30880		6.07e-16
	UK Born	13.0% (1399)	10779	0.71 (0.65, 0.77)	
Socio-economic status	1	14.4% (2317)	16131		1.05e-45
	2	19.6% (2469)	12621	1.48 (1.39, 1.58)	
	3	20.3% (1325)	6530	1.62 (1.50, 1.75)	
	4	17.0% (645)	3796	1.37 (1.24, 1.52)	
	5	13.1% (339)	2581	1.07 (0.94, 1.21)	

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

**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.

## 4.2. Data sources

---

Table 4.10: Results from a logistic regression model with data completeness (Complete/Missing) for date of ending treatment as an outcome, adjusted for: year, sex, age (grouped as grouped as 0-14 year olds, 15-44 year olds, 45-64 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 was used. Odds ratios shown are adjusted for all explanatory variables.

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

## Biases in the ETS

Routine observational datasets are subject to numerous potential biases, such as selection bias, recall bias, measurement bias, and unmeasured confounding.[53] 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 possible, 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 TB 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.[54] 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 the nearest year, month and day. These summary measures are displayed graphically using scatterplots (with trend lines) and violin plots. Violin plots are a compact method of plotting continuous distributions across multiple categories. They are effectively mirrored density plots and can be interpreted similarly. Here the underlying data has also been plotted.

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. It is also likely that some of the dates recorded are inaccurate or systematically biased. 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, reduces 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.)).

## 4.2. Data sources

---



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. 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% ( $3969/6044$ ) complete in 2000, 60.4% ( $4720/7809$ ) complete in 2008, and 87.7% ( $5677/6472$ ) complete 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.)). These biases may also be the result of the defaults used during data entry with the first of the month or the first month of the year being used when the exact date is not known. 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% ( $699/6044$ ) complete in 2000, to 89.4% ( $5786/6472$ ) complete 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.

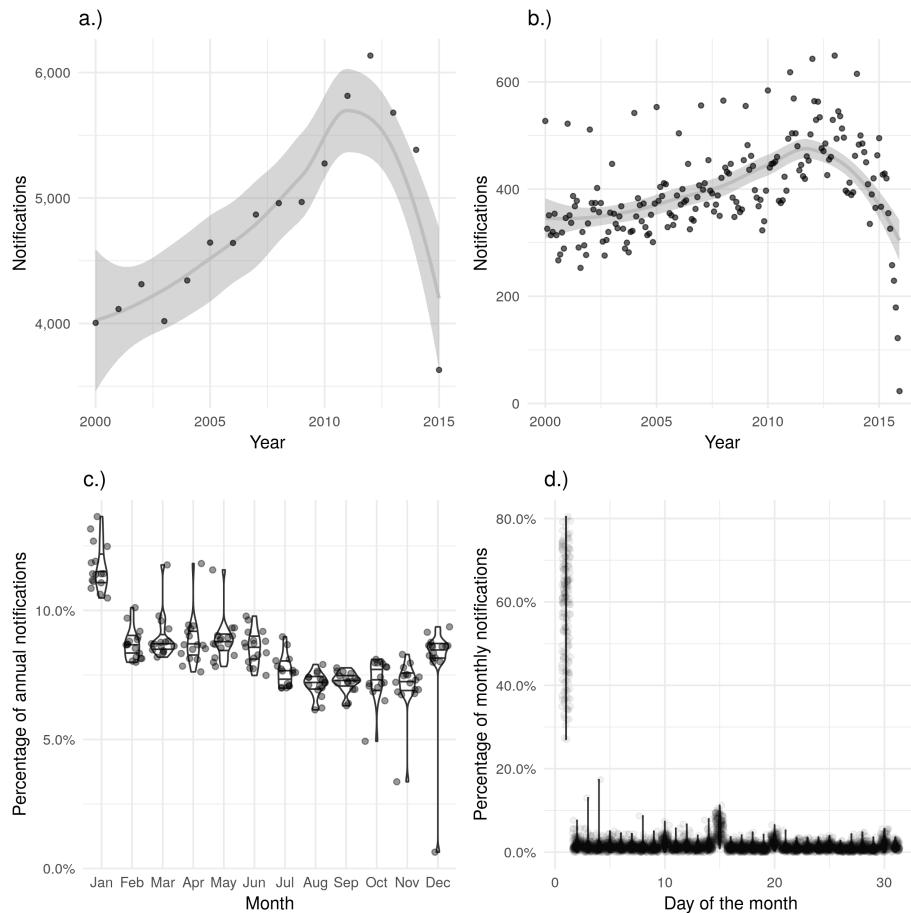


Figure 4.3: a.) and b.) show notifications over time by date of symptom onset in the ETS, with a.) aggregated by year and b.) aggregated by month. A trendline has been produced using a locally weighted regression model. Both of these plots show the same overall trend, but b.) contains a large amount 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 reporting 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

## 4.2. Data sources

---

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.)).

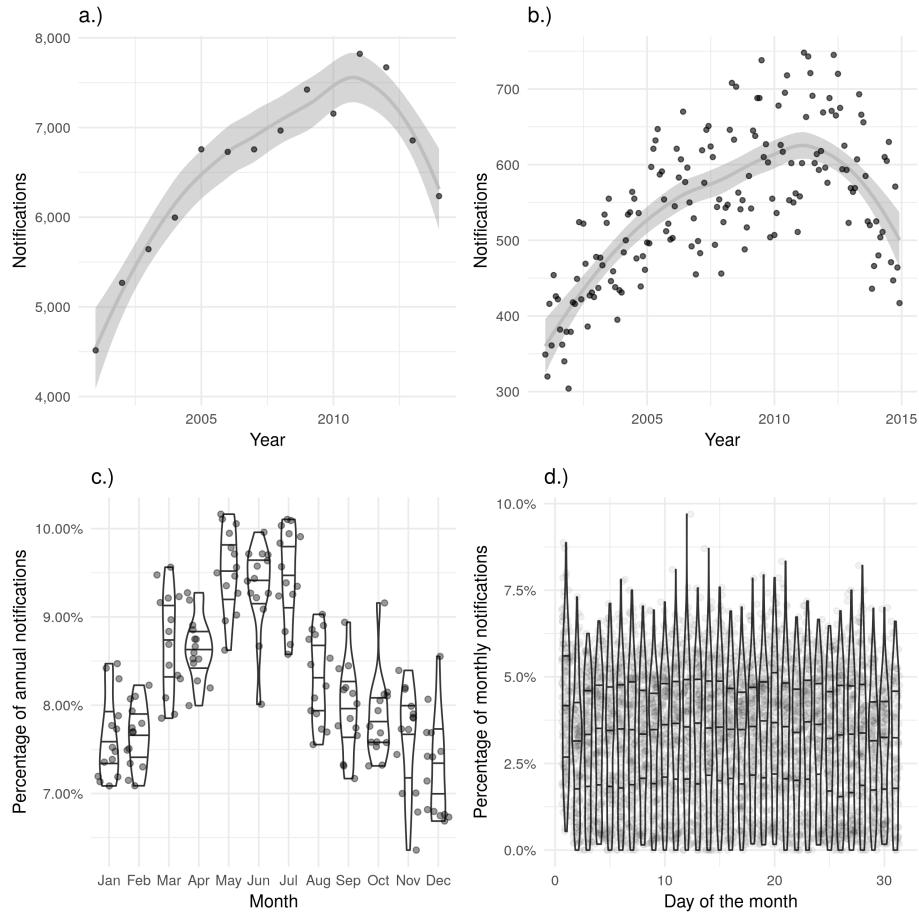


Figure 4.4: a.) and b.) show notifications over time by date of starting treatment in the ETS, with a.) aggregated by year and b.) aggregated by month. A trendline has been produced using a locally weighted regression model. Both of these plots show the same overall trend, but b.) contains a large amount 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.). Also as previously noted, these biases may also be the result of the defaults used during data entry. There were also several outlier months in which all notifications were reported as having their treatment on the same date. This is highly unlikely and may indicate an additional data quality issue. 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.

## 4.2. Data sources

---

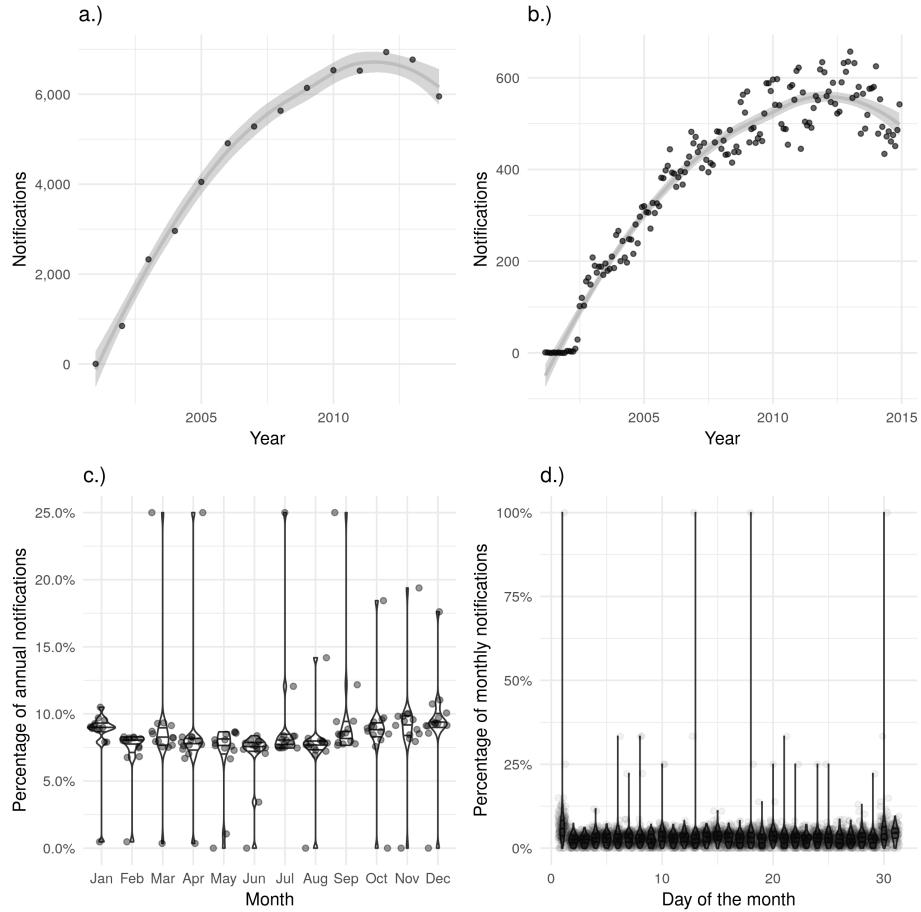


Figure 4.5: a.) and b.) show notifications over time by date of treatment ending in the ETS, with a.) aggregated by year and b.) aggregated by month. A trendline has been produced using a locally weighted regression model. Both of these plots show the same overall trend, but b.) contains a large amount 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. d.) also contains several clear outliers with data from some months indicating that 100% of notifications had their treatment on the same day. 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

## 4.2. Data sources

---

Force Survey (LFS) stratified by single year of age and UK birth status<sup>2</sup>. 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,[55] with the population estimates from 2000 being reformatted to match those from 2001 onwards. Data from the LFS was available by year, so each dataset was separately imported into R.[56] Reporting practices have changed with time so the appropriate variables for age, country of origin, country of birth, and survey weight (used to make survey responses representative of the general population) were extracted from each yearly extract, standardised, and combined into a single 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,[57] and the LFS are assessed for performance and quality elsewhere.[58,59] However, both have several failings 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 TB 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.

---

<sup>2</sup>Demographic data sources: <https://github.com/seabbs/tbinenglanddataclean>



Figure 4.7: Overall population estimates in England derived using ONS (Total) and LFS (Total (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 and non-UK born populations are estimated using the LFS data.

### 4.3. TB notifications

---



Figure 4.8: Percentage difference between ONS population estimates and estimates derived from the 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 TB 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 then there has been a 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 status 66.8%

(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.[2] 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 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.[2] 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

#### 4.4. Population Demographics in England

---

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 total yearly notification by 5 year age group in the ETS 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 conduct-

ing 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 proportion of the population in each 5 year age group stratified by UK birth status for 2000, 2008, and 2016.

## 4.5 TB 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 yearly PHE TB report,[2] they are limited in detail and do not report age stratified, or 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:

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

Age-standardised rates were calculated using the epiR package for R,[60] using the average age distribution of England from 2000-2015 as the standard population to allow comparison between years. Those aged 90+ were excluded as demographic data for this population were unreliable. The code used to calculate incidence rates is available online as an R package (`tbinenglanddataclean`<sup>3</sup>; see Chapter 1).

#### **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 has 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. This may be driven by a change in the composition of the non-UK born population or it may be the result of increased screening of those entering the UK. In comparison, incidence rates 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).

---

<sup>3</sup>`tbinenglanddataclean`: <https://github.com/seabbs/tbinenglanddataclean>

## 4.5. TB incidence rates

---



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 fallen 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.



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 highlighted by linking points with a line. Incidence rates declined overall in children over time. In adults incidence rates increased until 2011 and have since fallen. In older adults incidence rates consistently fell. In the non-UK born, incidence rate also fell in children but peaked earlier 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 have fallen. Incidence rates also increased over time in UK born adults until 2012 but has consistently fallen in UK born older adults.

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

#### *4.5. TB incidence rates*

---

(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.

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 2015, stratified by UK birth status. Point estimates are given along with 95% confidence intervals for each incidence rate estimate. Trends across age distributions are highlighted 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 available.

#### 4.5.5 Incidence rates in children (0-14 years old) as a proxy for TB transmission

Trends in incidence rates in UK born young children (0-14 years old) 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.[2] 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). The trend in UK born children diverging from that seen in the overall population may be interpreted as evidence that TB transmission increased from 2000 to 2008, and then decreased subsequently. 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,[2] the role of BCG vaccination has not previously been considered. There is some evidence that BCG vaccination may reduce all-cause mortality,[33–35] TB mortality,[28] and improve treatment outcomes.[36] 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` R package:

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

### 4.6.3 All-cause mortality

In 2015 fewer UK born cases died from any cause in the ETS than in 2000 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 than in the non-UK 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.

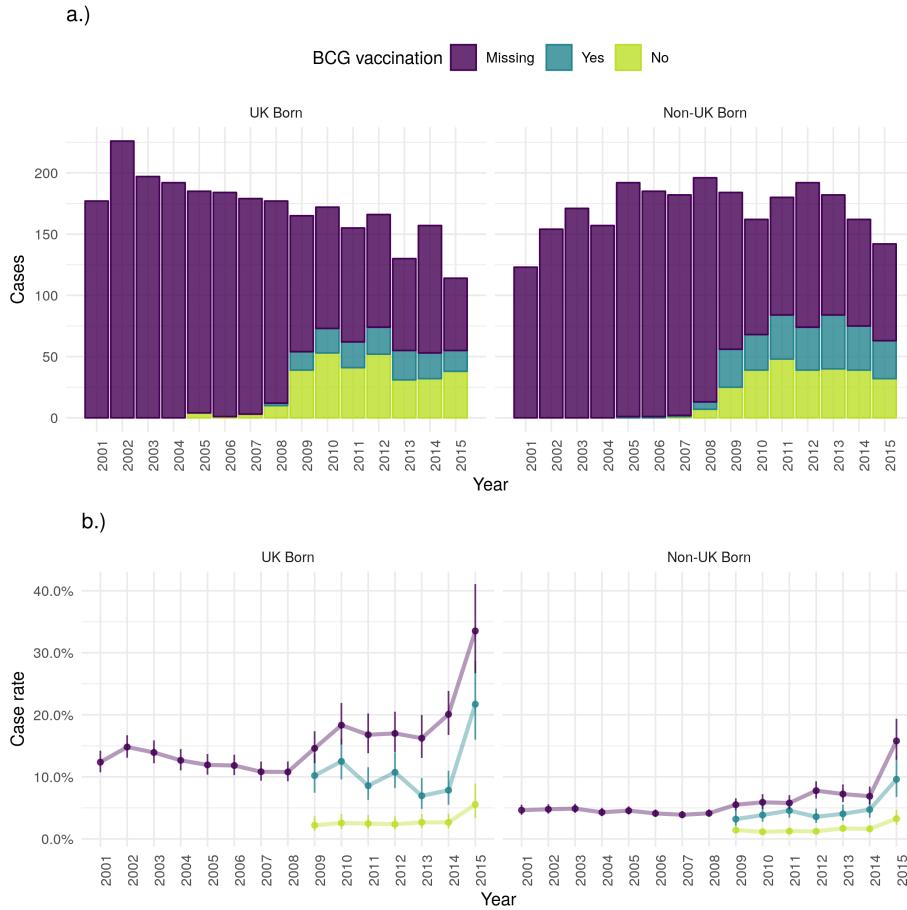


Figure 4.15: a.) Cases that died from any cause by year of notification stratified by UK birth and BCG status, b.) Case all-cause fatality rate stratified by UK birth and BCG status. Point estimates along with 95% confidence are shown for all estimates. All-cause mortality has reduced over time in the 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 likely to die from any cause. Data is incomplete for 2015, with cases that survived being potentially more likely to be missing than those that died. This may be the cause of the observed increase in uncertainty and may also have resulted in a biased mortality rate for 2015.



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. In both populations the case all-cause 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 TB related mortality rate in those unvaccinated 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.

#### 4.6. TB outcomes

---

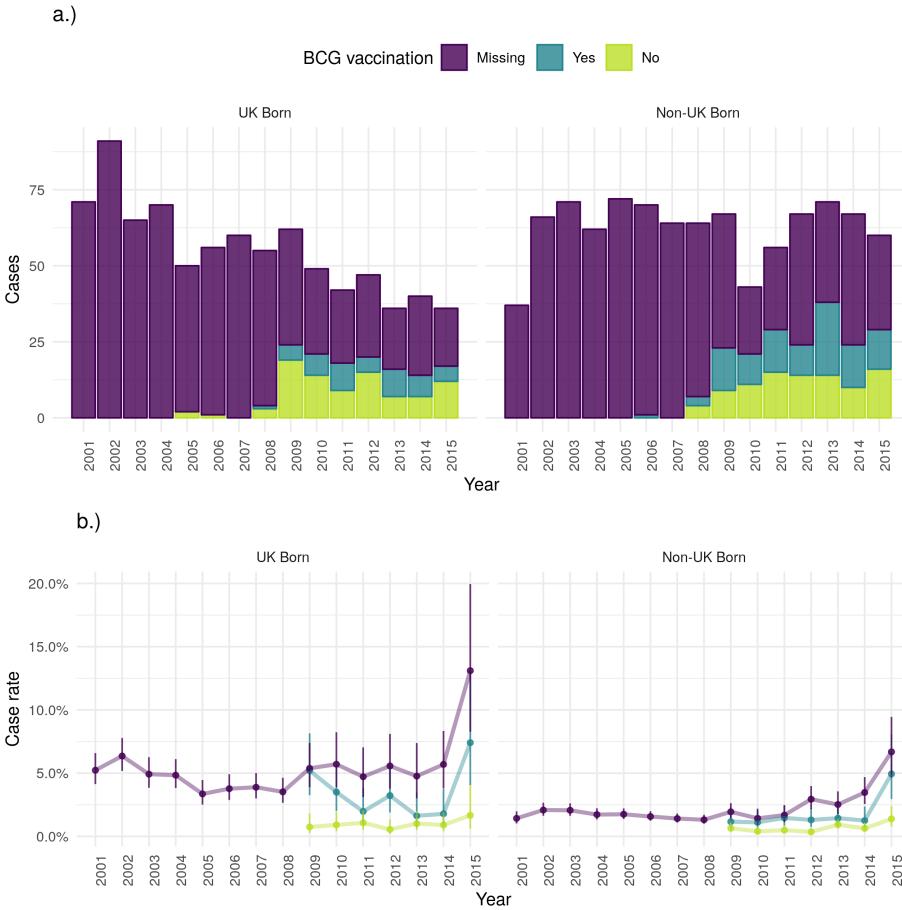


Figure 4.17: a.) Cases that died from TB by year of notification stratified by UK birth and BCG status, b.) Case TB fatality rate stratified by UK birth and BCG status. Point estimates along with 95% confidence are shown for all estimates. TB mortality has reduced over time in the UK born but remained 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 likely to die from TB. Data is incomplete for 2015, with cases that survived being potentially more likely to be missing than those that died. This may be the cause of the observed increase in uncertainty and may also have resulted in a biased TB fatality rate for 2015.



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. All estimates have a large degree of uncertainty making drawing conclusions difficult. There is no strong evidence to suggest a difference between those who 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).

#### 4.6. TB outcomes

---



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 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 case successful treatment rate varies by BCG status. Data is incomplete for 2015, with cases that were successfully treated being potentially less likely to be missing than those that were not. This may be the cause of the observed increase in uncertainty and may also have resulted in a biased successful treatment rate for 2015.



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. 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 appeared to be a particular issue in the non-UK born.

#### 4.6. TB outcomes

---



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. Data is incomplete for 2015, with cases that were lost to follow up being potentially less likely to be missing than those that were not. This may be the cause of the observed increase in uncertainty and may also have resulted in a biased lost to follow up rate for 2015.



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. 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 datasets, with a particular focus on the impact of missing data, the mechanisms underlying that missing data, 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 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

#### **4.7. Discussion**

---

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, there was a large degree of recall bias when aggregating by day or month. 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.

As reported elsewhere, I found that TB incidence initially increased from 2000 until 2011 but has since decreased. This was mainly driven by changing incidence in the non-UK born with a slight decrease in UK born incidence in recent years. Stratifying by age, I found that non-UK born cases were more likely to be 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 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, then increasing 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. These trends are not mirrored in the UK born with incidence rates initially increasing in children through to 2008 before beginning to decline. Incidence rates also increased 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 reliably used to proxy TB transmission.

Using case rates, I found that there was some evidence that cases who were not BCG vaccinated may be more likely to suffer from negative TB outcomes 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 9 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.
- TB 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 (`tbinenglanddataclean`; see Chapter 1), along with documentation of the required data sources and package functions. If interested in reproducing this work from the raw data 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 9).

# Chapter 5

## Reassessing the evidence for universal school-age 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 aim to update one of these studies.

I will recreate 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 investigate scenarios considered by the Joint Committee on Vaccination and Immunisation (JCVI), and explore new approaches using notification data (see Chapter 4). I will estimate 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 adapted from a preprint<sup>1</sup> supervised by Hannah Christensen and Ellen Brooks-Pollock.

### 5.2 Background

The Bacillus Calmette–Guérin (BCG) vaccine remains the only licensed vaccine for use against Tuberculosis (TB), although its use globally is controversial due to evidence of variable effectiveness,[25] and waning protection 10-15 years after vaccination.[28] Global usage of the BCG varies between no vaccination, universal vaccination, and high-risk group vaccination and may target either neonates or school-aged children.[5,61] The World Health Organization (WHO) recommends vaccination for all neonates as early as possible after birth in high burden settings, with vaccination in low burden settings being dependent on the country specific epidemiology of TB.[39] This recommendation is based on the strong

---

<sup>1</sup>Preprint: <https://doi.org/10.1101/624916>

evidence that the BCG is highly protective in children,[23,24] whilst its effectiveness has been shown to vary with latitude when given later in life.[40]

In England and Wales, universal school-aged (at 13-14 years old) vaccination (hereafter referred to as the BCG schools scheme) was introduced after a MRC trial in the 1950s estimated BCG's effectiveness at 78% in the ethnic White UK born population.[26] The policy remained in place until 2005, when England and Wales changed to targeted vaccination of high-risk neonates. The 2005 change in BCG vaccination policy was motivated by evidence of decreased transmission of TB, an increasing proportion of TB cases occurring in the non-UK born,[2] and modelling evidence that suggested stopping the BCG schools scheme would have minimal long term effects on incidence rates.[41] Due to the complex nature of both TB and the BCG vaccine, the ongoing impact of this change in policy is hard to directly estimate, with decision makers relying on expert opinion, evidence from surveillance data, and insight from modelling studies.

In 1987, an assessment of the school-age vaccination program was carried out in England and Wales.[41] This study was used, combined with a sensitivity analysis of notification rates, as supporting evidence by the Joint Committee on Vaccination and Immunisation (JCVI) BCG subgroup for the change in vaccination policy.[62,63] This chapter aims to re-evaluate this modelling, and re-estimate the predicted impact of stopping the schools scheme. Re-evaluating this work allows for the strength of the evidence used in decision making to be assessed and may highlight any issues with the approach used.

## 5.3 Methods

### 5.3.1 Modelling the impact of ending the BCG schools scheme

I implemented, and updated, Sutherland et al.'s model for estimating the impact of ending the BCG schools scheme, which is outlined briefly below.[41] This model was based on data from TB notification surveys conducted in 1973, 1978, and 1983.[64] These were used to estimate incidence rates, stratified by BCG vaccination status, in the ethnic White UK born population of England and Wales aged 15-19 years old, 20-24 years old and 25-29 years old. Future incidence rates were forecast by assuming an annual decrease in incidence rates, which was based on historic trends.[41,65] Primary impacts from ending the schools scheme, including the number of vaccines required to prevent a single notification, were estimated by calculating the difference in incidence rates between the vaccinated and unvaccinated populations. Additional notifications from TB transmission were then calculated using a transmission chain model and combined with the primary impact estimates, to calculate the number of annual additional notifications due to ending BCG vaccination. Based on data availability, the model used a 5-year time step.

#### Original estimation of notification rates

The effectiveness of the BCG vaccine was originally estimated by an MRC trial in 1953 at 78% in the United Kingdom.[26] As a follow up to this trial members of the MRC bio-statistics unit conducted a series of notification surveys attempting to ascertain any change in effectiveness, as well as acting as an estimate of notification rates across different demographics.[64] Surveys of those aged 15-24 years were carried out at 5-year intervals in 1973, 1978 and 1983 in England and Wales. For the 1983 survey records of BCG status,

### **5.3. Methods**

---

Tuberculin status and ethnicity were extracted from the records of notifying physicians and the records of the local health and education authorities. Total notifications across the study period were then aggregated for the following groups: Tuberculin negative and BCG vaccinated, Tuberculin negative and BCG unvaccinated, Tuberculin positive and not vaccinated and those who did not participate. These totals were then combined with the population estimates for each cohort at 13 years of age to estimate the ethnic make up of the population, and to construct notification rates for each category. Data were drawn from a range of sources including: Office of National Statistics data; annual local authority returns for total Tuberculin test results; BCG vaccinations in the schools scheme; and the Labour force survey (1983).

For 1983, there were 874 notifications recorded in 15-24 year old ethnic White UK born persons in England and Wales; survey participation was 80%. As the number of Tuberculin negative subjects not given BCG was unreported the number of notifications was estimated at 1.9% of those vaccinated with the BCG.[64] See [64] for full details of the survey and the additional assumptions used to give similar estimates for both the 1973 and 1978 surveys. The findings of these surveys were as follows: in the ethnic White population notification rates had fallen by an annual average of 9% and BCG efficacy had remained high.[41,65]

Evidence suggests that the BCG vaccine has a high efficacy for at least the first 15 years after vaccination, therefore Sutherland et al. extrapolated from the data on the 20-24 cohort to a theoretical 25-29 year old cohort. Data on the notifications in 25-29 year olds was available for the first 6 months of the 1983 survey and this was then scaled up to a yearly estimate using the ratio of notifications from this age group against the total number of notifications recorded in that year. Population estimates for the 25-29 year old cohort were based on data from the 20-24 cohort adjusted for all causes of mortality (0.34%). Migration was ignored. The Tuberculin positive cohort had a sharp decline in the previous two age cohorts, therefore it was assumed that this continued. Lastly, the efficacy was estimated as being that seen in the 20-24 cohort but with the same decline in protection seen between the last two cohorts. These assumptions allowed notification rates to be estimated for the 25-29 year old population, resulting in a complete cohort over the projected 15 years of BCG effectiveness.

#### **Original construction of forward estimates**

Based on these estimated notification rates, Sutherland et al. then sought to quantify the ongoing risk of developing notified TB, projected forward in time, for both the vaccinated and unvaccinated populations. To construct these estimates several key assumptions, based on the results seen in the previous surveys, were made. Firstly, it was assumed that efficacy was not degrading within the ethnic White population and therefore historic estimates would continue to apply into the future. Additionally, it was assumed that the annual decay of 9% in notification rates, across all ethnic White populations, would continue indefinitely.

These assumptions allowed the notification rates in both the BCG vaccinated and unvaccinated groups to be projected forward in time. By assuming that the schools scheme is responsible for the observed variation between vaccinated and unvaccinated rates the rate of prevented cases can then be estimated. By scaling this against a cohort of 100,000 13 year old individuals, the number of prevented cases over a 15-year period can be projected

for each cohort. By dividing the total number in a given cohort by the number of prevented cases the estimated number of vaccines required to prevent a single case in the 15-year period can then be calculated.

To estimate the total number of prevented notifications, for each cohort, in England and Wales the total number receiving the BCG and the coverage of the schools scheme was required. The coverage of the BCG schools scheme was estimated from annual reports of the Department of Health and Social Security (DHSS) and was assumed to be 75% for all years. The number of BCG vaccines given each year was estimated from the DHSS returns for the years 1967 to 1981, it was then taken as 75% of the estimated ethnic White population aged 13 years from 1982-1996, for each 5-year period thereafter it was assumed to be 2.1 million.

Using the data on BCG coverage, the number of vaccines given each year, and the projected differences between vaccinated and unvaccinated notification rates allowed the number of prevented notifications, due to vaccination, for each age group to be found for each year. These estimates can then be used to give the total number of prevented notifications for those aged between 15-29 years. To understand these estimates, estimates of the projected yearly notifications if the scheme continues were required. These totals were derived from the vaccinated and unvaccinated rates supplemented with similar projections from the Tuberculin positive or otherwise ineligible sourced from the 1983 BCG survey.[64]

### **Original transmission chain model**

Sutherland et al. defined their TB transmission model as follows:

1. The total expected number of secondary notifications ( $T$ ) arising from any single primary notification was estimated as,

$$T = (1 - d)^z < 1 \quad (5.1)$$

Where  $d$  is the percentage annual decay in notification rates, and  $z$  is the average interval between the notification of any individual and the notification of the patient who infected them.

2. The total expected number of secondary notifications arising from any single primary notification ( $T$ ) is related to the number of notifications in each generation using the relative generation size ( $x$ ) with the following power series,

$$T = x + x^2 + x^3 + \dots = \frac{x}{1 - x} \quad (5.2)$$

3. The expected average interval between each primary notification and all secondary notifications ( $Z$ ) is defined to be the sum of time to all notifications, weighted by the fraction in each generation, divided by the sum of all notifications.

$$Z = \frac{xz + x^22z + x^33z + \dots}{x + x^2 + x^3 + \dots} = \frac{z}{1 - x} \quad (5.3)$$

Both 2. and 3. are only valid when  $x < 1$ .

### **Updating the transmission chain model**

If we assume a constant decay rate of  $d\%$  per year and that the next generation of secondary cases are notified  $z$  years after the person who infected them, then 1 notification in year 0

### 5.3. Methods

---

will result in  $(1 - d)^z$  secondary notifications  $z$  years later. This is therefore the relative generation size ( $x$ ), not the *total* number of secondary notifications ( $T$ ), as (incorrectly) stated by Sutherland et al. Therefore, to correct this I revised 1. to the following,

$$x = (1 - d)^z < 1 \quad (5.4)$$

The Sutherland et al. model required several additional assumptions.[41] Firstly, as incidence rates for those ineligible for the BCG schools scheme are not published, I assumed that they were equal to those in the unvaccinated population. In addition, in order to reproduce the distribution of cases annually (rather than by generation or overall) reported in Sutherland et al. I introduced an additional model step and parameter; the proportion of secondary cases in the first generation that occurred in the first year ( $f$ ). This modelling step is only required to reproduce the final table from [41] and does not impact estimates of the impact of ending the BCG schools scheme. It is included only for validation purposes.

The annual distribution of secondary notifications ( $N$ ) was modelled by first estimating the number of secondary notifications that occurred in the current year ( $i$ ) due to primary notifications in that year ( $N^{\text{Current}}$ ) and then estimating how many secondary notifications occurred 5 years later ( $N^{\text{Projected}}$ ).  $N^{\text{Current}}$  was estimated using the number of primary notifications ( $P$ ) multiplied by the number of total expected number of secondary notifications per primary notification ( $T$ ), the proportion of secondary cases in the first generation that occurred in the first year ( $f$ ), and the relative size of the first generation ( $x$ ).  $N^{\text{Projected}}$  was then estimated by assuming that it was equal to the number of secondary notifications, minus notifications occurring in the first year, that occurred  $Z$  (the expected average interval between each primary notification and all secondary notifications) years ago. As this used the overall number of notifications from the previous time step a decay of  $(1 - d)^{5-Z}$  was applied. This approach can be summarised as follows,

$$N^{\text{Current}} = PTfx \quad (5.5)$$

$$N^{\text{Projected}} = (PT - N^{\text{Current}})(1 - d)^{5-Z} \quad (5.6)$$

$$N_i = N_i^{\text{Current}} + N_{i-1}^{\text{Projected}} \quad (5.7)$$

I fitted the proportion of secondary cases in the first generation that occurred in the first year ( $f$ ) using least squares to the original estimates of the total notifications due to ending the scheme under several scenarios, for several years. I validated the fitted model by comparing the results with those from the original implementation using the mean absolute percentage error, normalised by the original estimate, as the performance metric.

#### 5.3.2 Updating model parameter estimates and incorporating parameter uncertainty

Incidence rates were included as point estimates in [41]; in the updated model I included uncertainty in these rates. I did this by first estimating notifications for 1973, 1978, and 1983, using published incidence rates and population estimates. Samples were then generated using a Poisson distribution.[41,64] These samples were then used to estimate a distribution of incidence rates to replace the point estimates used in the original analysis. Sutherland et al. assumed a serial interval ( $z$ ) of 2 years between linked notifications. Using a newly

available literature source I updated this assumption with an estimate of 1.44 (95% CI 1.29 to 1.63) years.[66]

I considered the original assumption of a 9% annual decrease in incidence rates as well as three scenarios based on those considered by the JCVI BCG subgroup:[62,63] these were a 3.9% decrease, a 1.9% decrease, and no change annual in incidence rates. Data on the annual decrease in incidence rates in the ethnic White UK population were not available so I used two proxy measures. The first proxy measure was the annual change in notifications in England and Wales, which was estimated using data from Public Health England (PHE). The standard deviation of this measure was then calculated using the `prop.test` function in R.[56] The second proxy used was the annual decrease in the UK born age-specific incidence rates in the English population. These were calculated using notification data from the Enhanced TB surveillance system (ETS) and the June Labour Force Survey.[2] Incidence rates were estimated using the epiR package.[60] Uncertainty was incorporated by sampling from a normal distribution for both proxy measures. Data collection for the ETS began in 2000 and prior to this notification data was only available in years with notifications surveys (1973, 1978, and 1983). I therefore estimated incidence rates between 1984 and 1999, and for the years between notifications surveys (1974-1977 and 1979-1982), using locally estimated scatter plot smoothing (LOESS) regression fitted to incidence rates published in [11] and the estimated incidence rates from 2000 on-wards. LOESS is a local regression method that combines multiple regression models in a k-nearest neighbours meta-model.[67] This approach allows nonlinear trends to be fitted using a series of linear models. For years prior to 1973 the annual decreases were assumed to be the mean of the annual decreases from the previous 3 years. For both proxy measures the annual decreases in incidence rates post 2016 were assumed to be the average of the estimates in 2013-2015.

### 5.3.3 Statistical analysis

For each scenario, I ran the model for 69 years (1969-2028) with 10,000 parameter samples. I tested the difference between scenarios using the Mann-Whitney test for the number of vaccines needed to prevent a single case in 15 years after vaccination for a cohort aged 13-14 years old at vaccination. As in [41] a 15-year time horizon was used with 5-year intervals. The year closest to the year of the change in vaccination policy (2005), which had model estimates, was used as the baseline. The code for this analysis is available online<sup>2</sup>.

## 5.4 Results

The model produced results that were comparable with those from [41] (Table 5.1; Table 5.2). When estimating the total notifications from ending the BCG schools scheme at different times in ethnic White UK born adults aged 15-29 years old in England and Wales the model had a median absolute error of 2.03% (2.5, 97.5% Q: 0.00%, 3.72%) and a maximum absolute error of 3.91% when compared to [41]. I found that the percentage of cases in the first year was  $f = 0.764$  when fitted to the Sutherland et al. estimates using the least squares method.

---

<sup>2</sup>Code: <https://github.com/seabbs/AssessBCGPolicyChange>

Table 5.1: Comparison of results published by Sutherland et al. vs. the recreated model. This table shows the total notifications including primary and secondary effects from ending the BCG schools scheme at various times in ethnic White adults aged 15-29 years old in England and Wales.

Year of Ending Scheme	1988			1993			1998		
	Original	Recreated	Difference	Original	Recreated	Difference	Original	Recreated	Difference
1986	288	296	8	226	226	0	208	205	-3
1991	288	296	8	165	166	1	130	131	1
1996	288	296	8	165	166	1	90	91	1

Table 5.2: Continued: Comparison of results published by Sutherland et al. vs. the recreated model. This table shows the total notifications including primary and secondary effects from ending the BCG schools scheme at various times in ethnic White adults aged 15-29 years old in England and Wales.

Year of Ending Scheme	2003			2008			2013		
	Original	Recreated	Difference	Original	Recreated	Difference	Original	Recreated	Difference
1986	181	175	-6	128	123	-5	80	78	-2
1991	128	126	-2	115	111	-4	80	78	-2
1996	77	76	-1	80	80	0	72	70	-2

## 5.4. Results

---

### 5.4.1 Annual change in TB incidence rates

I found that the assumption of a 9% annual decrease in incidence rates in the ethnic White UK born was not comparable to estimates using either notification data or age-specific incidence rates in the time period studied (Figure 5.1). The median annual decrease estimated using notifications was 3.06% (2.5, 97.5% Quantiles (Q): -8.32%, 11.45%), with a maximum of 15.14% (2.5, 97.5% Q: 14.22%, 16.03%) in 1987 and a minimum of -10.17% (2.5, 97.5% Q: -10.82%, -9.52%) in 2005. Using age-specific incidence rates I estimated the median annual decrease in incidence rates for 15-19 year olds was 1.65% (2.5, 97.5% Q: -40.49%, 39.97%), 3.16% (2.5, 97.5% Q: -33.95%, 38.30%) for 20-24 year olds, and 2.63% (2.5, 97.5% Q: -36.28%, 37.17%) for 25-29 year olds. There was substantial variation between years and a high degree of uncertainty.

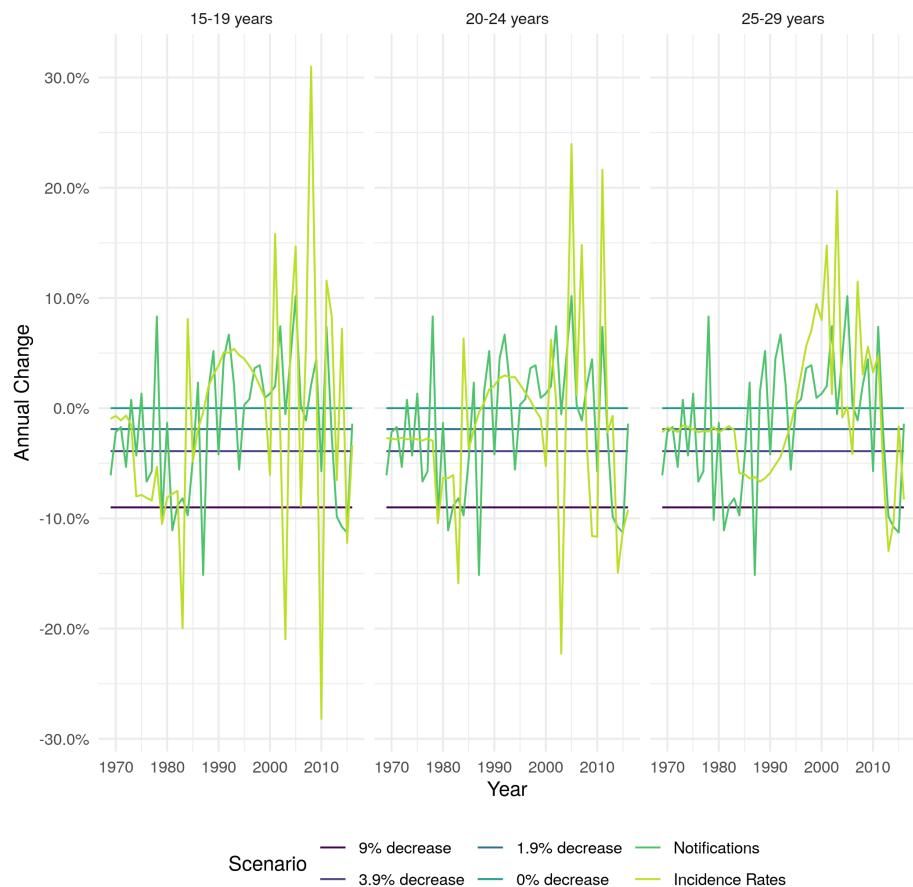


Figure 5.1: Annual percentage change in ethnic White UK born incidence rates for those aged 15-19, 20-24, and 25-29 years old under different scenarios. For the notification and incidence rate scenarios each line represents the median of 10,000 parameter samples.

## **5.5 Vaccines required to prevent a single notification**

I found that incorporating uncertainty, did not alter the number of vaccines required to prevent a single notification within 15 years in a cohort vaccinated at school-age, when the annual decrease in incidence rates was assumed to be 9% (Figure 5.2; Table 5.3). However, the updated estimate had a wide range (15000 (2.5, 97.5% Q: 12000, 19000) vaccines required in 2004). As the assumed annual decrease in incidence rates was reduced the number of vaccines required to prevent a single notification also reduced. Assuming an annual decrease of 1.9% (one of the scenarios evaluated by the JCVI) resulted in an estimate of 1600 (2.5, 97.5% Q: 1300, 2000) vaccines required to prevent a single notification in 2004. This assumption was the most comparable, although not equivalent, to estimates derived using notifications (1400 (2.5, 97.5% Q: 1100, 1700), P: 0.077) and age-specific incidence rates (1500 (2.5, 97.5% Q: 460, 4900), P: 0.083). The estimate using incidence rates had a high degree of uncertainty (Figure 5.2; Table 5.3). The number of vaccines required increased slightly over time with 1800 (2.5, 97.5% Q: 1500, 2200) required in 2009, 2000 (2.5, 97.5% Q: 1600, 2500) required in 2014, and 2200 (2.5, 97.5% Q: 1800, 2700) required in 2019 when an annual decrease of 1.9% in incidence rates was assumed.

## 5.5. Vaccines required to prevent a single notification

---

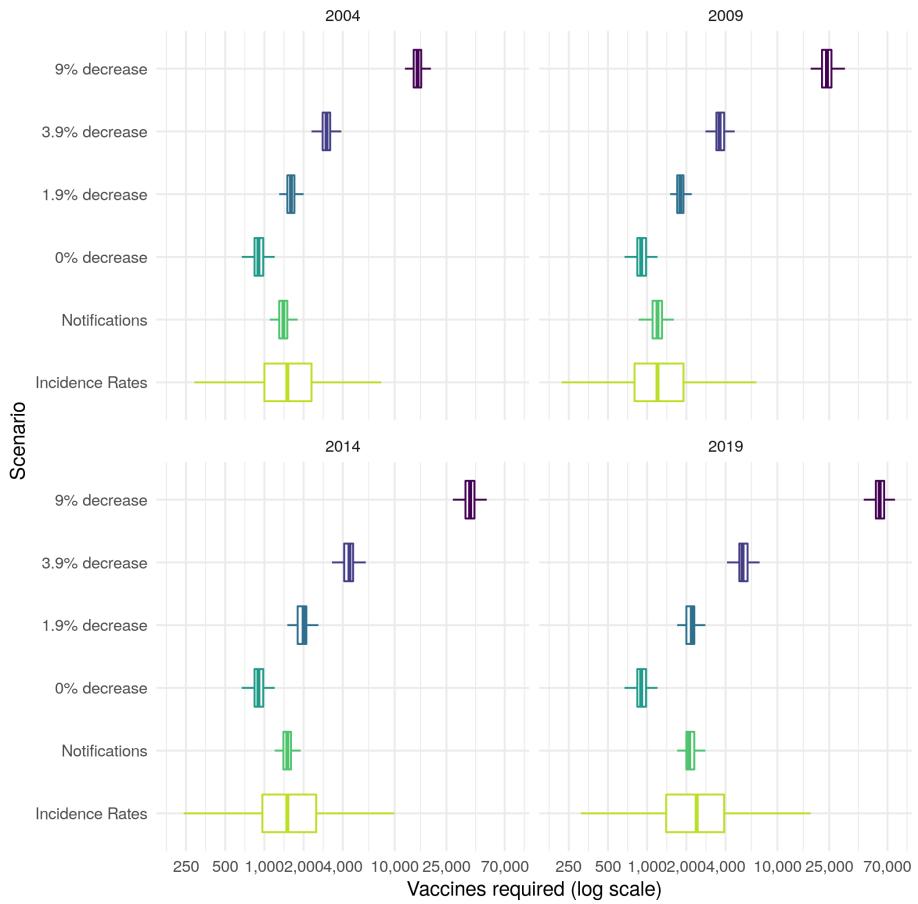


Figure 5.2: Vaccines required in a cohort of those vaccinated at school-age to prevent a single case of Tuberculosis within 15 years of vaccination in 2004, 2009, 2014, or 2019. The years presented were dictated by the 5-year timestep of the model. The percentage annual decrease scenarios considered were based on those considered by the JCVI BCG subgroup, with the addition of a scenario using aggregate notification data and a scenario using estimates of age-specific incidence rates in the UK born. Each boxplot summarises the output of 10,000 model simulations for each scenario.

Table 5.3: The median number (with the 2.5% and 97.5% quantiles) of vaccines required to prevent a single case of TB within 15 years in a ethnic White UK born adult vaccinated at 13 years old. The percentage annual decrease scenarios considered were based on those considered by the JCVI BCG subgroup, with the addition of a scenario using aggregate notification data and a scenario using estimates of age-specific incidence rates in the UK born.

Year of Vaccination	9% decrease (original parameters)	9% decrease	3.9% decrease	1.9% decrease	0% decrease	Notifications	Incidence Rates
1969	460	460 (390, 540)	460 (390, 540)	460 (390, 540)	460 (390, 540)	460 (390, 540)	460 (390, 540)
1974	940	940 (780, 1200)	880 (720, 1100)	850 (700, 1100)	830 (680, 1000)	860 (710, 1100)	860 (640, 1100)
1979	1400	1400 (1100, 1800)	1100 (910, 1400)	1000 (820, 1300)	900 (740, 1100)	1000 (860, 1300)	1100 (680, 1700)
1984	2200	2200 (1800, 2900)	1300 (1100, 1700)	1100 (900, 1400)	900 (740, 1100)	1600 (1300, 2000)	1400 (730, 2900)
1989	3600	3600 (2900, 4600)	1600 (1300, 2100)	1200 (990, 1500)	900 (740, 1100)	1800 (1500, 2200)	1700 (760, 3800)
1994	5800	5800 (4700, 7300)	2000 (1600, 2500)	1300 (1100, 1700)	900 (740, 1100)	1700 (1400, 2200)	1600 (640, 4200)
1999	9300	9200 (7600, 12000)	2500 (2000, 3100)	1500 (1200, 1900)	900 (740, 1100)	1600 (1300, 2000)	1500 (510, 4200)
2004	15000	15000 (12000, 19000)	3000 (2400, 3800)	1600 (1300, 2000)	900 (740, 1100)	1400 (1100, 1700)	1500 (460, 4900)
2009	24000	24000 (19000, 30000)	3600 (3000, 4600)	1800 (1500, 2200)	900 (740, 1100)	1200 (960, 1500)	1200 (350, 4300)
2014	38000	38000 (31000, 48000)	4500 (3600, 5600)	2000 (1600, 2500)	900 (740, 1100)	1500 (1200, 1900)	1500 (390, 6000)
2019	61000	61000 (50000, 78000)	5400 (4400, 6900)	2200 (1800, 2700)	900 (740, 1100)	2100 (1800, 2700)	2300 (470, 11000)
2024	98000	98000 (80000, 120000)	6600 (5400, 8400)	2400 (1900, 3000)	900 (740, 1100)	3200 (2600, 4100)	3300 (550, 18000)

### 5.5.1 Average annual additional cases from ending the BCG schools scheme at various dates

I found that updating parameter values, and incorporating uncertainty, did not alter the average annual primary additional notifications from stopping the BCG schools scheme when the annual decrease was assumed to be 9%. However, when these changes were combined with the updated transmission model we found that the impact of ending BCG vaccination was greater than previously reported with an increase in the number of estimated cases due to onwards transmission (Figure 5.3; Table 5.4). These estimates were uncertain with 94 (2.5, 97.5% Q: 72, 119) additional annual notifications if vaccination was stopped in 2001. As the assumed annual decrease in incidence rates was reduced the annual number of additional notifications increased with 6099 (2.5, 97.5% Q: 4691, 7719) notifications when the annual decrease was assumed to be 1.9% and vaccination stopped in 2001. The number of annual notifications reduced with time: 5314 (2.5, 97.5% Q: 4082, 6725) from ending vaccination in 2006; 4327 (2.5, 97.5% Q: 3315, 5496) from ending vaccination in 2011, and 2852 (2.5, 97.5% Q: 2074, 3741) from ending vaccination in 2016 (Figure 5.3; Table 5.4).

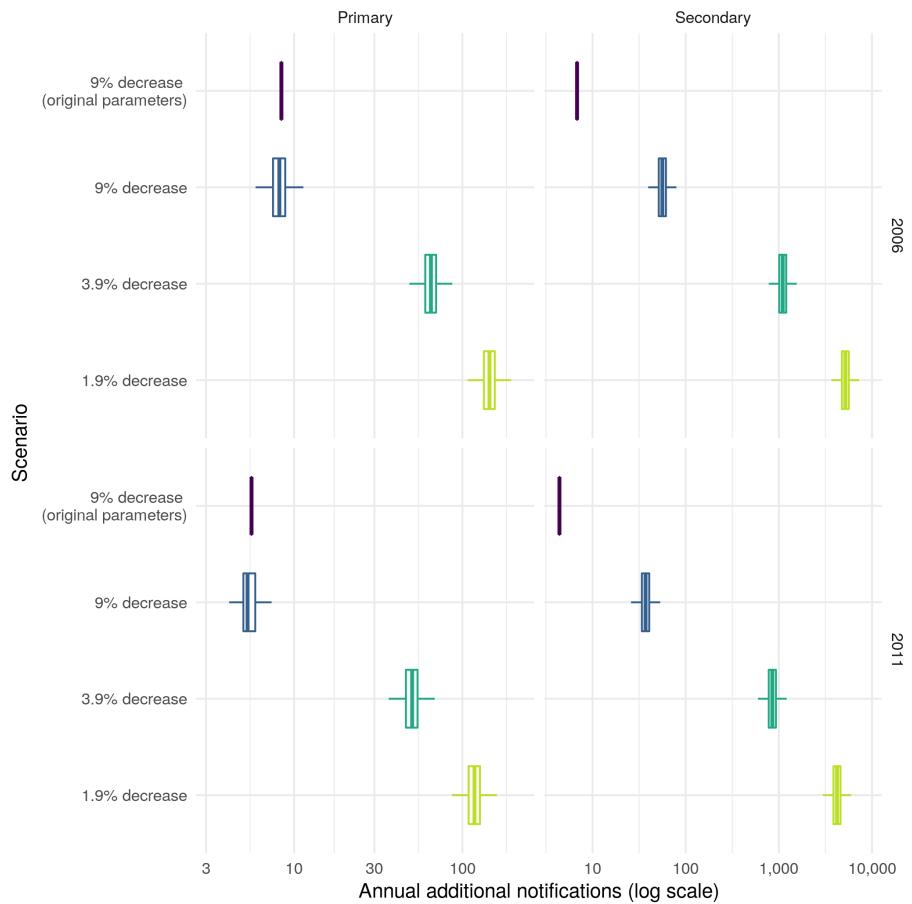


Figure 5.3: Annual additional (stratified into primary and secondary) notifications in 15-29 year olds from stopping the BCG schools scheme in 2006, and 2011 until 2028. The years presented were dictated by the 5-year timestep of the model. The percentage annual decrease scenarios considered were based on those considered by the JCVI BCG subgroup. Data based scenarios and the JCVI 0% decrease scenario were not presented here as the updated transmission model could not support these scenarios. Each boxplot summarises the output of 10,000 model simulations for each scenario. Secondary notifications are reported assuming they occurred in the same year as the primary notifications that caused them.

Table 5.4: The median number (with the 2.5% and 97.5% quantiles) of additional annual notifications due to ending the BCG schools scheme in selected years. The percentage annual decrease scenarios considered were based on those considered by the JCVI BCG subgroup. Data based scenarios and the JCVI 0% decrease scenario were not presented here as the updated transmission model could not support these scenarios.

Year Ending Scheme	9% decrease (original parameters)	9% decrease	3.9% decrease	1.9% decrease
1971	111	111 (88, 135)	204 (161, 249)	286 (226, 350)
1976	90	90 (71, 110)	187 (148, 229)	276 (219, 337)
1981	55	55 (44, 68)	152 (120, 185)	244 (193, 298)
1986	35	35 (28, 43)	125 (99, 152)	218 (172, 266)
1991	25	25 (19, 30)	109 (86, 133)	203 (161, 247)
1996	17	17 (13, 21)	93 (74, 114)	185 (147, 225)
2001	12	12 (9, 15)	79 (62, 96)	166 (132, 203)
2006	8	8 (6, 10)	65 (51, 79)	145 (115, 177)
2011	6	5 (4, 7)	50 (39, 62)	118 (92, 145)
2016	3	3 (2, 4)	32 (23, 40)	78 (58, 99)
2021	1	1 (1, 2)	15 (9, 21)	39 (24, 54)

## 5.6 Discussion

The existing method for estimating the impact of the BCG schools scheme produced uncertain estimates of the impact of ending the scheme in all years evaluated when parameter uncertainty was included. The approach used to estimate additional notifications due to transmission was found to be incorrect. Once corrected, the transmission model produced much higher estimates of additional notifications caused by ending BCG vaccination than previously reported. Further updating the model with the annual decrease in TB notifications based on both notifications and using age-specific incidence rates resulted in a decrease in the number of vaccines needed to prevent a single case in all years considered. A scenario with a 1.9% annual decrease in incidence rates was most comparable to the results based on notifications. Using this scenario, I found that the number of TB notifications arising from ending school age BCG vaccination was much greater than originally estimated using the scenario considered in Sutherland et al.

In addition to identifying that the public health impact of the change in vaccine policy was likely much larger than originally estimated, my results also provide new insights into the uncertainty of the previously published model predictions by including parameter uncertainty and measurement error and updates these predictions using newly available data. As historical data on incidence rates in the ethnic White UK born in England and Wales were not available, I considered two approaches to proxy them and investigated multiple scenarios based on those explored by the JCVI BCG subgroup. The simulation approach used here, although updated where possible, is not the most accurate method for estimating the impact of ending the BCG schools scheme as it relies on numerous assumptions based on the available knowledge in 1987 and does not account for the role of non-White and non-UK born cases. However, the strength of this work is that the estimates are based on the framework used to inform policy making. This allowed the strength of the model used in the decision-making process to be assessed once parameter uncertainty had been incorporated and for flaws in the model to be identified. This would not have been possible if the impact had been assessed using only the observed data or with an alternative model. It also allowed estimates based on updated data to be compared to historic estimates within the same framework. This would also not have been possible if a different framework had been used. As mentioned, a weakness of the model used in this study is that it did not include the whole population or age groups outside those directly affected by vaccination. Heterogeneous mixing between these groups is also likely to be an important consideration. The exclusion of these factors means that my results are likely to underestimate the impact of ending the BCG schools scheme. A final limitation is that this study only considers the impact of ending the BCG schools scheme and not the impact of the introduction of the targeted neonatal vaccination program. This should be considered when evaluating the change in policy as a whole.

Little work has been done to assess the impact of the 2005 change in BCG vaccination policy or to assess the quantitative evidence used in decision making. However, multiple studies have evaluated the cost effectiveness of various BCG programs and the impact of switching between them. A cluster-randomised trial in Brazil found that BCG vaccination of those at school-age was cheaper than treatment and would prevent one TB case per 381 vaccinations even with a vaccine effectiveness of only 34% (8-53%).[68] This is substantially fewer than my estimate of 2000 (2.5, 97.5% Q: 1600, 2500) (the most comparable year from my results). However, the same trial found that for regions close to the equator BCG effectiveness was

## 5.6. Discussion

---

low in school-age children but unchanged in neonates,[69] highlighting the importance of considering the BCG vaccines reduced effectiveness near the equator when determining vaccination policy.[70] There is also some research which supports universal re-vaccination of those at school-age, in countries with high incidence and universal vaccination of neonates, as it may be cost effective when BCG effectiveness is moderate to high.[69,71] There is some evidence that targeted vaccination of high risk neonates maybe more cost effective than universal vaccination of neonates.[72,73] However, a study in Sweden found that incidence rates in Swedish-born children increased slightly after universal vaccination of neonates was discontinued in favour of targeted vaccination.[74] In France, which switched from universal vaccination of neonates to targeted vaccination in 2007, it has also been shown that targeted vaccination reduced coverage in those most at risk.[75] Targeted vaccination may not be more cost effective than universal vaccination when possible reductions in transmission are considered. This chapter indicated that a substantial number of cases due to transmission may be preventable if universal school-age BCG vaccination was still in place. This result is dependent on the effectiveness of BCG vaccination when given later in life, for which there is good evidence in the ethnic White UK born.[26] I did not consider neonatal vaccination which would be less impacted by BCG's effectiveness reducing when given later in life, but may also be less likely to result in the same reductions in ongoing transmission.

This chapter indicates that some of the evidence used to justify the 2005 change in BCG vaccination policy may have depended on a methodologically flawed model, resulting in the impact of ending BCG vaccination being underestimated. Modelling evidence can often be complex and difficult to reproduce, it is important that policy makers, or those who work with them, have the skills to assess its quality. This study also highlights the importance of including both parameter and measurement error, as excluding these sources of variation may lead to spuriously precise results. Spurious precision is problematic for policy makers as the worst-case scenario often needs to be considered when making policy decisions. In addition, my exploration of the assumptions used to estimate the annual change in TB incidence rates in the ethnic White UK born illustrates the structural impact of assuming an annual decrease in TB incidence rates. More realistic estimates of the annual decrease in incidence rates resulted in a greatly increased impact of ending the BCG schools scheme. Policy makers should consider these updated estimates when assessing the role of BCG vaccination in those at school-age. However, decisions regarding vaccine policy in the UK require economic evaluation, which discounts costs and benefits in the future; discounting has not been applied in this study which estimates the epidemiological impact of vaccination only.

This chapter has reassessed some of the evidence previously used in decision making, correcting the transmission model used, and updating the approach with new data. However, as 15 years of detailed surveillance data have been collected since the ending of the BCG schools scheme it is now possible to use regression-based approaches to estimate the direct impact on incidence rates of ending the BCG schools scheme (see Chapter 7).[76] These approaches could also be used to estimate the impact of vaccinating high-risk neonates, which may outweigh any negative impacts of ending the BCG schools scheme. In addition, the development (see Chapter 8), and use, of a transmission dynamic model would allow the more accurate estimation of indirect effects and the forecasting of long-term impacts (see Chapter 10).

## 5.7 Summary

- This chapter reevaluates a key piece of the quantitative evidence used to motivate the change in BCG vaccination policy in 2005, correcting a methodological flaw in the original model.
- The inclusion of parameter uncertainty, and measurement error, allowed the uncertainty in the final estimates to be presented. Previously published estimates may have been spuriously precise.
- As this study used a historical approach the model used is not the most accurate method for assessing the impact of ending the BCG schools scheme. However, it provides an estimate that is based on the available data and on the framework used to inform policy making. This allowed the strength of some the quantitative evidence used in the decision-making process to be assessed.
- The impact of ending the BCG schools scheme was found to be greater than previously thought when notification data were used. These results highlight the importance of independent evaluations of modelling evidence, including uncertainty, and evaluating multiple scenarios 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.2583056<sup>3</sup>

---

<sup>3</sup>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 (ETS) 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 adapted from [77]<sup>1</sup> (also available as a preprint<sup>2</sup>) 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 paper.

---

<sup>1</sup>Paper: <https://doi.org/10.1016/j.vaccine.2019.06.056>

<sup>2</sup>Preprint: <https://doi.org/10.1101/366476>

## 6.2 Background

Bacillus Calmette–Guérin (BCG) is one of the mostly widely-used vaccines and the only vaccine that protects against TB disease. BCG was first used in humans in 1921 and was introduced into the WHO Expanded Program on Immunization in 1974.[38] 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.[5] 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).[25] 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.[69] Unlike in adults, BCG has consistently been shown to be highly protective against TB and TB meningitis in children.[23,24] For this reason the majority of countries that use BCG, vaccinate at birth.[27] 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.[30,31] There is some evidence that BCG vaccination induces innate immune responses which may provide non-specific protection,[32] TB patients with BCG scars were found to respond better to treatment with earlier sputum smear conversion,[36] and there is evidence to suggest that BCG vaccination is associated with reduced all-cause neonatal mortality[33,34] and both reduced TB[28] and all-cause[35] 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 TB Surveillance (ETS) system

I extracted all notifications from the 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

### **6.3. Method**

---

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;[28] 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 bimodal 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.[56] 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  $\chi^2$  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,[78–80] age,[81] Index of Multiple Deprivation (2010) categorised into five groups for England (IMD rank),[15] ethnicity,[78,82] UK birth status,[45,83] 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.[51] I imputed 50 data sets (for 20 iterations) 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,[84] and effect sizes compared with those from the main analysis. All code for this analysis is available online<sup>3</sup>.

## 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 born cases 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.

---

<sup>3</sup>Code: <https://github.com/seabbs/ExploreBCGOnOutcomes>

## 6.4. Results

---

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

Outcome	Total	BCG status		
		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 6.2: Potential confounders for individuals in England notified with TB between 2009-2015, stratified by BCG vaccination status.

Confounder	Total	BCG status		
		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

#### 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,

## 6.4. Results

---

compared to those vaccinated more recently (aOR: 12.19 (95% CI 3.48 to 42.64), (see 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).

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

Variable	Total	All-cause mortality	Univariable		Multivariable	
			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

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

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

Outcome	BCG vaccinated	Univariable				Multivariable			
		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  $\leq 10$  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

Outcome	Years since BCG	Univariable				Multivariable			
		Cases**	Cases with outcome (%)	OR (95% CI)	P-value	Cases***	Cases with outcome (%)	aOR (95% CI)	P-value
All-cause mortality	$\leq 10$	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*)	$\leq 10$	2	2 (100)	1	-	2	2 (100)	1	-
	11+	108	59 (55)	<i>Insufficient data</i>	-	98	53 (54)	<i>Insufficient data</i>	-
Recurrent TB	$\leq 10$	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	$\leq 10$	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	$\leq 10$	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.

Outcome	Age at BCG	Univariable				Multivariable			
		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)	
Pulmonary TB	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)	
	< 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)	
Sputum smear status - positive	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)	
	< 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 (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. Results

---

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

Outcome	Univariable			Multivariable		
	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

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)

Outcome	Univariable			Multivariable		
	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).

Outcome	Age group	Univariable			Multivariable		
		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

#### 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

---

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

Study population	Outcome	BCG	Univariable		Multivariable	
			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

## 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 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 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 compared to notification within 10 years. 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.[85] 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.[34,86,87] 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.[87] 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 (adjusted Hazard ratio (aHR): 0.58 (95% CI 0.39 to 0.85).[35] 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.[34] 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 in-

## 6.5. Discussion

---

tervals. Several meta-analyses have found evidence supporting this association,[24,28] with one meta-analysis estimating a 71% (Risk ratio (RR): 0.29 95% CI 0.16 to 0.53) reduction in deaths due to TB in individuals vaccinated with BCG.[24] 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).[28] 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,[36] 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.[32] For low incidence countries, where the reduction in TB cases has been used as evidence to scale back vaccination programs,[27] 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 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 beneficial 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 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 mortality 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<sup>4</sup>](https://doi.org/10.5281/zenodo.1213799)

---

<sup>4</sup>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 TB Surveillance (ETS) system, with demographic data from the Labour Force Survey (LFS) 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 adapted from [76]<sup>1</sup> supervised by Hannah Christensen and Ellen Brooks-Pollock.[76] Nicky Welton 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,[20,30,31] and high effectiveness of the BCG vaccine in children,[4,23,24] and variable effectiveness in adults.[27] 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,

---

<sup>1</sup>Paper: <https://doi.org/10.2807/1560-7917.ES.2019.24.49.1900220> Preprint: <https://doi.org/10.1101/567511>

either of those at school-age or neonates, to targeted vaccination of neonates considered at high-risk of TB (see Chapter 2).[5] In Sweden, which discontinued universal vaccination 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.[74] 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.[75]

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).[20] A recent study found that this reduction may be linked to improved TB interventions.[88] 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.[20] 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 onwards 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.[89] 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 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.[20]

I obtained yearly population estimates from the April to June 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 (see Section 4.2.2).

### **7.3.2 Constructing Retrospective cohorts**

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

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

## 7.4. Statistical methods overview

---

while aged between 14 and 19 years.

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

Cohort 2: individuals born between 2005 and 2010, who were notified with TB while aged 0 to 5 years.

Comparison cohort 2: individuals born between 2000 and 2004, who were notified with TB while aged 0 to 5 years.

Cohorts were stratified by vaccination programme using age criteria and then stratified further by whether the scheme was in place during the time period they entered the study. Each cohort was further 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 eight 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 series of increasingly complex models. I considered the following covariates: age,[20,27] incidence rates in both the UK born and non-UK born who were not in the age group of interest,[20] 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.[90] 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 improvement 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.[56] Reproducibility was ensured by using R package infrastructure<sup>2</sup>. Missing data imputation using a gradient boosting model (GBM) was implemented using the `h2o` package (see Section 7.4.2).[91] Incidence rates, with 95% confidence intervals, were calculated using the `epiR` package (see Chapter 4).[60] The `brms` package,[92] and Stan,[93] was used to perform Markov Chain Monte Carlo (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.[93] 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 more complex, high-performing models compared to those usually used for imputation, whilst also allowing the results to be validated using predictive modelling performance metrics. I included year of notification, sex, age, PHE 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

---

<sup>2</sup>Code: <https://github.com/seabbs/DirectEffBCGPolicyChange>

## 7.4. Statistical methods overview

---

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.

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  $1 \times 10^{-5}$ , 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.[91] Much like a random forest model they work by 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 (fits a logistic regression model to model predictions in order to return a probability) using the calibration dataset. The fitted imputation model had a logloss (the negative of the log likelihood) of 0.28 on the test set, with an area under the curve (AUC) of 0.93, both of which indicate robust performance on unseen data. 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.[88] 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).

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

Inclusion of imputed values for UK birth status should reduce bias caused by any missing 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 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.[92] 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 ( $C_0$ ) and then using the following equation,

$$C_P^i = C_0(1 - I^i), \text{ Where } i = e, l, u. \quad (7.1)$$

Where  $C_P^i$  is the predicted number of cases prevented using the median ( $e$ ), 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,

$$C_P^i = C_{NE}(1 - I^i), \text{ Where } i = e, l, u. \quad (7.2)$$

Where  $C_{NE}$  is the number of cases observed assuming that the cohort was not exposed to targeted high-risk neonatal vaccination. As from 2005 onwards this cohort were in fact exposed to this vaccination scheme an additional step was required. This first required calculating the number of cases that would be expected if the cohort had not been exposed to the scheme,

$$C_{NE} = \frac{C_0}{I^i} \quad (7.3)$$

Then combining this with the previous equation so that  $C_P^i$  can be estimated using observed data ( $C_0$ ),

$$C_P^i = \frac{C_0(1 - I^i)}{I^i}, \text{ Where } i = e, l, u. \quad (7.4)$$

## 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).

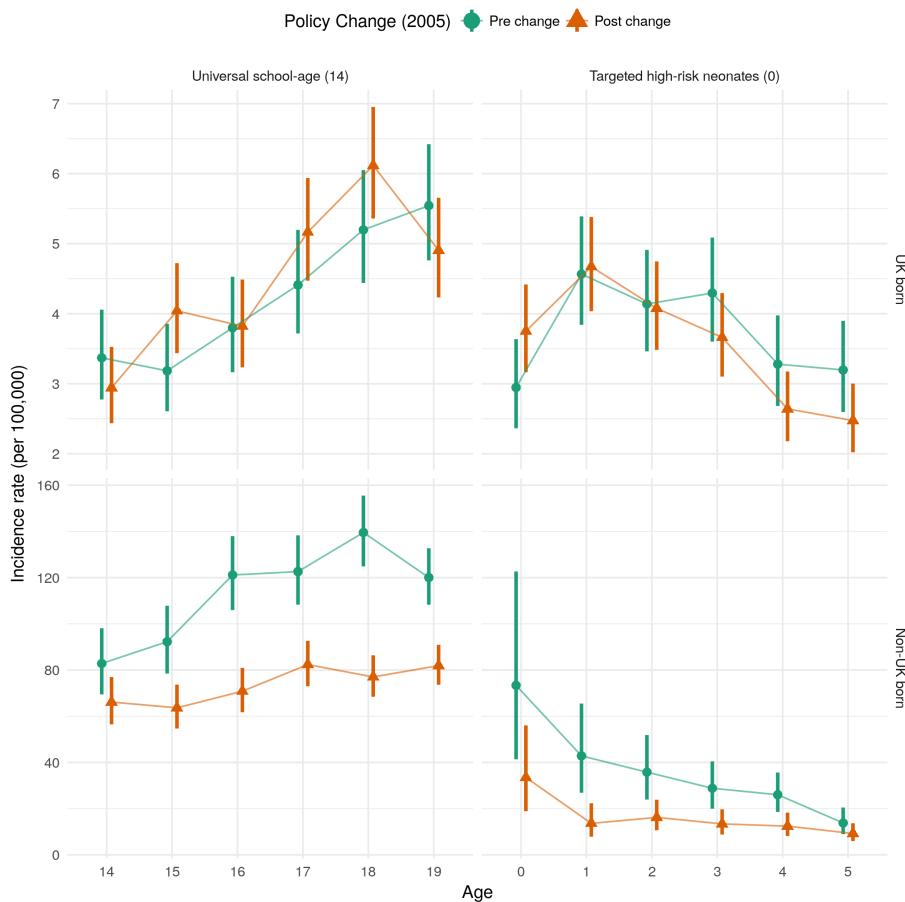


Figure 7.1: Mean incidence rates per 100,000, with 95% confidence intervals for each retrospective cohort, stratified by the vaccination policy and UK birth status. The top and bottom panels are on different scales in order to highlight trends in incidence rates over time.

Trends in incidence rates varied by age group and UK birth status. From 2000 until 2012

## 7.5. Results

---

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).

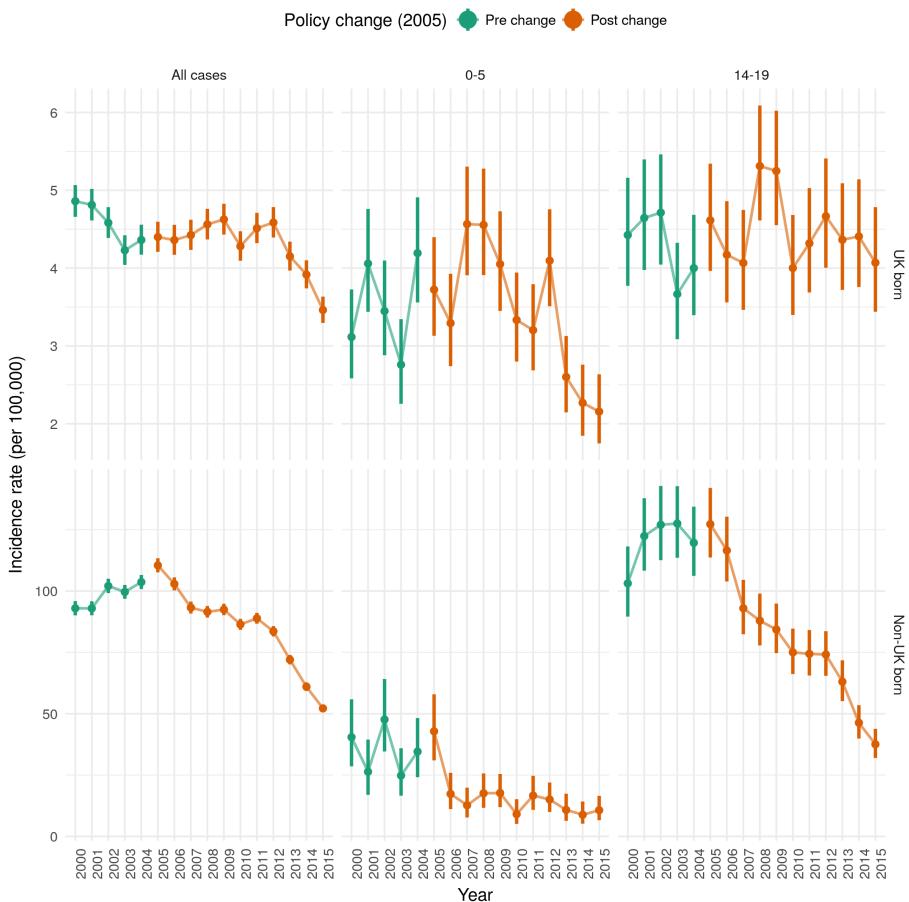


Figure 7.2: Incidence rates (with 95% confidence intervals) per 100,000 per year for UK born population and non-UK born population, aged 0-5 and therefore directly affected by the targeted neonatal vaccination programme, and aged 14-19 and therefore directly affected by the universal school-age scheme.

### **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 association 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.

Model	IRR (CI 95%)*	Variable						DoF**	LPD***	LOOIC (se)****
		Policy Change	Age	UK born rates	Non-UK born rates	Year of study entry				
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	1.00 (1.00, 1.00)	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	1.00 (1.00, 1.00)	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 (-).

Variable	IRR (95% CrI)*	
	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 eligibility, group level	-	
Intercept (standard deviation)	-	1.13 (1.05, 1.26)
Year of study eligibility, 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 for 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.

Model	IRR (CI 95%)*	Variable						DoF**	LPD***	LOOIC (se)****
		Policy Change	Age	UK born rates	Non-UK born rates	Year of study entry				
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	1.00 (1.00, 1.00)	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	1.00 (1.00, 1.00)	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 (see Section 7.3.2) the evidence of an association between the change in policy and TB incidence was mixed across all models 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 the 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 eligible for 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.

Model	IRR (CI 95%)*	Variable						DoF**	LPD***	LOOIC (se)****
		Policy Change	Age	UK born rates	Non-UK born rates	Year of study entry				
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	1.00 (1.00, 1.00)	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	1.00 (1.00, 1.00)	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 (-).

Variable	IRR (95% CrI)*	
	UK born	Non-UK born
Policy change**		
Pre-change	<i>Reference</i>	<i>Reference</i>
Post-change	0.96 (0.82, 1.14)	0.62 (0.44, 0.88)
Age		
0	<i>Reference</i>	<i>Reference</i>
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 eligibility, group level		-
Intercept (standard deviation)	1.13 (1.04, 1.26)	-
Year of study eligibility, 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 relevant 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.

Model	IRR (CI 95%)*	Variable						DoF**	LPD***	LOOIC (se)****
		Policy Change	Age	UK born rates	Non-UK born rates	Year of study entry				
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	1.00 (1.00, 1.00)	No	No	No	No	Yes	13	-150	330 (25)	
Model 1	1.00 (1.00, 1.00)	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 (2015) in the directly impacted populations with 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 (age)	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 2015, 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. TB 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).[25,28] 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 dataset (ETS), and a repeated survey (LFS) both of which

## 7.6. Discussion

---

may have introduced bias. Whilst the LFS is a robust data source, widely used in academic studies,[45,94,95] 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 MAR 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,[15,78] and ethnicity.[15,78,82] 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.[96] 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,[97] 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.[98] 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. These results should be interpreted carefully, especially in the non-UK born, as I could not fully rule out the impact of other TB control measures that may have been changed at the same time as vaccination policy. The severity of TB disease is known to differ across age groups with children having a higher incidence of TB meningitis, which can be severe, compared to other age groups.[20] This variation

should also be considered when evaluating these results.

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 chapter 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 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. 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, I 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](https://doi.org/10.5281/zenodo.2583056)<sup>3</sup>

---

<sup>3</sup>Alternatively code link: <https://github.com/seabbs/DirectEffBCGPolicyChange>

# Chapter 8

# Developing a dynamic transmission model of Tuberculosis

## 8.1 Introduction

In the previous chapter (Chapter 7) I estimated the impact of the change in BCG policy on the subset of the population who were directly impacted. Unfortunately, the time horizon of this estimate was limited by the available data. Additionally, 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 indirect impacts, via onwards transmission, not captured in this estimate. Both of these limitations can be overcome using a dynamic transmission model (see Chapter 1). This explicitly models the rate that individuals are infected using the mass action assumption.[7,8] A dynamic transmission model also allows estimates to be made of the long term impact of BCG policy changes, via model simulation.

This chapter presents the development and parameterisation of a 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 TB 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 (Chapter 9).

## 8.2 Previously published TB models

The model presented in this chapter is not the first dynamic model of TB transmission in England to be developed nor is it the first model including BCG vaccination. Vynnycky et al. developed a highly detailed - and rigorously parameterised - model of TB transmission in England and Wales.[99] However, unlike the model presented in this chapter, this model only included the white male population, it only modelled transmission until 1990, and it did not capture the full complexity of BCG vaccination. Harris et al. reviewed TB

models that explored the impact of potential future vaccines.[100] They found that model structures evolved over time, with only a few models including the variable infectiousness of active disease, vaccine waning, and age stratification - all of which are included in the model presented in this chapter. Few of the models they evaluated considered low incidence settings, like England. The model presented in this chapter also makes use of recent advances in TB modelling that indicate which features are key to reproducing TB transmission dynamics.[101,102] Several previous studies have evaluated the role of BCG vaccination and considered various different vaccination scenarios.[72,103–106] However, these studies either did not consider TB transmission or made use of simple transmission models that did not capture the full complexity of either TB transmission or BCG vaccination. Finally, the model presented in this chapter makes use of a highly detailed routine surveillance data set to provide setting appropriate prior distributions for key parameters that were not available for previous models. For further details of previous modelling approaches and their strengths and weaknesses in comparison to the model presented here please see the discussion section of this chapter.

## 8.3 Choice of model structure

When developing an infectious disease dynamic model there is a trade-off between reproducing reality and interpretability.[7] 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.[7] 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 ETS (Chapter 4) is used to support evidence from the literature. Further background information can be found in Chapter 2 and Chapter 4.

### 8.3.1 TB disease

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

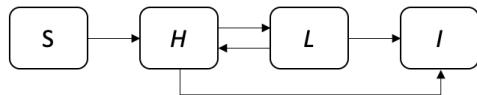
1. **Latency** - after an initial infection 5-10% of individuals develop symptomatic TB within 1-2 years. The majority of individuals enter a latent state in which they passively carry *TB mycobacteria* but are not symptomatic. Reactivation of the bacilli can then occur many years later due to a loss of immune control.[12] Simplistically latent TB may be modelled with a single latent compartment[107], more commonly an additional transition rate between the susceptible and active disease states is added.[108] 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.[101,108] More complex structures that are commonly used incorporate either parallel or serial latency (Figure 8.1). Both of these structures incorporate both slow and fast latent periods and have been shown to produce identical activation dynamics.[101] This is unfortunate as they represent two disparate biological mechanisms, with the serial assumption representing decreasing risk over time for individuals

### *8.3. Choice of model structure*

---

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.[101]

a.)



b.)

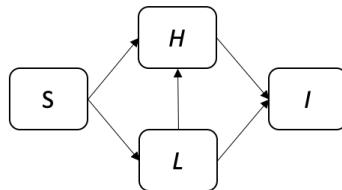


Figure 8.1: Flow diagrams of a.) the serial latency assumption and b.) the parallel latency assumption. The flow diagrams contain the following compartments; Susceptible ( $S$ ), high risk latent ( $H$ ), low risk latent ( $L$ ), and infected ( $I$ ). Solid arrows represent transition rates. Note that 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 variants 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,[13,109] 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 TB than other age-groups.[110] 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).[2] 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,[111] with smear negative cases contributing approximately 76% (95% CI 70%, 80%) less to transmission than smear positive cases.[112] The proportion of smear positive cases varies with age,[113] 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

### *8.3. Choice of model structure*

---

2000 and 2015. the model presented here includes sputum status via the force of infection.[7,8]

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.[114] 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%).[115] 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.[116,117] 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).[2,18] 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,[101] as does the proportion of cases that develop pulmonary TB,[110], 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 modelled 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 (see 8.5.2).

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.[2] 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 75-79 years old (62.9 per 100,000 in 2016) and those aged 25-29 years old (61.6 per 100,000 in 2016).[2] 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 a mixing parameter that controls the degree of contact between non-UK born cases and those born in the UK.[7,8]

### **8.3.2 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[4,23,24] but to have variable protection in adults ranging from 0-80%.[27] 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 1950s found that the BCG vaccine was highly effective.[26] There is little evidence to suggest that this has changed in the UK born population.
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,[28] and a recent study suggests that this protection may last later into adulthood in the UK born.[29]
3. **Protection from initial infection** - there is evidence that the BCG vaccine provides partial protection against initial infection.[4] 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,[23,25] although there is little evidence of this in England. In order to answer questions relevant to BCG vaccination, TB disease must be modelled in young children and young adults. To capture the waning of BCG protection age structure must be modelled beyond these age groups.[28]
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.[5] 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 modelled 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 model 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 to 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.[7,8] Non-UK born cases are included into the force of infection.[7,8]

### **Demographic model**

The model is stratified into 11 age groups with 5 year age groups from 0 to 49, a single age group from 50-69, and a single age group from 70 to 89. 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 modelled due to large amounts of uncertainty in the demographic data and because cases in this 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 were 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 modelled age group. The target age group can be varied to represent changing BCG vaccination policy. The vaccinated population is then modelled 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 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 modelled 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 [7,8]).

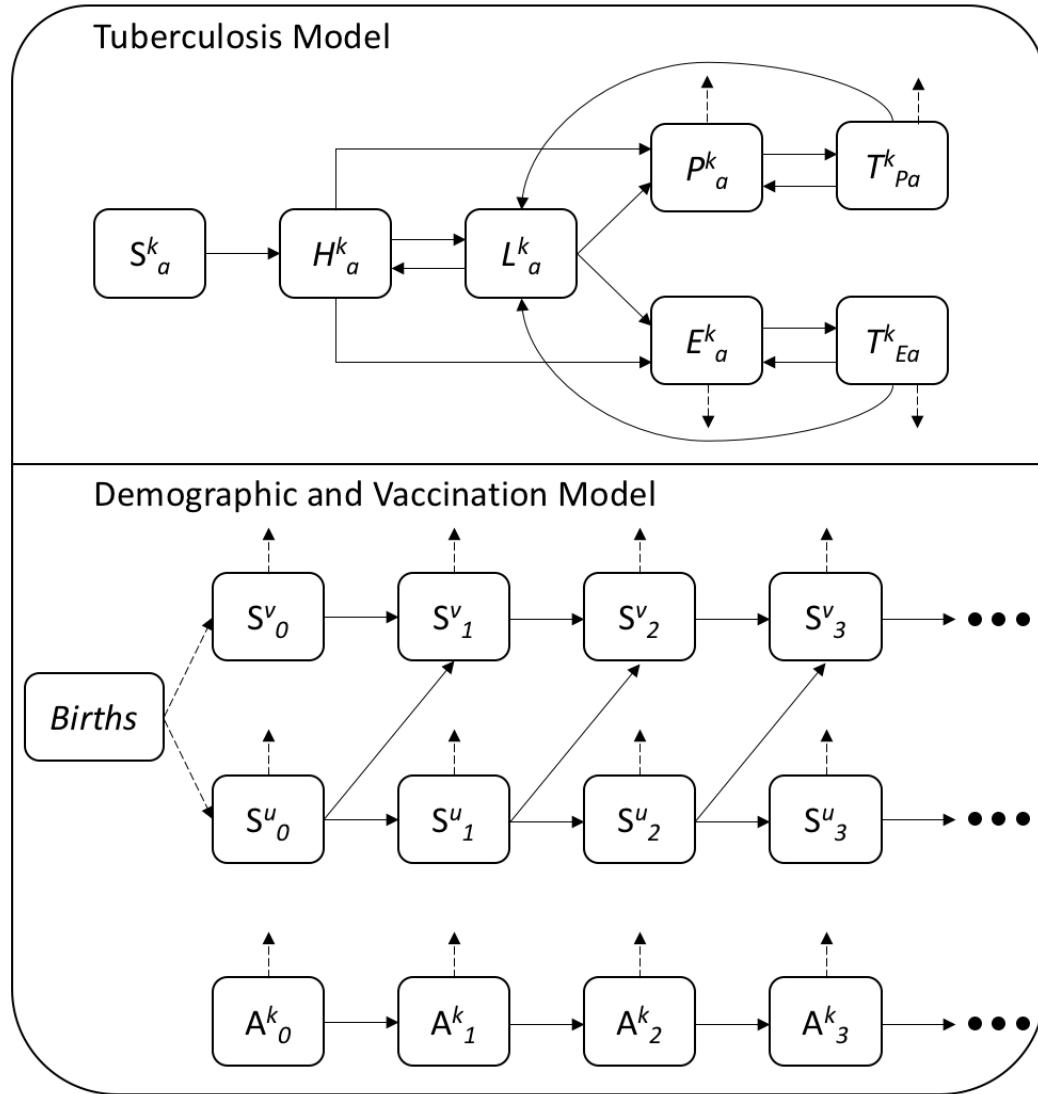


Figure 8.2: Flow diagram for the dynamic TB disease model with demographics and vaccination described. The TB model contains the following compartments; Susceptible ( $S$ ), high risk latent ( $H$ ), low risk latent ( $L$ ), active cases with pulmonary TB ( $P$ ), active cases with extra-pulmonary TB 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  $a = 1, 2, \dots, 11$ ) in the disease model and the  $0, 1, 2, 3$  subscripts in the demographic model. Each age groups spans 5 years (i.e  $0 - 4, 5 - 9, 10 - 14, \dots$ ) 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 vaccinated 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  $C = S, H, L, P, E, T_P, T_E$ ),

$$\frac{dC}{dt} = J_d + J_p \quad (8.1)$$

The disease model ( $J_d$ ) is then defined as,

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

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

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

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

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

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

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

Where the unvaccinated ( $u$ ) and vaccinated ( $v$ ) populations are represented by  $k = u, v$  and age groups are represented by  $a = 0, 1, 2, 3, \dots, 11$ . The disease model parameters are defined as follows:  $\lambda_a$  is the force of infection;  $\epsilon_a$  is the rate of activation from each latent population;  $\kappa_a$  is the rate of transition into the low risk latent population;  $\nu_a$  is the rate of starting treatment;  $\delta$  is the protection from re-infection conferred by prior latent infection;  $\Upsilon_a$  is the proportion of cases that develop pulmonary TB, with or without extra-pulmonary TB;  $\mu_a^{P,E}$  is the mortality from active pulmonary ( $P$ ) and extra-pulmonary ( $E$ ) TB;  $\zeta_a$  is the rate of treatment failure;  $\phi_a$  is the rate of successful treatment;  $\alpha_a$  is the effectiveness of the BCG vaccine at preventing active TB disease; and  $\chi_a^k$  is the protection inferred due to vaccination from initial infection. In the unvaccinated population (when  $k = u$ )  $\alpha_a^u = 0$  and  $\chi_a^u = 0$ . Parameters with an  $a$  subscript, or superscript, are age-stratified.

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

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

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

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

Where  $\omega(t)$  is the time varying number of births,  $\gamma_a$  is the age-specific proportion that are vaccinated,  $\theta_a$  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;

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

Code for this model is available online.<sup>1</sup>

#### 8.4.3 Force of infection

The force of infection ( $\lambda_a^k$ ) is the rate at which susceptible individuals are infected. Here it is defined using the law of mass action which assumes that infectious cases and susceptible cases randomly mix with the rate of mixing being determined by the fraction of the population that are susceptible, the transmission probability, and the contact rate.[7,8] It is age stratified ( $a, A = \max(a)$ ) by contact rates, by the proportion of cases that are smear positive in a given age group, by the transmission probability of each case, and by the rate of starting treatment. Stratification by vaccine status ( $k$ ) is introduced by the number of current pulmonary TB cases. It can be defined as follows,

$$\lambda_a = \frac{\beta_a}{N_a} \sum_{i=1}^A \rho_i C_{ai} \left( \frac{M_i \iota_i}{\nu_i^P} + \sum_{j=u,v} P_i^j \right) \quad (8.13)$$

Where  $\iota_i$  is the age stratified number of non-UK born pulmonary cases notified in a given year,  $P_i^{u,v}$  is the number of vaccinated, and unvaccinated, pulmonary TB cases,  $\rho_i$  is the age-specific proportion of cases that are smear positive,  $\nu_i^P$  is the age-specific rate of starting treatment for active pulmonary TB,  $C_{ai}$  is the age-stratified contact matrix (Section 8.5.2),  $\beta_a$  is the age-stratified transmission probability,  $N_a$  is the number of people in age group  $a$ , and  $M_i$  is the age-stratified mixing rate between the UK born and non-UK born population. Non-UK born cases pulmonary cases ( $\iota_i$ ) are included in the the force of infection, along with their own mixing rate ( $M_i$ ), and the duration of time spent infectious ( $\frac{1}{\nu_i^P}$ ) so that their external forcing on UK born cases is modelled (Section 8.3). A separate mixing rate ( $M_i$ ) is used as the extent of this forcing is not known. Both UK born and non-UK born pulmonary TB cases are multiplied by the proportion of cases that are smear positive ( $\rho_i$ ) as it is thought that these cases are responsible for the majority of onwards transmission (Section 8.3).

## 8.5 Parameterisation and data synthesis

Parameters distributions were either estimated from the available data, assumed based on common values found in the literature, or based on expert opinion. Parameter distributions were assumed to be normal unless otherwise stated. Where no comparable estimates were found in the literature, and data was not available, a largely uninformative distribution was used, informed when possible by expert opinion. Normal distributions were extrapolated based on published confidence/credible intervals for prior distributions based on literature sources. All prior distributions were truncated to be greater than or equal to 0 with proportions further truncated to be less than or equal to 1. Prior distributions are first summarised

<sup>1</sup>Model code: <https://github.com/seabbs/ModelTBBCGEngland/blob/master/inst/bi/BaseLineModel.bi>

in tables stratified by submodel, with additional details on their derivation given in the subsequent sections. Code for all prior derivations is available online<sup>2</sup>. The data sources used to estimate model parameters have been detailed although for the ETS and the LFS 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 demographic, clinical, and microbiological data; a yearly report is published detailing data collection, cleaning, and trends in TB incidence (Chapter 4).[2]

#### **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.

---

<sup>2</sup><https://github.com/seabbs/ModelTBBCGEngland/tree/master/data-raw>

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. All prior distributions were truncated to be greater than or equal to 0 with proportions further truncated to be less than or equal to 1. Additional detail for each prior derivation can be found in the following sections. 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
$C_{\text{eff}}$	The assumed effective number of contacts per infectious TB case.	$\mathcal{N}(1, 1)$	-	Estimated using a dynamic model of TB transmission in England which found an effective contact rate of 1 in 1990. A conservative interval has then been applied.	Vynnycky et al.
$C_{\text{eff}}^{\text{hist}}$	The assumed historic effective number of contacts per infectious TB case.	$\mathcal{U}(C_{\text{eff}}, 20)$	-	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. It has been assumed that the historic contact rate is bounded below by the current contact rate.	Vynnycky et al.
$C_{\text{eff}}^{\text{half-life}}$	It is assumed that the historic effective contact rate decays from 1935 to 1980 with a rate of $C_{\text{eff}}^{\text{half-life}}$ .	$\mathcal{N}(5, 5)$	-	The prior distribution is informed by historic TB notifications.	Assumption
$\beta_{\text{young-adult}}$	This parameter modifies the effective contact rate in scenarios when the transmission probability is modified for young adults (15-24).	$\mathcal{U}(0, 10)$	-	An uninformative prior has been used bounded above to restrict the transmission probability in young adults to be no greater than 10 times that in other age-groups.	Assumption
$\Upsilon$	The age-specific proportion of cases that have pulmonary TB	$\Upsilon_{0-14} = \mathcal{N}(0.629, 0.00101)$ , $\Upsilon_{15-59} = \mathcal{N}(0.706, 0.00411)$ , $\Upsilon_{60-89} = \mathcal{N}(0.750, 0.00569)$	Proportion	Estimated using the age-specific proportion of cases that had pulmonary TB in the ETS.	ETS
$\rho$	The age-specific proportion of pulmonary TB cases that are smear positive	$\rho_{0-14} = \mathcal{N}(0.302, 0.0189)$ , $\rho_{15-59} = \mathcal{N}(0.652, 0.00518)$ , $\rho_{60-89} = \mathcal{N}(0.536, 0.00845)$	Proportion	Estimated using the age-specific proportion of pulmonary TB cases that were smear positive in the ETS.	ETS

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. All prior distributions were truncated to be greater than or equal to 0 with proportions further truncated to be less than or equal to 1. Additional detail for each prior derivation can be found in the following sections. 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
$C$	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 provide a symmetric contact matrix.	Mossong et al.
$\iota(t)$	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.	ETS
$\iota_{\text{scale}}$	Scaling parameter for the importation of non-UK born cases between 1960 and 2000.	$\mathcal{U}(0, 200)$	-	This largely uninformative prior range was chosen so that non-UK born scaling can vary between approximately linear to approximately constant.	Assumption
$M$	The proportion of mixing between the UK born and non-UK born population.	$\mathcal{N}(1, 1)$	Proportion	Any degree of mixing is allowed as there is little data on which to base this estimate. Mixing greater than 1 is allowed as this is used to represent non-UK born cases being in some way more infectious than non-UK born cases.	Assumption
$M_{\text{young-adult}}$	This parameter modifies the non-UK born mixing rate in scenarios when $M$ is modified for young adults (15-24).	$\mathcal{U}(0, 10)$	-	An uninformative prior has been used bounded above to restrict non-UK born mixing in young adults to be no greater than 10 times that in other age-groups.	Assumption
$\chi$	Age-specific protection from infection with TB due to BCG vaccination	$\chi_i^v = \mathcal{N}(0.185, 0.0536)$ , where $i$ is the age group vaccinated.	Proportion	A meta-analysis of the protection from infection due to BCG vaccination in children. It has been assumed that there is no reduction in protection in UK born adults. Distribution derived by the assumption of a normal distribution based on published 95% confidence intervals.	Roy et al.

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. All prior distributions were truncated to be greater than or equal to 0 with proportions further truncated to be less than or equal to 1. Additional detail for each prior derivation can be found in the following sections. 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
$\epsilon_H$	The age-specific rate of transition to active disease during high risk latent period.	$\epsilon_H^{0-4} = \mathcal{N}(0.00695, 0.00130)$ , $\epsilon_H^{5-14} = \mathcal{N}(0.0028, 0.000561)$ , $\epsilon_H^{15-89} = \mathcal{N}(0.000335, 0.0000893)$	$days^{-1}$	From fitting a similar model to contact data in Australia, and Holland. Distribution derived by the assumption of a normal distribution based on published 95% credible intervals.	Ragonnet et al.
$\kappa$	The reciprocal of the age-specific average high risk latent period.	$\kappa^{0-4} = \mathcal{N}(0.0133, 0.00242)$ , $\kappa^{5-14} = \mathcal{N}(0.0120, 0.00207)$ , $\kappa^{15-89} = \mathcal{N}(0.00725, 0.00191)$	$days^{-1}$	From fitting a similar model to contact data in Australia, and Holland. Distribution derived by the assumption of a normal distribution based on published 95% credible intervals.	Ragonnet et al.
$\epsilon_L$	The reciprocal of the age-specific average low risk latent period.	$\epsilon_L^{0-4} = \mathcal{N}(8.00e^{-6}, 4.08e^{-6})$ , $\epsilon_L^{5-14} = \mathcal{N}(9.84e^{-6}, 4.67e^{-6})$ , $\epsilon_L^{15-89} = \mathcal{N}(5.95e^{-6}, 2.07e^{-6})$	$days^{-1}$	From fitting a similar model to contact data in Australia, and Holland. Distribution derived by the assumption of a normal distribution based on published 95% credible intervals.	Ragonnet et al.
$\epsilon_L^{\text{older-adult}}$	This parameter modifies the activation rate of low risk latent cases who are 70+ and reduces the activation rate for other adults so that the mean activation rate is $\epsilon_L^{15-89}$	$\mathcal{N}(2, 0.5)$	-	Evidence suggests that activation risk increases when individuals enter old age. A largely uninformative prior has been used centred around an increase in risk of double the average rate.	Horsburgh, Jr. et al.

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. All prior distributions were truncated to be greater than or equal to 0 with proportions further truncated to be less than or equal to 1. Additional detail for each prior derivation can be found in the following sections. 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
$\alpha_i^T$	The BCG vaccine effectiveness at preventing the development of active TB disease in a TB free population	$\alpha_j^T = 1 - e^{\alpha_j^{\ln(T)}}$ , where $\alpha_{j=i}^{\ln(T)} = \mathcal{N}(-1.86, 0.22)$ , $\alpha_{j=i+5}^{\ln(T)} = \mathcal{N}(-1.19, 0.24)$ , $\alpha_{j=i+10}^{\ln(T)} = \mathcal{N}(-0.84, 0.22)$ , $\alpha_{j=i+15}^{\ln(T)} = \mathcal{N}(-0.84, 0.2)$ , $\alpha_{j=i+20}^{\ln(T)} = \mathcal{N}(-0.28, 0.19)$ , $\alpha_{j=i+25}^{\ln(T)} = \mathcal{N}(-0.23, 0.29)$ ) and $i$ is the age group vaccinated	Proportion	Poisson regression used to calculate Risk Ratios from literature values. A distribution is then found using the log normal approximation. Effectiveness estimates are caculated using 1 minus the exponentiated log normal distribution.	Hart et al. and Mangtani et al.
$\delta$	Reduction in susceptibility to infection for low risk latent cases.	$\mathcal{N}(0.78, 0.0408)$	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. Distribution derived by the assumption of a normal distribution based on published 95% confidence intervals.	Andrews et al.
$\nu^{P,E}$	The reciprocal of the average infectious period	$\nu_{0-14}^P = \mathcal{N}(0.181, 0.310)^{-1}$ , $\nu_{15-89}^P = \mathcal{N}(0.328, 0.447)^{-1}$ , $\nu_{0-14}^E = \mathcal{N}(0.306, 0.602)^{-1}$ , $\nu_{15-89}^E = \mathcal{N}(0.480, 0.866)^{-1}$	$\text{years}^{-1}$	Estimated based on the time from initial symptoms to starting treatment.	ETS
$\phi$	The reciprocal of the time to successful treatment completion	$\phi_{0-14} = \mathcal{N}(0.606, 0.237)^{-1}$ , $\phi_{15-69} = \mathcal{N}(0.645, 0.290)^{-1}$ , $\phi_{70-89} = \mathcal{N}(0.616, 0.265)^{-1}$	$\text{years}^{-1}$	Estimated based on the time from starting treatment to treatment completion.	ETS
$\mu$	Rate of age-specific pulmonary/extra-pulmonary TB mortality	$\mu_{0-14} = \mathcal{N}(0.0039, 0.018)$ , $\mu_{15-59} = \mathcal{N}(0.0226, 0.00787)$ , $\mu_{60-89} = \mathcal{N}(0.117, 0.0165)$	$\text{years}^{-1}$	Estimated based on outcomes at 12 months where cause of death was known, including all-cause deaths in the denominator.	ETS

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. All prior distributions were truncated to be greater than or equal to 0 with proportions further truncated to be less than or equal to 1. Additional detail for each prior derivation can be found in the following sections. 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
$\zeta$	Rate of loss to follow up	$\zeta_{0-14} = \mathcal{N}(0.00976, 0.0179)$ , $\zeta_{15-59} = \mathcal{N}(0.0304, 0.00764)$ , $\zeta_{60-89} = \mathcal{N}(0.00614, 0.0159)$	$years^{-1}$	Estimated based on outcomes at 12 months for TB cases	ETS

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
$\iota(t), \mu, \nu^{P,E}, \phi,$ $\rho, \Upsilon, \zeta$	-	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
$\mu^{\text{all-cause}}(t), \omega(t)$	-	England	-	The Office for National Statistics (ONS) compiles demographic, health, economic, and social data for the United Kingdom	ONS
$C_{\text{eff}}, C_{\text{eff}}^{\text{hist}}$	Dynamic mod-elling study	England	Up to 1990	Used a dynamic model of tuberculosis, robustly parameterised to the available evidence and including realistic population demographics to estimate the effective contact rate of TB over time until the 1990's in the UK born white male population.	[118]
$C$	Contact survey	Europe - includ-ing 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.	[119]

## 8.5. Parameterisation and data synthesis

---

Parameters	Study Type	Setting	Year	Description	Source
$\chi$	Systematic review	Global and meta-analysis	Up to 2014	A meta-analysis; conducted with the aim of determining whether BCG vaccination protects against tuberculosis infection as assessed by interferon $\gamma$ release assays (IGRA) in children. Estimated both protection from initial latent infection and active TB disease.	[4]
$\epsilon_H, \epsilon_L, \kappa$	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.	[101]
$\alpha_i^T$	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 vaccine in 2.5 year intervals from vaccination.	[26]
$\alpha_i^T$	Population based case-control study	England	2002-2014	Recruited UK-born White subjects with TB and randomly sampled 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.	[29]

Parameters	Study	Type	Setting	Year	Description	Source
$\delta$	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.	[115]
$\beta_{\text{young-adult}}, \gamma, M_{\text{young-adult}}, M$	-	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 was estimated using the ETS for each age-group included in the model from 2000 until 2015. Prior to 2000, incidence in the non-UK born are unavailable and the relationship to transmission in UK-born cases is unknown. To account for this, importation of non-UK born cases in the model begins in 1960 and then is scaled up through to 2000. As the form of this relationship is unknown the following functional form was used to scale cases based on those observed in 2000,

$$\text{Non UK-born cases (time } t) = \left( \frac{\exp\left(\frac{t-1960}{\ln(2)(-1)\tau_{\text{scale}}}\right) - 1}{\exp\left(\frac{2000-1960}{\ln(2)(-1)\tau_{\text{scale}}}\right) - 1} \right) \text{Non UK-born cases (} t = 2000\text{)} \quad (8.14)$$

This functional form was chosen as it is flexible enough to represent exponential growth, bounded growth and near linear growth depending on the choice of  $\tau_{\text{scale}}$ . This allows the scale up of non-UK born cases to be fitted to the available data during the model fitting stage (Chapter 9). To incorporate the uncertainty in the number of observed non-UK born cases a normal distribution was used, with the standard deviation and mean 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 ( $\beta_a$ ) can be redefined in terms of effective contacts (the number of contacts that are infected from each infected case;  $C_{\text{eff}}$ ), historic effective contacts ( $C_{\text{eff}}^{\text{hist}}$ ), actual average yearly total contacts ( $C_{\text{actual}}$ ), the average period of time infectious ( $\frac{1}{\nu_{\text{avg}}^P}$ ), and the average mortality rate ( $\mu_{\text{avg}}$ ) as follows,

$$\beta_a = \frac{(\nu_{\text{avg}}^P + \mu_{\text{avg}})C_{\text{eff}}^{\text{scaled}}}{C_{\text{actual}}} \quad (8.15)$$

$$C_{\text{eff}}^{\text{scaled}} = \begin{cases} C_{\text{eff}}^{\text{hist}} & \text{if } t < 1935, \\ C_{\text{eff}} + \left( C_{\text{eff}}^{\text{hist}} - C_{\text{eff}} \right) \left( \frac{t-1935}{C_{\text{eff}}^{\text{half-life}}} \right)^{1/2} & \text{if } 1935 \leq t \leq 1980, \\ C_{\text{eff}} & \text{if } t > 1980. \end{cases} \quad (8.16)$$

Vynnycky et al. found that the effective contact rate for TB was approximately 22 in 1900 and fell to approximately 1 in 1990.[118] Incidence rates have increased since the early 1980s and it is unclear what impact this has had on the effective contact rate. I have assumed that the effective contact rate is normally distributed with a mean of 1 and a standard deviation of 0.5. For the historic effective contact rate I have assumed a uniform distribution with a lower bound of the current effective contact rate and an upper bound of 20. I have also assumed that the historic contact rate declines over time, starting in 1935, reducing to the current effective contact rate in 1980. The speed of this decay is set by defining a decay parameter ( $C_{\text{eff}}^{\text{half-life}}$ ). The prior for the decay parameter is assumed to be normally distributed with a mean of 5 years and a standard deviation of 5 (truncated to be greater than 0). This prior is based on the observed trend in notifications. Age stratification of  $\beta_a$  is explored by including modifiers for certain age-groups. The baseline scenario is that no modification is required, with variation explored for young adults (15-29;  $\beta_{\text{young adult}}$ ) as a scenario. The prior for this modifier was assumed to be uniform, bounded by 0 and 10. 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 the reciprocal of the time with active, untreated, disease from 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/extrapulmonary 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 2 years, representing natural recovery or other removal from the infectious population. From 1952 to 1990 the time with active TB is assumed to decrease linearly.

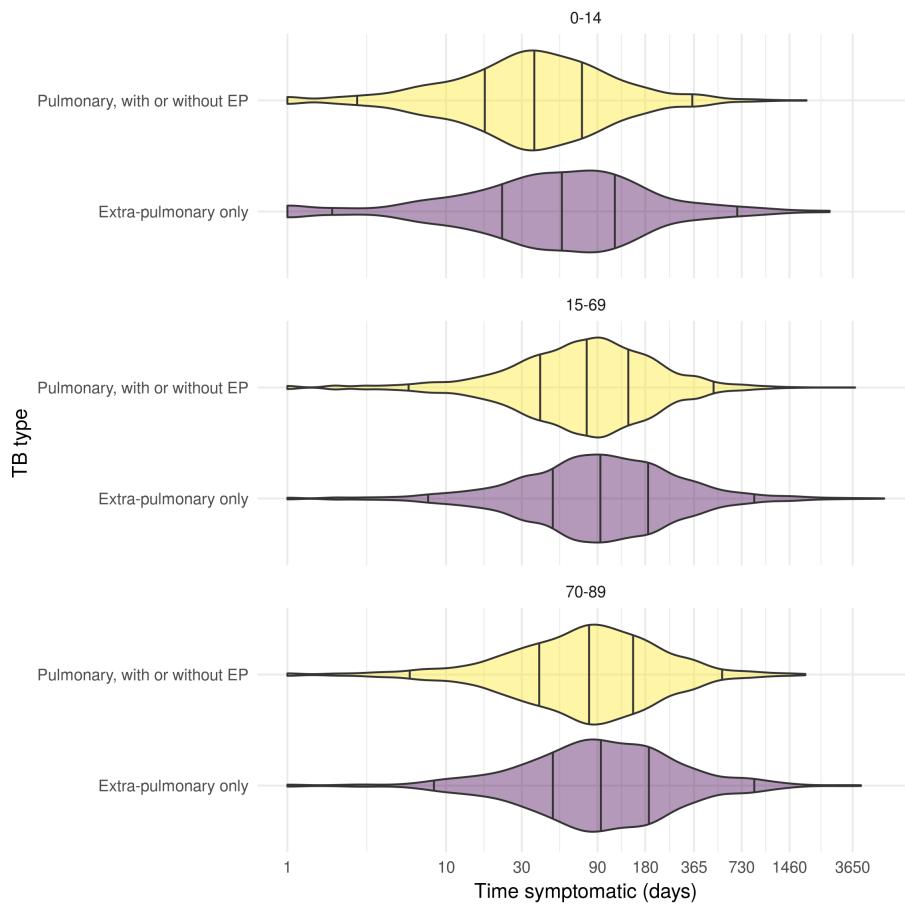


Figure 8.3: Distribution of time to treatment (days) from the date of reported symptom onset until the date started treatment for the UK born, stratified by age group and pulmonary/extrapulmonary TB status in the ETS system for notifications between 2000 and 2012. Age is stratified into three groups; children (0-14), adults (15-69) older adults (70-89). 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. Vertical lines indicate the 2.5%, 25%, 50%, 75%, and 97.5% quantiles.

**Rate of successful treatment** The rate of successful treatment was estimated as the reciprocal of 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.

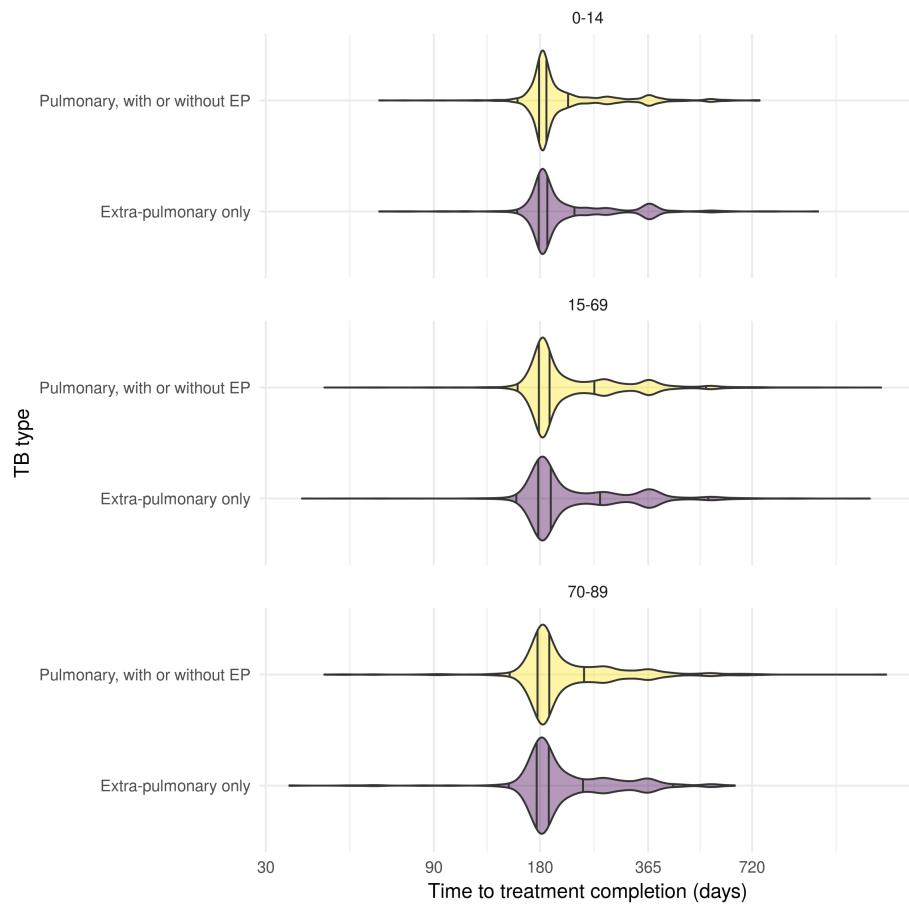


Figure 8.4: Distribution of time to treatment completion in the UK born successfully treated (days), stratified by age group and pulmonary/extrapulmonary TB status in the ETS for notifications between 2000 and 2012. Age is stratified into three groups; children (0-14), adults (15-69) older adults (70-89). 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. Vertical lines indicate the 2.5%, 25%, 50%, 75%, and 97.5% quantiles.

**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.[119]

**The POLYMOD contact survey** The POLYMOD survey,[119] 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. Survey 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 lead to over-fitting without a strong a priori hypothesis. In high and medium burden countries it has been shown that within household transmission is not a major driver of overall transmission.[109] 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 7.7% (1849/24,060) of cases in England between 2010 and 2012 lived in a household with another case.[120] 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.[109] 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 across age groups.

**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<sup>3</sup> 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 num-

---

<sup>3</sup>**socialmixr**: <https://github.com/sbfnk/socialmixr>

ber 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,

$$C_{ij}N_i = C_{ji}N_j \quad (8.17)$$

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,

$$C'_{ij} = \frac{C_{ij}N_i + C_{ji}N_j}{N_i + N_j} \quad (8.18)$$

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.[119] 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.

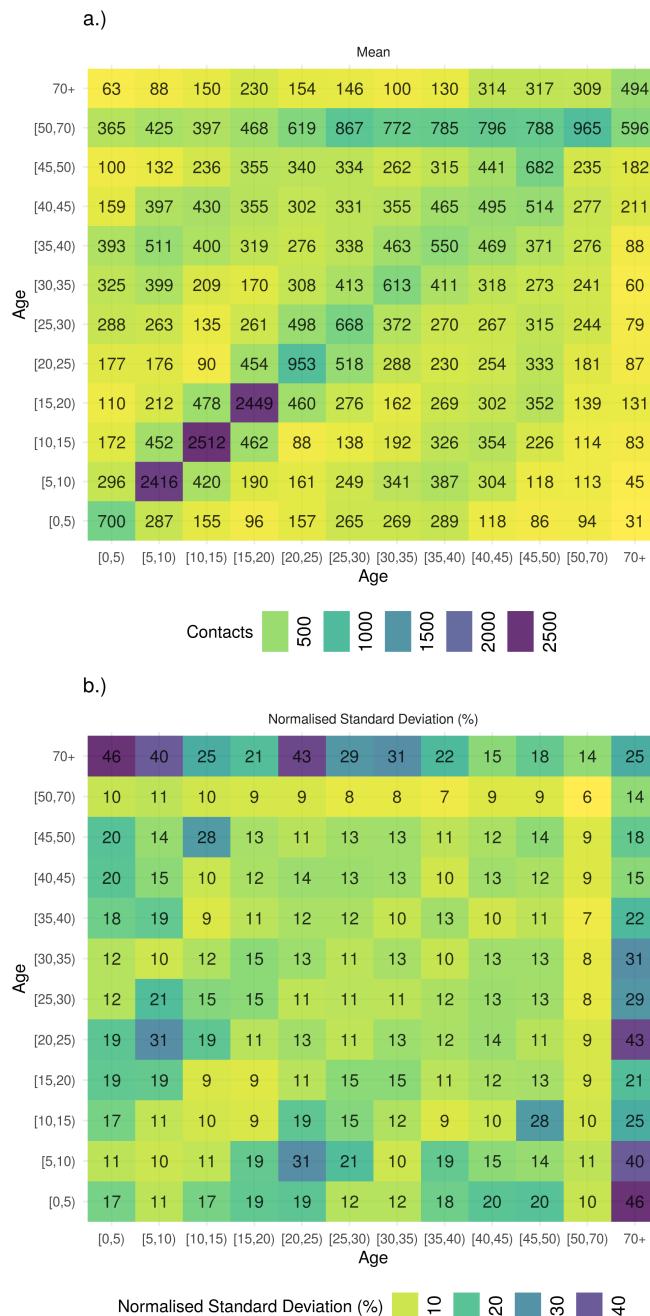


Figure 8.5: a.) Mean contacts (non-unique social contacts per year) and the b.) normalised standard deviation (%) of 1000 bootstrapped samples of social contacts from the POLYMOD social contact survey using 5 year age groups up 49 years old and then a single group for 50-69 year olds. Mixing is highly assortative by age with children and young adults representing the majority of contacts. There is also evidence of mixing between children and middle age adults with older children mixing with progressively older adults. Contact rates in older adults are highly uncertain, with the most uncertainty in mixing between older adults and young children.

### Vaccination model parameters

An overview of the vaccination 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 in the following section.

**Effectiveness of the BCG vaccine at preventing active TB** The effectiveness of the BCG vaccine is usually estimated using its 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.[4] The overall effectiveness ( $\alpha_a^T$ ) of the BCG vaccine can be estimated from the combined effectiveness at preventing initial infection ( $\chi_a^v$ ) and then preventing activation in latently infected individuals ( $\alpha_a^v$ ) using the following equation,

$$\alpha_a^T = \chi_a^v + (1 - \chi_a^v)\alpha_a^v \quad (8.19)$$

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

$$\alpha_a^v = \frac{\alpha_a^T - \chi_a^v}{1 - \chi_a^v} \quad (8.20)$$

There is strong evidence that the overall effectiveness of the BCG vaccine reduces over time.[28,29] For this reason the effectiveness of the BCG vaccination overall ( $\alpha_a^T$ ) 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 0-4, and 5-9, years since vaccination estimates of the effectiveness of the BCG vaccine were extracted from the MRC trial.[26] Using 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.[29] Table 8.3 details the estimated effectiveness for each five yearly band after initial vaccination from these data sources. I have assumed that the BCG vaccine is equally effective regardless of the age at vaccination 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 - which can be sampled from - for the overall effectiveness of the BCG vaccine ( $\alpha_a^T$ ). This can be summarised by the following equation,

$$\alpha_a^T \sim 1 - e^{(\mathcal{N}(\ln(RR_{t_v}), SE_{t_v}))} \quad (8.21)$$

Where  $RR_{t_v}$  is logged risk ratio and the  $SE_{t_v}$  is the standard error of the logged Risk Ratio with both being dependent on the time since vaccination ( $t_v$ ). 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.

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.[4] 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 that 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,[28,29] 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 ( $\chi_j$ ) reduces over time with the same functional form as for the overall effectiveness of BCG vaccination ( $\alpha_j^T$ ). This relation can be formalised using the following equation,

$$\chi_j^v = \frac{\alpha_j^T \chi_i^v}{\alpha_i^T} \quad (8.22)$$

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

### Demographic model parameters

The demographic model parameters 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). All prior distributions were truncated to be greater than or equal to 0 with proportions further truncated to be less than or equal to 1. Additional detail for each prior derivation can be found in the following sections.  $\mathcal{N}$  = Normal 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.	ONS
$\gamma$	BCG vaccination coverage	$\gamma_i = \mathcal{N}(0.75, 0.05)$ Where $i$ is the age group vaccinated.	Proportion	England has a robust national health service and an established system for providing BCG vaccination.	Assumption
$\theta$	Rate of ageing	-	$\text{years}^{-1}$	Defined as the reciprocal of the width of the modelled age groups.	Model defined
$\mu^{\text{all-cause}}(t)$	Time varying all-cause age-specific mortality rate	-	$\text{years}^{-1}$	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.	ONS

**Age-stratified population estimates** Age-stratified and UK birth stratified population estimates for England were estimated using the LFS (Section 4.5). 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 (45-49 years old) 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 modelled age groups.

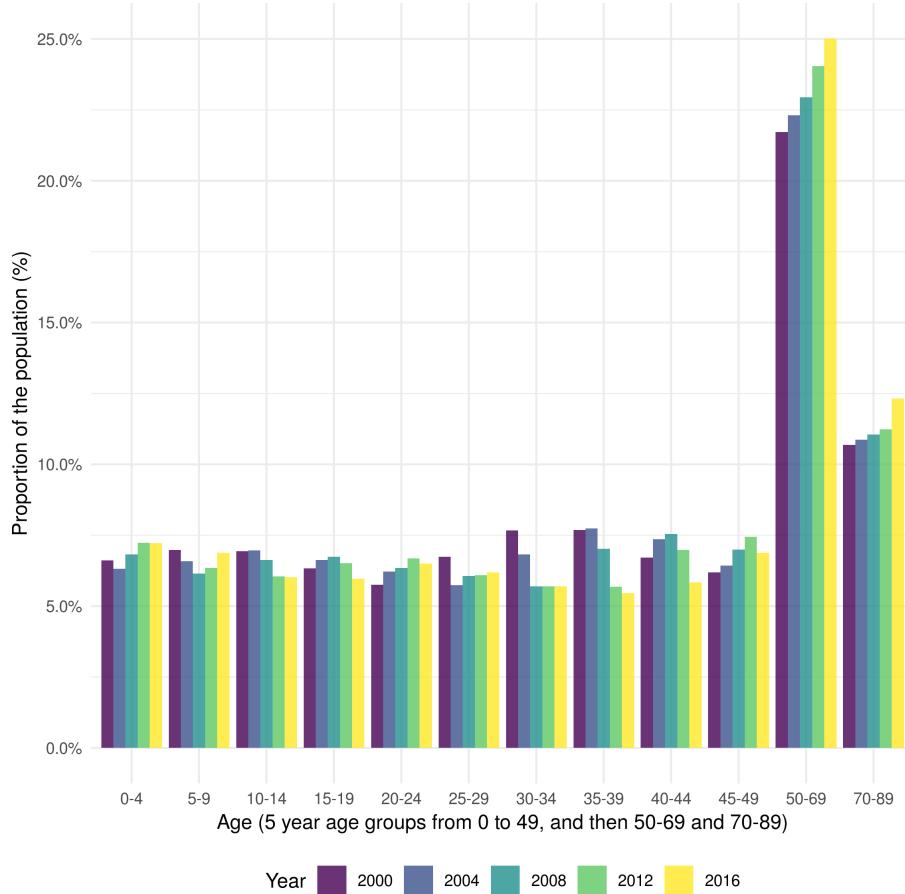


Figure 8.6: Distribution of the UK born population of England in 2000, 2004, 2008, and 2012. Age is grouped into 5 year age groups from 0 to 49, from 50-69, and from 70 to 89. Those aged 90+ are excluded due to low quality data. The age groups used here represent those used in the model. The figure indicates 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 modelled. The ONS publishes the recorded number of births in England each year

starting from 1929 through to 2015, with projections available through to 2101 (Figure 8.7). 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.

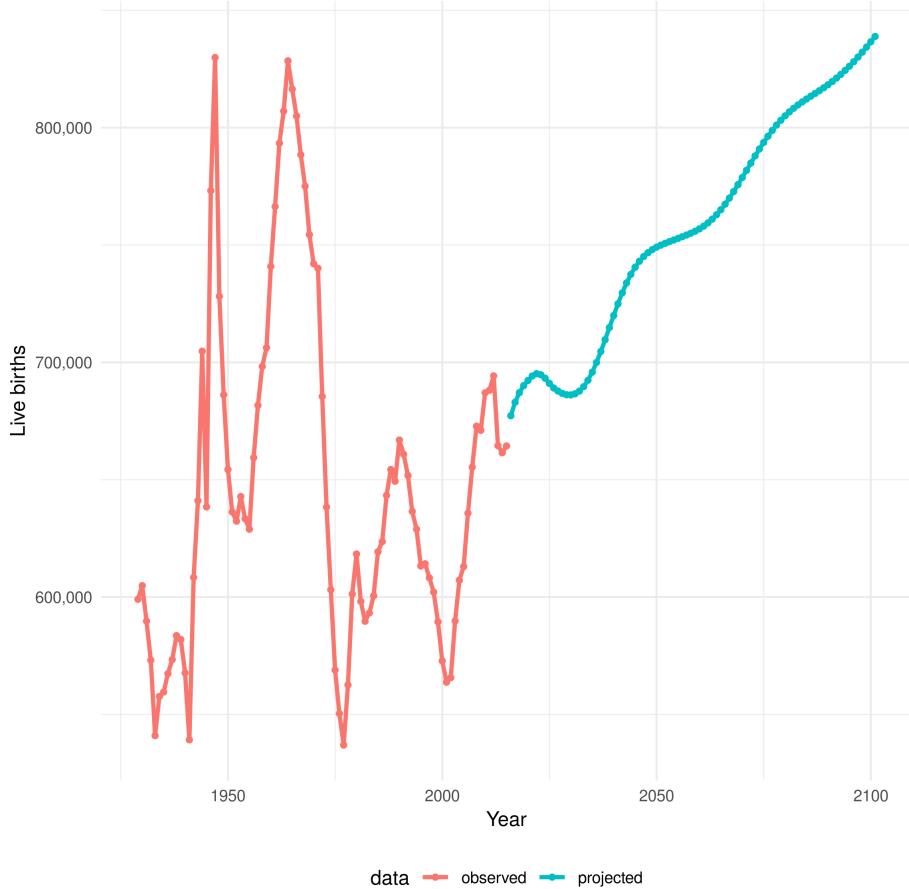


Figure 8.7: Estimated and projected live births in England from 1929 until 2101. The red line indicates estimated data and the blue line indicates projected data. Data is sourced from the ONS.

**Age-specific mortality rates** The time varying, age-specific, noisy, all-cause mortality rates ( $\mu_a^{\text{all-cause}}(t)$ ) 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 I forecast rates using an age-stratified exponential model (Figure 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 ( $\mu_a(t)$ ), excluding deaths from, or related to, TB the following equation was used,

$$\mu_a(t) = \mu_a^{\text{all-cause}}(t) - \left( \frac{\mu_a^P(P_a + T_{Pa}) + \mu_a^E(E_a + T_{Ea})}{N_a} \right) \quad (8.23)$$

Where  $\mu_a(t)$  is constrained to be greater than or equal to zero,  $\mu_a^P$  and  $\mu_a^E$  are the age stratified deaths rates in pulmonary ( $P_a$ ) and extra-pulmonary TB cases ( $E_a$ ), and  $N_a$  is the age stratified population.

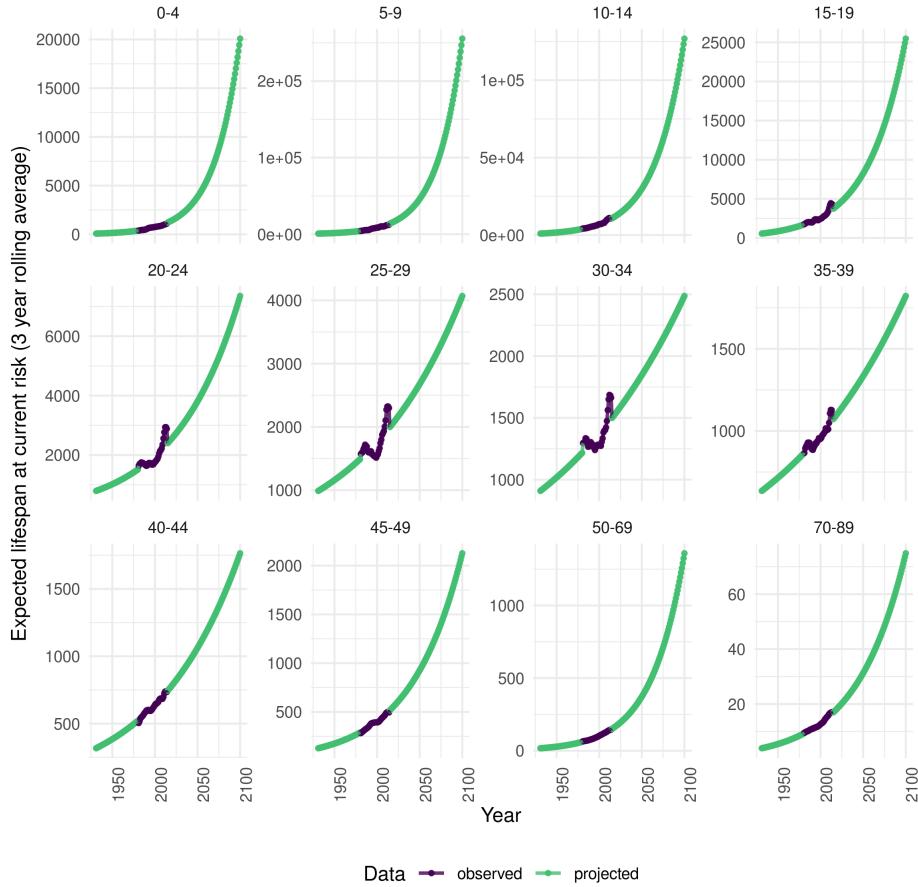


Figure 8.8: Three year rolling average expected remaining lifespan stratified by age group in England from 2000 to 2014. Age is grouped into 5 year age groups from 0 to 49, from 50-69, and from 70 to 89. Those aged 90+ are excluded due to low quality data. The age groups used here represent those used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

## 8.6 Initialisation

Dynamics transmission models are affected by the conditions under which they are initialised.[7] For models of endemic disease this can be problematic as the full disease outbreak can often not be modelled, 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.[7] Models that include demographic processes are more complex to burn-in as demographic data is typi-

## 8.7. Scenarios

---

cally required to initialise the model so that it has the demographics observed during the period of time modelled.

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

Model simulations are initiated in 1931 due to the availability of population data from the 1931 census and because data on live births is only available from 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 recovery from active TB takes 2 years. TB treatment is assumed to begin in 1952 with the discovery of isoniazid and BCG vaccination begins at school-age (15 years old) in 1953. BCG vaccination coverage is assumed to vary randomly over the time horizon of the model but to have the same distribution at all time points. The assumed distribution is normal with a mean of 75% 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.

### 8.6.2 Initial disease distribution

The model is initialised with the number of pulmonary and extra-pulmonary cases reported in 1931. The high risk latent population is initialised by scaling the number of observed cases in 1931 by the proportion of high risk latent cases that develop active TB, the duration that these cases are high risk and then dividing by the infectious period. The low risk latent population is then initialised by scaling the high risk latent population by the cumulative sum of the age distribution of UK born cases in 2000, reduced by 50% to account for mortality (approximately 5% of the population). Finally, the initial susceptible population is based on the population estimate from the 1931 census minus the assumed initial latent cases.

All initial disease compartments, excepting the low risk latent compartment, are distributed based on the age distribution of observed UK born cases in 2000. To account for possible measurement error a normal distribution is sampled around the assumed population in each compartment with a standard deviation of 5% of the reported cases.

## 8.7 Scenarios

All dynamic transmission models require a series of assumptions to be made. These assumptions fall into two categories: structural assumptions and parameter assumptions.[121] 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.

During model fitting, I will consider the evidence for modifying the transmission probability, and non-UK born mixing, by age using three distinct scenarios (Table 8.5). These

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
$\beta_a$ - transmission probability	Constant across all age groups
$M$ - Non-UK born mixing	Variable in young adults (15-29) Constant across all age groups Variable in young adults (15-29)

scenarios aim to test some of the key modelling assumptions made here. Identifying if the transmission probablity, or non-UK born mixing, 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.

## 8.8 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 model that met these requirements and made use of the data available. I defined prior distributions for each model parameter and initialisation conditions. I then 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.

This chapter has outlined a realistic dynamic transmission model of TB that includes the key features required to investigate BCG vaccination policy and is robustly parameterised from an extensive, and previously unused in a TB model, routine surveillance dataset. 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. 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.[47] 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. This has not been possible as there is little data from which to extrapolate either the number of high risk

notifications in the ETS,[21] 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.[2] However, drug resistant TB cases are known to make up a small fraction of TB cases in England in the UK born, and 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 modelled 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 dataset; 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. Manissero 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. [103,104] 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.[105] 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.[72] 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.[100,106,122] 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.[100] 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.[101] 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.[102] 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.[122] Bhunu et al. developed a dynamic transmission model of TB that in order to investigate the effects of pre- and post-exposure vaccines for TB control. Again their model did not include multiple latent compartments or age structure unlike the model presented here.[106]

Vynnycky et al. modelled 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.[99] 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 modelled 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 may be 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,[118] 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,[101] 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 modelled 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 assumption 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 where 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 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. Both of these scenarios are explored in the final chapter of this thesis.

## 8.9 Summary

- This chapter presents a 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. Code for this model is available online.<sup>4</sup>
- The model has been robustly parameterised to a rich routine surveillance data set, which has allowed more complex features to be modelled 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

---

<sup>4</sup>Model code: <https://github.com/seabbs/ModelTBBCGEngland/blob/master/inst/bi/BaseLineModel.bi>

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 it's structure and parameterisation. The majority of this uncertainty relates to the amount of TB transmission occurring in England. In order to use this model to understand TB transmission, and the impact of different BCG vaccination policies (Chapter 10) this uncertainty needs to be reduced and the parameter space tightened to reflect more realistic ranges. 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 quantitative way “better” than other parameter sets. An alternative to model fitting is using a model parameterised with expert knowledge only for inference. This approach is not appropriate here due to the large amount of uncertainty for many of the model parameters. Any inference based on just the parametrisation from the previous chapter would have large credible intervals, reflect reality poorly, and likely be biased in multiple areas.

This chapter details an approach to fitting 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, the parameters that are fitted, and the parameters that are modelled stochastically. 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, ad hoc techniques used to improve 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 ob-

served continuous/discrete time process.[123] 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. I specify such an observational model in Section 9.2.2.

### **9.2.1 Observed data**

The primary data source for the model is the reported, age-stratified, UK born TB notifications from 2000 to 2004 as recorded in the Enhanced TB Surveillance (ETS) system (see Chapter 4). 2000 to 2004 are the years for which notifications are stratified by UK birth status with universal school-age BCG vaccination. Data were grouped using the age groups present in the dynamic model (5 year age groups up to 49 and then a group from 50-69 and a 70+ group). Using age-stratified incidence data, versus aggregated data, allows for more complex trends to be identified. Non-UK born TB notifications were extracted from the same source for use as an input to the models force of infection (Chapter 8).

Additional datasets were considered during initial model fitting and during later model calibration. These were a condensed version of the age stratified data discussed above with a reduced number of age groups (children (aged 0-14 years old), adults (15-69 years old), and older adults (70-89 years old)) and a dataset of historic pulmonary TB notifications. The advantage of condensing age groups was that the number of notifications in each group increased. This reduced the impact of stochastic noise, making fitting the model easier as trends in the data are more consistent. Secondly, reducing the number of data points, whilst still capturing the important age dynamics, reduces the compute requirements of the model (see Section 9.3.1). As will be discussed in Section 9.3 this was a major consideration as the model fitting approach used was highly compute intensive. The downside of this approach is that potentially important information may be lost when data is condensed into fewer groups. Historic pulmonary TB notifications (including both UK born and non-UK born cases) from 1990, 1994, and 1998 were considered as using data from the decade prior to the time period of interest allows the long term trends to be fitted to. A subset of the available data was used as this limited the impact on the compute time of the fitting pipeline. These data were originally collected in the Statutory Notifications of Infectious Diseases (NOIDS) dataset with notifications from 1913 to 1999. These data were sourced from Public Health England,[2] and made available in R using `tbinenglanddataclean`<sup>1</sup>. Downsides of using these data are that reporting standards may have changed over time so a single measurement model may not be appropriate and non-UK born cases are included in these data making fitting to this data dependent on the number of non-UK born cases pre 2000 which are themselves estimated during model fitting.

### **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. I assumed that all

---

<sup>1</sup>Historic TB notification data via `tbinenglanddataclean`: <https://www.samabbott.co.uk/tbinenglanddataclean/>

reporting errors are Gaussian and that there are no time variable reporting errors. This model was used for all data fitted to. The reporting model can be defined as follows,

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

Where  $O$  are the observed notifications,  $A$  are the incident cases of disease as forecast by the disease transmission model,  $E_{\text{syst}}$  is the systematic reporting error,  $E_{\text{noise}}$  is the reporting noise, and  $\mathcal{N}$  represents the Gaussian (normal) distribution. The priors for the model are defined in Table 9.1. The prior for systematic reporting error is based on the assumption that underreporting is more likely than over-reporting. The prior for the reporting noise is based on the observed variation between years. This observation model is also used when incorporating non-UK born incidence rates into the models force of infection (Chapter 8). This allows the uncertainty in these observations to be properly accounted for in the incidence estimates produced by the fitted model.

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. A potential improvement to this model would be to introduce separate systematic reporting errors for notifications pre- and post- the introduction of the ETS. However, this may result in over-fitting and so has not been implemented here.

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
$E_{\text{syst}}$	Systematic reporting error of incident TB cases	$\mathcal{N}(0.9, 0.05)$ truncated to be greater than 0.8 and lower than 1.	Proportion	Assumption is that underreporting of TB cases is likely with no overreporting.	Assumption
$E_{\text{noise}}$	Magnitude of reporting noise for incidence TB cases.	$\mathcal{U}(0, 0.025)$ .	Proportion	It is likely that reporting accuracy varies each year. An upper bound of 2.5	

### 9.2.3 Fitted parameters

The model outlined in 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 practice 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 transmission related information in the observed data may be lost due to over-fitting from other variables.

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 match their prior distributions. Parameters that have updated posterior distributions based on the data are,

- Mixing rate between UK born and non-UK born ( $M$ ).
- Scaling on non-UK born cases ( $\iota_{\text{scale}}$ ).
- Effective contact rate ( $c_{\text{eff}}$ ).
- Historic effective contact rate ( $c_{\text{eff}}^{\text{hist}}$ ).
- Half life of the effective contact rate ( $c_{\text{half}}^{\text{hist}}$ ).
- Low risk latent activation rate modifier for older adults (70+) ( $\epsilon_L^{\text{older-adult}}$ ).
- Systematic reporting error ( $E_{\text{syst}}$ ).
- Reporting noise ( $E_{\text{noise}}$ ).

In addition for scenarios with age variable transmission probabilities or non-UK born mixing the following parameters may also be fitted to,

- Transmission probability modifier for young adults ( $\beta_{\text{young adult}}$ ).
- Non-UK born mixing modifier for young adults ( $M_{\text{young adult}}$ ).

### 9.2.4 Stochastic parameters

Several key data inputs such as incidence in the non-UK born population, coverage of the BCG vaccination program, births, deaths and the contact rate are not perfectly observed, or recorded, and may vary across time. For these reasons, these parameters are included in the model developed in the last chapter as noise terms. This means that they are resampled for each timestep and so vary stochastically over time. This results in a model that is semi-stochastic rather than being fully deterministic (Chapter 1). A semi-stochastic model can be defined as a deterministic model that incorporates stochastic elements but that is still solved as a deterministic system within a given timestep. It is a modelling approach that has been used previously in the literature when key parameters are uncertain and potentially time varying.[9] For further details of the stochastic parameters included in the model see Chapter 8.

## 9.3 Model fitting pipeline

Fitting dynamic transmission models is complex and requires the use of specialist statistical techniques. There are a variety of these tools available. Ranging from tried and tested to cutting edge. Historically many modellers have used maximum likelihood methods to fit deterministic 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 (or interval estimates).[123] Unfortunately, many of these methods also require tuning prior to use. This section outlines the theoretical justification, and implementation details, of an automated model fitting pipeline used to fit the previously detailed state space model.

**LibBi** was used for all model fitting.[123] **LibBi** is a software package for state-space modelling and Bayesian inference. It uses a domain specific language for model specification, which is then optimised and compiled to provide highly efficient model code. It focuses on full information model fitting approaches including: particle Markov chain Monte Carlo (PMCMC), and SMC-SMC methods for parameter estimation. All fitting algorithms are highly efficient and scalable across multiple CPUs or GPUs. The **rbi** and **rbi.helpers** packages were used to interface with **LibBi** from R.[124,125] **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 (using CPU voltage manipulation), with these compute requirements in mind<sup>2</sup>. Whilst a cluster was theoretically available, in practise the hardware available was limited, installing **LibBi** was challenging, and run times were constrained by fair access. All model fitting code is available on GitHub as an R package<sup>3</sup>.

### 9.3.1 The particle filter

In order to fit a model to data it is necessary 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.[123] 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 [123] for a more technical discussion of the bootstrap particle filter.

1. *Sampling*: 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 each observed data point, the particle filter is then advanced through a series of *propagation*, *weighting*, and *re-sampling* steps.

---

<sup>2</sup>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/>

<sup>3</sup>Model fitting pipeline R package: <https://github.com/seabbs/ModelTBBCGEngland>

- *Propagation:* For each particle the model is simulated, producing a forecast of the observed 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.
  - *Re-sampling:* The particle stock is restored to equal weights by re-sampling particles, with replacement, with the probability of each sample being drawn being proportional to its weight.
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.

### 9.3.2 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 *re-sampling* 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 *re-sampling* 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 may be used instead.[123] 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 Carlo - Sequential Monte Carlo (SMC-SMC).[123] This algorithm is used for all dynamic model fitting in this thesis.
  - *Re-sampling:* The parameter particle stock is restored to equal weights by re-sampling particles, with replacement, with the probability of each sample being drawn being proportional to its weight.
  - *Rejuvenation:* *Re-sampling* 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 diversified 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 re-sampling 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 minimize unnecessary rejuvenation an effective sample size threshold can be used. This only triggers rejuvenation when particle diversity has decreased below the target effective sample size threshold.

### Marginal Metropolis-Hastings

The Metropolis-Hastings step may be used as a model fitting approach in its 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.[123] 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,  $\theta$  is the current parameter set,  $\theta'$  is some proposed parameter set sampled from some proposal distribution  $q(\theta'|\theta)$ .[123] By construction, samples drawn using this rule are ergodic to the posterior distribution. This means that after convergence, samples drawn using this rule may be considered as samples from the posterior distribution.

#### 9.3.3 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 may be the most efficient 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 necessary 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.[125] 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 threshold, commonly 1.[125]

I have implemented this as a two step process for each fitted scenario. Firstly, I used the Nelder-Mead simplex method, via `LibBi`,[123] to find a parameter set that optimised the

maximum likelihood. I then initialised a 1000 step PMCMC chain with this parameter set, using 1024 particles in the particle filter. I then used `rbi.helpers`,[125] 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 distribution, using 250 samples per step and starting with 16 particles. I initially planned to repeat this process for multiple draws from the posterior distribution but this proved to be in-feasible 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. Additionally 1024 was specified as the maximum number of feasible particles to use in the particle filter.

#### Proposal calibration

When using an MCMC algorithm a proposal distribution is required to provide new parameter samples to evaluate. For SMC-SMC a proposal distribution is required to inform the MCMC sampler that is run during the rejuvenation step. By default if no proposal distribution is provided `LibBi` uses the prior distribution.[123] The prior distribution can be an inefficient proposal distribution as it is likely to have a low acceptance rate (from the MCMC sampler).[123] 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 examining how to optimise the proposal distribution to achieve an good acceptance rate. In practice it has been shown that a rate of between 10% and 20% is optimal for upwards of 5 parameters.[123] 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 acceptance rate. Based on the acceptance 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,[125] 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). 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 Gaussian distributions for each parameter, truncated by the range of the prior, with the mean based on the current parameter value. The standard deviation for each parameter was assumed to be the standard deviation of the prior if it was Gaussian and otherwise assumed to be the range of the prior if it was uniform. For each iteration I halved the size of the standard

deviation of each parameter. 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 1024 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 methods 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 difficult to debug. Finally, for SMC-SMC efficient exploration of the proposal distribution is less important than when using MCMC alone as SMC-SMC 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 rejuvenation 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.

### 9.3.4 Model comparison

In the previous chapter multiple scenarios were outlined, each of which could be valid based on theoretical considerations. The observed data can be used to identify which of these scenarios 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.[126]

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 over-fitting.[126] 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  $\theta$ , with  $\theta$  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 complex models. It can be estimated as follows,[126]

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

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 it's 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 limitation 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 models.[126]

In this chapter I have used the DIC, as estimated by the DIC function from `rbi.helpers`,[125] to evaluate the various model structures outlined in the previous chapter.

### 9.3.5 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.[127] Multiple techniques exist that can globally study a multi-dimensional parameter space but the partial rank correlation coefficient method (PRCC) that I will discuss, and implement, here has been shown to be both reliable and efficient.[127]

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 parameter correlation or are just difficult to interpret with multiple dimensions.[127] 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 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,  $\sigma_X$  is the standard deviation of  $X$ , and  $\sigma_Y$  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 whilst it can capture non-linear relationships between outputs and inputs these relationships must be monotonic.[127] For relationships that are non-monotonic methods that rely on the decomposition of model output variance, such as the extended Fourier amplitude sensitivity test,[127] are more appropriate. However, these

approaches are computationally demanding as they typically require multiple 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 algorithm. This means that they cannot be used for “free” (i.e with negligible 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<sup>4</sup>,[60] using the samples from the posterior distribution of the model calculated during the SMC-SMC step. Parameter sensitivity measures such as PRCC must be calculated separately for each output time point. I calculated the PRCC for each fitted parameter, at the final time point with fitted data (2004), for overall TB incidence rates. These results are then summarised by plotting the absolute PRCC values, indicating the direction of correlation using colour<sup>5</sup>.

### 9.3.6 Pipeline overview

The full model fitting pipeline can be summarised as follows<sup>6</sup>:

1. Model initialisation using minimal error checking and single precision computation. Implemented using the `disable-assert` and `enable-single` flags in `LibBi`.[123] Outputs are only given for times with observed data and a subset of parameters are recorded for final reporting<sup>7</sup>.
2. 1000 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`.[123] This approximates the mean of the posterior distribution.
4. 1000 PMCMC steps, with 1024 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.[125] A minimum of 64 particles and a maximum of 1024 particles are assessed 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. It is assumed that the proposal for each parameter is Gaussian, truncated by the range of the prior with the mean based on the current parameter value. The standard deviation 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.

---

<sup>4</sup>Sensitivity code: [https://github.com/seabbs/ModelTBBCGEngland/blob/master/R/test\\_sensitivity.R](https://github.com/seabbs/ModelTBBCGEngland/blob/master/R/test_sensitivity.R)

<sup>5</sup>Sensitivity plotting code: [https://github.com/seabbs/ModelTBBCGEngland/blob/master/R/plot\\_sensitivity.R](https://github.com/seabbs/ModelTBBCGEngland/blob/master/R/plot_sensitivity.R)

<sup>6</sup>Model fitting code: [https://github.com/seabbs/ModelTBBCGEngland/blob/master/R/fit\\_model.R](https://github.com/seabbs/ModelTBBCGEngland/blob/master/R/fit_model.R)

<sup>7</sup>Model code: <https://github.com/seabbs/ModelTBBCGEngland/blob/master/inst/bi/BaseLineModel.bi>

7. SMC-SMC model fitting with 1000 initial parameter particles. Particle rejuvenation was set to trigger when the effective sample size decreased below 25%, with 10 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`.[125] 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 coefficient (PRCC) for each model parameter, for the final time point fitted to (2004), for overall TB incidence rates. 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.[60]

## 9.4 Results

The pipeline outlined above resulted in a poor fit to the observed data. The SMC-SMC algorithm had a low effective sample size for each iteration, and a low acceptance rate for particle rejuvenation steps in all scenarios evaluated. This resulted in spuriously tight posterior distributions. Ad hoc calibration (detailed in the following section) failed to improve the quality of this fit or find a subset of the model - or parameters - that fit the observed data to an acceptable degree whilst remaining computationally feasible. All results presented in the following section are based on posterior distributions produced by the model fitting pipeline using the prior distributions specified in the previous chapter. The results are preliminary and indicative only.

### 9.4.1 Ad hoc calibration

Minor alterations to the model fitting pipeline had little impact on the quality of the fit. To attempt to improve the quality of the model fit I used a combination of ad hoc approaches. As a first step I introduced a calibration model with variation allowed only in the fitted parameters with all other parameters using point estimates. This allowed a reduced number of particles to be used to estimate the marginal likelihood and hence dramatically decreased compute cost and run-times. This reduced model was then used for the following tests:

- Increasing the number of particles used in the outer SMC loop at the expense of reducing the number of particles used for marginal likelihood estimation.
- Increasing the number of particles used for marginal likelihood estimation.
- Sequentially decreasing model complexity by fixing fitted parameters to manually tuned point estimates.
- Increasing the number of parameters fitted to rather than used as fixed distributions.
- Varying the size of the proposal distribution, rejuvenation threshold, number of particles and rejuvenation steps.
- Varying the fitted observed data. This took 3 main forms:
  - Aggregation: fitting to overall incidence only; fitting to incidence grouped by large age groups (i.e children, adults, older adults); fitting to only age groups of interest (i.e children).

- Reducing the time-span of the fitted data. This included simplifying down to a single year of data but also included using various combinations of time points.
- Exploring additional observed data. This included attempting to fit to observed pulmonary TB case from 1980 on-wards. This approach sought to constrain the parameter space to give a more realistic age distribution of cases.
- Changing the functional form for the decay in the historic contact rate. This involved exploring linear decay with the decay gradient dictated by the year that the current contact rate takes effect and the year that the historic contact rate began to decay.
- Exploring using time dependent modifiers on the transmission probability and non-UK born mixing.
- Exploring using modifiers for children, adults, and older adults for both the transmission probability and non-UK born mixing. This was essentially an extension of the original scenarios considered using the model fitting pipeline.
- Exploring widening and narrowing the prior distributions of fitted parameters beyond realistic ranges.
- Exploring varying the size of the initial high, and low risk latent populations. This included starting with no latent cases, starting with a reduced proportion of latent cases and starting with a much larger latent population to simulate a historically more widespread disease.

None of these approaches dramatically improved model fit to the point that more robust inference could be drawn. Reducing the number of parameters, and time points, fitted to decreased the computational cost of the pipeline and improved acceptance rates. However, model fits remained poor until a single time-point and aggregated incidence were fitted to using a single varying parameter (effective contact rate) with all others manually tuned. Unfortunately, this simplified the model to the extent that it could not be used to generate meaningful results. The introduction of multiple time-points led to poor model fits regardless of the observational data used. This effect may be attributed to particle degradation but was not resolved by the addition of more particles in either the marginal likelihood estimation or the outer SMC step.[123] The use of pre-ETS pulmonary TB data worsened the quality of the model fit. This may be attributed to the data including UK born cases and therefore making the model fit more sensitive to the assumption used for the number of non-UK born notifications prior to 2000. Using manual prior tuning, transmission and mixing modifiers allowed a relatively close fit to the observed data but additional parameters, beyond those specified in the model fitting pipeline did little to improve on this. The quality of the model fit using the model fitting pipeline was poor regardless of the number of modifier parameters used. Varying the initial latent populations resulted in higher than previously estimated historic effective contacts but again did little to improve the quality of the model fit.

Murray et al. suggests that increasing the number of particles used in SMC may improve the quality of the model fit.[123] Unfortunately as LibBi stores SMC particle paths in Random access memory (RAM) the number of particles was restricted. An additional limitation is that the current rejuvenation step also need to be stored in RAM. Attempts to increase the amount of available RAM (64 GB) using a 500GB SWAP (virtual memory) drive increased the upper limit on the number of particles but gains from this were restricted due to thrashing<sup>8</sup>. Thrashing occurs when too much data is written to SWAP memory in a

---

<sup>8</sup>Thrashing: [https://en.wikipedia.org/wiki/Thrashing\\_\(computer\\_science\)](https://en.wikipedia.org/wiki/Thrashing_(computer_science))

short period of time and usually results in a system crash. A major cause of the high RAM requirements of the model was that `LibBi` stores all parameters defined as initial conditions across every time point in the model, regardless of the settings used. Attempting to reduce the RAM footprint by fitting a greater number of parameters allowed for a greater number of particles to be used but also increased the model degrees of freedom and hence upped the required number of particles required by a greater amount than provided by fitting the parameter. Reducing the number of particles used to estimate the marginal likelihood allowed a greater number of particles to be used in the outer SMC step but this resulted in highly inaccurate estimates of the likelihood. These highly inaccurate likelihood estimates resulted in the SMC-SMC algorithm focusing on parameter sets that had a low likelihood estimate yet fit the data poorly - ultimately leading to poorer model fits.

Varying the proposal size, rejuvenation threshold, and number of rejuvenation steps showed some promise at improving the quality of the model fit but ultimately computational constraints limited how much progress could be made using this approach. It is possible that a much longer run time could result in an improved fit to data with no other model changes. Alternatively these results may be driven by the model being too complex for an SMC-SMC approach to be viable.

### 9.4.2 Particle and proposal calibration

After development of the model fitting pipeline but before results could be produced both the `optimise` (from `rbi`) and `adapt_particle` (from `rbi.helpers`) began to error with multiple, transient, `LibBi` error messages. This meant that steps 3-5 of the model fitting pipeline could not be used with the final model. As a work around the maximum permitted number of particles (1024) was used for all model fitting (increasing the number of particles was also explored as detailed in the previous section).

As discussed, there were multiple issues with fitting the final model and this made it difficult to determine what the mean of posterior distribution was. This made manually tuning the proposal distribution challenging and so instead a standard deviation of 1% was used. This value was chosen as it increased the acceptance rate by limiting each rejuvenation step to a relatively small subset of the prior distribution whilst not preventing the exploration of new parameter space in scenarios when model fits from the initial particle sample were poor.

### 9.4.3 Model comparison

Whilst none of the scenarios fitted the data well, scenarios that included variable transmission probability in young adults fitted the data much better than those that did not (Table 9.2). When considered on its own, allowing non-UK born mixing to vary for young adults had only a small impact on the quality of the model fit in comparison to allowing the transmission probability to vary. However, the scenario that allowed both the transmission probability and non-UK born mixing to vary fit the data much better than any other scenario considered (Table 9.2). In the following sections only the results from this scenario will be discussed.

Table 9.2: DIC values for each scenario evaluated during model fitting - arranged from best to worst model fit. Using 1000 samples from the posterior distribution of the fitted model for each scenario.

Scenario	DIC
Transmission variable in young adults (15-29) and non-UK born mixing variable in young adults (15-29)	4559
Transmission variable in young adults (15-29)	6170
Non-UK born mixing variable in young adults (15-29)	20464
Baseline	20693

#### 9.4.4 Model Fit to TB incidence from the ETS

The fitted model consistently under-predicted overall TB cases for all years with data (Table 9.3). It also failed to capture the overall trend in TB incidence with the forecast incidence increasing year-on-year in comparison to the observed data which showed greater variation. Stratifying by age shows that the model also failed to captured the age distribution of TB incidence (Figure 9.1). Whilst the model under predicted TB incidence in all age-groups it was more accurate for older adults, implying that even if the magnitude of cases had been better predicted the distribution would still not match the observed data.

Table 9.3: Observed versus predicted overall TB cases for years that the model was fitted to. (95% CrI): 95% credible interval estimated using the 2.5% quantile and the 97.5% quantile. Using 1000 samples from the posterior distribution of the fitted model for the scenario with variability in both transmission and non-UK born mixing.

Year	Observed Cases	Predicted Cases (95% CrI)
2000	1803	716 (270, 1427)
2001	1866	713 (272, 1397)
2002	1833	724 (283, 1373)
2003	1685	742 (314, 1376)
2004	1776	747 (321, 1369)

## 9.4. Results

---

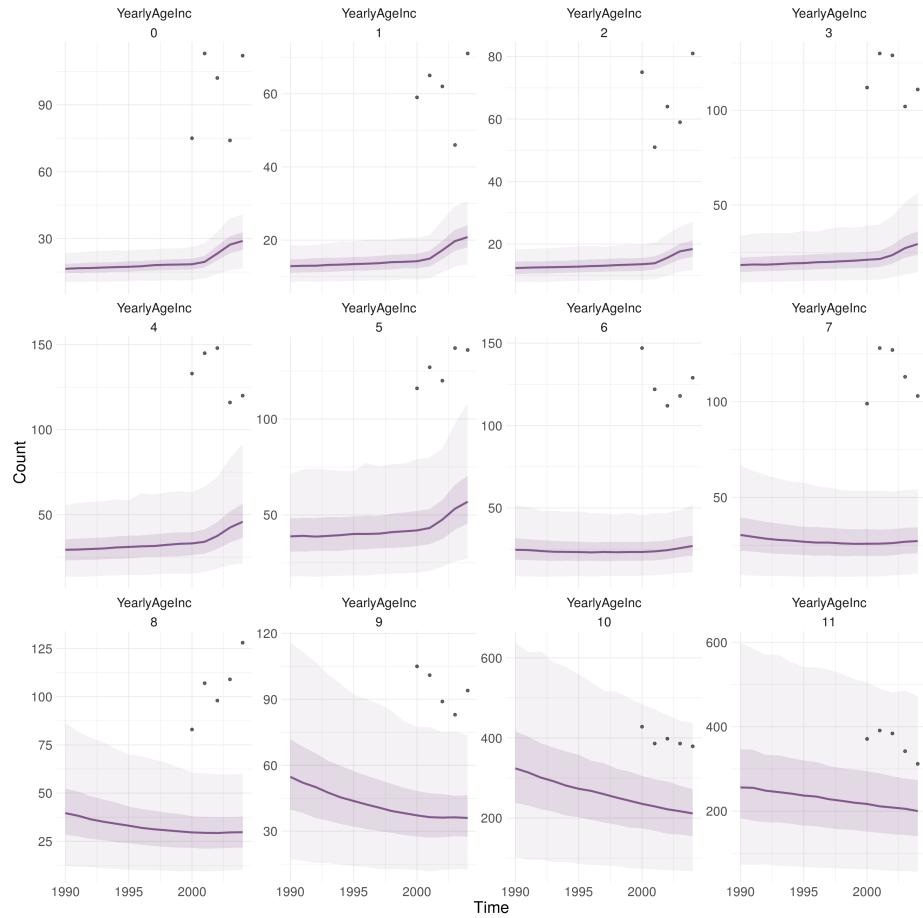


Figure 9.1: Observed and predicted annual TB incidence stratified by model age group (0-11). 0-9 refers to 5 year age groups from 0-4 years old to 45-49 years old. 10 refers to those aged between 50 and 69 and 11 refers to those aged 70+. The darker ribbon identifies the interquartile range, whilst the lighter ribbon indicates the 2.5% and 97.5% quantiles. The line represents the median. Using 1000 samples from the posterior distribution of the fitted model for the scenario with variability in both transmission and non-UK born mixing.

### 9.4.5 Posterior parameter distributions

The fitted model had a low effective sample size manifesting in spuriously tight posterior distributions (Figure 9.2; Table 9.4). This indicates that model fitting did not allow for a full exploration of the parameter space. Given that a relatively tight proposal distribution was used this may indicate that an independent proposal is insufficient. It is likely that the key factor behind the model's consistent under-prediction is the selection of a low value for the current effective contact rate and the historic effective contact rate. It is difficult to interpret these findings further due the low quality of the posterior distribution.

Table 9.4: Prior and posterior medians for fitted model parameters. (95% CrI): 95% credible interval estimated using the 2.5% quantile and the 97.5% quantile. Using 1000 samples from the posterior distribution of the fitted model for the scenario with variability in both transmission and non-UK born mixing.

Parameter	Prior (95%CrI)	Posterior (95%CrI)
$\beta_{\text{young adult}}$	5.03 (0.21, 9.68)	7.58 (7.58, 7.58)
$c_{\text{eff}}$	1.11 (0.08, 3.04)	0.53 (0.53, 0.53)
$c_{\text{eff}}^{\text{hist}}$	10.46 (1.47, 19.41)	1.86 (1.86, 1.86)
$c_{\text{half}}^{\text{hist}}$	6.06 (0.48, 15.41)	6.61 (6.61, 6.61)
$M$	1.19 (0.09, 3.02)	0.39 (0.39, 0.39)
$M_{\text{young adult}}$	5.41 (0.33, 9.77)	8.98 (8.98, 8.98)
$E_{\text{syst}}$	0.90 (0.82, 0.98)	0.83 (0.83, 0.83)
$E_{\text{noise}}$	0.01 (0.00, 0.02)	0.02 (0.02, 0.02)
$\iota_{\text{scale}}$	99.37 (5.55, 194.43)	80.13 (80.13, 80.13)
$\epsilon_L^{\text{older-adult}}$	2.01 (1.14, 2.99)	2.11 (2.11, 2.11)

## 9.4. Results

---

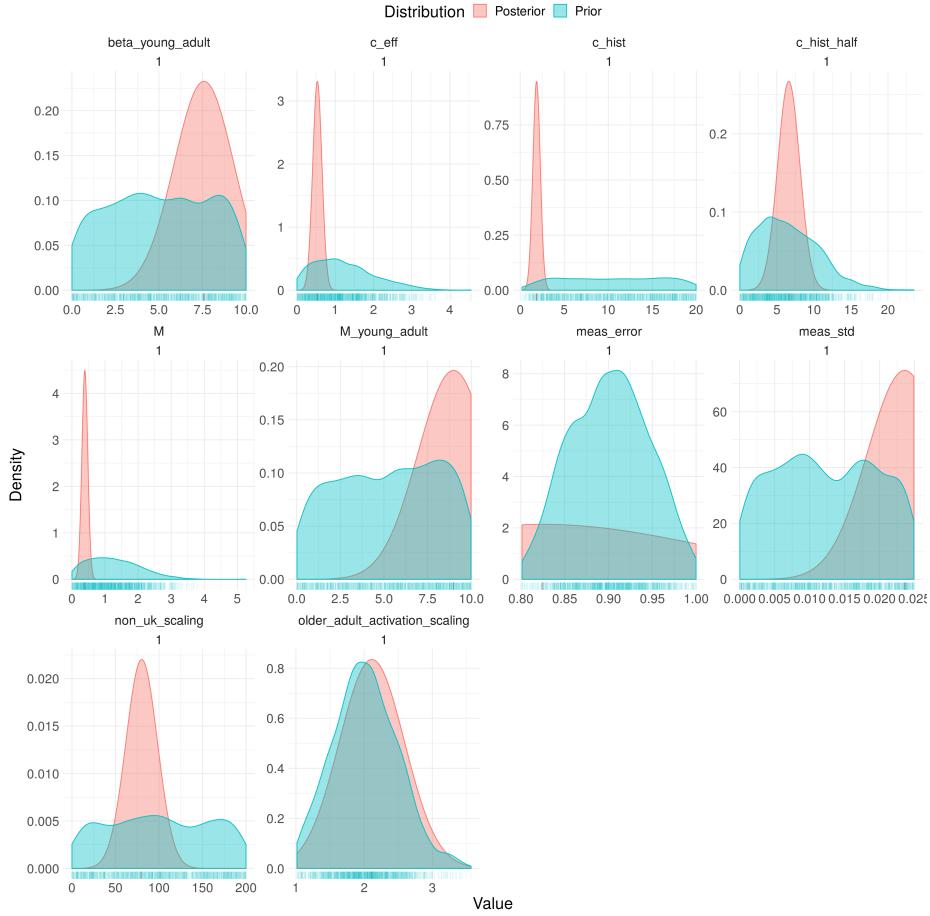


Figure 9.2: Prior and posterior distributions for fitted model parameters. Parameter names in figure are in their coded form. They can be interpreted as the following parameters:  $\beta_{\text{young adult}}$ ,  $c_{\text{eff}}$ ,  $c_{\text{hist}}^{\text{hist}}$ ,  $c_{\text{half}}^{\text{hist}}$ ,  $M$ ,  $M_{\text{young adult}}$ ,  $E_{\text{syst}}$ ,  $E_{\text{noise}}$ ,  $\iota_{\text{scale}}$ , and  $\epsilon_L^{\text{older-adult}}$ . Using 1000 samples from the posterior distribution of the fitted model for the scenario with variability in both transmission and non-UK born mixing.

### 9.4.6 Parameter Sensitivity

Figure 9.3 shows the partial rank correlation coefficients for each parameter that was fitted to. It indicates that variation in the effective contact rate and non-UK born mixing lead to the greatest variation in TB incidence. Based on the model structure this makes sense as these parameters are directly linked to modern day transmission. The parameters that modify TB transmission and non-UK born mixing in young adults also lead to significant variation in TB incidence. This corresponds to the findings from the scenario analysis discussed above in which the introduction of these parameters resulted in a greatly improved model fit. The lack of diversity in the posterior distribution seen above means that these findings are indicative only.

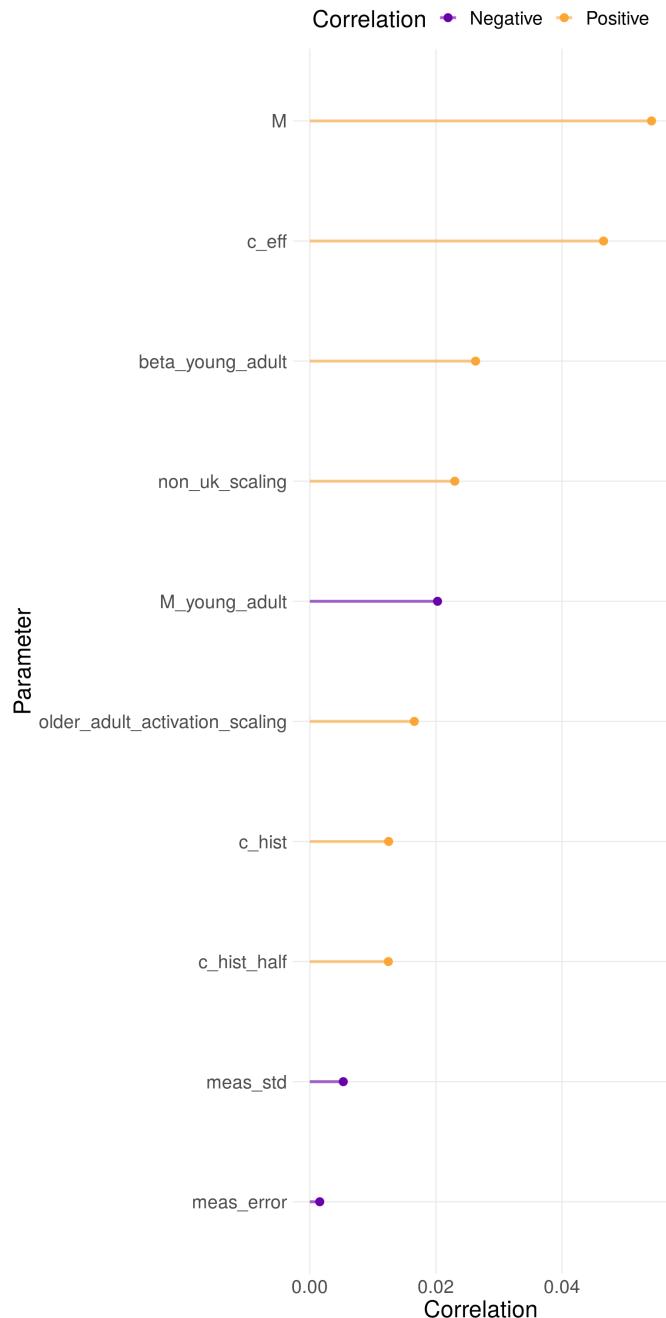


Figure 9.3: Partial rank correlation coefficients for each parameter fitted too. Parameter names in figure are in their coded form. They can be interpreted as the following parameters:  $\beta_{\text{young adult}}$ ,  $c_{\text{eff}}$ ,  $c_{\text{hist}}$ ,  $c_{\text{half}}$ ,  $M$ ,  $M_{\text{young adult}}$ ,  $E_{\text{syst}}$ ,  $E_{\text{noise}}$ ,  $\iota_{\text{scale}}$ , and  $\epsilon_L^{\text{older-adult}}$ . Using 1000 samples from the posterior distribution of the fitted model for the scenario with variability in both transmission and non-UK born mixing.

## 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, discussed the approaches taken to try and improve the quality of the model fit, and presented preliminary results from the model fitting pipeline. The model fit the observed data poorly using the approach laid out here. Whilst multiple ad hoc approaches were used to try and improve the quality of the model fit these did little to improve it. The model consistently under-predicted overall TB incidence and also failed to reproduce the observed age distribution. There was little evidence of a good fit to trends over time in the observed data. Model comparison showed that models that included modifying parameters for the transmission probability, and non-UK born mixing in young adults fitted the observed data better than those that did not. The estimated posterior distributions for all parameters were spuriously tight in comparison to the prior distributions. This may be evidence of poor mixing meaning that any conclusions drawn about the parameter sets selected may be incorrect. Parameters that contributed to recent TB transmission dominated those that contributed across the time-span of the model in the parameter sensitivity analysis.

As discussed earlier in this chapter, SMC-SMC was used for all model fitting. This uses an external SMC step to estimate the posterior distribution as well as an internal SMC step to estimate the marginal likelihood. In general, Bayesian model fitting approaches are beneficial as they allow prior information to be fully incorporated into model fitting and they produce posterior distribution estimates rather than point estimates.[123,128] There are two main families of approaches, SMC and MCMC. Using SMC to estimate the posterior distribution has numerous advantages over MCMC. The first of these is that MCMC approaches 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 convergence 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 re-sampling 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,[123] 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 comparison, each SMC parameter particle can have its marginal likelihood computed separately. Although the re-sampling step remains a bottleneck as it can only be completed once all marginal likelihoods have been computed. On the other hand SMC is less interactive than MCMC meaning that model fitting is harder to inspect when it is in progress. This is because SMC cannot be inspected sequentially, 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 rejuvenation steps required, and the expected total run time. SMC will also have a variable run time based on the effective sample

size as rejuvenation only happens when parameter particles have been depleted beyond a certain point. An additional benefit of SMC approaches are that they can theoretically be extended to model selection as well as parameter posterior estimation - effectively estimating posteriors for candidate model structures.[128] Beyond SMC and MCMC there are multiple other model fitting algorithms that each have their own strengths and benefits - discussion of these is outside the scope of this thesis.

The particle filter has been shown to provide an unbiased estimate of the likelihood for arbitrary state-space models.[123] As a full information technique the particle filter provides a more accurate estimate of the likelihood than other approximate techniques with relatively little tuning or user interaction.[123] The major downside of the particle filter 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. In addition, the bootstrap particle filter may become depleted (i.e variance between particles is reduce to such an extent that the effective sample size becomes small) when the model being fitted is a poor fit for the observed data.[123] Whilst this can be resolved using additional particles, or rejuvenation steps, this may not be computationally tractable. There are several alternative particle filters that seek to mitigate these issues but many of them are highly complex and do not significantly reduce the required compute - see [123] for a detailed discussion of some of these alternatives. An alternative approach to the particle filter is to use an approximate technique, such as approximate Bayesian computation (ABC). ABC can be used to avoid having to estimate the likelihood by comparing observed and simulated data.[128] This dramatically reduces the required compute as multiple model simulations per parameter set are no longer required. The comparison between the observed and simulated data is facilitated using a distance function with parameter sets being accepted if they are within some threshold distance. This threshold can be be tuned to produce a good estimate of the posterior distribution. Developing a distance function for all of the observed data can often be challenging so instead the distance is often calculated using a set of summary features.[128] In principle, ABC can be used with a wide variety of algorithms (including a simple rejection approach, MCMC, and SMC) to estimate the posterior distribution but in practice it has been shown that ABC-SMC generally performs better than alternative ABC approaches.[128] The two major limitations of ABC compared to the use of a particle filter is that in most cases summary statistics must be used rather than calculating the distance from the observed data and a function for calculating the distance must be chosen.[128,129] Whilst some techniques exist for evaluating summary statistics the choice is often subjective, relying on domain knowledge.[129] Chosen summary statistics rarely capture the information contained in the observed data fully and can inflate posterior distribution and in the worst cases introduce bias to the estimates.[130] The choice of distance function may also influence the estimates of the posterior distribution.[129]

This chapter showcased the use of `LibBi`, `rbi`, and `rbi.helpers` for fitting a complex transmission model. Unfortunately, the quality of the model fit was poor and the complexity of `LibBi` makes understanding the root cause difficult. It is possible that the model developed in the last chapter is too complex to be fitted using this approach - at least without the use of several orders of magnitude greater compute resources (or compute time) than available.[123] However, it is also possible that even with these resources this pipeline may not have produced a high quality model fit as the model developed in the previous chapter was clearly much more complex than those envisioned by the developers of the software.[123]

Alternatively, the model itself may not be identifiable with the available data.[129] `LibBi`, `rbi` and `rbi.helpers` have great potential as a standardised toolbox for modellers. However, in their current state they are difficult to use beyond relatively simple use cases. This difficulty is compounded by sporadic, inconsistent, documentation of both the underlying software and it's R interface. On top of this, both `LibBi` itself and the R libraries that support it have multiple apparent bugs that can be frustrating to debug due to the layered nature of the software. `rbi` and `rbi.helpers` are under development and it is likely that many of these issues will be dealt with over time. However, `LibBi` itself has not had a major release since 2016, the community around it is largely inactive, and it's complex nature makes it difficult for newcomers to contribute towards it's development. It is possible that the fitting issues outlined here may be due to errors in the code used to implement the model from the previous chapter. This is unlikely as manual prior tuning has shown that a relatively good fit to the observed data can be achieved with the current model but these results could not be reproduced using the model fitting pipeline. However, this does not mean that the modelling code is bug free. Model bugs, if present, would bias results from the fitted model rather than preventing a good fit to the data or would prevent the model from being manually tuned to fit the data.

There are many other tools available within the R ecosystem for fitting infectious disease models and detailing them all is beyond the scope of this thesis. However, `LibBi`, `rbi`, and `rbi.helpers` represent an attempt to provide a complete modelling framework rather than being a simple toolbox for model fitting. The `pomp` package has a similar aim making a comparison worthwhile.[10] `pomp` defines models using a similar structure to that presented here and used within `Libbi`. However, rather than using it's own modelling language it relies on the use of either R or C code. This has several advantages in that `pomp` models can more easily be generalized, can be understood and implemented by users new to the package and can make use of packages from the wider R and C ecosystem. The downside to this approach is that for complex models the use of C is essentially a requirement for efficiency reasons and implementing complex models in C can be an error prone and time consuming process. `pomp` offers support for PMCMC, iterated filtering, and ABC-MCMC but does not support SMC based approaches (such as SMC-SMC and ABC-SMC).[10] In developing the work presented here, and in the previous chapter, a simpler model was developed using `pomp`<sup>9</sup>. This work was not included in this thesis as it was not sufficiently developed. However, it highlight similar issues with the `pomp` package to those observed when using `LibBi`. Whilst `pomp`'s documentation, stability, and testing were much improved over `LibBi` it had similar limitations when it came too complex models. This was to such an extent that I developed numerous helper functions to deal with both the input and output of `pomp` models<sup>10</sup>. `pomp`'s documentation also has a heavy focus on iterated filtering over other model fitting techniques.

Whilst the results presented here are not encouraging it may be the case that with a greatly extended run-time a better fitting model may be found. The first step to testing this is to rerun the pipeline using several orders of magnitude more rejuvenation steps - as these have no RAM overhead. Even if the resulting model fits are still poor this may indicate areas for improvement - in either the model fitting pipeline or the model itself. Reducing model complexity may also help via decreasing the computational burden, allowing more

---

<sup>9</sup>pomp module code: <https://gist.github.com/seabbs/f08a8a46b1342b8649df963ac015ea31>

<sup>10</sup>idmodelr: <https://github.com/seabbs/idmodelr>

particles to be used, and also reducing the size of the potential parameter space. Possible simplifications are to remove features, such as the treatment population, that do little to alter the dynamics but are present due to use case concerns or switch from a continuous framework to a discrete one. An alternative option is to re-code the model developed in the last chapter into a more generic form (i.e C) and to attempt to fit the model using other techniques that are less compute intensive and more robust - such as SMC-ABC. Another option would be to re-implement the model in the form required by another modelling package - such as `pomp`. However, this would mean running the risk of again getting stuck within a framework that is difficult to debug and may not be working as expected. A final option would be to reduce the complexity of the model and potentially fit to less complex data. This may be the only solution if the current model is not identifiable.

The model fitting pipeline developed here was theoretically robust, and highly reproducible, but in practice did not produce a high quality model fit. It is difficult to determine whether this was caused by the complexity of the model combined with the high compute requirements of SMC-SMC or if the software implementation itself was at fault. However, `LibBi`'s high barrier to entry and difficulty of use made both implementing the model, and assessing whether the model fitting was working as expected more difficult. It is likely that a more generic model implementation coupled with a less compute intensive fitting approach would produce more useful results. This work does still have some merit as it pushed both `LibBi` and SMC-SMC to its limits helping to define what the limitations of this approach, and specialised modelling packages more generally, may be. It is possible that with an extended run-time model fits may become more reliable and hence more usable.

## 9.6 Summary

- Defined the disease transmission model from the previous chapter as a state-space model, outlined the available data to be used for fitting it, and detailed a measurement model to link the observed data with the dynamic TB model.
- Developed a model fitting pipeline based on SMC-SMC to fit the previously defined state-space model to TB notification data. The theoretical background for this approach was outlined as well as the key steps for implementing it in practise.
- As the quality of model fit achieved was poor ad hoc calibration approaches, and the results they gave, were discussed.
- The scenarios outlined in the previous chapter were evaluated using DIC and the implications of the findings were discussed.
- Model forecasts from the best fitting scenario, as established using the DIC, were compared to observed data. Posterior distributions from this model fit were then contrasted with the prior distributions. Finally, parameter sensitivity was estimated using the posterior distributions.
- Discussed the strengths and limitations of the work presented here, as well as outlining potential further work.

# Chapter 10

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

### 10.1 Introduction

In the previous chapter I outlined a model fitting pipeline and discussed the results from using it on the model developed in Chapter 8. Whilst this fitted model may be used to explore the epidemiology of tuberculosis (TB) in the early 2000's it does not - as currently stands - explore the impact of the 2005 change in BCG vaccination policy (Chapter 2). Models are useful in this context as dynamic model forecasts can be derived for multiple scenarios that may only exist on paper and so have little to no data to support them. These forecasts may then be used by policy makers as indicators of the likely impact of these scenarios.

This chapter details the approach used to extrapolate 1000 samples from the posterior distribution of the fitted model from the previous chapter (for the best fitting scenario with variability in both transmission and non-UK born mixing) beyond the change in BCG policy in 2005 and into the future. It first outlines the scenarios considered, then details the assumptions used to expand the time horizon of the model. Finally the impact of each scenario is explored over multiple time horizons. As discussed in the previous chapter these findings are preliminary in nature, meaning quantitative conclusions cannot be drawn and qualitative conclusions must be appropriately caveated.

### 10.2 Methods

#### 10.2.1 Scenarios considered

I considered three scenarios from 2005 on-wards. These were:

- Universal BCG vaccination of those at school-age continued with the same coverage

as previously.

- Universal BCG vaccination of those at school-age was phased out in 2005 and replaced with universal BCG vaccination of neonates with the same coverage levels as assumed for the BCG schools scheme.
- Universal BCG vaccination of those at school-age was phased out in 2005 (i.e no vaccination post 2004).

The BCG policy change in 2005 was from universal school-age BCG vaccination to targeted vaccination of high risk neonates. However, here universal vaccination of neonates is used as a proxy for targeted vaccination of high risk neonates. This was necessary because the high risk population was not modelled in the model developed in Chapter 8 due to the lack of data on which to base key assumptions. No vaccination was used as a baseline in order to explore the absolute impact of vaccination. Vaccination coverage was assumed to be constant across all scenarios as there was little data on which to base between assumption variation. Regardless of the scenario considered it was assumed that school-age vaccination was in place from 1953 through to 2004.

### **10.2.2 Forecasting assumptions**

Data on non-UK born cases, which were imported into the model via the force of infection (Chapter 8), were not available beyond 2015. To account for this an, age and year adjusted, Poisson regression model was used to forecast future TB incidence in the non-UK born with age treated as a categorical variable. As for years with data, uncertainty was introduced into these forecasts by assuming that non-UK born incidence rates were scaled using the fitted measurement error and normally distributed with a standard distribution based on the fitted measurement standard error (Chapter 9).

As outlined in Chapter 8, births from 2015 on-wards were based on projections from the Office for National Statistics (ONS). Age-specific mortality rates were estimated for 2016 on-wards using ONS estimates from 1981-2015, and an exponential model (Chapter 8). Both births and age-specific mortality rates were assumed to have a normal distribution with a standard deviation of 5% of the predicted value. It was assumed that all other parameters were unchanged from the values estimated for 2000-2004 (Chapter 9).

### **10.2.3 Analytical methods**

Estimated age-stratified, and aggregated, TB incidence, and mortality were compared both visually and numerically from 2005 through to 2040 for each scenario. Multiple time horizons were evaluated across this timespan as initially the impact of any policy change may be masked by the large reservoir of vaccinated individuals in the population and because of the impact of the assumed decrease in non-UK born cases over time.

## **10.3 Results**

All results presented in the following section are based on 1000 samples from the posterior distribution of the fitted model from the previous chapter (for the best fitting scenario with variability in both transmission and non-UK born mixing) extended beyond the change in BCG policy in 2005 using the assumptions detailed in the previous section. These results should be considered preliminary because of the low quality of fit achieved in Chapter 9.

## 10.3. Results

---

This means that quantitative estimates are unlikely to be accurate. However, the underlying changes in dynamics caused by vaccination may be used for insight into the likely impact of each scenario and therefore there is still some value in exploring these results.

### 10.3.1 Forecasting the long-term impact of each vaccination scenario.

Continuing with school-age vaccination resulted in the fewest number of cases regardless of the time-span considered (Table 10.1). However, the difference between vaccination scenarios was consistently small when compared to the overall number of cases. In all scenarios TB incidence was forecast to decrease over time in line with the decreases assumed in non-UK born TB incidence. The lower bounds for each scenario were relatively comparable, with the upper bounds being higher for both neonatal BCG vaccination and no BCG vaccination when compared to school-age BCG vaccination.

Table 10.1: Forecast of overall TB incidence for each scenario evaluated from 2005 to 2040. For brevity only 5 year intervals are shown. (95% CrI): 95% credible interval estimated using the 2.5% quantile and the 97.5% quantile. Using 1000 samples from the posterior distribution of the fitted model for the scenario with variability in both transmission and non-UK born mixing.

Year	School-age BCG (95% CrI)	Neonatal BCG (95% CrI)	No BCG (95% CrI)
2005	766 (347, 1419)	757 (321, 1420)	757 (326, 1552)
2010	735 (344, 1369)	856 (376, 1516)	908 (415, 1729)
2015	649 (283, 1234)	757 (311, 1379)	800 (343, 1547)
2020	594 (239, 1124)	691 (277, 1260)	737 (297, 1425)
2025	537 (203, 995)	614 (236, 1144)	659 (250, 1268)
2030	488 (178, 925)	554 (198, 1029)	594 (213, 1171)
2035	442 (155, 845)	501 (157, 926)	538 (177, 1034)
2040	403 (135, 775)	453 (134, 849)	486 (155, 961)

As expected neonatal vaccination resulted in a rapid decline in TB incidence in 0-4 year olds and a smaller but still large reduction in 5-9 year olds (Figure 10.1). There was a slight reduction in 10-15 year olds. School-age vaccination resulted in lower incidence in all adult populations when compared to any other scenario, except in adults over 45 where all scenarios were comparable. Neonatal vaccination resulted in a slight decrease in TB incidence rates when compared to no vaccination in young adults.

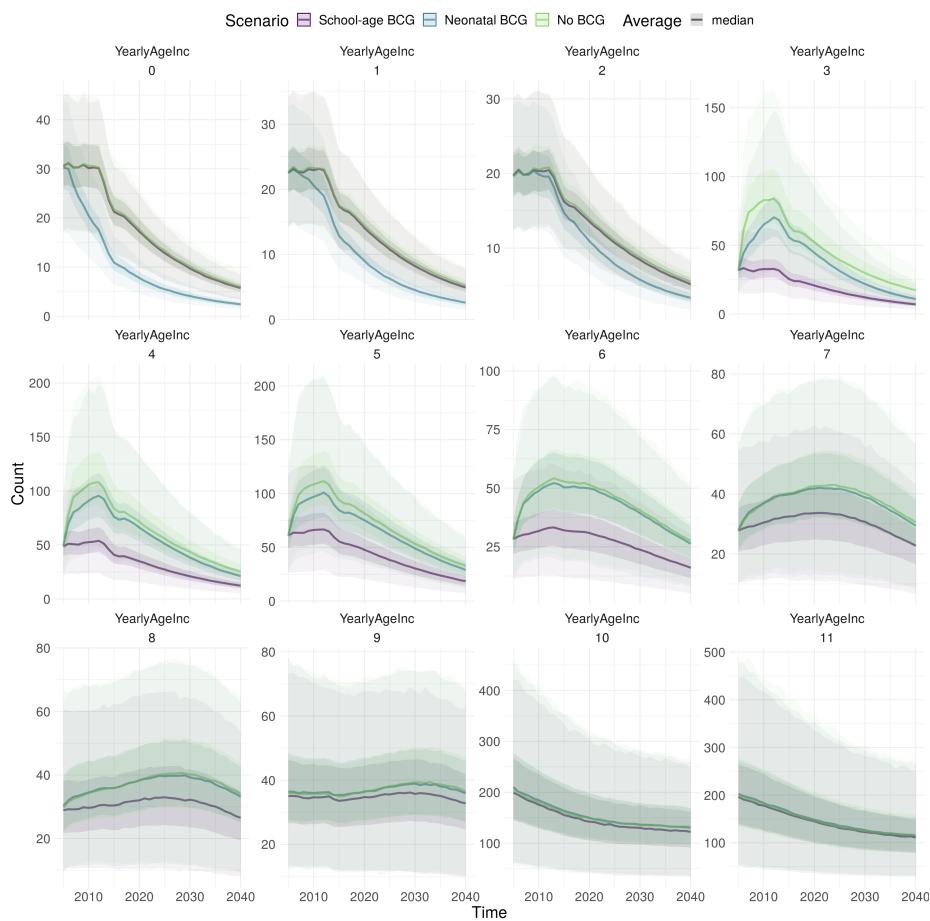


Figure 10.1: Forecast of TB incidence for each scenario evaluated from 2005 to 2040, stratified by age group (0-11). 0-9 refers to 5 year age groups from 0-4 years old to 45-49 years old. 10 refers to those aged between 50 and 69 and 11 refers to those aged 70+. Scenarios are differentiated by colour. The darker ribbon for each colour identifies the interquartile range, whilst the lighter ribbon indicates the 2.5% and 97.5% quantiles. The line represents the median. Using 1000 samples from the posterior distribution of the fitted model for the scenario with variability in both transmission and non-UK born mixing.

### 10.3. Results

---

As estimates of TB mortality were low (significantly lower than estimated using the observed data) the impact of any vaccination scenario was minimal (Table 10.2; Figure 10.2). Continuing school-age BCG vaccination resulted in a very small reduction in TB mortality compared to any other scenario (Table 10.2). The age distributed impact of each scenario on TB mortality was comparable to that observed for TB incidence (Figure 10.2).

Table 10.2: Forecast of overall TB mortality for each scenario evaluated from 2005 to 2040. For brevity only 5 year intervals are shown.(95% CrI): 95% credible interval estimated using the 2.5% quantile and the 97.5% quantile. Using 1000 samples from the posterior distribution of the fitted model for the scenario with variability in both transmission and non-UK born mixing.

Year	School-age BCG (95% CrI)	Neonatal BCG (95% CrI)	No BCG (95% CrI)
2005	51 (16, 119)	51 (14, 128)	50 (15, 136)
2010	48 (15, 110)	52 (16, 130)	53 (17, 146)
2015	43 (13, 101)	47 (13, 119)	48 (15, 133)
2020	39 (11, 92)	43 (12, 107)	43 (13, 120)
2025	35 (10, 84)	39 (10, 96)	39 (11, 108)
2030	32 (9, 79)	35 (9, 89)	36 (9, 99)
2035	30 (8, 74)	33 (8, 82)	33 (8, 90)
2040	28 (8, 70)	31 (8, 76)	31 (7, 83)

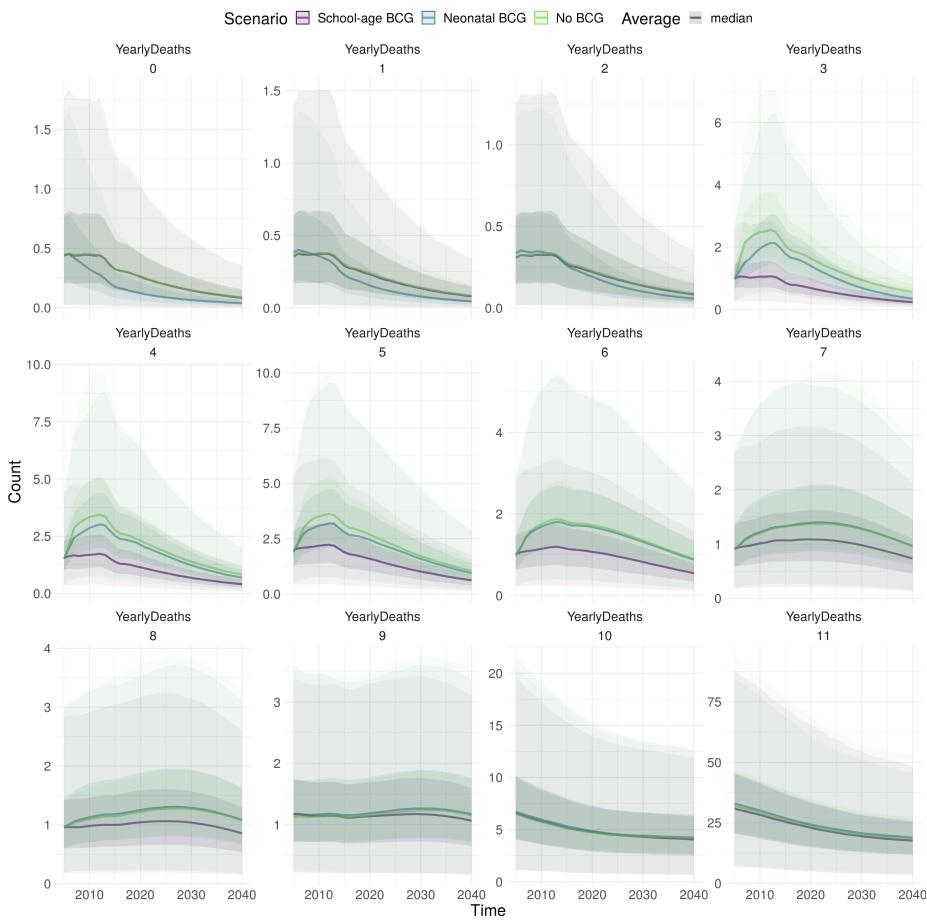


Figure 10.2: Forecast of TB mortality for each scenario evaluated from 2005 to 2040, stratified by age group (0-11). 0-9 refers to 5 year age groups from 0-4 years old to 45-49 years old. 10 refers to those aged between 50 and 69 and 11 refers to those aged 70+. Scenarios are differentiated by colour. The darker ribbon for each colour identifies the interquartile range, whilst the lighter ribbon indicates the 2.5% and 97.5% quantiles. The line represents the median. Using 1000 samples from the posterior distribution of the fitted model for the scenario with variability in both transmission and non-UK born mixing.

## 10.4 Discussion

In this chapter I outlined 3 vaccination scenarios to explore: continuing school-age BCG vaccination; universal neonatal vaccination; and no further vaccination. For each scenario I forecast TB incidence and TB mortality from 2005 through to 2040 assuming that non-UK born incidence rates would follow the same age stratified trends observed between 2010 and 2015. Although the results presented here are only preliminary, due to the quality of the model fit, it appears that continuing school-age BCG vaccination would have resulted in slightly reduced TB incidence across all time-points considered compared to any other scenario. Neonatal BCG vaccination resulted in reduced TB incidence and mortality in young children but had little impact later in life with a comparable effect to no vaccination. School-age vaccination had no impact on young children but did reduce TB incidence and mortality for all adults up to 45 years old. Beyond 45 years old no scenario impacted TB incidence or mortality.

The model developed in Chapter 8 was motivated by existing theory and robustly parameterised to the available data. It represents the only (known) open-source model of TB transmission and BCG vaccination. However, the model fitting pipeline developed in Chapter 9 did not produce a good fit to the observed data. Ad-hoc model calibration (as discussed in Chapter 9) failed to significantly improve on this fit. This means that the findings presented in this chapter can be considered as indicative only. However, these findings still represent the only modelling study of TB dynamics after a large scale change in BCG vaccination policy. The lack of data to support modelling the high-risk TB population population meant that targeted vaccination of high-risk neonates could not be considered as a scenario. This means that the results presented in this chapter do not contain the vaccination policy that is currently in place and so findings from the model cannot be directly compared to observed incidence data. However, only considering scenarios that alter the age of those vaccinated, rather than both the age of those vaccinated and the targeting of the vaccine, make understanding the impact of changes in vaccination policy easier to determine. The forecasts presented in this chapter are highly sensitive to the forecasted number of non-UK born TB cases. Whilst the regression method outlined in this chapter extrapolates based on current age-stratified trends it may be the case that this extrapolation breaks down over the long - or short - term. To a lesser extent the forecasts presented here are sensitive to the projected number of births and mortality rates. This is particularly the case for incidence rates in neonates and in older adults.

To my knowledge, there are no other dynamic modelling studies evaluating the use of the BCG vaccine in low burden settings that include a comparable level of detail and that are robustly parameterised based on the latest evidence. Harris et al. recently reviewed mathematical models that explored the epidemiological impacts of future TB vaccines.[100] 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. These findings agree with those presented in this Chapter, with fewer overall cases observed when vaccination continued in those at school-age, compared to neonatal vaccination. However vaccination in neonates did lead to a decrease in incidence in children both over the long - and short term - in comparison to vaccination at school-age.

The results presented in this chapter generally agree with other findings from this thesis.

In Chapter 7, which estimated the impact of the change in vaccination policy in those directly impacted by it, there was some evidence that changing to neonatal vaccination was associated with a small increase in incidence rates in those who were school-age. There was less evidence of a reduction in incidence rates in UK born neonates. The first of these results matches the findings presented here. However, this chapter estimated a rapid reduction in TB incidence in young children which was not seen in the previous work. Chapter 5, which recreated a previously published transmission chain model estimated an initial impact from ending school-age BCG vaccination but that this impact would decline with time. The results presented here agree with this findings as long as it is assumed that non-UK born incidence will decrease over time. However, here the impact of the scheme was estimated to continue beyond the 15 year time horizon estimated in Chapter 5. In Chapter 6 I found some evidence that BCG vaccination may decrease all-cause mortality in TB cases and some evidence that indicated that this may be related to reduced TB mortality. If either of these associations were causal then they would increase the benefit of vaccination scenarios compared to no vaccination but would not alter the trade-offs between neonatal and school-age vaccination.

Further work is need to improve the fit of the model to observed data (Chapter 9). This will result in improved forecasts and more reliable results. In addition, other vaccination coverage scenarios could be considered that explored the impact of vaccinating a reduced proportion of both those at school-age and neonates. If additional data becomes available, or if the appropriate assumptions are used, the inclusion of the high-risk population into the model would allow the evaluation of targeted high-risk neonatal vaccination in comparison to the other scenarios considered here. The extrapolation of the trend in non-UK born cases is a limitation of this model and as such should be further explored using other assumptions such as constant non-UK born cases, incidence rates based on expert opinion and estimates based on other global modelling studies.

The results presented here indicate that changing from a school-age BCG vaccination programme to a neonatal BCG programme lead to an overall increase in TB incidence, with increases concentrated in the young adults, and to a lesser degree, in older adults. Neonatal vaccination led to a decrease in TB incidence in children both in the short - and long term. This indicates that direct vaccination provides the best protection for children rather than indirect protection via reduced transmission. This finding is likely to be dependent on the degree of background transmission and so further modelling studies are needed in diverse settings before conclusions can be generalised. No vaccination was shown to lead to increased incidence in all age-groups when compared to school-age vaccination and in children only when compared to neonatal vaccination. The impact of any vaccination programme on older adults was small. These results are preliminary in nature as the model on which they are based fitted poorly to the observed data. However, they do indicate some of the trade-offs involved in setting BCG policy. If reducing childhood incidence was a goal of the 2005 policy change then these results indicate a clear success. On the other hand if reducing overall TB incidence was the goal then stopping shool-age vaccination has not been a success.

## 10.5 Summary

- Continuing school-age vaccination results in lower overall incidence rates compared to both neonatal vaccination, and no vaccination.
- Neonatal vaccination resulted in low incidence in children compared to any other scenario.
- No vaccination led to higher incidence in all age groups when compared to school-age vaccination and in children only when compared to neonatal vaccination.
- The impact of any vaccination on cases in older adults (50+) was small.
- These results are indicative only due to poor quality of model fit achieved in the previous chapter (Chapter 9).



# **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, the results, and the strengths and weaknesses of the findings, for that chapter. Consequently, the aim of this chapter is to summarise and discuss the findings from this thesis as a whole.

### **11.1 Principal findings**

In Chapter 4, I explored TB epidemiology in England with a focus on BCG vaccination and data completeness. I found that there was some evidence that negative outcomes were more frequent in Tuberculosis (TB) cases not BCG vaccinated than in those that were. I also found that missingness in routine surveillance sources of TB data was associated with multiple risk factors. In Chapter 6, I used logistic regression to estimate associations between BCG vaccination and TB outcomes. I found supporting evidence that BCG vaccination was associated with reduced all-cause mortality 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 modelling evidence that was used by policy makers to assess the impact of ending school-age universal BCG vaccination. I found that the previous approach was methodologically flawed and had underestimated the amount of uncertainty surrounding the effect estimates. Using newly available data, I also found that ending universal school-age BCG vaccination 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 TB incidence rates in the target populations, using Poisson and negative binomial models. 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. Using this approach, I was unable to rule out an unrelated policy change as the cause of this reduction. Finally, in Chapter 10, I forecast the impact of various vaccination scenarios using a dynamic TB model that was developed and fitted in Chapter 8 and Chapter 9. I found that the BCG schools scheme was

projected to reduce UK born TB incidence compared to both neonatal and no vaccination over a range of time horizons. However, neonatal vaccination reduced incidence in children compared to any other scenario, although it had little impact in any other age group. No vaccination programme evaluated had an impact on incidence in older adults. The results from Chapter 10 are preliminary as the model on which they were based was a very poor fit to the observed data (Chapter 9).

## 11.2 Strengths and limitations

This thesis has used multiple methods, and data sources, to explore the impact of the 2005 change in BCG policy. This multi-method approach 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 a single surveillance dataset. 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. Datasets that would have added value include: notification data from other countries that also changed BCG policy, regional datasets with more reliable data on BCG status and year of BCG, and data on BCG coverage and eligibility. However, to my knowledge, no similarly thorough use of an equivalent data source exists for TB. Another limitation of the work in this thesis is the very poor fit of the dynamic model, developed in Chapter 8, to the observed data. This made drawing conclusions from it difficult. On the other hand, the model fitting presented in Chapter 9 represented one of the only open source examples of fitting a very complex model to data using robust bayesian, plug and play, approaches. The lack of success is itself a useful result from which much can be learned about the usability of the fitting tools and the upper limits on model complexity. 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 tools as a by-product of the main research question.

## 11.3 Implications for policy makers

This thesis has highlighted the trade-off between vaccinating those at school-age and neonates in a setting 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.[3] Future BCG policy should consider these factors. In addition, new TB vaccines are in development that may be less susceptible to waning effectiveness when given later in life over a greater geographic area.[22] The findings from this thesis may be applicable to these new vaccines in areas where the BCG vaccine is currently known to be ineffective when given later in life. 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 mortality in TB cases may add additional weight to the argument that wider vaccination maybe

### *11.3. Implications for policy makers*

---

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 that the impact of the BCG policy has varied depending on UK birth status. This may 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.

Policy makers in the UK in particular should consider the implications for BCG vaccination of the results from this thesis. Of particular importance is the finding that the previously published approach for estimating the impact of ending BCG vaccination in the UK was methodologically flawed, spuriously precise and, when updated with newly available data, produced much larger estimates of the impact of ending BCG vaccination than previously thought. This finding highlights the importance of independent validation of modelling studies by subject area experts. To counterbalance this, the finding that overall the change in BCG vaccination policy was associated with decreased incidence rates indicate that the change in policy may have been justified. However, the benefit of the policy change was focussed on the non-UK born population who may have been impacted by other policy changes. Considering UK born cases alone, this thesis provides evidence that TB cases increased, with cases decreasing in young children. The modelling evidence, although preliminary, supported the finding that the change in policy would lead to increased overall incidence rates in the UK born population but reduce incidence rates in young children compared to school-age vaccination. On top of these specific findings, the points made in the previous paragraph also apply to the UK.

Unfortunately, definite recommendations cannot be made to policy makers based on the findings from this thesis. This is due to several limitations of the evidence and due to the inherent complexity of both TB and the BCG vaccine. A particular limitation is the difficulty in identifying who - post the change in BCG vaccination policy - should have been vaccinated and who should not have been. This meant that targeted high-risk neonatal vaccination could not be modelled and therefore could not be compared to school-age BCG vaccination. It also meant that population-level studies had to be used when studying the impact of change in policy. As the impact of the targeted scheme was likely focussed on the groups it targeted this may have diluted the impact of this policy. An additional issue is that whilst a trade-off was identified between reducing the number of TB cases in young children at the cost of a larger number of adult TB cases the impact of TB symptoms on children compared to adults was not considered. To be able to make more precise recommendations additional data sources are required. The first of these is a measure of the impact of TB on individuals at various ages, such as a quality-adjusted life-year (QALY). This would ideally be arrived at by assessing expert opinion and could then be used to identify what level of trade-off between childhood and adult TB cases resulting in the lowest overall QALY burden. Finally, data is required on who should have been vaccinated under the targeted high-risk neonatal vaccination scheme and of these individuals how many of them then went on to develop TB. This data would allow more precise statistical estimates to be made of the overall impact of the change in policy and would also enable targeted high-risk neonatal vaccination to be included in the dynamic TB model. Inclusion in the dynamic model would then allow the impact of various targeted vaccination scenarios to be considered over an extended time-frame and compared to universal vaccination scenarios.

## 11.4 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<sup>1</sup>, with a formatted version available on my personal site<sup>2</sup>. This thesis relies on data from the Enhanced TB Surveillance system and the Labour Force Survey. The cleaning and munging of this data has been standardised as an R package, `tbinenglanddataclean`<sup>3</sup>, 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`<sup>4</sup>, 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 details of the open source projects developed as part of this thesis.

## 11.5 Public engagement

Public engagement has been a constant theme throughout my doctoral work. 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 to 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<sup>5</sup>. 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. 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<sup>6</sup>.

## 11.6 Future research

The finding that BCG vaccination may reduce mortality in TB cases from Chapter 6 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 did not fit the observed data well. Additional compute time is needed to diagnosis whether this is a limitation of the fitting

<sup>1</sup>Thesis GitHub: <https://github.com/seabbs/thesis>

<sup>2</sup>Thesis website: <https://www.samabbott.co.uk/thesis>

<sup>3</sup>`tbinenglanddataclean`: <https://www.samabbott.co.uk/tbinenglanddataclean/>

<sup>4</sup>`prettypublisher`: <https://github.com/seabbs/prettypublisher>

<sup>5</sup>See my personal site: <https://www.samabbott.co.uk>

<sup>6</sup>Available here: <https://github.com/seabbs/pebblegame>

technique used, if the model itself was a poor fit for the data, or if the model was overly complex. Additional strategies for exploring this issue are discussed in Chapter 9. In addition this model 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 model 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 active 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.

## 11.7 Conclusions

This thesis has provided new evidence regarding the use of BCG vaccination 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. Although this model was a poor fit to the observed data there was some evidence that continuing school-age vaccination would have resulted in fewer UK born TB cases but that neonatal vaccination reduced TB incidence in UK born children compared to continuing school-age vaccination. There was also some evidence that none of the vaccination strategies evaluated impacted incidence rates in older adults. 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 than at neonates. However, this depends on the potential duration of protection conferred by vaccination. In addition, they highlight the trade-off between neonatal and school-age vaccination with school-age vaccination reducing overall TB incidence but neonatal vaccination reducing incidence in young children where more severe outcomes are more common. 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 engagement, and as part of other research projects.



# References

- 1 Stone AC, Wilbur AK, Buikstra JE *et al.* Tuberculosis and leprosy in perspective. *American Journal of Physical Anthropology* 2009;140:66–94. doi:10.1002/ajpa.21185
- 2 Public Health England. Tuberculosis in England 2017 report ( presenting data to end of 2016 ) About Public Health England. 2017.
- 3 The World Health Organization. BCG vaccine:WHO position paper. *Weekly epidemiological record* 2018;1–24.
- 4 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.
- 5 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.
- 6 McElreath, Richard. *Statistical Rethinking*. 1st ed. Chapman; Hall/CRC 2018.
- 7 Anderson RM, May RM. Infectious Diseases of Humans: Dynamics and Control (Oxford Univ. Press, Oxford 1991.
- 8 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.
- 9 Funk S, Camacho A, Kucharski AJ *et al.* Real-time forecasting of infectious disease dynamics with a stochastic semi-mechanistic model. *Epidemics* 2016;22:56–61.
- 10 King AA, Nguyen D, Ionides EL. Statistical Inference for Partially Observed Markov Processes via the R Package pomp. *Journal Of Statistical Software* 2016;69:1–43.
- 11 Houben RMGJ, Dodd PJ. The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling. *PLoS medicine* 2016;13:e1002152–13.
- 12 Gideon HP, Flynn JL. Latent tuberculosis: What the host "sees"? *Immunologic Research* 2011;50:202–12.
- 13 Sepkowitz K. How contagious is tuberculosis? *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 1996;23:954–62.
- 14 Rottenberg ME, Pawlowski A, Jansson M *et al.* Tuberculosis and HIV Co-Infection.

- PLoS Pathogens* 2012;8:e1002464.
- 15 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.
  - 16 Narasimhan P, Wood J, Macintyre CR *et al.* Risk Factors for Tuberculosis. 2013;2013.
  - 17 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.
  - 18 World Health Organization. Global Tuberculosis Report. 2016.
  - 19 World Health Organisation. Global Tuberculosis Report. 2018.
  - 20 PHE. Tuberculosis in England 2016 Report (presenting data to end of 2015). *Public Health England* 2016;Version 1.:173.
  - 21 Public Health England. Tuberculosis in England: 2018. 2019;1–218.
  - 22 Medicine C. History of BCG Vaccine. 2013;8:53–8.
  - 23 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.
  - 24 Colditz GA, Brewer TF, Berkey CS *et al.* Efficacy of BCG Vaccine in the Prevention of Tuberculosis. *JAMA* 1994;271:698.
  - 25 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.
  - 26 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.
  - 27 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.
  - 28 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.
  - 29 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.
  - 30 Fine P. Stopping routine vaccination for tuberculosis in schools. *BMJ (Clinical research ed)* 2005;331:647–8.
  - 31 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.
  - 32 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.
- 33 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.
- 34 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.
- 35 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.
- 36 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.
- 37 Zimmermann, Petra, Finn, Adam, Curtis, Nigel. Does BCG Vaccination Protect Against Nontuberculous Mycobacterial Infection? A Systematic Review and Meta-Analysis. *J Infect Dis* 2018;218:679–87.
- 38 The World Health Organization. BCG Vaccine. *Weekly epidemiological record* 2004;79:27–48.
- 39 World Health Organization. *Global Tuberculosis Report*. 2017.
- 40 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.
- 41 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.
- 42 Schrager LK, Harris RC, Vekemans J. Research and development of new tuberculosis vaccines: a review. *F1000Research* 2018;7:1732–32.
- 43 R Core Team. R: a language and environment for statistical computing. 2019.
- 44 Abbott S. getTBinR: an R package for accessing and summarising the World Health Organisation Tuberculosis data. *Journal of Open Source Software* 2019;4:1260. doi:10.21105/joss.01260
- 45 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.
- 46 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.
- 47 Public Health England. The Green Book. 2013;391–409.

- 48 Sterne JAC, White IR, Carlin JB *et al.* Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj* 2009;338:b2393–3.
- 49 Pillaye J, Clarke A. An evaluation of completeness of tuberculosis notification in the United Kingdom. *BMC Public Health* 2003;3:31.
- 50 PHE. Tuberculosis in England 2016 Report (presenting data to end of 2015). 2016.
- 51 van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software* 2011;45:1–67. <https://www.jstatsoft.org/v45/i03/>
- 52 Sterne JA, Davey Smith G. Sifting the evidence—what's wrong with significance tests? *Bmj* 2001;322:226–31.
- 53 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.
- 54 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.
- 55 Wickham H. Tidy Data. *Journal of Statistical Software* 2014;59:1–23.
- 56 R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: 2016.
- 57 Office for National Statistics. Accuracy of official high-age population estimates, in England and Wales: an evaluation. 2016. <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/ageing/methodologies/accuracyofficialhighagepopulationestimatesinenglandandwalesanevaluation>
- 58 NHS Trafford CCG Governing Body. Performance and Quality Report. 2015;1–47.
- 59 Office for National Statistics. A comparison of the 2011 Census and the Labour Force Survey ( LFS ) labour market indicators. 2012. <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/acomparisonofthe2011censusandthelabourforcesurveylfslabourmarketindicators/2012-12-11>
- 60 Stevenson M, Nunes T, Heuer C *et al.* *epiR: Tools for the Analysis of Epidemiological Data*. 2017.
- 61 Pilger D, Nguipdop-Djomo P, Abubakar I *et al.* BCG vaccination in England since 2005: A survey of policy and practice. *BMJ Open* 2012;2:e001303.
- 62 Joint Committee on Vaccination, panel IB. Minutes. 2002.
- 63 Joint Committee on Vaccination, panel IB. Minutes. 2003.
- 64 Sutherland I. Effectiveness of BCG vaccination in England and Wales in 1983. *Tubercle* 1987;68:81–92.
- 65 Springett VH, Derbyshire JH, Nunn AJ *et al.* Changes in tuberculosis notification rates

- 
- in the white ethnic group in England and Wales between 1953 and 1983. *Journal of epidemiology and community health* 1988; **42**:370–6.
- 66 Borgdorff MW, Hof S van den, Kalisvaart N *et al.* Influence of sampling on clustering and associations with risk factors in the molecular epidemiology of tuberculosis. *Am J Epidemiol* 2011; **174**:243–51.
- 67 Fox J, Weisberg S. *An R companion to applied regression*. Third. Thousand Oaks CA:: Sage 2019. <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>
- 68 Pereira SM, Barreto ML, Pilger D *et al.* Effectiveness and cost-effectiveness of first BCG vaccination against tuberculosis in school-age children without previous tuberculin test (BCG-REVAC trial): A cluster-randomised trial. *Lancet Infectious Diseases* 2012; **12**:300–6.
- 69 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.
- 70 Fine PEM. Variation in protection by BCG: implications of and for heterologous immunity. *Lancet* 1995; **346**:1339–45.
- 71 Dye C. Making wider use of the world's most widely used vaccine: Bacille Calmette-Guerin revaccination reconsidered. *Journal of The Royal Society Interface* 2013; **10**:20130365–5.
- 72 Usher C, Adams R, Schmitz S *et al.* Evaluating the neonatal BCG vaccination programme in Ireland. *Archives of Public Health* 2016; **74**:1–12.
- 73 Hersh AL, Tala-Heikkila M, Tala E *et al.* A cost-effectiveness analysis of universal versus selective immunization with *Mycobacterium bovis* bacille Calmette-Guérin in Finland. *Int J Tuberc Lung Dis* 2003; **7**:22–9.
- 74 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.
- 75 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.
- 76 Abbott S, Christensen H, Welton NJ *et al.* Estimating the effect of the 2005 change in bcg policy in england: A retrospective cohort study, 2000 to 2015. *Eurosurveillance* 2019; **24**:1900220. doi:10.2807/1560-7917.ES.2019.24.49.1900220
- 77 Abbott S, Christensen H, Lalor MK *et al.* Exploring the effects of BCG vaccination in patients diagnosed with tuberculosis: Observational study using the Enhanced Tuberculosis Surveillance system. *Vaccine* 2019;1–6.
- 78 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.
- 79 Roth A, Sodemann M, Jensen H *et al.* Tuberculin reaction, BCG scar, and lower female

- mortality. *Epidemiology (Cambridge, Mass)* 2006;17:562–8.
- 80 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.
- 81 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.
- 82 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.
- 83 Djuretic T, Herbert J, Drobniowski F *et al.* Antibiotic resistant tuberculosis in the United Kingdom : 2002;477–82.
- 84 Barnard J, Rubin DB. Miscellanea. Small-sample degrees of freedom with multiple imputation. *Biometrika* 1999;86:948–55.
- 85 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.
- 86 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.
- 87 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.
- 88 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.
- 89 Parikh SR, Andrews NJ, Beebejaun K *et al.* Effectiveness 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.
- 90 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.
- 91 ai H. R Interface for H2O. 2018.
- 92 Bürkner P-C. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* 2017;80:1–28. doi:10.18637/jss.v080.i01
- 93 Carpenter B, Gelman A, Hoffman M *et al.* Stan: A probabilistic programming language. *Journal of Statistical Software, Articles* 2017;76:1–32. doi:10.18637/jss.v076.i01
- 94 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.

- 
- 95 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.
- 96 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.
- 97 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.
- 98 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.
- 99 Vynnycky E, Fine PE. The annual risk of infection with Mycobacterium tuberculosis in England and Wales since 1901. *Int J Tuberc Lung Dis* 1997;1:389–96.
- 100 Harris RC, Dodd PJ, White RG. The potential impact of BCG vaccine supply shortages on global paediatric tuberculosis mortality. *BMC Med* 2016;14:138.
- 101 Ragonnet R, Trauer JM, Scott N *et al.* Optimally capturing latency dynamics in models of tuberculosis transmission. *Epidemics* 2017;21:39–47.
- 102 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.
- 103 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.
- 104 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.
- 105 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.
- 106 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.
- 107 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.
- 108 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.
- 109 Mathema B, Andrews JR, Cohen T *et al.* Drivers of Tuberculosis Transmission. *J Infect Dis* 2018;216:S644–53.

- 110 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.
- 111 Shaw JB, Wynn-Williams N. Infectivity of pulmonary tuberculosis in relation to sputum status. *American review of tuberculosis* 1954;**69**:724–32.
- 112 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.
- 113 Piccini P, Chiappini E, Tortoli E *et al.* Clinical peculiarities of tuberculosis. *BMC Infect Dis* 2014;**14 Suppl 1**:S4.
- 114 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.
- 115 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.
- 116 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.
- 117 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.
- 118 Vynnycky E, Fine PE. Interpreting the decline in tuberculosis: the role of secular trends in effective contact. *Int J Epidemiol* 1999;**28**:327–34.
- 119 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.
- 120 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.
- 121 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.
- 122 Egbetade S, Polytechnic T, Ibrahim MO. Modelling The Impact of BCG Vaccines on Tuberculosis Epidemics. 2011;**1**:49–55.
- 123 Murray LM. Bayesian State-Space Modelling on high-performance hardware using LibBi. *Journal of Statistical Software* 2015;**67**:1–36.
- 124 Jacob PE, Funk S. *Rbi: R interface to libbi*. 2019. <https://CRAN.R-project.org/package=rbi>
- 125 Funk S. RBI.helpers. Published Online First: 2019.<https://github.com/sbfnk/RBI.helpers>
- 126 Gelman A. *Bayesian Data Analysis: Second Edition*. 2004.

- 
- 127 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.
- 128 Toni T, Welch D, Strelkowa N *et al.* Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface* 2009;6:187–202.
- 129 Lintusaari J, Gutmann MU, Kaski S *et al.* On the identifiability of transmission dynamic models for infectious diseases. *Genetics* 2016;201:911–8.
- 130 Busetto AG, Numminen E, Corander J *et al.* Approximate bayesian computation. *PLoS computational biology* 2013;9:e1002803.