

**A Thesis Submitted for the Degree of PhD at the University of Warwick**

**Permanent WRAP URL:**

<http://wrap.warwick.ac.uk/135007>

**Copyright and reuse:**

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it.

Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: [wrap@warwick.ac.uk](mailto:wrap@warwick.ac.uk)

**Persistence of tuberculosis inferred from case  
and contact networks in Birmingham, UK**

by

Melinda Lea Munang  
BMedSci BMBS MRCP (UK)

A thesis submitted in partial fulfilment of the requirements for the degree of  
Doctor of Philosophy

University of Warwick  
School of Life Sciences

June 2019

# Table of Contents

Acknowledgements.....	4
Declaration.....	5
Abstract.....	6
Abbreviations.....	7
List of Tables.....	8
List of Figures .....	10
1 Introduction.....	14
1.1 Overview.....	14
1.2 The global plan to end TB .....	15
1.3 TB in low incidence countries .....	16
1.4 TB control .....	17
1.5 Control strategy in England .....	18
1.6 Modern contact investigation strategies.....	21
1.7 The Birmingham TB register .....	22
1.8 Rationale and research objectives.....	23
1.9 Thesis outline .....	23
2 Background to the research dataset .....	24
2.1 Introduction.....	24
2.2 The Birmingham Tuberculosis Register .....	24
2.3 The clinical database.....	25
2.4 Routine data collection .....	26
2.4.1 Catchment population.....	26
2.4.2 Case table.....	28
2.4.2.1 Active TB cases .....	28
2.4.2.2 Latent TB cases.....	30
2.4.3 Contact table.....	32
2.4.4 Demographic and clinical data .....	34
2.4.5 Laboratory methods.....	34
2.5 The research dataset.....	35
2.5.1 Deduplication of the clinical database.....	35
2.5.2 Pseudonymisation.....	36
2.6 Ethical approval .....	36
2.7 Summary.....	36
3 Time-dependent risks of recurrent tuberculosis treatment episodes.....	37
3.1 Introduction.....	37
3.2 Objective .....	38
3.3 Method .....	38
3.3.1 Study design and setting.....	38
3.3.2 Study population.....	39
3.3.3 Identification of individuals with repeat TB episodes.....	39
3.3.4 Definitions .....	39
3.3.5 Laboratory methods.....	40
3.3.6 Statistical analysis.....	40
3.4 Results.....	42
3.4.1 Overall description of repeat TB treatment episodes .....	42
3.4.2 Demographic and clinical characteristics of TB cases by calendar time.....	44

3.4.3	Risks and predictors for active TB after latent infection treatment.....	46
3.4.4	Risks and predictors of recurrent active TB .....	47
3.4.4.1	Strain typing in recurrent active TB .....	48
3.4.5	Risks of latent infection treatment after active TB or latent infection treatment.....	48
3.5	Discussion .....	52
4	What happens to tuberculosis contacts? Competing risks application to estimate risks of infection with single versus repeat exposures. ....	55
4.1	Introduction.....	55
4.2	Objectives .....	58
4.3	Methods.....	58
4.3.1	Study population.....	58
4.3.2	Contact investigation and management.....	59
4.3.3	Contact outcomes .....	60
4.3.4	Analysis of risks and predictors of disease, infection and repeat TB exposure .....	60
4.4	Results.....	62
4.4.1	Overview of all contact episodes and outcomes.....	62
4.4.2	Study population.....	63
4.4.3	Risk of infection and disease with single versus two contact episodes..	63
4.4.4	Risk factors for disease, infection and second contact after the first contact episode .....	67
4.5	Discussion .....	72
5	The contact tracing network: global structure, local properties and extent of tuberculosis transmission .....	74
5.1	Introduction.....	74
5.2	Objectives .....	77
5.3	Method .....	77
5.3.1	Terminology .....	77
5.3.2	Network visualisation.....	80
5.3.3	Descriptive comparison of components in the contact tracing network.	80
5.3.4	Evaluation of degree distribution .....	80
5.3.5	Relating network metrics to number of infected contacts .....	81
5.3.6	Comparison of the contact tracing network and molecular network.....	81
5.4	Results.....	82
5.4.1	Overview of the static contact tracing network .....	82
5.4.2	Components in the static contact tracing network.....	83
5.4.2.1	Distance .....	83
5.4.2.2	Cohesion .....	85
5.4.2.3	Degree distribution .....	88
5.4.2.4	Mixing patterns .....	90
5.4.2.5	Network metrics and number of infected diagnosed .....	91
5.4.3	MIRU-VNTR typing .....	94
5.5	Discussion .....	97
6	The largest component.....	100
6.1	Introduction.....	100
6.2	Objectives .....	101
6.3	Method .....	102

6.3.1	Terminology .....	102
6.3.2	Visualisation .....	105
6.3.3	Network metrics .....	106
6.3.4	Subgroup detection .....	106
6.3.5	Genotypic data .....	107
6.4	Results .....	108
6.4.1	Overview of the (static) largest component.....	108
6.4.2	Evolution of the largest component with time.....	110
6.4.3	Connections in the largest component versus medium and large components .....	113
6.4.4	Groups in the largest component .....	116
6.4.5	Important nodes in the largest component: centrality measures.....	119
6.4.6	Description of groups containing the most central nodes (node 12063, group C and node 122256, group H).....	121
6.4.6.1	Group C – a school-based super-spreading event.....	121
6.4.6.2	Group H .....	124
6.4.7	Molecular epidemiology of the largest component .....	125
6.5	Discussion .....	133
7	Measuring growth in the contact tracing network .....	136
7.1	Introduction.....	136
7.2	Objectives .....	139
7.3	Survival analysis approach .....	140
7.3.1	Method.....	142
7.3.1.1	Terminology .....	142
7.3.1.2	Outdegree to alter .....	143
7.3.1.3	Indegree from alter .....	144
7.3.2	Results .....	145
7.3.2.1	Outdegree to alter .....	145
7.3.2.2	Indegree from alter .....	148
7.4	Growth curve model .....	151
7.4.1	Method.....	151
7.4.2	Results .....	153
7.4.2.1	Empirical and parametric component growth trajectories over time	153
7.4.2.2	Exploratory analysis of relationship between component growth and static component predictors.....	156
7.4.2.3	Unconditional means model for component growth .....	157
7.4.2.4	Unconditional growth model to evaluate the fixed and random effects of time.....	157
7.5	Discussion .....	160
8	Conclusions.....	161
9	Appendix.....	166
9.1	Case table fields in the Birmingham TB Register, 1980 - 2011 .....	166
9.2	Contact table fields in the Birmingham TB Register, 1980 - 2011 .....	169
10	Bibliography .....	171

## Acknowledgements

I am indebted to the following individuals for their contributions to, and support of, this dissertation:

**Professor Deirdre Hollingsworth**, Senior Group Leader, Big Data Institute, University of Oxford – Academic Supervisor

**Professor Graham Medley**, Professor of Infectious Diseases Modelling, London School of Hygiene and Tropical Medicine – Academic Supervisor

**Dr Martin Dedicoat**, Consultant in Infectious Diseases, Birmingham Heartlands Hospital and clinical lead for the Birmingham TB control board – Clinical Supervisor

**Mrs Zeitun Afzal, Mrs Jane Stewart, Mrs Eva Howes, Mrs Maxine Osborne and Miss Paula Richards**, TB Aftercare Department, Birmingham Chest Clinic – for maintaining the Birmingham TB register and help with preparation of the research dataset

**Dr Ron Crump**, Post-doctoral Researcher, Mathematics Institute, University of Warwick – for guidance on developing a multilevel model

**Dr Jason Evans**, Lead Scientist, Wales Centre for Mycobacteria – for guidance on interpretation of TB MIRU-VNTR typing during preparation of the research dataset

**Mrs Cathy Browne**, TB specialist nurse and cluster investigator, Birmingham Chest Clinic – for help and discussion during preparation of the research dataset and insight on clinical interpretation of large components

**Mrs Mary Mannion, Dr Mark Bailey and Dr Neil Jenkins**, Department of Infection and Tropical Medicine, Birmingham Heartlands Hospital – for administrative and IT support

**The R community** – for extensive availability of online discussion for *R* tasks

I would also like to thank my family, MC and CMG, for their patience and love.

## **Declaration**

I declare that this thesis has been composed by myself under the supervision of Professor Deirdre Hollingsworth, Professor Graham Medley and Dr Martin Dedicat. The research it describes has been done by me, except where acknowledged. This thesis, or any work included in this thesis, has not been submitted for any other degree. All external sources of information are acknowledged in the references.

Melinda Munang

June 2019

## **Abstract**

Tuberculosis (TB) is a public health priority in urban cities of high income countries such as the UK, where the local incidence can be several times that of the national incidence. A large dataset of 63,620 individuals registered at a single centre in Birmingham, UK captured all TB treatment episodes (active disease and latent infection) and contained individual level detail on contacts of TB cases from 1980 to 2011. Exploratory analysis of the pseudonymised research dataset revealed clusters of individuals presenting as cases and contacts through time. Repeated cases originated from the pool of known individuals treated for active or latent TB with a probability of 1.5% at five years and 2.7% at ten years, but routine recording of latent TB treatment episodes is not widespread to estimate the future burden of retreatment TB. When repeated contacts were examined, their probability of being diagnosed as a case was twice that of non-repeated contacts (3.9% versus 1.6% for active disease and 10.7% versus 3.7% for latent infection) at one year. Contact repetition should be recognised but consistent recording of patient identity is lacking. In further evaluation of the role of contact structure in case detection, only the eigenvector centrality (connections to other highly scoring individuals) was associated with at least one case detection in the local network. Because networks were viewed from a static approach and network metrics may reflect effect rather than cause of the contact tracing process, further interpretation was difficult. However network visualisation identified a large cluster of 3,148 individuals, who entered the dataset at all times in the study period, that were linked through a superspreading event. Evaluating transmission was limited by a small sample of patients with mycobacterial repetitive unit-variable number tandem repeats (MIRU-VNTR) typing but data available suggested that the superspreading event was nested within a risk network rather than a transmission network.

## **Abbreviations**

CSHR	Cause-specific hazard ratio
CI	Confidence interval
DOT	Directly observed therapy
ETS	Enhanced tuberculosis surveillance
HR	Hazard ratio
IQR	Interquartile range
MIRU-VNTR	Mycobacterial repetitive unit-variable number tandem repeats
NHS	National Health Service
SD	Standard deviation
TB	Tuberculosis
UK	United Kingdom

## List of Tables

Table 1. Number of individuals and their TB treatment episodes from 14,610 notified active and latent cases in Birmingham and Solihull, UK, 1980-2011.	43
Table 2. Comparison of demography and clinical characteristics of initial TB treatment episodes in Period 1 (1980-1999) and Period 2 (2000-2011).	45
Table 3. Demography and Cox model to determine predictors of active TB after a first treatment episode for latent TB infection.	50
Table 5. Demographic characteristics for 52,383 contacts and nature of their first TB contact exposure.	64
Table 6. Numbers at risk of events (disease, infection or further contact episode) and event occurrences over time following a first and second contact episode.	65
Table 8. Univariable analysis of cause-specific hazard ratios for disease, infection and second contact episode after a first contact episode.	69
Table 9. Multivariable analysis of cause-specific hazard ratios for disease, infection and second contact episode after the first contact episode.	70
Table 10. Definition of terms defining individuals in the contact tracing data.	78
Table 11. Definition of network terminology adapted from Hanneman and Riddle (2005).	78
Table 12. Definition of network metrics adapted from the <i>R</i> package igraph (version 1.1.2) (Csardi and Nepusz, 2006).	79
Table 13. Descriptive statistics for 5,728 connected components in the static contact tracing network. The largest component with diameter 23 was excluded.	87
Table 14. Outdegree distribution for nodes with pulmonary disease (n=5,293), non-pulmonary disease (N=2,140) and latent infection (N=2,290), by type of edge.	88
Table 15. Mixing patterns between pairs of nodes linked by household, close (non-household) and casual edges.	90
Table 16. Genotypic and epidemiological links for 1184 diseased nodes involved in contact tracing networks evolving from 2004 onwards. Numbers denote counts of (unordered) node pairs.	95
Table 17. Sensitivity and specificity of MIRU-VNTR typing in predicting epidemiological linkage, by ethnicity, birthplace, age and disease extent.	96
Table 18. Definition of terms defining individuals in the contact tracing data.	103
Table 19. Definition of network terminology adapted from Hanneman and Riddle (2005).	103
Table 20. Definition of component metrics adapted from the <i>R</i> package igraph (version 1.1.2) (Csardi and Nepusz, 2006).	104
Table 21. Centrality scores for key nodes in the static largest component.	121
Table 22. Definition of common terms used in the survival approach to measuring growth in the contact tracing network.	143
Table 23. Probability (Aalen-Johansen estimator) for a first generation infected to have contacts traced.	147
Table 24. Probability (Aalen-Johansen estimator) for infection in a second generation contact.	148
Table 25. Univariable Cox regression for time to second generation contact tracing.	148

Table 26. Cox regression for variables associated with time to indegree from alter.  
.....151

Table 27. Results of fitting preliminary multi-level models of component growth. 160

## List of Figures

Figure 1. TB incidence rates in England and Wales, 1913 to 2016. Number of cases per year was taken from historical case notification data (Public Health England, 2017) and population estimates extrapolated from census data (Office for National Statistics, 2017 and University of Portsmouth, 2017). .....	15
Figure 2. Schematic for epidemiologic basis of TB control, adapted from Rieder (1995 and 1999). .....	18
Figure 3. Geographic distribution of TB control boards in England (which are aligned to Public Health England centres) and 2016 case notification rates by board. Image from Public Health England, Tuberculosis in England 2017 report (London: Public Health England, 2017). .....	20
Figure 4. Geographical area corresponding to local authority boundaries included in the Birmingham TB register shaded in yellow. Numbers denote the three-year average (2014-2016) incidence rates for each local authority (Public Health England, 2017a). .....	27
Figure 5. Number of TB disease case notifications in the Birmingham TB Register, 1980 to 2011.....	29
Figure 6. Annual TB disease incidence in Birmingham and Solihull, 1980 to 2011. Population estimates were extrapolated from census data (Office for National Statistics, 1997; 2000; 2005; 2016). Vertical bars denote 95% confidence interval calculated assuming a Poisson distribution.....	29
Figure 7. Number of treated latent TB infection case notifications in the Birmingham TB register, 1980 to 2011.....	31
Figure 8. Annual treated latent TB case incidence in Birmingham and Solihull, 1980 to 2011. Population estimates were extrapolated from census data (Office for National Statistics, 1997; 2000; 2005; 2016). Vertical bars denote 95% confidence interval calculated assuming a Poisson distribution. ....	31
Figure 9. Number of household and close (non-household) contacts named by TB cases, 1980 to 2011. ....	33
Figure 10. Active and latent TB treatment episodes notified in Birmingham and Solihull, UK, 1980-2011 and final study population included in analysis. N = number of individuals. ....	43
Figure 11. Cumulative hazard for any repeat TB treatment episode (A) and cumulative hazard by year of initial treatment episode (B). Period 1, first treatment episode notified in years 1980-1999; Period 2, first treatment episode notified in years 2000-2011. Dotted lines denote 95% confidence intervals.....	44
Figure 12. Cumulative hazard of repeat TB treatment episodes by episode types. Dotted lines denote 95% confidence intervals. Period 1, first treatment episode notified in years 1980-1999; Period 2, first treatment episode notified in years 2000-2011. ....	49
Figure 13. TB contact, disease and infection model to investigate the importance of second contact episodes. ....	61
Figure 14. Overview of TB contact episodes and outcomes for 52,383 contacts (individuals). ....	62
Figure 15. Nelson-Aalen estimates of the cumulative hazard for diagnosis of TB disease, latent TB infection and further contact episode after a first or second contact episode. Dotted lines denote log-transformed 95% confidence intervals. ....	65

Figure 16. Aalen-Johansen estimates of the cumulative incidence function for all event types following a 1st and 2nd contact episode. ....	66
Figure 17. Predicted cumulative incidence function (Aalen-Johansen estimates) for disease in male, household contacts of culture-positive pulmonary TB patients by ethnic group and first quartile, median and third quartile age of diseased contacts. ....	71
Figure 18. Predicted cumulative incidence function (Aalen-Johansen estimates) for .....	71
Figure 19. Status of 64,334 individuals in the case and contact database. Shaded areas denote those involved in contact tracing networks. ....	82
Figure 20. Undirected and unweighted diameters of 5,729 connected components in the static contact tracing network. The directed and weighted (different edge types) subgraphs are shown to illustrate evolution of the contact tracing network. Node labels denote the year of diagnosis. ....	84
Figure 21. Pairwise distance distribution between 9,247 infected nodes in the directed and undirected contact tracing network. Numbers above bars denote actual number of pairs with the path length. The largest component was excluded. ....	85
Figure 22. Indegree distribution (household, close and casual) for nodes in the contact tracing network, by node infection type. ....	89
Figure 23. Box plots of the number of infected contacts for pulmonary (N=5,293) and non-pulmonary (N=2,140) index cases (with at least one outdegree), by edge type and number of outdegrees. Median number is denoted by band in the centre of boxes, first and third quartile by box edges, maximum number within 1.5 times the interquartile range by the uppermost whisker, outliers by circles. ....	92
Figure 24. Distribution of centrality measures for nodes with no infected contacts (solid line) versus egos with at least one infected contact (dotted line). ....	93
Figure 25. Distribution of time between infection diagnosis in an index case and contact by disease. ....	93
Figure 26. No. of typed isolates compared to the total burden of disease by year. ...	94
Figure 27. The simplified largest component, with 2,415 contact nodes with only one incident edge removed. ....	109
Figure 28. Number of cases and contacts added per year in the largest component. ....	110
Figure 29. Evolution of the simplified largest component from year 1990 to 2011. ....	112
Figure 30. Outdegree distribution in pulmonary (A, B, C) and non-pulmonary (D, E, F) nodes in the largest component versus 536 medium and large (other) components in the contact tracing network, by type of edge. ....	114
Figure 32. Indegree distribution in pulmonary (A), non-pulmonary (B) and latent case (C) nodes in the largest component versus 536 medium and large (other) components in the contact tracing network. ....	115
Figure 32. The simplified largest component grouped, in coloured polygons, as delineated by the Clauset-Newman algorithm. Nodes are not coloured to simplify the diagram. ....	117
Figure 33. The simplified largest component contracted into groups. Groups are referenced by a letter (label). Size of each node is proportional to the sum of case and contact nodes with at least 2 edges in each group. Edges between	

groups were simplified and do not reflect the actual number of links between groups.....	118
Figure 34. Subgraph of group C (containing node 12063 circled in red) and group H (containing node 122256 circled in green) in the largest component. Labels denote the year of diagnosis (09=2009, 99=1999). Nodes who did not become cases and had only one incident edge were removed. Only nodes up to 2 (undirected) path lengths away were included. A microepidemic of 11 secondary disease and 57 latent infection cases were detected on contact tracing.....	123
Figure 35. Reduced graph of the largest component showing culture-positive and culture-negative nodes and their MIRU-VNTR typing availability. All case nodes were retained in this graph. Latent and contact only nodes were excluded if they did not result in an increase of the number of connected components. Labels are depicted for nodes with MIRU-VNTR typing information with the alphabet referencing group membership (see Figure 33) and unique number in each group referencing an individual node. ....	127
Figure 36. Phylogenetic distribution of 63 nodes with 15- or 24-loci MIRU-VNTR sequences in the largest component. Numbers highlight the root at which more detailed tree analyses was done with a sample of unique MIRU-VNTR sequences in the wider database included for reference (Figures 37 – 40). Nodes were labelled with an alphabet referencing epidemiological group membership (see Figure 26) and unique number in each group referencing an individual node, and were coloured according to their epidemiological group. ....	128
Figure 37. Phylogenetic distribution of nodes within root 1 (Figure 36) in comparison to the wider MIRU-VNTR sequences in Birmingham.....	129
Figure 38. Phylogenetic distribution nodes within root 4 (Figure 36) in comparison to the wider MIRU-VNTR sequences in Birmingham. ....	130
Figure 39. Phylogenetic distribution of nodes within root 2 (Figure 36) in comparison to the wider MIRU-VNTR sequences in Birmingham.....	131
Figure 40. Phylogenetic distribution of nodes within root 3 (Figure 36) in comparison to the wider MIRU-VNTR sequences in Birmingham.....	132
Figure 41. Examples of events causing component growth. A contact is diagnosed with TB and secondary contact tracing occurs (A) or a contact is named twice by different cases (B). Node labels denote the year of diagnosis. ....	138
Figure 42. Schematic depicting growth events in a network. Circles are individuals in the contact tracing network. ....	140
Figure 43. Event histories for (directed) case-contact pairs. Blue node denotes backward traced case, red node denotes forward traced case. ....	141
Figure 44. Transition states in a survival model to evaluate second generation contact tracing and its yield in case diagnosis. ....	143
Figure 45. Nelson-Aalen estimate of the cumulative hazard for a case-case pair to have second generation contact tracing followed by infection in a secondary contact. ....	146
Figure 46. Kaplan-Meier estimator of duration remaining without an indegree from an alter for 51,496 case-contact pairs. Dotted lines denote 95% confidence intervals. ....	149
Figure 47. Ordinary least squares fitted growth trajectories for components, by calendar year of first initial infected case. The average change change trajectory in for all components within each time period is shown in blue.....	153

Figure 48. Empirical and average growth trajectories of 1,416 connected components with more than 1 infected case in the contact tracing network. The largest component was excluded. The smooth, non-parametric average change trajectory for all connected components is shown in blue. ....155

Figure 49. Linear change trajectories, residual variance and R2 statistic resulting from fitting separate ordinary least squares regression models for 1,416 connected components with more than one infected case in the contact tracing network. The average change trajectory for all components is shown in blue.156

Figure 50. Ordinary least squares fitted trajectories for component growth, by levels of selected predictors. The average change trajectory for all components is shown in blue. ....158

## 1 Introduction

“If fully used the combination of diagnostic facilities and effective chemotherapy now available in this country could result in tuberculosis becoming truly rare in England and Wales by the end of this century; on present trends there are likely to be a few hundred new notification of tuberculosis in the year 2000 A.D.”  
(Springett, 1972)

### 1.1 Overview

Tuberculosis (TB) has been present throughout human history, from its detection in Egyptian mummies (Cave, 1939; Nerlich *et al.*, 1997) to the epidemic that swept across Europe during the industrial revolution (Dubos and Dubos, 1952). Since the turn of the 20<sup>th</sup> century there has been declining TB incidence in the Western world (Long *et al.*, 1999; Raviglione *et al.*, 1993; Schneider and Castro, 2003) as illustrated by historical TB data in England and Wales in Figure 1. This trend prompted predictions that TB would soon be a disease of the past (Springett, 1972).

The 1990s saw a resurgence in cases with multiple factors such as the HIV epidemic, increased migration and downgrading of TB control programs thought to be contributory (Grange and Zumla, 2002; Snider, 1997). In 1993 TB was declared a global health emergency (WHO Global Tuberculosis Programme, 1994) and yet still today it is the single largest cause of death worldwide from an infectious disease with an estimated 1.3 million deaths out of the 10.4 million people developing active disease in 2016 (World Health Organization, 2017).

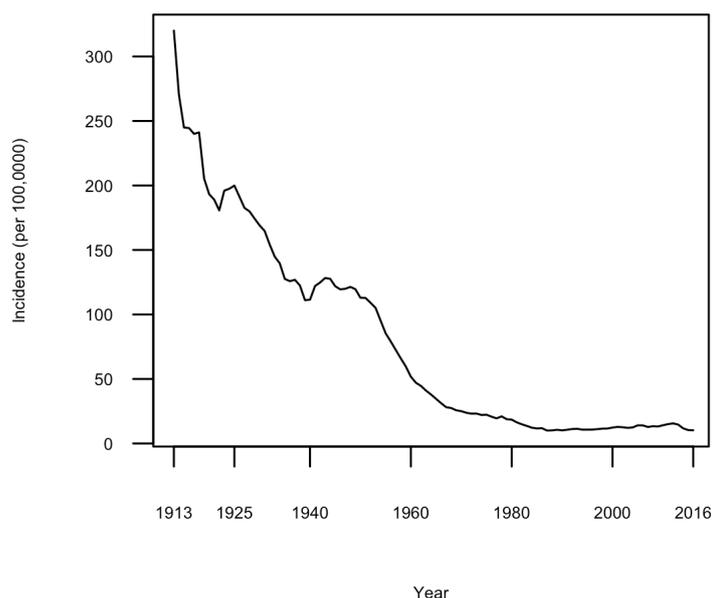


Figure 1. TB incidence rates in England and Wales, 1913 to 2016. Number of cases per year was taken from historical case notification data (Public Health England, 2017) and population estimates extrapolated from census data (Office for National Statistics, 2017 and University of Portsmouth, 2017).

## 1.2 The global plan to end TB

Global targets for TB elimination (defined as less than one case of TB disease per million population per year) by 2050 were set in 2010 (World Health Organization and Stop TB Partnership, 2010). Concerted efforts in active case management with early diagnosis and provision of directly-observed therapy in low and middle-income countries has resulted in approximately a 2% yearly decline in incidence, but this reduction rate needs to double for the incidence trajectory to reach the milestone of less than 85 cases per 100,000 per year by 2020 set by the World Health Organization (World Health Organization, 2017). Ten very high burden countries with large estimated proportion of untreated cases (India, China, the Russian Federation, Indonesia, the Philippines, Pakistan, Nigeria, Ukraine, Myanmar and Uzbekistan) are important targets for improvement (World Health Organization, 2017).

### 1.3 TB in low incidence countries

Low incidence countries (<100 cases per million population) have the greatest potential to reach TB elimination. In these settings TB is heterogeneously distributed, with the burden of disease in urban areas (Hayward *et al.*, 2003; Oren *et al.*, 2011), amongst migrants (Cain *et al.*, 2007; Gaudette and Ellis, 1993; Rose *et al.*, 2001), the poor (Bhatti *et al.*, 1995; Hawker *et al.*, 1999) and other disadvantaged groups such as homeless persons, prisoners and substance misusers (Haddad *et al.*, 2005; Story *et al.*, 2007). Thus the epidemiology can be characterised by concentration in hard-to-reach risk groups and a low rate of transmission in the general population, sporadic outbreaks and active disease mainly generated by latent infection acquired remotely rather than recently/locally (Lonnroth *et al.*, 2015). This epidemiological profile has been referred to as TB in big cities, a unique situation that requires innovative responses to enable transition into the pre-elimination phase (de Vries *et al.*, 2014). Birmingham, UK, falls into this category.

In 2014 the World Health Organization identified 33<sup>1</sup> countries with a national TB incidence of <100 cases per million to energise the progress towards TB pre-elimination in these countries (Lonnroth *et al.*, 2015). Some cities in these low incidence countries however have up to twice the national incidence rate (de Vries *et al.*, 2014).

---

<sup>1</sup> Australia, Austria, Bahamas, Belgium, Canada, Costa Rica, Cuba, Cyprus, Czech Republic, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Israel, Italy, Jamaica, Jordan, Luxembourg, Malta, Netherlands, New Zealand, Norway, Puerto Rico, Slovakia, Slovenia, Sweden, Switzerland, United Arab Emirates, United States of America, West Bank and Gaza Strip

## 1.4 TB control

TB transmission relies on inhalation of airborne droplet nuclei containing *Mycobacterium tuberculosis* (Loudon and Roberts, 1967; Riley *et al.*, 1995), a slow growing intracellular bacilli that has a lipid-rich cell wall and outer membrane that acts as an effective barrier to phagocytosis and antimicrobials (Osoba, 2004). Infection depends first on exposure to sources emitting infectious droplet nuclei. The pool of infectious cases in the community, duration of infectiousness and patterns of interactions between susceptible and infectious individuals are therefore primary drivers of observed TB epidemiology (Figure 2). Secular trends such as improvements in living standards and social conditions were widely credited with the sharp decline in TB incidence at the turn of the 20<sup>th</sup> century by reducing the number of successful transmissions, or effective contact rate, even prior to the availability of effective drug treatment (Vynnycky and Fine, 1999). However early diagnosis and treatment of active disease to reduce community exposure to TB remains at the core TB control and is the dominant intervention employed in the Stop TB Strategy (World Health Organization and Stop TB Partnership, 2006).

Infection following exposure is dependent on the intensity of the bacillary load (O’Shea *et al.*, 2014; Riley *et al.*, 1995) and host immune response, particularly macrophage function (Schluger and Rom, 1998) (Figure 2). Interventions to modify these factors could include prioritisation of contacts of more infectious cases (e.g. sputum smear-positive cases), and attention to ventilation in congregate settings where TB is endemic (Taylor *et al.*, 2016). Thereafter risk of progression to active disease, whether infectious or non-infectious, is higher within the first few years of infection (Sloot *et al.*, 2014; Trauer *et al.*, 2016), impaired host immune response (Harries, 1990; Keane *et al.*, 2001) and pathogen virulence (Newton *et al.*, 2006; Yang *et al.*, 2012) (Figure 2). These factors are less amenable to epidemiologic control. However, chemotherapy for latent infection is protective against future disease and infectiousness (Smieja *et al.*, 2000). Reducing the reservoir of infection by finding and treating individuals with latent infection is an essential requirement for TB elimination (Dye *et al.*, 2013).

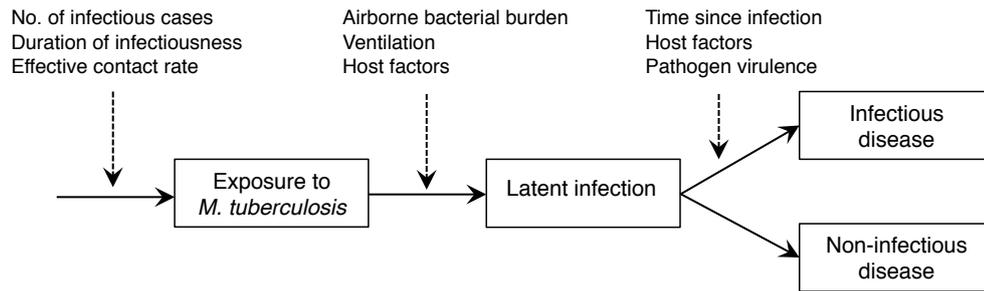


Figure 2. Schematic for epidemiologic basis of TB control, adapted from Rieder (1995 and 1999).

### 1.5 Control strategy in England

England had one of the highest rates of tuberculosis among industrialised countries (14.4 per 100,000 per year) in 2012 (European Centre for Disease Prevention and Control/WHO Regional Office for Europe, 2018). The problem was attributed to relatively larger number of migrants from high TB burden countries with over 70% of TB cases diagnosed in the foreign-born (Gilbert *et al.*, 2009), variable implementation of migrant screening programmes (Pareek *et al.*, 2011) and inadequate numbers of specialist staff (Bothamley *et al.*, 2011).

Recently the two decades of increasing TB incidence in England has reversed. Since 2012, an 8% decline in the number of cases annually has been observed with the annual incidence dropping to 10.2 per 100,000 in 2016 (Public Health England, 2017b). A recent analysis has suggested declines in TB rates in all populations (UK and non UK-born (except EU countries) and regardless of time since arrival in the UK) contributed more to this trend rather than change in migration patterns (Thomas *et al.*, 2018). Interventions that may have contributed to the incidence decline include provision of outreach services for hard-to-reach groups (Daly *et al.*, 2016; Jit *et al.*, 2011) and strengthened governance of TB control activities (Anderson *et al.*, 2014).

Another major control strategy is pulmonary TB screening in visa applicants from high incidence countries ( $\geq 40$  per 100,000) prior to UK entry (Public Health England, 2017c). This was rolled out in 2014 following successful pilot studies from 2005. In 2016, the incidence of pulmonary TB observed was 100 per 100,000 applicants (Public Health England, 2017c).

Following on from the pre-entry screening programme a national tuberculosis strategy for England was published in 2015 (Public Health England, 2015). This signaled renewed political commitment to the TB elimination agenda and provision of leadership and funding. A major component of this strategy is the establishment of seven multi-agency TB control boards (Figure 3) to deliver co-ordinated TB control activities within each area.

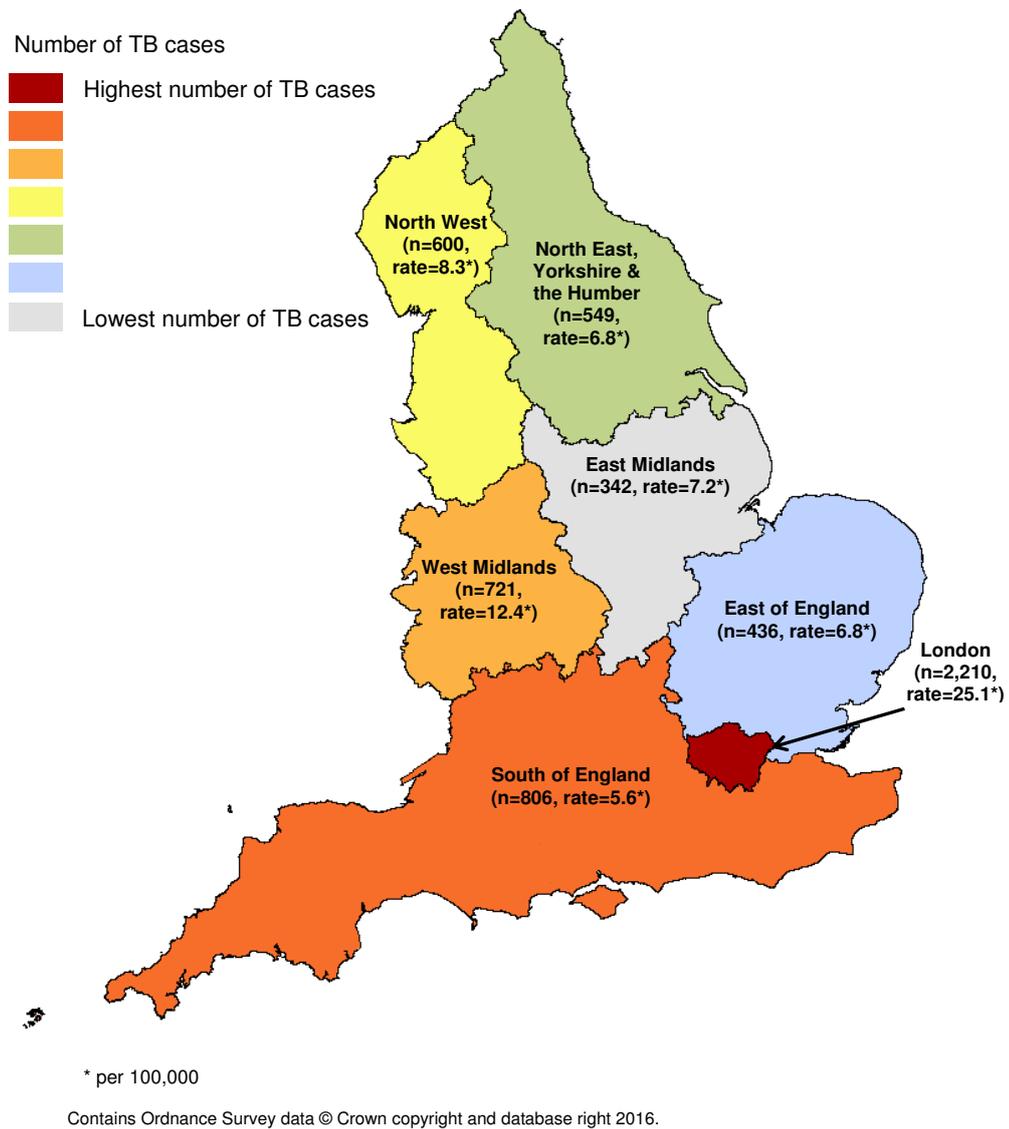


Figure 3. Geographic distribution of TB control boards in England (which are aligned to Public Health England centres) and 2016 case notification rates by board. Image from Public Health England, Tuberculosis in England 2017 report (London: Public Health England, 2017).

## 1.6 Modern contact investigation strategies

In low-incidence countries active case finding through contact investigation is a principal activity of TB control programmes. The prevailing approach is the concentric circle or “stone-in-pond” method. The underlying principle is that those in closest physical proximity (i.e. shared air space over the most time) to an infectious case are most likely to be infected, and if they are not then the yield from contact tracing those less close is minimal (Veen, 1992). This approach is pragmatic particularly when dealing with outbreaks in congregate settings.

The advent of a polymerase chain reaction-based test to enumerate variable number tandem repeats in the mycobacterial genome and distinguish relatedness has made universal genotyping routine (Bolotin *et al.*, 2010; Lambregts-van Weezenbeek *et al.*, 2003; Public Health England, 2014). Genotyping has been useful in highlighting casual contact/location-based transmission, among other things such as detection of false-positive TB laboratory results (Frieden *et al.*, 1996). Thus there is an increasing interest in utilising network data to identify relevant contacts (Cook *et al.*, 2012). Post hoc examination of location data after a molecular outbreak has been helpful in identifying the congregate setting where transmission occurred (Gardy *et al.*, 2011). However lack of data systems that could utilise prospectively collected location data in such a way is a limiting factor.

Geospatial scanning has also been proposed as a further adjunct to identify hotspots of TB in conjunction with genotyping information (Moonan *et al.*, 2012; Smith *et al.*, 2017). These hotspots could be targets for intervention but to date no practical application of its utility in the field has been demonstrated.

## 1.7 The Birmingham TB register

Birmingham is the second most populous UK city with approximately 1.1 million residents in 2016 (Office for National Statistics, 2017) and is among five UK cities<sup>2</sup> with a high TB burden ( $\geq 40$  per 100,000 during 2006-2008) (Kruijshaar *et al.*, 2012). Therefore outside of London, the Birmingham TB Control Board has a significant task in the elimination agenda.

Case reporting to the national surveillance database for the city, and the adjacent urban borough of Solihull, is done centrally from the Birmingham Chest Clinic, a purpose-built outpatient facility for TB and other lung diseases in operation since 1932. National surveillance reporting was paper-based until 2008/9, but a local electronic database of case notifications and contact investigation for clinical use has been in existence at the Birmingham Chest Clinic since 1980.

This comprehensive population level data spanning three decades has never been systematically analysed. While subject to drawbacks such as reliability of data quality the Birmingham TB Register presents a unique opportunity to investigate the local epidemiology and inform service planning. Specifically, the large amount of TB contact data linked to cases may generate new knowledge and understanding about the transmission of TB in this setting. A general aim of the research in this thesis was to go beyond traditional demographic and clinical factors to prioritise contacts and communities, and identify innovative criteria based on contact patterns.

---

<sup>2</sup> The other cities are London, Leicester, Slough and Luton

## **1.8 Rationale and research objectives**

TB in big cities is a particular challenge and requires setting-specific research to understand its epidemiology. The availability of this large electronic register of cases and contacts in Birmingham, a high-burden city can be exploited for this purpose. Implementation of modern approaches to contact investigation as described in section 1.6 will require increased resources but their feasibility for routine use needs further study. Findings could inform priorities and resource planning not only for the local TB Board, but are likely transferable to other settings of urban TB.

Research objectives are outlined separately in each chapter of the thesis.

## **1.9 Thesis outline**

The next chapter is an overview of the Birmingham TB Register's background, data collection methods and conversion into the research dataset. The subsequent five chapters explore the data set in different ways. Chapter 3 is a time-to-event analysis of recurrent TB treatment episodes. Chapter 4 investigates transitions between being named as a case and being named as contact in the dataset using competing risks methodology. Chapter 5 is a descriptive analysis of the network formed by cases and contacts combined. Chapter 6 is a further descriptive analysis of a very large group of connected cases and contacts in the network. Finally chapter 7 is an exploratory investigation of how clusters of cases and contacts grow in size using survival methods and an individual growth model.

Molecular epidemiology based on mycobacterial repetitive unit-variable number tandem repeats (MIRU-VNTR) was available for a subset of cases and is explored throughout the chapters where relevant.

Although location/geographic data was available for cases and contacts in the dataset this theme is not investigated in this thesis.

## **2 Background to the research dataset**

### **2.1 Introduction**

The data used throughout this research is routine clinical information contained in a database owned and managed by the Heart of England NHS Foundation Trust, a large organisation that runs general and specialist clinical services in the West Midlands region of the UK. This chapter will describe the historical background to the clinical database, how clinical data was collected and known temporal changes that have occurred, data structure and finally methods used to create the research dataset.

### **2.2 The Birmingham Tuberculosis Register**

There has been a long history of routinely collected data related to tuberculosis (TB) at Birmingham Heartlands Hospital, part of the Heart of England NHS Foundation Trust. The municipal sanatorium for the city of Birmingham was established near the present day hospital site following legislation entrusting sanitation and control of infectious diseases to the local authority (Public Health Act, 1875). While initially an isolation facility for smallpox cases due to its location away from the general public at the time, in 1895 two per thousand population died of tuberculosis leading to the site evolving into a TB sanatorium by 1910 (Ayres *et al.*, 1995).

Prior to statutory requirement to notify cases of clinical tuberculosis in 1911 (Public Health (Tuberculosis) Regulations, 1911), reporting was voluntary but encouraged by a payment for each case notified. Birmingham City Council recognised early the importance of establishing how many cases they had to deal with in order to control TB and introduced mandatory notification as early as 1904 (Ayres *et al.*, 1995). Hand-written registers for all TB cases within the city boundaries dating back to 1903 are in hospital archives detailing name, date of birth, sex, address, date of admissions and discharges or death (Martin Dedicoat 2017, personal communication, 2 October).

Early records in the Birmingham TB register collected the minimum data necessary to inform trend analyses and public health action. The hospital continued to maintain the city-wide register following nationalisation in 1948 (National Health Service Act, 1946). In 1978, five-yearly national TB surveys were initiated with more detailed clinical and demographic data collected (MRC Tuberculosis and Chest Diseases Unit, 1980). This additional information was gradually incorporated into the Birmingham TB register to enable reporting. From 1999 onwards, the Health Protection Agency (now Public Health England) established an enhanced tuberculosis surveillance (ETS) system in England and Wales, which routinely records additional demographic, clinical and microbiological information (Public Health England, 2013). The Birmingham TB register now mirrors ETS data to streamline upstream reporting.

### **2.3 The clinical database**

The Birmingham TB register became digital in 1980. With just over 600 TB cases treated at the Birmingham Chest Clinic (an outpatient centre for tuberculosis care established by city council in 1933 and absorbed into the national health service as part of the Heart of England NHS Foundation Trust) in 1980, the need for an electronic tool to co-ordinate clinical appointments for case management was evident. Dr John Innes, a chest physician at the Birmingham Chest Clinic wrote a relational database in dBase for this purpose (Martin Dediccoat 2017, personal communication, 2 October). Because patients treated for latent TB infection also required clinical follow-up, they were included the database. From 1987 onwards, the large numbers of TB contacts assessed were also entered to track management and follow-up. Contacts had less demographic and clinical data recorded compared to cases, but linkage to index cases was preserved and type of relationship, where known, was entered. No other known UK dataset contains information about latent TB cases and TB contacts over this time frame.

Thus while historically a surveillance register, the electronic Birmingham TB register is primarily an administrative tool that networks with other NHS clinical systems, for example patient registration data and appointment bookings. It also enables outcomes assessment that informs commissioning evaluation locally. All data were collected prospectively in the course of normal healthcare provision by the clinical team.

Local case notification data continues to be reported nationally through the web-based ETS. Data from the Birmingham TB register is entered separately into ETS because within Public Health England's national database system no data flow to or from NHS clinical systems are possible.

## **2.4 Routine data collection**

### **2.4.1 Catchment population**

The electronic Birmingham TB register holds records of all cases of active TB, treated latent TB infection and TB contacts residing within the Birmingham and Solihull local authority boundaries since 1980 (Figure 4). The geographic size of the area is approximately 167 square miles with a population of 1.34 million in 2016 (Office for National Statistics, 2017). There have been minor additions to the catchment boundaries in 1982 (borough of Sutton Coldfield), 1995 (Frankley and Kitwell estates in Bromsgrove) and 2004 (New Hall Ward/Walmley area).

Individuals moving in to the area on TB therapy after an initial diagnosis elsewhere may not be entered into the register if they did not attend TB services locally i.e. they chose to continue follow-up with the care provider at their previous address. Individuals who were diagnosed and started on TB therapy in the study area but who subsequently moved out were included and their transfer to a different care provider was recorded as an outcome in the register.

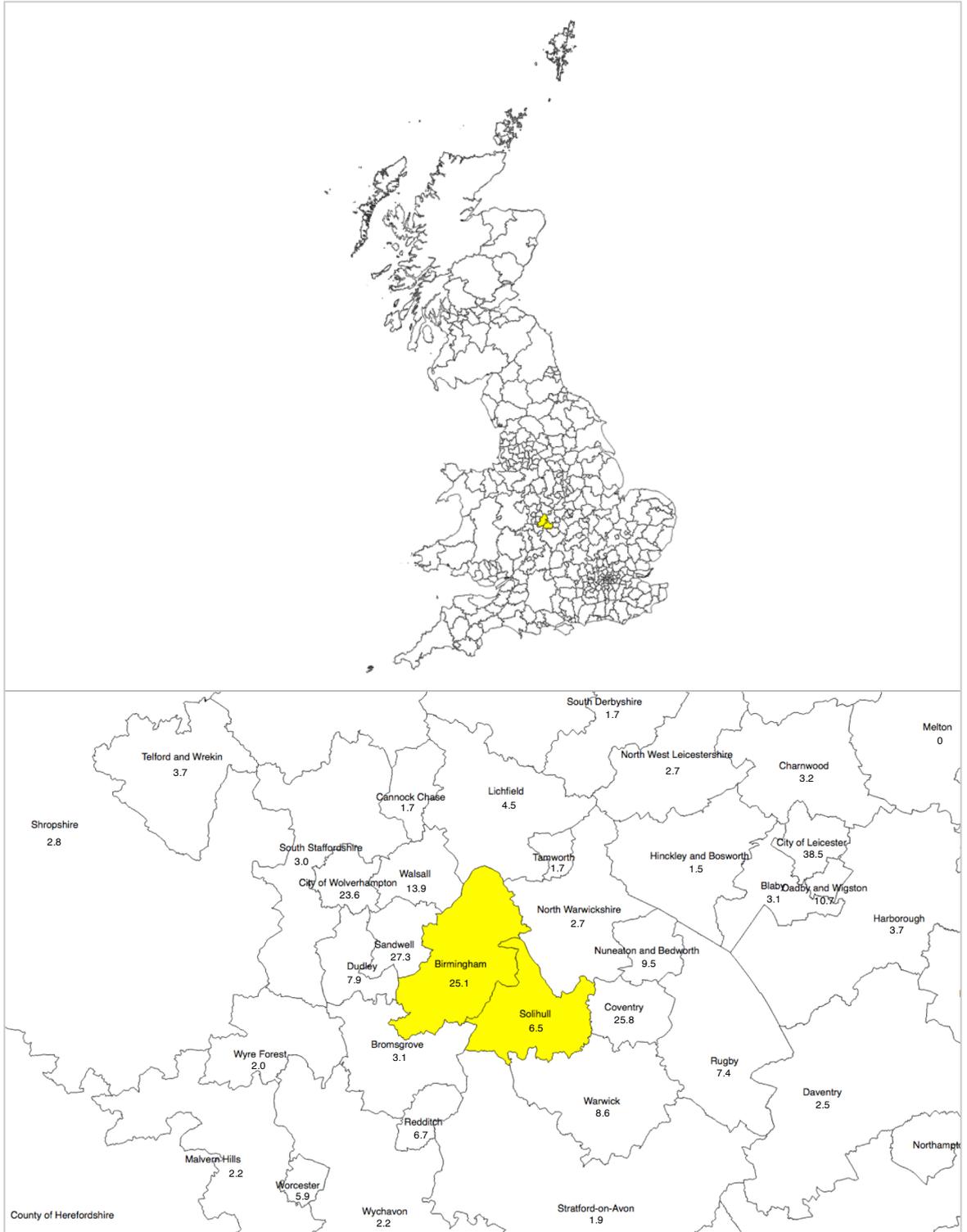


Figure 4. Geographical area corresponding to local authority boundaries included in the Birmingham TB register shaded in yellow. Numbers denote the three-year average (2014-2016) incidence rates for each local authority (Public Health England, 2017a).

## **2.4.2 Case table**

All patients treated for active or latent TB (both culture-confirmed and unconfirmed) residing in the catchment area were entered into a single case table. Data fields included in the case table, and completeness of data entry, are detailed in Appendix 1.1.

### **2.4.2.1 Active TB cases**

Notification of active TB cases for entry into the Birmingham TB register was received from treating clinicians or the laboratory following culture confirmation. The Birmingham Chest Clinic has historically held all city-wide notifications of TB since statutory requirements for reporting clinical disease began. Thus the case table likely reflects accurate ascertainment of all active TB cases in the area, including post-mortem diagnoses which are notified by the laboratory if culture positive. Post-mortem cases with only histological diagnoses may have escaped registration and statutory notification. Cases that were subsequently found to have an alternative diagnosis and denotified were retained in the case database, but could be identified as denotified.

There were 300 to 400 case notifications of TB disease annually (Figure 5). Annual incidence declined from 50 per 100,000 (95% confidence interval, CI 46, 55) in 1980 to a low of 21 per 100,000 (95% CI 18, 23) in 1988. In the next decade the incidence fluctuated between 20 to 30 per 100,000 but since late 1990s the incidence has been trending upwards reaching a peak of 41 per 100,000 (95% CI 37, 44) in 2009 (Figure 6).

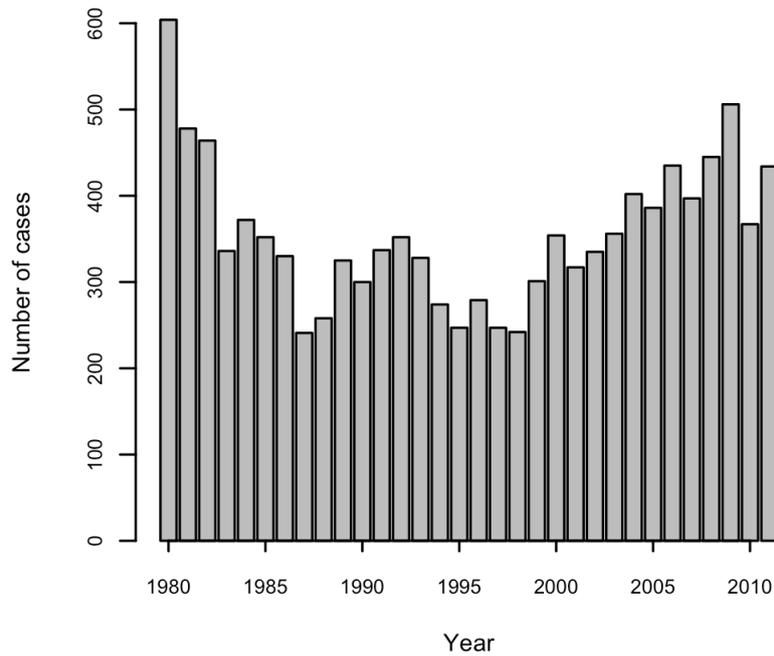


Figure 5. Number of TB disease case notifications in the Birmingham TB Register, 1980 to 2011.

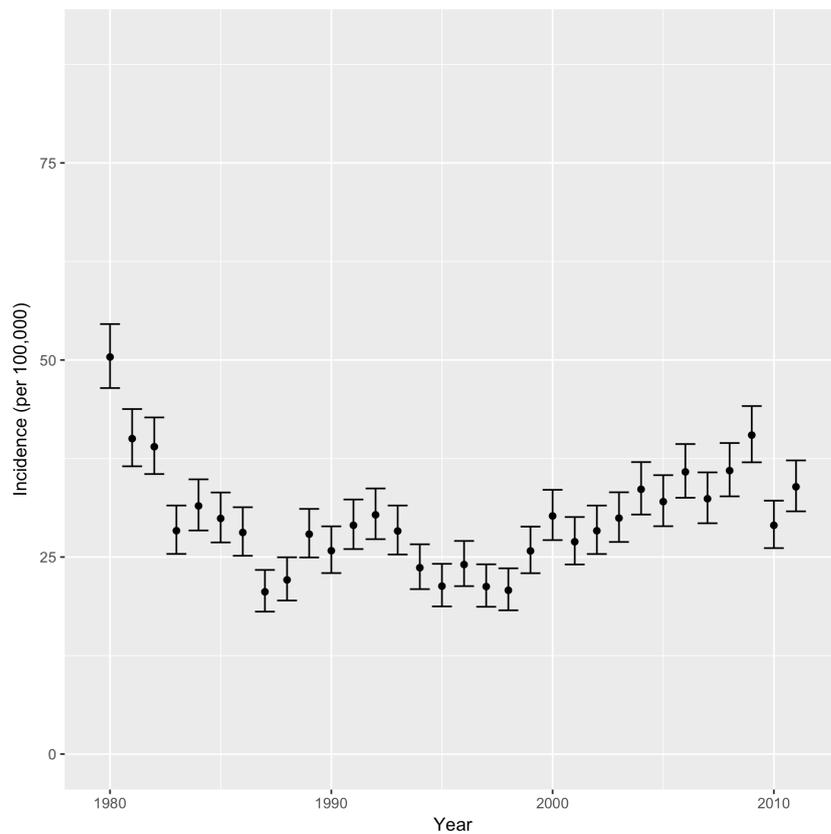


Figure 6. Annual TB disease incidence in Birmingham and Solihull, 1980 to 2011. Population estimates were extrapolated from census data (Office for National Statistics, 1997; 2000; 2005; 2016). Vertical bars denote 95% confidence interval calculated assuming a Poisson distribution.

#### 2.4.2.2 Latent TB cases

Latent cases are not statutorily reported but as all treatment monitoring in the community were undertaken by TB nurses based at the Birmingham Chest Clinic, the case database likely includes all treated latent TB cases in the area.

The case database does not however reflect complete ascertainment of the burden of latent TB infection. Firstly individuals found to have latent infection, but were not given or declined treatment, were not recorded. Secondly testing for latent TB infection was only offered to contacts in certain age groups. Up to 2006 only children under 16 were assessed for latent infection. Locally latent TB testing was expanded to those aged 35 and under from around 2007 (Zeitun Afzal 2014, personal communication, 12 April).

Where testing was offered this was largely using tuberculin skin testing. Reaction of grade 3 or 4 in BCG-vaccinated individuals or grade 2 in unvaccinated individuals following a Heaf test was considered positive. Mantoux testing largely replaced Heaf testing in the late 2000s. Induration of 15mm or more for BCG-vaccinated individuals or 6mm or more for non-vaccinated individuals were considered positive. At these cut-offs the estimated sensitivity and specificity of the tuberculin skin testing is high for BCG-vaccinated children (up to age 15) (90% to 100% and 83% to 96% respectively) but lower for BCG-vaccinated children (66% to 87% and 79% to 89%) (Seddon *et al.*, 2016). In adults sensitivity is similar but specificity ranges from 35% to 79% in vaccinated adults (Pai *et al.*, 2008). From 2011 onwards, positive Mantoux tests were confirmed with interferon gamma release assays before definitive diagnosis. This assay has high specificity (86% to 100%) (Pai *et al.*, 2008).

Up to 2012 almost all active case finding efforts were concentrated among TB contacts with only *ad hoc* cases found through screening in other at risk groups. Hence the majority of latent cases were diagnosed as a result of exposure to an index case.

There were between 50 to 100 cases of treated latent TB infection annually in the study area (Figure 7). The numbers treated have increased dramatically since 2000s, with an incidence of 9 per 100,000 in 2000 (95% CI 7, 11) rising to 23 per 100,000 in 2011 (95% CI 20, 25) (Figure 8).

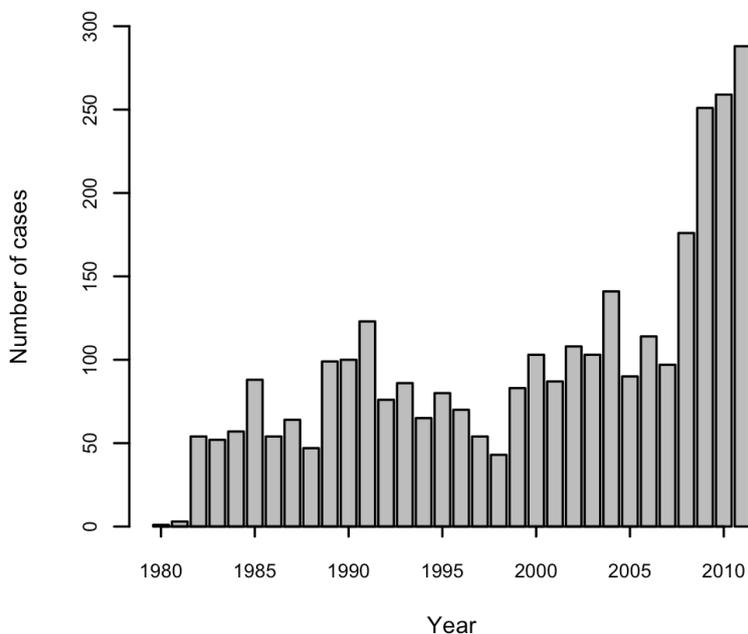


Figure 7. Number of treated latent TB infection case notifications in the Birmingham TB register, 1980 to 2011.

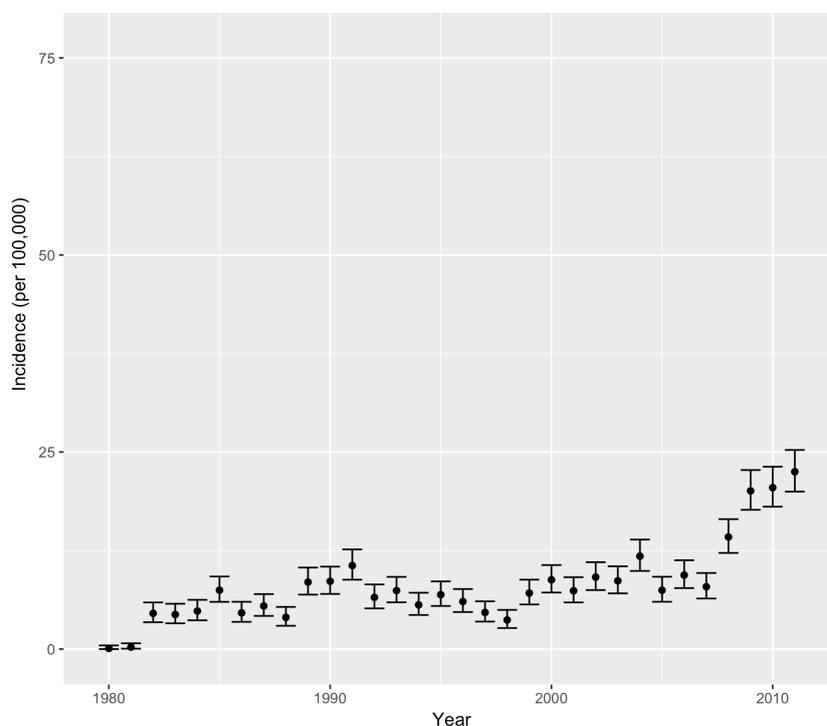


Figure 8. Annual treated latent TB case incidence in Birmingham and Solihull, 1980 to 2011. Population estimates were extrapolated from census data (Office for National Statistics, 1997; 2000; 2005; 2016). Vertical bars denote 95% confidence interval calculated assuming a Poisson distribution.

### **2.4.3 Contact table**

Contact tracing has always been performed in Birmingham and Solihull but contacts were only included in the TB register, in a separate table from cases, from 1987 onwards. In brief, each TB case was interviewed by a TB nurse within five working days of their diagnosis. The interview was usually undertaken within the home and occasionally in the clinic. Cases were asked to volunteer the names of individuals they had regular contact with during and 3 months prior to the infectious period. Named contacts residing at the same residential address as the index case were categorised as household contact. Those residing at a different address were categorised as a close contact. Contacts who required tracing due to exposure at a congregate setting e.g. workplace or school, but not named as contacts at interview, were categorised as casual contacts. For non-infectious cases only household members were contact traced.

Collection of casual contact data was done on a case-by-case basis. If an infectious case was known to have regular attendance within an indoor congregate setting (e.g. school or workplace) a risk assessment was done by the clinical and public health team. This usually involved a visit to the congregate setting location. Casual contacts were then sought depending on perceived infectiousness of the index case, e.g. known transmission to household and contacts already examined, duration of symptoms, site of TB especially if laryngeal, duration of infectious period and vulnerability of individuals exposed at the congregate setting. This process was subjective but directed by established guidance (National Institute for Health and Clinical Excellence, 2016). Sequential contact tracing was usually the norm, starting with those in closest proximity to a case e.g. sharing a class and progressing to those that are presumed to have less intimate contact e.g. being in the same class year but not sharing a class.

Contacts had their index case details recorded in the contacts table. Occasionally, contacts were referred in from TB programmes outwith Birmingham because their index case resided there while the contact resided in the Birmingham catchment area.

Each contact was screened for symptoms by interview and further evaluated for latent TB (see section 1.4.2.2). Contacts who were uncontactable, refused screening or deemed not to need screening by the treating physician would still be entered in the contact table and incomplete follow-up documented.

A contact diagnosed with active or latent TB would be entered in both the contact and case tables. Data fields in the contact table are detailed in Appendix 1.2.

The median number of household and close, non-household contacts named by a TB case was between 1 and 3 individuals (Figure 9). Less contacts were named over time (median 2, interquartile range, IQR 0 – 7 in 1995 compared to 2, IQR 0 – 4 in 2010).

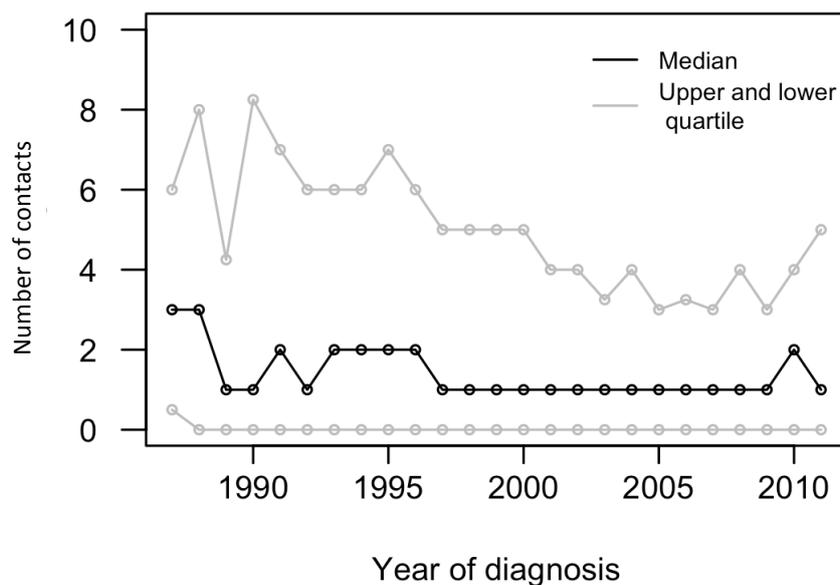


Figure 9. Number of household and close (non-household) contacts named by TB cases, 1980 to 2011.

#### 2.4.4 Demographic and clinical data

Ethnicity and birthplace were self-reported. Address was also self-reported; multiple addresses were not entered. However school and work location was recorded if known but this is not uniformly elicited (Catherine Brown, 2013, personal communication, 22 February). Social risk factors such as homelessness, alcohol history, smoking history and prison history were not always known and variably entered into the database. Medical history was also not usually recorded except when HIV co-infection was present (< 1% of cases per year).

#### 2.4.5 Laboratory methods

All microbiological specimens were processed at Public Health England Regional Centre for Mycobacteriology, Birmingham. Drug susceptibility testing was performed using the resistance ratio or proportion method (Drobniewski *et al.*, 2007). Since 2003, strain typing by 15-loci mycobacterial repetitive unit-variable number tandem repeats (MIRU-VNTR) was done prospectively on all culture-positive specimens. This typing method uses the polymerase chain reaction to quantify the number of repetitive sequences, or tandem repeats, at multiple independent genetic loci across the *M. tuberculosis* genome (Barnes and Cave, 2003; P Supply *et al.*, 2001). The number of repeats at each loci is variable and distinguishes relatedness between strains.

Typing using 12 loci was shown to be very stable and reproducible, but did not provide sufficient differentiation for isolates in the Beijing clade (Kremer *et al.*, 2004; Supply *et al.*, 2006). Additional VNTR loci can provide higher resolution, and subsequently 24-loci MIRU-VNTR typing has become the international standard (Nikolayevskyy *et al.*, 2016; Supply *et al.*, 2006). From 2010 onwards, 24-loci typing became routine in Birmingham. Selected isolates prior to 2003 were re-typed retrospectively if they potentially belonged to a cluster, defined as having 2 or more patients with indistinguishable typing at all loci, with up to one missing loci allowed (Public Health England, 2014).

Typing data, which consisted of concatenated numbers representing copy numbers at each loci, were entered manually into the Birmingham TB register from laboratory reports and therefore subject to errors.

## **2.5 The research dataset**

Rationale for the creation of a research dataset from the electronic Birmingham TB register is discussed in Chapter 1. This section describes methods used in preparing the research dataset.

All data from the January 1980 to December 2011, including personal identifiers, were extracted for the research dataset. Data after December 2011 were not included for pragmatic reasons i.e. to enable sufficient time for ethical approval of the research prior to the start of the research period for the thesis in October 2012.

Data were manipulated using Microsoft Access 2010 and *R* version 2.14.2 (R Core Team, 2017).

### **2.5.1 Deduplication of the clinical database**

Individuals entered more than once in each of the case and contact tables could not be recognised as such in the clinical database. Therefore deduplication across both the case and contact tables was performed. A deterministic match using first name, last name, date of birth and one of either hospital number, NHS registration number, address or telephone number was employed to identify individuals within the data. The following minor mismatches were tolerated: up to 2 character differences in names; single difference in either day or month in date of birth (except in cases where the date of birth was entered as 1<sup>st</sup> January or was blank, an individual was matched if the year of birth, calculated from age, matched); difference in house number of address; or if the fourth required field was blank or unknown. For individuals with agreement on first name, last name and date of birth but who had no available hospital or NHS registration number and were discrepant on address or telephone number, a match was also made if further confirmation by either case note review or discussion with the clinical team was obtained. This procedure was

designed to ensure that matches were accurate at the expense of potentially missing matches, although no formal estimates of specificity and sensitivity are possible.

### **2.5.2 Pseudonymisation**

The final tables contained 14,909 case episodes and 56,615 contact episodes. Unique individuals (N=63,620) were given new numerical identifiers. All personal identifiers were removed. All dates apart from notification and contact dates were reduced to years. The last two letters of postcodes were removed.

### **2.6 Ethical approval**

Ethical approval to pseudonymise and analyse the Birmingham TB register was granted by the National Research Ethics Service, NHS Health Research Authority (reference number 13/EM/0126).

### **2.7 Summary**

The large clinical dataset of TB cases and contacts in Birmingham, UK is a register of all individuals treated for TB (disease and latent infection) or named as a TB contact since 1980 in a defined geographical region. Data was censored at end 2011 to create a pseudonymised research dataset used in this thesis. A conservative approach to identify repeat individuals was taken to ensure accuracy at the expense of missing some individuals.

We can be reasonably certain of the register's completeness with regards to the number of active TB cases but there are several sources of detection bias in ascertainment of latent TB infection cases and TB contacts. Tests for latent TB infection was only offered to select age groups and those diagnosed but not treated were not always recorded. Naming of contacts via contact tracing was voluntary and the extent of contacts asked for depended on case infectiousness but also directed by subjective clinical assessment on a case-by-case basis. Changes in clinical protocols over time have likely further impacted on latent TB infection and TB contact recognition and the time points at which these changes occurred were not always clear. Furthermore demographic and clinical data entered had variable completeness and formal estimates of precision in data entry are available.

### **3 Time-dependent risks of recurrent tuberculosis treatment episodes**

#### **3.1 Introduction**

The standard treatment regimen for tuberculosis (TB) since drug trials began in 1946 has remained unchanged until today. To achieve durable cure six months of four drugs (isoniazid, rifampicin, ethambutol and pyrazinamide) are given for two months, followed by two drugs for a further four months. Global expansion of treatment by directly observed therapy (DOT), necessary to avoid incomplete treatment (due to the prolonged course of therapy) and the risk of resistance, has been a key strategy in TB control that has led to a 50% decline in global incidence rates compared to 1990 and nearly halving of mortality rates in 2015 (World Health Organization, 2017).

Retreatment cases attributable to both relapse and reinfection may signal gaps in control programmes and therefore deserve investigation. Under trial conditions 3.4% relapse after five years in established treatment regimens (Hong Kong Chest Service and British Medical Research Council, 1987) . Current understanding of retreatment relies largely on patient recall of previous episodes with 10-20% and 4-13% reporting at least one previous diagnosis/treatment in low and high-income countries respectively (Kim *et al.*, 2013; Mak *et al.*, 2008; World Health Organization, 2017)

In low-incidence countries (defined as <20 cases per 100,000) such as the UK, active case finding in at risk groups including contacts of TB cases and new migrants is an established additional control measure (National Institute for Health and Clinical Excellence, 2016). Testing and treatment of latent TB infection has been a routine component of this measure since the late 1960s for children (American Thoracic Society, 2000; Joint Tuberculosis Committee of the British Thoracic and Tuberculosis Association, 1973) and in the UK was extended to young adults under the age of 35 since 1990 (Subcommittee of the Joint Tuberculosis Committee of the

British Thoracic Society, 1990). Progression to active disease is estimated to occur in around 4% despite treatment (Ena and Valls, 2005) and overall effectiveness is only 60% with incomplete treatment adherence (Smieja *et al.*, 2000). These estimates were pooled from studies with variable follow-up times. Systematic evaluation of long-term outcomes for treated latent cases including time since treatment is thus also valuable for programme evaluation.

In this chapter we consider repeated episodes of latent and active TB diagnosis and treatment (referred to as a TB episode) to understand their contribution to TB epidemiology and its related risk factors in an urban UK setting.

## **3.2 Objective**

By identifying repeat individuals in a longitudinal TB register spanning 31-years and applying a survival approach

- Estimate short and long-term of risks of rediagnosis/retreatment after a first TB treatment episode (latent infection and active TB).
- Analyse predictors of recurrent TB treatment episodes.
- Infer the contribution of reinfection in recurrent TB treatment episodes by assessing available molecular typing data.

## **3.3 Method**

### **3.3.1 Study design and setting**

We retrospectively analysed all TB treatment episodes diagnosed between the calendar years 1980 and 2011 in the city of Birmingham and urban borough of Solihull, UK: of 3223 latent and 11,387 active cases. Cases were extracted from the research dataset created from the Birmingham TB register as described in Chapter 2. All data were collected prospectively in the course of normal healthcare provision.

### 3.3.2 Study population

The study population was all patients with a treatment episode for latent or active TB (both culture-confirmed and unconfirmed) residing in the study area. Patients that were subsequently found to have an alternative diagnosis and denotified were excluded. Patients with a previous TB treatment episode noted in the register (field 16, Appendix 1.1) but who did not have that treatment episode treated in the study area were included; the treatment episode in the study area was considered as the first. No knowledge of treatment episodes outside the study area after a first TB treatment episode in the study area was possible.

During the study period almost all active case finding effort were concentrated in contacts with only *ad hoc* cases found through screening in other at risk groups, hence the majority of latent cases were diagnosed as a result of contact to an index case. Latent cases are not statutorily reported but as all treatment were managed by specialist nurses based at the Birmingham Chest Clinic, the latent data are also likely to represent near comprehensive surveillance. Cases found to have latent infection but not given treatment were not registered and therefore could not be included in the study. However uptake of latent TB treatment was over 99% (Martin Dediccoat 2017, personal communication, 2 October).

### 3.3.3 Identification of individuals with repeat TB episodes

Repeat individuals were searched for in the dataset using a minimum of four personal identifiers: first name, last name, date of birth and one of either hospital number, NHS registration number, address or telephone number was employed to identify individuals within the data. Further details of the deduplication process are described in Chapter 2 (section 2.5.1).

### 3.3.4 Definitions

A repeat TB episode was defined as any treatment episode for latent or active TB in the same individual that was registered at least 12 months after the first episode. Pulmonary TB was defined as TB affecting the lungs, pleural cavity, mediastinal

lymph nodes or larynx. Resistant TB was defined as an isolate with documented resistance to any first-line drug, i.e. rifampicin, isoniazid, pyrazinamide or ethambutol. Ethnicity was self-reported and Indian subcontinent ethnicity included those of Pakistani, Indian or Bangladeshi origin. Treatment outcomes at 12 months from diagnosis were defined as: treatment completed if patients completed a full course of prescribed treatment and had documented culture conversion or were discharged by their attending physician, treatment ongoing if a patient remained on treatment 12 months after diagnosis for any reason, defaulted for any patient lost to follow-up and died or transferred out if death from any cause occurred or if the patient moved out of the study area and care was formally transferred elsewhere. These definitions have not changed during the time course of the data collection.

### **3.3.5 Laboratory methods**

All microbiological specimens were processed at Public Health England Regional Centre for Mycobacteriology, Birmingham. Drug susceptibility testing was performed using the resistance ratio or proportion method (Drobniewski *et al.*, 2007). Strain typing by 15 or 24-loci mycobacterial repetitive unit-variable number tandem repeats (MIRU-VNTR) (available from 2003 and 2011 for 15- and 24-loci respectively) was performed as described elsewhere (P Supply *et al.*, 2001).

### **3.3.6 Statistical analysis**

We estimated the rate of a second TB treatment episode by survival analysis. Time periods at risk for each person were defined in years as the time between date of first episode notification and date of second episode notification or censor date (date of death, transfer out of city or end of study period, 31 December 2011). A small number of individuals (17, 0.1%) had more than two TB episodes but these episodes were not further evaluated. Cumulative hazard functions for each pair of sequential episode types (active or latent infection followed by latent infection, active infection followed by active infection and latent infection followed by active infection) were plotted for time to the second TB episode only using the Nelson-Aalen estimator.

The following independent variables were fitted in a Cox proportional hazards model with stratification on TB episode number: age at first TB episode, sex, ethnic group and whether UK or non-UK born, type of first infection episode (latent; pulmonary or non-pulmonary culture-negative, culture-positive fully-sensitive or culture-positive resistant TB), treatment outcome of first episode and year of first treatment episode diagnosis. The Efron approximation was used for tied events (i.e. where individuals had a second TB treatment episode at the same time and their individual contribution to the partial likelihood for covariates could not be evaluated). Multiple observations from the same individual were adjusted for using the robust variance estimator. Ethnicity and birthplace and type of infection episodes were grouped as described as exploratory analysis revealed the individual covariates to be highly collinear. Other variables such as history of previous TB, BCG status, risk factors for TB including alcohol or drug abuse, homelessness, history of incarceration and HIV status were available but data were missing for 25%-90% of individuals and therefore the influence of these variables was not assessed.

The multivariable Cox model, stratified as above, was constructed by purposeful selection and proportionality of hazards was examined graphically from log-log hazard plots and from Schonfeld residuals. Proportionality was violated for the notification year of first episode, and in an extended Cox model controlling for all other variables the relative hazard for notification in 2000-2011 (Period 2) was twice that of notification in 1980-1999 (Period 1). Period of first notification did not alter the hazard coefficients for other covariates in the model and its effect was therefore likely due to poorer repeat episode ascertainment in the earlier half of the study period when quality of data recording was poorer. Separate models were therefore constructed for Period 1 and Period 2.

All analyses were performed using *R* version 3.1.2 (R Core Team, 2017) and the *R* survival package (Therneau, 2015).

## **3.4 Results**

### **3.4.1 Overall description of repeat TB treatment episodes**

From 14,798 registered latent and active TB episodes, 14,610 (99%) were included in the final analysis: 162 were excluded due to denotification, 14 due to unknown treatment episode type, 12 due to repeated treatment episode occurring within 12 months (Figure 10). A total of 14,312 unique individuals were identified, of whom 318 individuals (2%) reported a preceding TB treatment episode outside the study area. Overall a repeat TB episode occurred in 280 individuals (2%) (Table 1).

Figure 11A shows the cumulative hazard for any repeat TB treatment episode. The risk of any repeat TB treatment episode was highest in the first five years after the initial episode and then declined with time. This risk varied by calendar year of initial treatment episode (Figure 11B). Those initially notified in the last decade of the register (2000-2011, Period 1) had a higher risk of a second treatment episode compared to those notified in the earlier part of the register (1980-1999, Period 2) (5-year transition probability 1.7% (95% confidence interval, CI 1.4%, 2.2%) versus 0.7% (95% CI 0.6%, 0.9%); 10-year transition probability 3.1% (95% CI 2.5%, 3.8%) versus 1.4% (95% CI 1.2%, 1.7%). The risk for any repeat episode in an individual whose first episode was notified in Period 2, adjusted for age, sex, ethnicity and birthplace, type of first TB episode and treatment outcome, was twice that of an individual notified in Period 1 (adjusted HR 2.07, 95% CI 1.55, 2.77,  $p < 0.001$ ). Further analysis of repeat treatment episodes was therefore split by the notification period of the first episode.

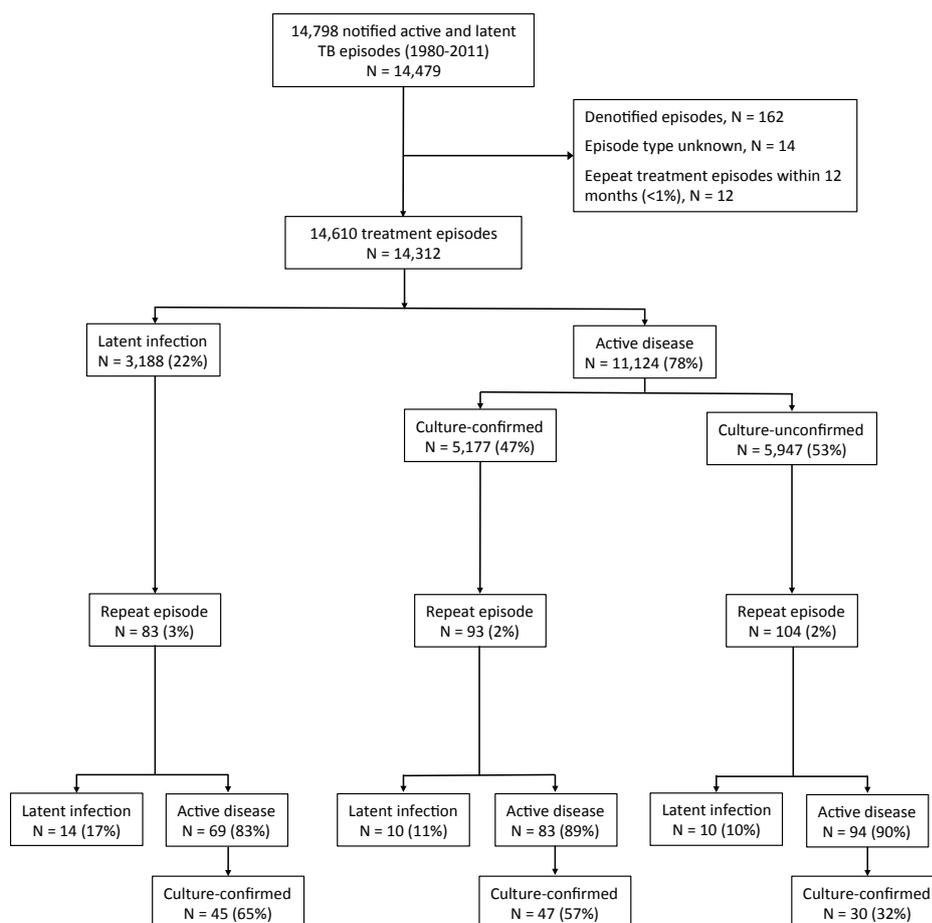


Figure 10. Active and latent TB treatment episodes notified in Birmingham and Solihull, UK, 1980-2011 and final study population included in analysis. N = number of individuals.

No. of TB treatment episodes	No. of individuals (N=14,312)
1	14,032
2	263
3	16
4	1

Table 1. Number of individuals and their TB treatment episodes from 14,610 notified active and latent cases in Birmingham and Solihull, UK, 1980-2011.

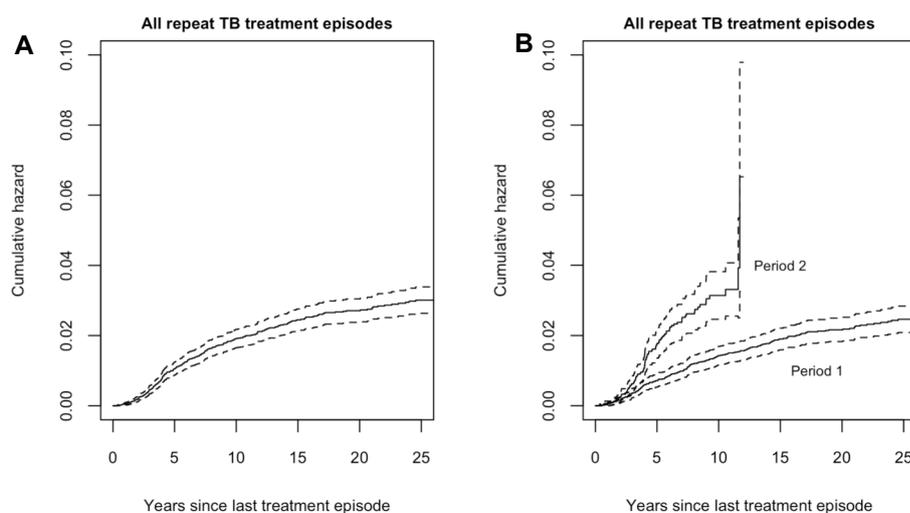


Figure 11. Cumulative hazard for any repeat TB treatment episode (A) and cumulative hazard by year of initial treatment episode (B). Period 1, first treatment episode notified in years 1980-1999; Period 2, first treatment episode notified in years 2000-2011. Dotted lines denote 95% confidence intervals.

### 3.4.2 Demographic and clinical characteristics of TB cases by calendar time

Significant changes in patient demography and the clinical characteristics of TB cases occurred over the prolonged study period (1980-2011) (Table 2). Compared to Period 1 (1980-1999) there was a large increase in the proportion of cases of Black African ethnicity from abroad (95% CI for difference in proportions 13.6%, 15.5%) in Period 2 (2000-2011), while the proportion of cases in the White, UK-born population decreased (95% CI for difference in proportions -8.2%, -5.9%). There was also marked increase in the burden of latent treatment cases (95% CI for difference in proportions 11.5%, 14.3%) and a reduction in the proportion of culture-negative cases (95% CI for difference in proportions for pulmonary cases -17.5%, -14.7%) in Period 2 compared to Period 1. There was a small increase in the proportion of isolates with resistance to any first-line drug (95% CI for difference in proportion 1.2%, 2%). The proportion completing treatment within 12 months was also lower in Period 2 (95% CI for difference in proportions -6.8%, -5.1%) with a small increase in patients defaulting treatment (95% CI for difference in proportions 2.1%, 3.1%). The proportion of missing demographic data (age, sex, ethnicity and birthplace) reduced in Period 2 compared to Period 1.

Chapter 3: Time-dependent risks of recurrent tuberculosis treatment episodes

Variable	Period 1 (1980-1999) (N=7,854), N (% or range)	Period 2 (2000-2011) (N=6,458), N (% or range)	95% CI for difference in proportions (%)	P <sup>a</sup>
Age at 1 <sup>st</sup> episode (median years, IQR)	27 (12-31)	29 (17-33)	0.99, 2 (years)	<0.001 <sup>b</sup>
Sex				
Male	4,004 (51)	3,364 (52)	-0.5, 2.6	0.186
Female	3,843 (49)	3,094 (48)	-2.7, 0.6	0.224
Not known	7 (<1)	0	-0.2,-0.02	0.016
Ethnicity and birthplace				
White, UK	1,475 (19)	759 (12)	-8.2, -5.9	<0.001
White, abroad	239 (3)	54 (1)	-2.6, -1.8	<0.001
Black Caribbean, UK	407 (5)	302 (5)	-1.2, 0.2	0.165
Black Caribbean, abroad	171 (2)	91 (1)	-1.2, -0.3	<0.001
Indian subcontinent, UK	1,167 (15)	1,348 (21)	4.8, 7.3	<0.001
Indian subcontinent, abroad	3,529 (45)	2,379 (37)	-9.7, -6.5	<0.001
Black African, UK	11 (<1)	75 (1)	0.7, 1.3	<0.001
Black African, abroad	95 (1)	1,019 (16)	13.6, 15.5	<0.001
Other, UK	61 (1)	67 (1)	-0.1, 0.6	0.099
Other, abroad	244 (3)	363 (6)	1.8, 3.2	<0.001
Not known	455 (6)	1 (<1)	-6.3, -5.3	<0.001
Type of 1 <sup>st</sup> episode				
Latent	1,292 (16)	1,896 (29)	11.5, 14.3	<0.001
NP or P with any resistance	26 (<1)	125 (2)	1.2, 2	<0.001
NP, sensitive	410 (5)	721 (11)	5, 6.9	<0.001
P, sensitive	2,112 (27)	1,783 (28)	-0.8, 2.2	0.337
NP, culture negative	1,533 (20)	736 (11)	-9.3, -6.9	<0.001
P, culture negative	2,481 (32)	1,197 (19)	-17.5, -14.7	<0.001
Treatment completed	7,514 (96)	5,796 (90)	-6.8, -5.1	<0.001
Treatment ongoing	29 (<1)	70 (1)	0.4, 1	<0.001
Defaulted	53 (1)	212 (3)	2.1, 3.1	<0.001
Died or transferred out	240 (3)	357 (6)	1.8, 3.1	<0.001
Not known	18 (<1)	23 (<1)	-0.1, 0.3	0.157

IQR, interquartile range; NP, non-pulmonary TB; P, pulmonary TB.

<sup>a</sup>Chi-square test, unless otherwise stated.

<sup>b</sup>Wilcoxon rank sums test.

Table 2. Comparison of demography and clinical characteristics of initial TB treatment episodes in Period 1 (1980-1999) and Period 2 (2000-2011).

### **3.4.3 Risks and predictors for active TB after latent infection treatment**

Sixty-nine of 3,188 individuals (2.2%) treated for latent infection developed active TB (Table 1). The median time between treatment episodes was 5.7 years (interquartile range, IQR 3.3, 7.1 years). Figure 12A shows the cumulative hazard plot for active disease after treatment for latent infection. This risk of developing active disease accumulated at a constant rate for up to 10 years following latent infection treatment (5 and 10-year transition probability in Period 1 1.2% (95% CI 0.7%, 1.9%) and 2.6% (95% CI 1.9%, 3.6%) respectively). The risk for those who were treated for latent infection in Period 2 was no different (5 and 10-year transition probability in Period 2 1.6% (95% CI 0.9%, 2.5%) and 2.3% (95% CI 1.4%, 3.7%) respectively).

The demographic profile of patients treated for latent TB infection is shown in Table 3. Age, being of Black African (born outside the UK) ethnicity and year of first latent infection treatment episode were significant predictors on univariable analysis. In the multivariable Cox model only the year of first episode predicted development of active TB after latent infection treatment (adjusted hazard ratio, HR 1.05 (95% CI 1.02, 1.09)).

### **3.4.4 Risks and predictors of recurrent active TB**

A total of 177 of 11,124 individuals (1.6%) treated for active TB developed a further episode of active TB (Table 1). The median time to this event was 5.5 years (IQR 3.3, 7.3 years). The cumulative hazard for recurrent active TB for individuals with a first treatment episode in Period 2 was similar to the risk of developing active TB following latent infection treatment (Figure 12B) (5 and 10-year transition probability 1.4% (95% CI 1%, 1.9%) and 2.7% (95% CI 2%, 3.5%) respectively). However individuals with a first treatment episode in Period 1 had a much lower risk of recurrent active TB (5 and 10-year transition probability 0.6% (95% CI 0.4%, 0.8%) and 1.1% (95% CI 0.8%, 1.4%) respectively).

Table 4 shows the demographic and clinical characteristics of patients with a single active TB treatment episode versus recurrent active TB treatment. On univariable analysis, patients with recurrent active TB were more likely to be older, of Indian subcontinent ethnicity (both UK and non-UK born) or other ethnicity (non-UK born), have a drug-resistant isolate (pulmonary or non-pulmonary disease) and either defaulted or still on treatment at 12 months from diagnosis compared to patients a single treatment episode. Recurrent active TB was also more likely if the first treatment episode was later in the study period. The final multivariable Cox model included all the above variables except being of other ethnicity (non-UK born) and treatment default at 12 months. There were no significant interaction effects between the variables. Being on treatment at 12 months after diagnosis of a first episode had the highest risk for recurrent active TB (adjusted hazard ratio 5.27, 95% CI 2.48, 11.2). Compared to the White, UK-born population ethnic groups from the Indian subcontinent had an adjusted hazard ratio of 2.67 (95% CI 1.38, 5.15) and 1.79 (95% CI 1.07, 2.99) if UK-born and non-UK born respectively. The adjusted hazard ratios for age at first episode and year of first diagnosis were lower (1.06 (95% CI 1.03, 1.09) and 1.04 (95% CI 1.01, 1.06) respectively).

#### **3.4.4.1 Strain typing in recurrent active TB**

Of 177 patients with repeat treatment for active TB, 47 (27%) had culture-confirmation in both episodes. Paired MIRU-VNTR strain typing was available in 17/47 (36%) of patients. Typing was identical for 15/17 (88%) with an average of 4 years (range 2-7 years) between relapse. Two pairs were identical at 24-loci, the rest were identical at 15-loci. Five of these cases clustered in three strain-type clusters in which the majority of individuals had social risk factors (alcohol, drugs and imprisonment). A further five clustered in a single 15-loci strain-type cluster associated with migrants from the Indian subcontinent and the remaining five were unique strain-type clusters. Of the two reinfection cases (different strain types in repeat episodes at 15-loci), one cultured 2 of the 3 strain-types associated with alcohol, drugs and imprisonment and the other had two different unique strain-types. The time to second episode was seven and four years respectively.

#### **3.4.5 Risks of latent infection treatment after active TB or latent infection treatment**

Only a small number of individuals (34/14,312, 0.2%) were treated for latent infection after an initial treatment episode for either latent infection or active TB (Table 1). The median time to this event was 7.7 years (IQR 4, 11.7 years). The cumulative hazard for this event was different for individuals with first treatment episode in Period 1 and Period 2 (Figure 12C). After any treatment episode in Period 1, the 5- and 10-year transition probability to latent infection treatment was 0.03% (95% CI <0.01%, 0.1%) and 0.1% (95% CI 0.06%, 0.2%) respectively. For those with a first treatment episode in Period 2, transition probabilities at five and 10 years were higher at 0.4% (95% CI 0.2%, 0.7%) and 0.7% (95% CI 0.4%, 1.2%).

In the univariable Cox model the only predictor of this transition was year of the first treatment episode (unadjusted hazard ratio 1.09 (95% CI 1.05, 1.13)).

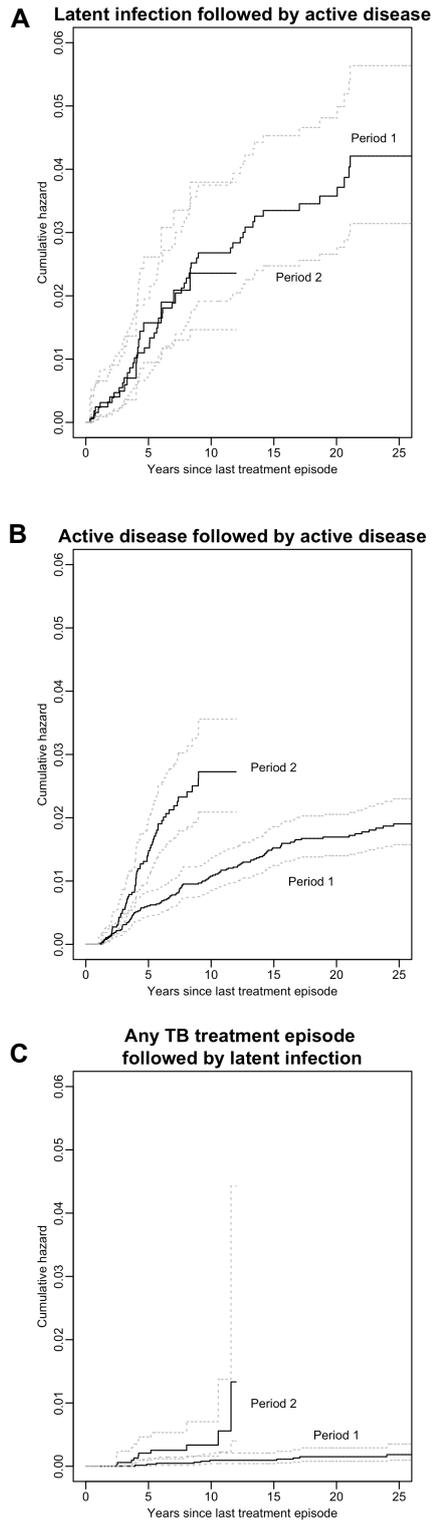


Figure 12. Cumulative hazard of repeat TB treatment episodes by episode types. Dotted lines denote 95% confidence intervals. Period 1, first treatment episode notified in years 1980-1999; Period 2, first treatment episode notified in years 2000-2011.

Chapter 3: Time-dependent risks of recurrent tuberculosis treatment episodes

Variable	Single latent treatment episode, N (% or range)	Recurrent active TB treatment, N (% or range)	Unadjusted HR (95% CI)	P	Adjusted HR (95% CI)	P	
Age at 1 <sup>st</sup> treatment episode (median years, IQR)	3,105 13 (6 – 24)	69 13 (7 – 18)					
Sex	Male	32 (46.4)	1.06 (1.01, 1.1)	0.01	0.99 (0.94, 1.05)	0.774	
	Female	1,556 (50.1)	37 (3.6)	1.15 (0.72, 1.84)	1.13 (0.7, 1.81)	0.611	
Ethnicity and birthplace	Not known	2 (<0.1)	0	0 (0, Infinite)	0 (0, Infinite)	0.999	
	White, UK	480 (15.5)	10 (14.5)	1	1	1	
	White, abroad	21 (0.7)	0	0 (0, Infinite)	0 (0, Infinite)	0.998	
	Black Caribbean, UK	185 (6)	5 (7.3)	1.07 (0.36, 3.16)	0.902	0.98 (0.33, 2.89)	0.969
	Black Caribbean, abroad	12 (0.4)	0	0 (0, Infinite)	0.998	0 (0, Infinite)	0.998
	Indian subcontinent, UK	1,132 (36.5)	28 (40.6)	1.59 (0.76, 3.32)	0.217	1.36 (0.65, 2.86)	0.414
	Indian subcontinent, abroad	684 (22)	17 (24.6)	0.88 (0.4, 2)	0.76	0.95 (0.42, 2.13)	0.903
	Black African, UK	53 (1.7)	0	0 (0, Infinite)	0.998	0 (0, Infinite)	0.998
	Black African, abroad	288 (9.3)	8 (11.6)	3.57 (1.4, 9.14)	0.008	2.08 (0.77, 5.6)	0.146
	Other, UK	62 (2)	1 (1.5)	1.17 (0.15, 9.2)	0.870	0.97 (0.12, 7.65)	0.975
Year of 1 <sup>st</sup> treatment episode (median, IQR)	Other, abroad	167 (5.4)	0	0 (0, Infinite)	0 (0, Infinite)	0.996	
	Not known	20 (0.6)	0	0 (0, Infinite)	0 (0, Infinite)	0.994	
Year of 1 <sup>st</sup> treatment episode (median, IQR)	2003 (1992 – 2009)	1995 (1991 – 2001)	1.07 (1.04, 1.1)	<0.001	1.05 (1.02, 1.09)	0.004	

IQR, interquartile range; HR, hazard ratio; NR, non-pulmonary TB; R, pulmonary TB.

Table 3. Demography and Cox model to determine predictors of active TB after a first treatment episode for latent TB infection.

Chapter 3: Time-dependent risks of recurrent tuberculosis treatment episodes

Variable	Single treatment episode, N (% or range)	Recurrent active TB treatment, N (% or range)	Unadjusted HR (95% CI)	P	Adjusted HR (95% CI)	P
Age at 1 <sup>st</sup> treatment episode (median years, IQR)	N=10,927 32 (20 – 53)	N=177 33 (22 – 50)	1.09 (1.06, 1.11)	0.014	1.06 (1.03, 1.09)	<0.001
Sex	Male	5,698 (52.1)	1	1	1	0.096
	Female	5,224 (47.8)	98 (55.4)	1.29 (0.95, 1.73)	1.29 (0.95, 1.75)	0.994
	Not known	5 (<1)	0	0 (0, Infinite)	0 (0, Infinite)	
Ethnicity and birthplace	White, UK	1,723 (15.8)	18 (10.2)	1	1	0.954
	White, abroad	269 (2.5)	3 (1.7)	0.87 (0.26, 2.98)	0.97 (0.28, 3.3)	0.163
	Black Caribbean, UK	508 (4.7)	10 (5.7)	1.92 (0.87, 4.24)	1.77 (0.79, 3.83)	0.552
	Black Caribbean, abroad	246 (2.3)	4 (2.3)	1.4 (0.47, 4.16)	1.39 (0.47, 4.13)	0.003
	Indian subcontinent, UK	1,309 (12)	28 (15.8)	2.41 (1.28, 4.55)	2.67 (1.38, 5.15)	0.027
	Indian subcontinent, abroad	5,095 (46.6)	101 (57.1)	1.86 (1.12, 3.09)	1.79 (1.07, 2.99)	0.998
	Black African, UK	33 (0.3)	0	0 (0, Infinite)	0 (0, Infinite)	0.271
	Black African, abroad	813 (7.4)	2 (1.7)	0.79 (0.23, 2.72)	0.5 (0.14, 1.73)	0.998
	Other, UK	65 (0.6)	0	0 (0, Infinite)	0 (0, Infinite)	0.068
	Other, abroad	431 (3.9)	9 (5.1)	2.28 (1.02, 5.09)	2.13 (0.95, 4.81)	0.318
Type of 1 <sup>st</sup> treatment episode	Not known	435 (4)	1 (0.6)	0.21 (0.03, 1.61)	0.35 (0.05, 2.71)	
	P, culture positive, no resistance	3,820 (34)	66 (37.3)	1	1	0.73
	P, culture negative	3,615 (33.1)	56 (31.6)	0.79 (0.55, 1.13)	0.201	0.108
	NP, culture positive, no resistance	1,118 (10.2)	12 (6.8)	0.82 (0.44, 1.52)	0.53	0.418
	NP, culture negative	2,228 (20.4)	38 (21.5)	0.8 (0.54, 1.2)	0.279	0.367
	NP or P with any resistance	146 (1.3)	5 (2.8)	3.78 (1.52, 9.41)	0.004	
	Treatment completed	9,960 (91.2)	161 (91)		1	
12-month outcome of 1 <sup>st</sup> treatment episode	Treatment ongoing	90 (0.8)	8 (4.5)	8.82 (4.33, 17.9)	5.27 (2.48, 11.2)	<0.001
	Defaulted	248 (2.3)	6 (3.4)	3.07 (1.36, 6.94)	1.72 (0.75, 3.9)	0.2
	Died or transferred out	589 (5.4)	1 (0.6)	0.56 (0.08, 4.04)	0.34 (0.05, 2.4)	0.274
	Not known	40 (0.4)	1 (0.6)	2.09 (0.29, 14.94)	0.464	0.772
	Year of 1 <sup>st</sup> treatment episode (median, IQR)	1996 (1986 – 2005)	1994 (1987 – 2002)	1.06 (1.05, 1.08)	<0.001	1.04 (1.01, 1.06)

IQR, interquartile range; HR, hazard ratio; NR, non-pulmonary TB; R, pulmonary TB.

Table 4. Demography, clinical characteristics and Cox model to determine predictors of recurrent active TB treatment episodes.

### 3.5 Discussion

This work adds to a contemporary understanding of retreatment TB in a setting where the disease is variably distributed in the population, particularly in migrants and those with social risk factors.

Incomplete ascertainment of repeat individuals was a major limitation of our study. This would explain the higher risk of recurrent active TB episodes in Period 2 compared to Period 1. Our findings are therefore the absolute minimum estimates of risk with true estimates likely to be closer to findings from Period 2, i.e. twice as high as Period 1. The analysis also misses individuals rediagnosed/retreated outside the study area. Similar observational/surveillance data recording elsewhere, in particular of latent TB diagnosis or treatment currently not included in national surveillance would be highly beneficial in refining these estimates. Data were poor for previous TB and vaccination status, indicating that patient recall is unreliable for estimation of repeat episodes.

The estimated rate of recurrent active TB in our area is low, consistent with findings from other high-income country settings (Bang *et al.*, 2010; Driver *et al.*, 2001; Pascopella *et al.*, 2011). Previous estimates were limited to culture-positive TB only and may have underestimated risk in a real clinical setting. The limited molecular fingerprinting data presented suggests that relapse (i.e. treatment failure) is more common than reinfection in our setting unless reinfection is from a common source. With strain typing routinely available, further increase in the culture rate for TB diagnosis is needed to confirm this finding. However data from low-incidence settings as well as some higher-incidence settings are similar (Bang *et al.*, 2010; Bryant *et al.*, 2013).

There was no differential risk for active TB after latent infection in Period 1 and Period 2, perhaps due to increased ascertainment in Period 1 for children who comprised the bulk of latent infection cases at the time (in practice treatment was extended to those up to age 35 only in 2007 in Birmingham). Thus the burden of retreatment TB in similar settings is likely to come from both latent and active cases. Upscale of latent TB diagnosis and treatment is an increasing priority (Pareek *et al.*,

2011) and our estimates of rates of progression to active disease, with the highest risk persisting up to 10 years after treatment, should inform clinicians and patients so that any subsequent active disease can be diagnosed at the earliest opportunity. DOT is only utilised for those with risk factors for non-adherence in the UK (National Institute for Health and Clinical Excellence, 2016) so adherence in latent treatment may be less complete than if DOT was routine. Using clinical data to distinguish recent from remote infection may enable prioritisation of those at higher risk of progression to active disease that would benefit from augmented treatment with weekly DOT using the simplified 12-dose rifapentine and isoniazid regime (Sterling *et al.*, 2011). Progression to active disease after latent treatment may also be host-dependent as differential patterns of host gene expression (modulating the immune response to infection) have been observed in patients with recurrent active TB compared to those with only a single episode of disease or latent infection (Mistry *et al.*, 2007; Tufariello *et al.*, 2003). Current diagnostic tools do not allow distinction between active disease and latent infection and disease may be subclinical. Computed tomography imaging has been proposed as an adjunct to better detect subclinical disease (Fujikawa *et al.*, 2014) but will have cost implications that need further evaluation.

The risk of repeat treatment for latent infection after latent or active TB was much lower. Data from Period 1 suggests that there is minimal incremental risk over time. Estimates from Period 2 are higher but longer follow-up time is required to draw conclusions about continuing risk beyond 10 years after the first episode.

Continuing treatment beyond 12 months from diagnosis had 4-5 times the hazard ratio for a future active TB episode compared to treatment completion. There may be a need for clinical vigilance in such cases (Kim *et al.*, 2013) with targeted interventions to ensure both treatment completion and selected extended follow-up, particularly as some of these patients will be those with a drug resistant isolate. Being of Indian subcontinent ethnicity independently increased the risk of recurrent active TB. The reason for this is unclear and no data in the literature suggests a biological basis for this. Thus one explanation for a differential risk based on ethnicity could be different re-exposure risks.

Age at first active TB treatment episode was also significant predictor in the multivariable Cox model for recurrent active TB. Risk of infection as well as disease following infection is strongly age-related (Rieder *et al.*, 1990; Stead and Lofgren, 1983). However a modelling study based on UK notification data from 1953 to 1988 found that in adolescents and adults the risk of developing disease following reinfection was lower than the risk of developing disease after the first infection, suggesting that previous infection provides some protection (Vynnycky and Fine, 1997). Based on the limited molecular data available recurrent active TB in our setting appears to be mostly related to relapse rather than re-infection and therefore the protective effect of previous infection may not apply.

Successful strategies for TB elimination will rely on an integrated approach of vaccination and diagnosis and treatment for both active disease and latent infection (Dye *et al.*, 2013). The relative contribution of each strategy to achieve this goal is likely to vary with the changing epidemiology of the epidemic in different settings and setting-specific modeling is crucial to refine efforts in control. Quantifying risks for repeated TB episodes may further contribute to model development. Such analyses require good epidemiological data gathering over extended periods. We have demonstrated the need to recognise repeat individuals over time to better understand the natural history of TB in our setting. Retreatment TB cases will arise from latent as well as active cases but existing national surveillance does not include data on latent infection diagnosis/treatment. We suggest that both the recording of patient identity and latent TB cases are important.

## **4 What happens to tuberculosis contacts? Competing risks application to estimate risks of infection with single versus repeat exposures.**

### **4.1 Introduction**

In the United Kingdom (UK) tuberculosis (TB) diagnosis and treatment is readily accessible through comprehensive primary and secondary care health services. Besides early diagnosis and treatment, the mainstay of TB control is active case finding in at-risk groups, namely contacts of TB patients and new entrants to the UK (National Institute for Health and Clinical Excellence, 2016).

Contact tracing is a key strategy of infectious diseases control. By identifying, isolating and/or treating individuals exposed to an infectious disease, further chains of infection are interrupted and disease spread curtailed. Contact tracing was first used successfully in sexually transmitted diseases (Wigfield, 1972) and is a recommended policy for controlling new or re-emerging infections such as severe acute respiratory syndrome (Donnelly *et al.*, 2003) and Ebola (Rivers *et al.*, 2014; World Health Organization, 2014).

Contact tracing in tuberculosis emerged with the availability of effective drug regimens for both latent infection and disease in the 1950s (Ferebee and Mount, 1962; Hsu, 1963). The prevailing approach follows the “-in-the-pond” principle (Veen, 1992) originally described by Van Geuns (Van Geuns *et al.*, 1950) who noted that more infected individuals were found with increasing intimacy of contact to the investigated case. In settings where the number of infectious cases is low and thus the overall risk of infection in the community is also low, the observed prevalence of infection or disease detected in individuals forming the most intimate contact ring of a case can be compared to that expected in the general population for the same age. Less intimate rings of contacts should be investigated until the observed prevalence in contacts is similar to the expected prevalence. This staged, directed process aims to identify all possible concurrent sources of infection in a

microepidemic, with new waves of concentric circles of contact tracing initiated from each case found. In Veen's original description, contacts diagnosed with latent infection should not have their source of infection assigned to the known index case until their own contacts are assessed.

Multiple factors determine disease transmission to contacts of infectious TB cases, i.e. pulmonary or laryngeal cases that generate airborne droplet nuclei containing TB bacilli. Concentration of droplet nuclei is affected by airflow which depends on ventilation and air circulation where contact occurs (Houk *et al.*, 1968). The amount of exposure, i.e. frequency and duration of contact is also important. In practice, contacts are usually considered at risk if the cumulative exposure time is 8 or more hours; this duration was derived from observations of higher risk of transmission for airline passengers in the same or adjoining row (Driver *et al.*, 1994; Kenyon *et al.*, 1996). The relative contributions of proximity to the infectious case and length of exposure whether brief but recurrent or prolonged to transmission is not well understood, nor is it well quantified in practice.

Certain characteristics of the infectious case affect the likelihood of transmission. The burden of infection, as measured independently by sputum-smear positivity, radiographic observation of lung cavities and shorter time to growth of *M. tuberculosis* (<9 days) in culture, correlates well with transmission (Bailey *et al.*, 2002; Marks *et al.*, 2000; O'Shea *et al.*, 2014). Once infected, the risk of progression to disease is age dependent being highest for children under 5 years and post-pubertal adolescents (Comstock *et al.*, 1974) and in those with impaired immunity for example due to HIV infection (Antonucci *et al.*, 1995), systemic diseases such as diabetes (Dillon *et al.*, 1952) or use of immunosuppressants such as anti-tumour necrosis factor agents (Gardam *et al.*, 2003).

The number of cases detected by contact tracing is higher in settings where there is a greater overall burden of disease with approximately 50% of contacts found to have infection in low and middle-income countries vs. 30% in high-income countries (Fox *et al.*, 2013). Contact tracing may find cases both related and unrelated to the known source case. For example, contacts of non-infectious cases are also found to have

infection (Mandal *et al.*, 2012; Underwood *et al.*, 2003). Transmission in these cases could have been from remote exposure or recent exposure to an unknown infectious case. Inferring transmission from the known source case requires matching TB isolates by typing in both case and contact or documenting conversion of a previously negative TB-specific cellular immune response to positive. Up to 40% of treated TB cases are culture-negative (Public Health England, 2017b) and therefore unavailable for molecular typing. Furthermore an estimated 75-95% of contacts will already be infected by the time of tracing but few (1-3%) have had the time to manifest disease (Kasaie *et al.*, 2014). Latent infection can be tested for but requires recent seroconversion to deduce likely transmission and this is seldom confirmed. Measuring linked transmission in contacts is therefore problematic. However because contact tracing accesses those at risk regardless of whether transmission occurred directly from known sources, studying contact characteristics and their long-term interactions, for example recurrent exposures may provide insight into network structure and ways to improve case detection (Read *et al.*, 2008; Smieszek *et al.*, 2009).

The overall yield of case finding through contact tracing is easier to measure. However reported evaluations may be underestimated due to incomplete follow-up of contacts (up to 40%) (Cavany *et al.*, 2017; Grinsdale *et al.*, 2011; Saunders *et al.*, 2014) or limited follow-up time. Very few longitudinal studies of contact outcomes are published (Morán-Mendoza *et al.*, 2007; Sloom *et al.*, 2014) although a meta-analysis has estimated a high annual incidence of disease at 5 years following exposure (171 per 100,000) (Fox *et al.*, 2013). A further consideration when assessing yield in contact studies is that diagnosis of latent infection is only sought in younger individuals due to the excess risk of treatment-related hepatotoxicity with age (Stead *et al.*, 1987). In the UK a test for latent infection was generally only considered for those aged 35 and under until 2016 (National Institute for Health and Clinical Excellence, 2016). Thus previous UK studies of contact tracing with shorter term follow-up (Rosser *et al.*, 2018) may have underestimated the rate of infection/disease in older contacts.

TB registry data captures exposures, infection and disease events in a population over a long time course and may allow more comprehensive understanding of contact outcomes including impact of repeat exposures/contact episodes.

## **4.2 Objectives**

Using an event history approach we aim to

- Estimate the risks and timing of disease and infection in contacts with a single versus repeat contact episodes.
- Investigate risk factors for outcomes after a single contact episode.

## **4.3 Methods**

### **4.3.1 Study population**

Contacts named by all notified TB cases from 1987 to 2011 who were residing in Birmingham city and the urban borough of Solihull were included in this analysis. Cases were extracted from the Birmingham TB register as outlined in Chapter 2. Contacts were excluded if they were previously notified as a case. Contacts registered more than once within a period of 12 months and linked to the same index case were considered as duplicate registrations and only the first episode was included. TB cases were interviewed by TB nurses within five working days of their notification to elicit the names of persons they may have had regular contact with over the infectious period, defined as the date of onset of pulmonary symptoms, if known, or three months before diagnosis. For non-pulmonary cases only household contacts, defined as those who share a bedroom, kitchen, bathroom or sitting room with the case were elicited. For pulmonary cases with contacts identified at a congregate setting e.g. a workplace or school further contacts not directly named by the patient but known to have shared the enclosed space with the case were sometimes recruited if considered at risk after an on-site clinical assessment.

### **4.3.2 Contact investigation and management**

All contacts were invited to attend investigations at the local TB clinic. If numerous contacts required investigation in a congregate setting, investigations took place within the congregate setting.

All contacts were screened for active TB by history taking, clinical examination, chest radiograph or other imaging as indicated and specimens for culture were collected where possible. For asymptomatic contacts aged 35 and under with normal chest radiography, latent infection was diagnosed by a positive Mantoux test as described elsewhere (National Institute for Health and Clinical Excellence, 2016; Public Health England and Department of Health, 2013) and/or positive interferon- $\gamma$  release assay (QuantiFERON-TB Gold In-Tube, Carnegie, Cellestis, Australia or T-SPOT.TB, Oxford Immunotec, Oxford, UK) performed at least 6 weeks after their last known exposure to the source case. Prior to 2007, only those aged 16 and under were investigated for latent infection. Contacts over 35 years but who had increased susceptibility to infection were tested for latent infection at clinician's discretion.

Contacts who failed to attend any part of their investigation were given two further appointments before referral back to their primary care provider advising that the individual had been exposed to TB and should be investigated for TB if symptoms develop. Disease or latent infection was managed according to national guidelines (National Institute for Health and Clinical Excellence, 2016).

### **4.3.3 Contact outcomes**

Contacts were searched for repeat entries in the contact dataset and case registry to determine if disease, latent infection or repeated exposure to TB cases/contact episode occurred. Repeat individuals were identified as described previously. The contact dataset did record outcomes at the end of the contact investigation protocol including no infection after examination, examination incomplete, latent infection or disease but data were unreliable. For example, 869 contacts documented as either having infection or disease were not subsequently notified as cases, whereas 296 contacts documented as having no infection after examination or examination incomplete were registered as cases within a year of contact. Thus it was not possible to ascertain reliably which contacts had a negative test of infection during contact tracing. Additionally, contacts diagnosed with latent infection but not treated were not included as they were not registered as cases. Because the dataset only recorded those resident in the study area, cases that moved out were also not captured.

### **4.3.4 Analysis of risks and predictors of disease, infection and repeat TB exposure**

Contact histories were described in a multi-state model (Figure 13). The initial state was the first contact exposure (C1). From this state contacts were at risk of a second contact exposure (C2), TB disease (D1) or latent infection (L1). Following a first occurrence of D1 or L1 no further transitions were modeled, i.e. they were absorbing states. Contacts transitioning to C2 were at further risk of a third contact exposure (C3), disease (D2) or latent infection (L2). D2 and L2 were different states from D1 and L1. Further contact episodes after C3 were not modeled due to small numbers (n=23). Follow-up time was calculated from the notification date of the source case for the last contact episode. Follow-up was censored at the end of the study period (31 December 2011) or at the date of death or transfer out-of-city if recorded. Cumulative hazard rates for transition between states was enumerated using the Nelson-Aalen estimator and cumulative incidence functions for each end-point were compared using the Aalen-Johansen estimator. The Efron method was used for handling tied events.

The effect of the following covariates for each event were investigated in a cause-specific hazards analysis: age at contact episode, sex, ethnicity, whether UK or non-UK born, household or non-household exposure, index case disease (denotified or latent infection, pulmonary and non-pulmonary and whether culture positive or negative) and calendar year of contact episode. Ethnicity was self-reported and Indian subcontinent ethnicity included ethnic groups from India, Pakistan and Bangladesh. Age, household or non-household exposure and index case disease were *a priori* included in the multivariable Cox model. The final model was constructed by purposeful selection of other variables. Proportionality of hazards was examined graphically from log-log hazard plots and from Schonfield residuals.

Analyses were performed in R version 3.1.2 (R Core Team, 2017) and the following packages: Biograph version 2.0.4 (Willekens, 2016) for data preparation and formatting; survival 2.37-7 (Therneau, 2015), etm version 0.6-2 (Allignol *et al.*, 2011) and mvna version 1.2-3 (Allignol *et al.*, 2008) for hazards analyses and model selection and mstate version 0.2.7 (de Wreede *et al.*, 2011) for prediction of time-to-event for select contact characteristics.

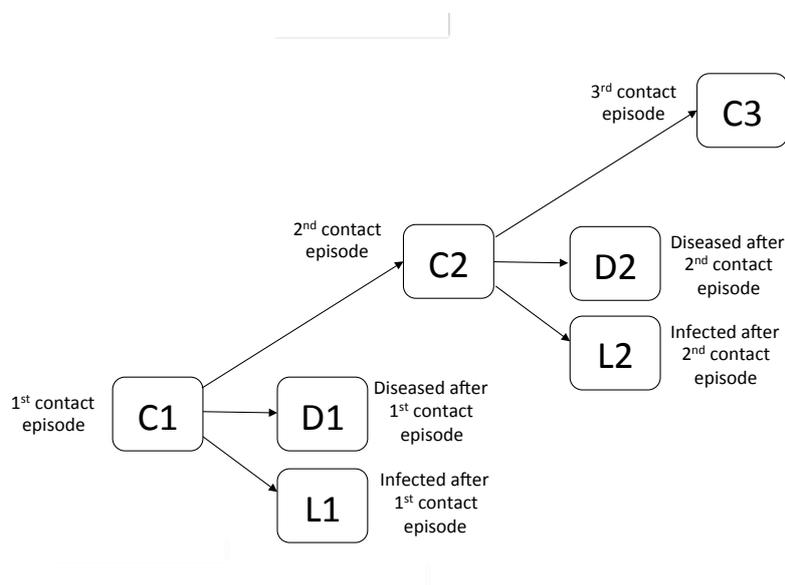


Figure 13. TB contact, disease and infection model to investigate the importance of second contact episodes.

## 4.4 Results

### 4.4.1 Overview of all contact episodes and outcomes

From January 1987 to December 2011 there were 56,615 contact episodes involving 52,776 individuals (contacts). Less than 2% of contact episodes were excluded either due to duplicate registration (n=53) or the contact had a documented previous infection from the disease register (n=961) (Figure 14). Six per cent (2913/52,383) of contacts had more than one contact episode. The proportion of contacts found infected or diseased following a first, second and third contact episode was 6%, 8% and 13% respectively.

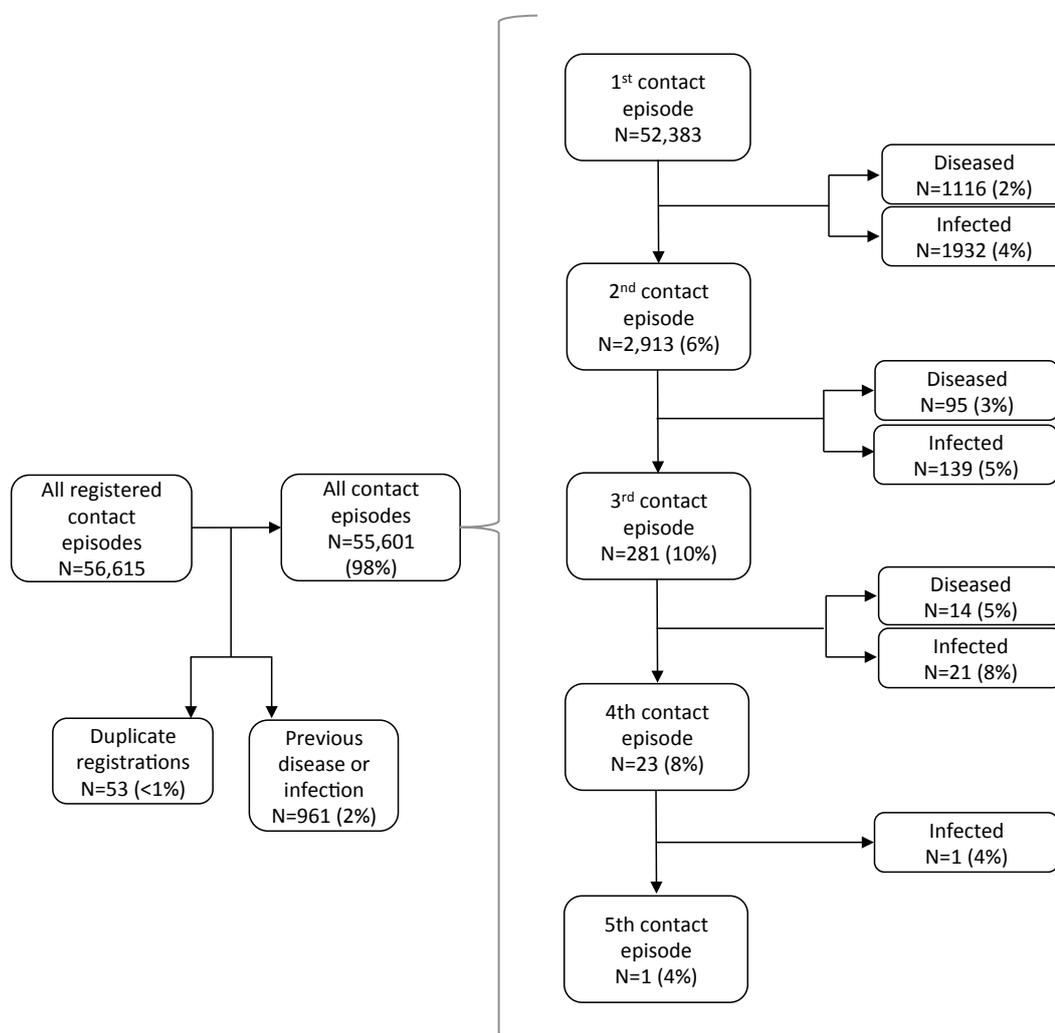


Figure 14. Overview of TB contact episodes and outcomes for 52,383 contacts (individuals).

#### **4.4.2 Study population**

There were 6530 index case notifications for the 52,383 contacts with up to two contacts episodes. The median number of contacts per index case was 5 (interquartile range, IQR, 3-9). Contacts were young with 18% aged 5 and under, 22% aged 6 to 15 and 35% aged 16 to 35 (Table 5). Fifty-one per cent were female. Country of birth and ethnicity were not recorded in 90% and 26% of contacts respectively. In the first contact episode, 56% were non-household contacts. Seventy-four per cent of contacts were contacts of pulmonary TB index cases, but only 48% of these index cases were culture-confirmed.

#### **4.4.3 Risk of infection and disease with single versus two contact episodes**

Table 6 shows the numbers at risk after a first (C1) or second (C2) contact episode and numbers transitioning to either disease (D1/D2), infection (L1/L2) or further contact episodes (C2/C3) over time. From C1, risk of C2 was lower than D1 or L1 in the first five years after contact but persisted up to 20 years such that the cumulative risk for C2 was highest overall (Figure 3 and Figure 4). Transition probabilities for these events with time are shown in Table 3. The risk of disease and infection from C2 was at least twice that from C1 (Figure 15 and Figure 16). One year after the contact episode, the probability for disease and infection from C2 were 3.9% and 10.7% (95% CI 2.9%, 5% and 9%, 12.6% respectively) versus 1.6% and 3.7% (95% CI 1.5%, 1.7% and 3.5%, 3.9%) from C1 (Table 7). Risk of disease continued for a longer time period after C2 compared to C1 (Figure 15). After 10 years, the probability of D2 was 7% (95% CI 5.7%, 8.5%) versus 2.3% (95% CI 2.1%, 2.4%) for D1 (Figure 16 and Table 7).

Chapter 4: What happens to tuberculosis contacts? Competing risks application to estimate risks of infection with single versus repeat exposures.

Variable	Transition types									
	1 <sup>st</sup> contact → no event		1 <sup>st</sup> contact → diseased		1 <sup>st</sup> contact → infected		1 <sup>st</sup> contact → 2 <sup>nd</sup> contact			
	N	%	N	%	N	%	N	%	N	%
Age (median years, IQR)	21	9-37	18	9-31	11	5-20	18	7-34		
Sex										
Male	22516	88.9	488	1.9	935	3.7	1392	5.5		
Female	23377	88.2	628	2.4	997	3.8	1519	5.7		
Unknown	529	99.6	0	0.0	0	0.0	2	0.4		
Country of birth										
UK	95	4.1	688	29.7	1333	57.4	206	8.9		
Non-UK	1555	55.9	427	15.3	597	21.5	204	7.3		
Unknown	44772	94.7	1	0.0	2	0.0	2503	5.3		
Ethnic group										
White	5058	88.6	139	2.4	311	5.4	204	3.6		
Black Caribbean	1479	79.8	128	6.9	141	7.6	112	6.0		
Indian subcontinent	23235	84.8	715	2.6	1140	4.2	2304	8.4		
Black African	1954	83.7	93	4.0	211	9.0	76	3.3		
Other	1426	86.1	39	2.4	127	7.7	64	3.9		
Unknown ethnicity	13270	98.8	2	0.0	2	0.0	153	1.1		
Household contact										
Yes	20043	86.9	608	2.6	890	3.9	1520	6.6		
No	26379	90.0	508	1.7	1042	3.6	1393	4.8		
Index case disease										
Denotified or latent infection	2255	93.2	24	1.0	46	1.9	95	3.9		
Pulmonary, culture positive	21787	86.1	796	3.2	1327	5.2	1389	5.5		
Pulmonary, culture negative	12097	91.4	181	1.4	338	2.6	615	4.7		
Non-pulmonary, culture positive	3328	90.3	47	1.3	98	2.7	212	5.8		
Non-pulmonary, culture negative	6183	93.1	60	0.9	99	1.5	300	4.5		
Unknown	772	69.8	8	0.7	24	2.2	302	27.3		
Year of contact (median, IQR)	1999	1992-2006	2000	1993-2006	2005	1998-2009	1991	1995-2001		
IQR, interquartile range										

Table 5. Demographic characteristics for 52,383 contacts and nature of their first TB contact exposure.

Years since last contact episode	<1	1-2	3-5	6-10	11-15	16-24
At risk from 1 <sup>st</sup> contact	52,383	46,814	44,321	37,396	28,162	20,281
Disease	797	126	106	62	23	2
Infection	1872	43	11	3	3	0
At risk from 2 <sup>nd</sup> contact	161	358	1,100	852	306	136
At risk from 2 <sup>nd</sup> contact	2,913	2,494	2,273	1,744	1,121	590
Disease	48	18	16	11	2	0
Infection	133	2	4	0	0	0
3 <sup>rd</sup> contact	16	37	122	78	23	5

Table 6. Numbers at risk of events (disease, infection or further contact episode) and event occurrences over time following a first and second contact episode.

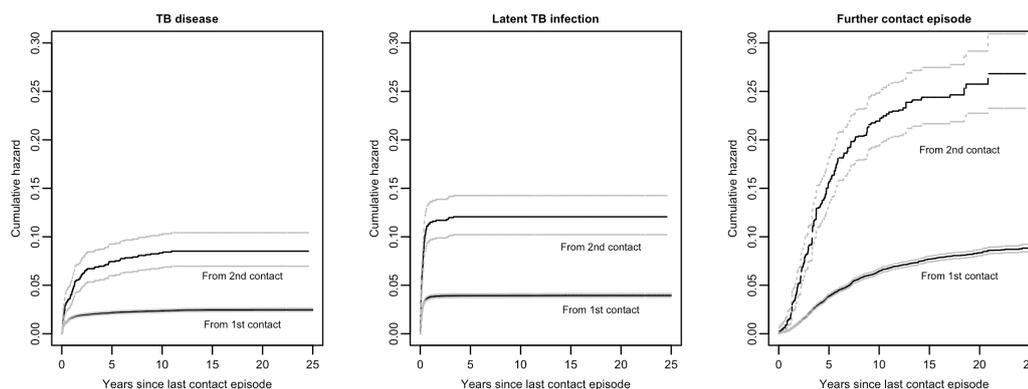


Figure 15. Nelson-Aalen estimates of the cumulative hazard for diagnosis of TB disease, latent TB infection and further contact episode after a first or second contact episode. Dotted lines denote log-transformed 95% confidence intervals.

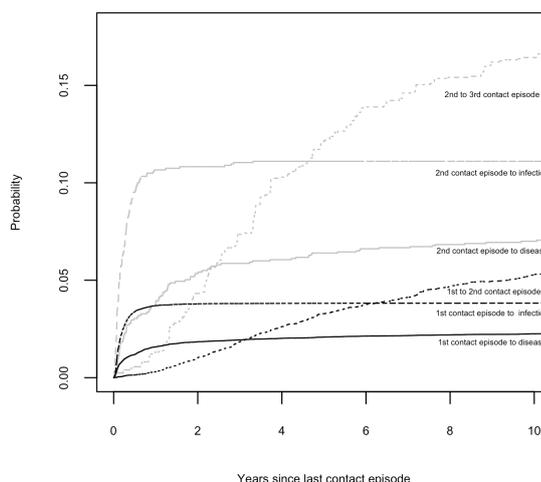


Figure 16. Aalen-Johansen estimates of the cumulative incidence function for all event types following a 1<sup>st</sup> and 2<sup>nd</sup> contact episode.

Years since last contact episode	Disease			
	After 1st contact episode		After 2nd contact episodes	
	Probability	95% CI	Probability	95% CI
1	0.016	0.015, 0.017	0.039	0.029, 0.05
2	0.018	0.017, 0.02	0.054	0.042, 0.067
5	0.021	0.02, 0.022	0.064	0.051, 0.078
10	0.023	0.021, 0.024	0.07	0.057, 0.085
15	0.023	0.022, 0.025	0.071	0.058, 0.086
20	0.023	0.022, 0.025	0.071	0.058, 0.086
25	0.023	0.022, 0.025	0.071	0.058, 0.086
	Infection			
	After 1st contact episode		After 2nd contact episodes	
	Probability	95% CI	Probability	95% CI
1	0.037	0.035, 0.039	0.107	0.09, 0.126
2	0.038	0.036, 0.04	0.108	0.092, 0.126
5	0.038	0.037, 0.04	0.111	0.094, 0.129
10	0.038	0.037, 0.04	0.111	0.094, 0.129
15	0.038	0.037, 0.04	0.111	0.094, 0.129
20	0.038	0.037, 0.04	0.111	0.094, 0.129
25	0.038	0.037, 0.04	0.111	0.094, 0.129
	Further contact episode			
	After 1st contact episode		After 2nd contact episodes	
	Probability	95% CI	Probability	95% CI
1	0.003	0.002, 0.004	0.013	0.008, 0.021
2	0.01	0.009, 0.01	0.043	0.033, 0.056
5	0.033	0.031, 0.035	0.122	0.105, 0.139
10	0.053	0.051, 0.055	0.164	0.146, 0.183
15	0.062	0.06, 0.065	0.18	0.161, 0.2
20	0.067	0.064, 0.07	0.189	0.168, 0.21
25	0.07	0.067, 0.074	0.196	0.172, 0.22

Table 7. Transition probabilities (Aalen-Johansen estimates) for event types after first and second contact episode.

#### **4.4.4 Risk factors for disease, infection and second contact after the first contact episode**

Detailed information on the cause-specific hazard ratios (CSHR) of risk factors on univariable analysis are shown in Table 8.

On multivariable analysis country of birth, ethnic group and index case disease type were highly collinear. As country of birth was unrecorded for less than 10% of contacts this variable was excluded from the multivariable analysis. For time-to-D1 from C1 the final multivariable model included all variables (including an interaction between age and ethnicity) except year of contact (Table 9). CSHR was highest for exposure to culture-confirmed pulmonary index cases (CSHR=4.61); culture-negative pulmonary index cases had a lower impact (CSHR=1.79). Disease risk with exposure to non-pulmonary index cases was no different compared with exposure to denotified or latently infected cases. Black Caribbean ethnicity had a CSHR of 2.7 compared to Whites (Table 9); this effect reduced with increasing age (Figure 17). In contrast, the CSHR for Indian subcontinent ethnicity increased from 0.54 to 1.03 with increasing age (Figure 17). The main effect of age had a small reducing effect (CHSR=0.97). Other significant risk factors for disease were household contact (CHSR=1.54) and female sex (CSHR=1.23).

For time-to-L1 from C1 the final model was similar to time-to-disease but excluded sex. Year of contact was a significant variable (CSHR=1.06, 95% CI 1.06, 1.07) but the survival curves for different calendar year of contact had relative hazards that were not constant with time i.e. they were non-proportional. Therefore the model was stratified by year of contact. The impact of pulmonary index case and household contact was lower compared to time-to-D1 (CHSR=2.88 for culture-confirmed and 1.42 for culture-unconfirmed cases, CSHR=1.14 for household contact). The main effect of age had a small reducing effect (CHSR=0.93). Indian subcontinent, Black African and other ethnicity had a slightly higher risk for infection compared to White ethnicity with increasing age (CHSR 1.03, 1.04 and 1.04 respectively) (Table 9 and Figure 18).

The final model for time-to-C2 from C1 included age, sex, ethnicity and index case disease type only (Table 9). Age had a reducing effect on the time-to-C2 (CSHR=0.99) while female sex had a slightly increased effect (CHSR=1.1). Indian subcontinent, Black African and Black Caribbean ethnicity had a greater risk of time-to-C2 (CSHR 2.56, 2.21 and 1.83 respectively) while other or unknown ethnicity had a lower risk (CHSR 0.28 for both) compared to White ethnicity. Pulmonary (both culture-positive and negative) and non-pulmonary (culture-positive only) index cases also had a greater risk compared to denotified or latently infected index cases (CSHR 2.09, 1.42 and 1.57 for pulmonary culture-positive, pulmonary culture-negative and non-pulmonary culture-negative respectively).

Chapter 4: What happens to tuberculosis contacts? Competing risks application to estimate risks of infection with single versus repeat exposures.

Risk factor	Disease			Infection			2 <sup>nd</sup> contact		
	CSHR	95% CI	P	CSHR	95% CI	P	CSHR	95% CI	P
Age	0.99	0.99, 1	<0.01	0.96	0.96, 0.96	<0.01	0.99	0.99, 1	<0.01
Female sex	1.23	1.1, 1.39	<0.01	1.02	0.93, 1.11	0.68	1.07	0.99, 1.16	0.08
Country of birth									
Non-UK	1			1			1		
UK	4.51	3.98, 5.1	<0.01	4.56	4.13, 5.03	<0.01	4.33	3.52, 5.33	<0.01
Unknown	7.5x10 <sup>-5</sup>	1.1x10 <sup>-5</sup> 5.4x10 <sup>-4</sup>	<0.01	1.6x10 <sup>-4</sup>	4x10 <sup>-4</sup> 6.4x10 <sup>-4</sup>	<0.01	0.27	0.23, 0.32	<0.01
Ethnic group									
White	1			1			1		
Black Caribbean	2.90	2.28, 3.69	<0.01	1.42	1.16, 1.73	<0.01	1.78	1.4, 2.25	<0.01
Indian subcontinent	1.07	0.89, 1.28	0.49	0.75	0.67, 0.86	<0.01	2.32	2, 2.69	<0.01
Black African	1.91	1.46, 2.48	<0.01	1.71	1.43, 2.03	<0.01	2.04	1.56, 2.67	<0.01
Other	1.00	0.7, 1.43	1.00	1.40	1.14, 1.73	<0.01	1.32	0.99, 1.76	<0.01
Unknown	0.01	0, 0.02	<0.01	<0.01	0.67, 0.86	<0.01	0.28	0.22, 0.34	<0.01
Household contact	1.52	1.35, 1.71	<0.01	1.09	0.99, 1.19	0.07	1.43	1.32, 1.54	<0.01
Index case disease									
Denotified or latent infection	1			1			1		
Pulmonary, culture positive	3.38	2.25, 5.08	<0.01	2.86	2.13, 3.84	<0.01	1.65	1.34, 2.03	<0.01
Pulmonary, culture negative	1.39	0.91, 2.13	0.13	1.35	0.99, 1.83	0.06	1.23	0.99, 1.53	0.06
Non-pulmonary, culture positive	1.35	0.82, 2.2	0.24	1.42	1, 2.01	0.05	1.75	1.37, 2.23	<0.01
Non-pulmonary, culture negative	0.91	0.57, 1.46	0.69	0.79	0.55, 1.11	0.18	1.14	0.9, 1.43	0.27
Unknown	1.04	0.47, 2.31	0.93	1.53	0.93, 2.5	0.09	0.97	0.6, 1.55	0.89
Year of contact	1.03	1.02, 1.04	<0.01	1.08	1.07, 1.09	<0.01	1.02	1.01, 1.02	<0.01

CSHR, cause-specific hazard ratio; CI, confidence interval.

Table 8. Univariable analysis of cause-specific hazard ratios for disease, infection and second contact episode after a first contact episode.

Chapter 4: What happens to tuberculosis contacts? Competing risks application to estimate risks of infection with single versus repeat exposures.

Risk factor	Disease			Infection <sup>1</sup>			2 <sup>nd</sup> contact		
	CSHR	95% CI	P	CSHR	95% CI	P	CSHR	95% CI	P
Age	0.97	0.96, 0.98	<0.01	0.93	0.92, 0.94	<0.01	0.99	0.99 to 1	<0.01
Female sex	1.27	1.13, 1.43	<0.01				1.1	1.02, 1.18	0.02
Ethnicity									
White	1			1			1		
Black Caribbean	2.66	1.85, 3.83	<0.01	0.88	0.66, 1.17	0.38	1.83	1.44, 2.32	<0.01
Indian subcontinent	0.54	0.41, 0.72	<0.01	0.4	0.33, 0.48	<0.01	2.56	2.2, 2.98	<0.01
Black African	1.42	0.94, 2.15	0.09	0.52	0.39, 0.68	<0.01	2.21	1.69, 2.9	<0.01
Other	0.83	0.47, 1.46	0.51	0.57	0.42, 0.79	<0.01	0.28	0.23, 0.35	<0.01
Unknown	8.3x10 <sup>-4</sup>	4.2x10 <sup>-5</sup> , 0.02	<0.01	7.7x10 <sup>-4</sup>	7.9x10 <sup>-5</sup> , 7.6x10 <sup>-3</sup>	<0.01	0.28	0.22, 0.34	<0.01
Household contact	1.48	1.31, 1.68	<0.01	1.14	1.04, 1.25	<0.01			
Index case disease									
Denotified or latent infection	1			1			1		
Pulmonary, culture positive	4.61	3.06, 6.93	<0.01	2.88	2.13, 3.87	<0.01	2.09	1.7, 2.58	<0.01
Pulmonary, culture negative	1.79	1.17, 2.75	<0.01	1.42	1.04, 1.94	0.03	1.42	1.15, 1.77	<0.01
Non-pulmonary, culture positive	1.25	0.77, 2.05	0.37	1.05	0.74, 1.5	0.77	1.57	1.24, 2	<0.01
Non-pulmonary, culture negative	0.97	0.6, 1.56	0.89	0.74	0.52, 1.04	0.09	1.09	0.87, 1.34	0.44
Unknown	1.69	0.47, 2.31	0.2	1.24	0.75, 2.05	0.4	1.33	0.6, 1.55	0.24
Age*Black Caribbean	1	0.98, 1.01	0.64	1.02	0.99, 1.03	0.08			
Age*Indian subcontinent	1.03	1.02, 1.04	<0.01	1.03	1.02, 1.04	<0.01			
Age*Black African	1.01	0.99, 1.03	0.35	1.04	1.03, 1.05	<0.01			
Age*Other	1.01	0.99, 1.03	0.49	1.04	1.02, 1.06	<0.01			
Age*Unknown ethnicity	1.07	1, 1.13	0.04	1.07	0.99, 1.15	0.07			

CSHR, cause-specific hazard ratio; CI, confidence interval.

<sup>1</sup>Analysis was stratified by year of contact

Table 9. Multivariable analysis of cause-specific hazard ratios for disease, infection and second contact episode after the first contact episode.

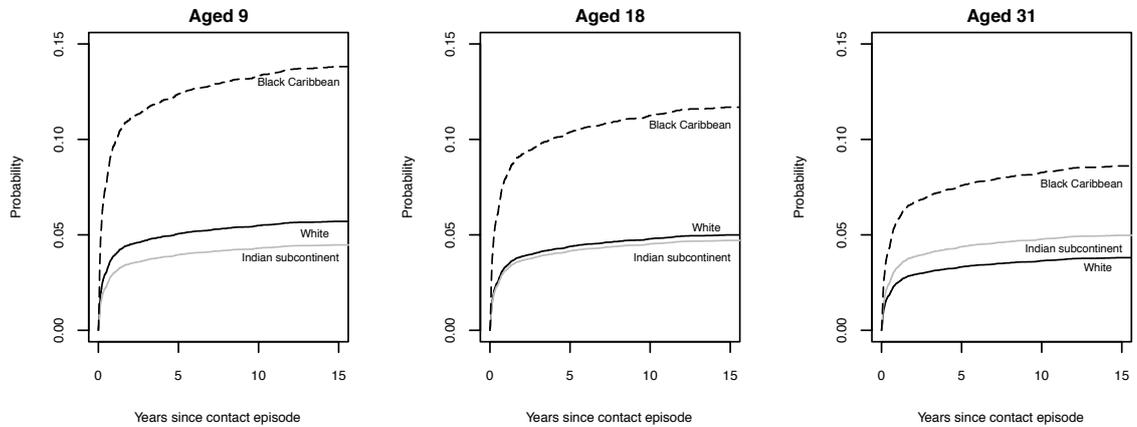


Figure 17. Predicted cumulative incidence function (Aalen-Johansen estimates) for disease in male, household contacts of culture-positive pulmonary TB patients by ethnic group and first quartile, median and third quartile age of diseased contacts.

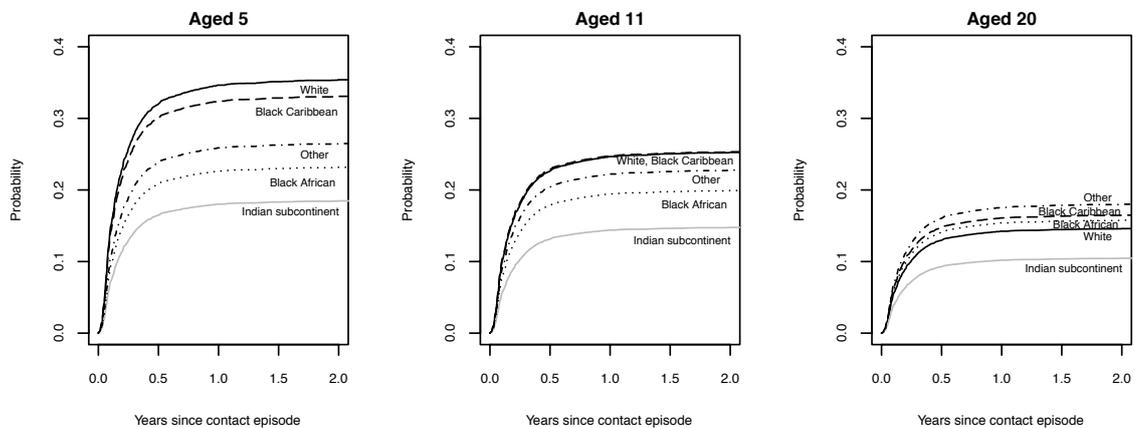


Figure 18. Predicted cumulative incidence function (Aalen-Johansen estimates) for latent infection in male, household contacts of culture-positive pulmonary TB patients by ethnic group and first quartile, median and third quartile age of infected contacts.

## **4.5 Discussion**

This analysis has shown empirically that a recurrent contact is a more common event than disease or infection within 10 years of the first contact and that recurrent contacts are at very high risk of both disease and infection. Contact repetition, an easily recognised entity, is thus an important indicator for TB case detection. A biological explanation is possible i.e. increased cumulative exposure following repeated contact to TB cases. Another possible explanation is that repeated contacts are part of a risk group with high prevalence of infection acquired remotely. For example infectiousness of the first index case appears to be less important for time-to-C2 compared to time-to-D1 or -L1. Indian subcontinent and Black African ethnicity also had increased risk for time-to-C2 at all ages in contrast to time-to-D1 or -L1. This suggests that C2 contacts may not be involved in direct chains of transmission. Instead contact repetition may signal community networks with a higher prevalence of TB compared to the general population.

The yield of contact tracing in finding TB disease at one year was similar to that estimated for high-income settings in a previous meta-analysis (~1.4%) (Fox *et al.*, 2013). Utilising record linkage and a survival approach in this analysis has enabled a closer evaluation of time-to-event. Thirty per cent of diseased contacts present a year after the index case was diagnosed. It is likely that these cases were not actively found through contact tracing which typically takes six to nine months to complete. A Dutch study using the same approach for 10-year registry data had similar results for contacts of pulmonary TB cases with 0.8% (74/9332) diseased within 180 days after the index case diagnosis and a further 0.4% (36/9213) diagnosed after 180 days (Sloot *et al.*, 2014). A proportion of the diseased contacts diagnosed late in our study could be due to their ineligibility for latent infection screening. However the median age for this ‘missed’ group was 26 years (IQR 18-41). Thus a considerable proportion of cases may still have been preventable with consistent follow-up. Fewer cases of infection in contacts were found in this analysis compared to pooled estimates of 30% for similar high-income countries (Fox *et al.*, 2013). This difference is likely attributable to the heterogeneity in the way latent infection is diagnosed (Erkens *et al.*, 2010).

Age-related risks of infection and progression to disease following infection are well documented (Rieder *et al.*, 1990; Stead and Lofgren, 1983) and consistent with findings here. Female susceptibility to disease is also well described (Comstock *et al.*, 1974; Rieder *et al.*, 1990) although the underlying cause is not clear. Differential risk from ethnicity and its interaction with age is more difficult to explain. Black Caribbean ethnicity was an independent predictor of molecular clustering in a London study (Maguire *et al.*, 2002) suggesting increased genetic susceptibility although the impact of infecting TB clade was not studied.

UK guidelines previously advocated the tracing of contacts from non-pulmonary index cases but since 2016 this group does not require further intervention (National Institute for Health and Clinical Excellence, 2016). This analysis has shown that these contacts do not have a greater risk of infection or disease compared to denotified or latently infected cases. However in the UK setting the yield from contact tracing this non-infectious group is not zero (Underwood *et al.*, 2003) and therefore without any other route of intervention opportunities may be lost.

There are several limitations to this analysis. The rate of infection after contact is likely an underestimate due to systematic screening of contacts under a certain age only and those infected but not treated were not included. Incomplete ascertainment of repeated contacts and linkages across the contact and case registry may also underestimate the transition rates between states. Nevertheless a competing risk approach has conclusively shown the increased risk of disease and infection in those with recurrent contact episodes. These contacts should be a high priority group in contact tracing procedures and control programmes should develop the means to recognise recurrent contacts.

## **5 The contact tracing network: global structure, local properties and extent of tuberculosis transmission**

### **5.1 Introduction**

In chapter 4 we established the rates at which individual contacts were diagnosed as cases. While demographic and clinical factors of the index case and contacts themselves influenced these transitions, the main finding i.e. that contacts with repeated exposure to tuberculosis (TB) cases were at higher risk of disease/infection suggests that the pattern of interactions in a community at risk is important in understanding disease distribution. Much understanding about the consequences of contact patterns evolved from mathematical models investigating the spread of sexually transmitted diseases. In such studies identification of a core-group of individuals with high partner change rate who associate with each other (“like with like” mixing), for example, appear to be important in the timing and size of an epidemic (R. M. Anderson *et al.*, 1990; Ghani *et al.*, 1997). More recently, contact patterns have been empirically fitted to transmission of airborne diseases. Spread of childhood infections such as parvovirus B19 and varicella correlate well with age-specific contact rates but also intimacy of contact (e.g. household, long-duration and frequent) (Melegaro *et al.*, 2011). In an investigation of the 2009 H1N1 influenza pandemic that began in a primary school, students social grouping into classes and grades were more predictive of infection transmission rather than physical proximity to an infectious case (Cauchemez *et al.*, 2011).

Thus understanding of mixing patterns can help target those at highest risk, over and above known traditional host factors (which may be related to both biological and social structure/contact rate patterns). Direct examination of contacts have been undertaken to better inform contact rates that underpin theoretical transmission models, the largest of which was a diary-based cross-sectional survey across eight European countries (Mossong *et al.*, 2008). Participants were asked to record their contacts over the course of a single day and the nature of contact, e.g. home/work, long/short, physical/non-physical. TB contact investigations similarly ask index cases regarding contacts and the type of contact had with these individuals during his/her infectious period. Not all contacts are recorded because generally only contacts that shared air space over a significant time period are enquired about, but contact investigation data may be a useful proxy to delineate subgroups of higher risk individuals by virtue of their contact patterns and refine estimates of contact rates between such subgroups. Mixing patterns uniquely important to TB transmission may exist due the inherent differences between the natural history of TB and other air-borne pathogens, primarily the long and variable incubation period (time between infection and onset of disease/diagnosis). Observation of contacts within short time intervals may therefore be limiting.

Interest in knowing contact structure in TB grew from technological advances in the 1990s that enabled genotyping of isolates to ascertain relatedness. Epidemiologically unlinked cases were thought to be in the same transmission link based on molecular similarity of the copy number and location of insertion sequence (IS)6110 (a mobile genetic element) (Bifani *et al.*, 1996; Genewein *et al.*, 1993). While these earlier molecular epidemiological studies suggested severe limitations in information gained through conventional contact tracing, whole genome sequencing has subsequently discriminated closely-related strains as different thereby helping rule out recent transmission where no epidemiological link could be found (Jajou *et al.*, 2018; Roetzer *et al.*, 2013; Walker *et al.*, 2013). Where true transmission trees have been signaled by whole-genome sequencing and subsequently epidemiologically confirmed, there has been increasing appreciation that social structure, in particular

information about shared locations is important (Gardy *et al.*, 2011). Increasingly formal investigation of social structures relevant to TB has exploited network theory to characterise and relate its role in transmission. In the first description of network analysis to inform TB transmission, (Klovdahl *et al.*, 2001) found that individuals and places/locations (the nodes) that were more centrally placed in a network (a set of nodes connected by links) played an important role in an outbreak of 37 cases linked by IS6110 fingerprinting. Similar post-hoc analyses have further confirmed the utility of network measures in prioritisation of contacts with higher risk of infection (Andre *et al.*, 2007; Kawatsu *et al.*, 2015).

This chapter is a descriptive analysis of the contact data of TB cases in Birmingham in a network context. The first section provides a comprehensive overview of the network, as it has never been systematically examined. The second section explores the relationship between whole network metrics and final component size: do certain network structures result in larger components? Leading on from this the third section attempts to learn if local metrics for an index case (egocentric approach) impacts on the number of infected contacts in his/her immediate circle. Finally, the fourth section combines available molecular (mycobacterial repetitive unit-variable number tandem repeats or MIRU-VNTR typing) data and the contact tracing network to determine the extent to which they overlap.

The work presented differs from previous TB network studies in two ways (i) consideration of networks at a population level rather than limited to “outbreaks” or presumed transmission events linked by molecular epidemiology; and (ii) prolonged observation period of contact links relevant to the slow natural history of TB. Similar to published network studies in TB and other infectious diseases, we consider the data within a static network i.e. where links between people are fixed once formed and do not change with time. In reality links are likely to be more fluid with time, albeit with some more stable connections between family for example. However theory on temporal or dynamic networks is still developing and are not applied here.

## **5.2 Objectives**

- Quantitatively describe the contact tracing network using network metrics
- Examine network metrics that impact on the final, static network size
- Explore the feasibility of using egocentric metrics to identify cases with a higher number of infected contacts.
- Assess the utility of molecular typing (MIRU-VNTR) when interpreting contact tracing networks

## **5.3 Method**

The contact tracing network refers to the total TB dataset of cases and contacts in Birmingham and Solihull, 1980 – 2011. Creation of this dataset and collection of contact data is described fully in Chapter 2. We included denotified cases (N=162) and their contacts, these individuals were labelled as a contact only for this purpose.

All data preparation and analyses were conducted in *R* version 3.3.3 (R Core Team, 2017).

### **5.3.1 Terminology**

Similar to previous chapters infected denotes an individual infected with either latent TB or active disease. Individuals were also identified as either an index case or contact. A contact is further characterised as either household, close or casual. Table 10 summarises the definition for these terms, in alphabetical order.

Standard network terminology was adapted from (Hanneman and Riddle, 2005) and explained in Table 11. Network metrics were calculated using the *R* package *igraph* (version 1.1.2) (Csardi and Nepusz, 2006). Table 12 provides an explanation for network-level and node-level metrics used in this chapter. Edges between nodes were distinguished as household, close or casual according to relationship between the nodes, but were not weighted.

<b>Term</b>	<b>Definition</b>
Casual contact	An individual not named directly by an index case but identified to be present at the same non-residential address at the time of the infectious period of the index case
Close contact	An individual named by an index case but does not reside at a mutual residential address
Household contact	An individual named by an index case and resides at a mutual residential address
Index case	An infected individual with at least one contact
Infected	Infected with either latent TB or (active) disease

Table 10. Definition of terms defining individuals in the contact tracing data.

<b>Term</b>	<b>Explanation</b>
Clique	An undirected component where every distinct pair of nodes are linked to each other
Component	A set of nodes that are all reachable from every other
Degree	Number of edges associated with a node Indegree is the number of edges the node receives Outdegree is the number of edges the node sends
Directed edge	Edges have a direction, <u>from</u> the case <u>to</u> the contact
Edge	The relationship or link between nodes. Two nodes are adjacent if an edge exists between them.
Egocentric	Analysis on an individual node, with each node treated as a separate unit/case
Network	The total collection of edges between nodes. Some nodes may not have edges to another node
Node	An individual
Sociocentric	Analysis on whole network, with each network being a separate unit/case
Triplet	A component of three nodes
Undirected edge	Nodes are not classified as a sender or receiver of an edge

Table 11. Definition of network terminology adapted from Hanneman and Riddle (2005).

Term	Calculation
Network/sociocentric level	
Assortativity coefficient	The degree to which nodes associated with those similar to themselves, calculated according to (Newman, 2002). This coefficient ranges from -1 to 1 with coefficients >0 denoting homophily and <0 denoting assortative mixing
Density	Number of links expressed as a proportion of all possible links between nodes ( $N*(N-1)/2$ )
Diameter	The maximum shortest path length between any two nodes in a component, assuming an undirected and unweighted network.
Clustering coefficient	The number of closed triplets/the number of triplets, with edge direction and weights ignored. In addition components with no triplets of nodes were not included in the global average
Reciprocity	The number of (unordered) node pairs that had reciprocal edges divided by the number of (unordered) node pairs that had no edge between them plus the number of node pairs that had non-reciprocal edges (a directed network was assumed)
Node/Egocentric level	
Adjacent	Two nodes are adjacent if linked by an edge
Betweenness centrality	The sum of the number of paths that pass through it divided by the total number of shortest paths for each pair of nodes in the graph. This score was normalised by $(N-1)(N-2)/2$ (considering an undirected path) as the score scale with the number of pairs of nodes in the graph. Nodes with high betweenness centrality scores act as bridges between other nodes in the network.
Closeness centrality	The reciprocal of the average length of the shortest path between a node and all other nodes in the graph. Nodes with higher closeness centrality scores reach other nodes in the network quicker.
Degree centrality	The number of incident edges on a node, normalised by the maximum number of edges possible $(N-1)$ . Edge direction is ignored.
Eigenvector centrality	The weighted sum of the centrality of the nodes linked to the node in question.
Shortest path length	The number of edges, between a pair of nodes in a component, which traverses the minimal number of edges.

Table 12. Definition of network metrics adapted from the R package igraph (version 1.1.2) (Csardi and Nepusz, 2006).

### **5.3.2 Network visualisation**

Network visualisation was performed using the *R* package *igraph* (version 1.1.2) (Csardi and Nepusz, 2006). An individual with multiple infection events was labeled according to their earliest infection status only. Contacts who later became cases were labeled according to their infection status alone. Contacts never infected (i.e. never diagnosed as a case) during the study period were labeled as a contact.

### **5.3.3 Descriptive comparison of components in the contact tracing network**

Components were grouped as small, medium or large based on their diameter (1 – 2, 3 – 4, 5 or more respectively). An outlying component with diameter 23 was excluded from further comparison due to its extreme size. This component will be discussed separately in Chapter 6.

The total number of nodes, infected nodes and proportion infected in each component was calculated. The following network metrics were evaluated: component density, clustering coefficient and reciprocity. Triplets of two diseased nodes and one non-infected node (who will not have an outdegree) were excluded when clustering coefficient was considered. The two-proportions *z* test was used to compare proportion infected between components of different sizes. The Wilcoxon-rank-sum or the Kruskal-Wallis test was used to compare network metrics because values were not normally distributed.

### **5.3.4 Evaluation of degree distribution**

Because contact investigation differed for pulmonary (household and close contacts assessed), non-pulmonary (household contacts only assessed) and latent-infected (no contacts assessed) cases, degree distribution was enumerated by infection type in each node. Median outdegree between node disease types were compared using the Wilcoxon-rank-sum test.

### **5.3.5 Relating network metrics to number of infected contacts**

Egocentric networks for each pulmonary and non-pulmonary index node who named at least one contact were examined to evaluate whether their position in the network at the point of diagnosis (i.e. before they had named contacts) influenced the number of infected contacts they had.

### **5.3.6 Comparison of the contact tracing network and molecular network**

To assess the utility of contact tracing and typing in inferring disease transmission, all typed, culture-positive cases from 2004 onwards were assessed for a match on 15- or 24- mycobacterial repetitive unit-variable number tandem repeats (MIRU-VNTR) loci. All genotyping was performed prospectively at the Public Health England Regional Centre for Mycobacteriology, Birmingham using methods described by (Supply *et al.*, 2001). Strain typing using 24-loci was only performed from 2011 onwards.

MIRU-VNTR profiles were considered identical if it matched on all loci or had up to two missing loci but matched on all other available loci. Cases with identical MIRU-VNTR profiles were considered epidemiologically linked if they were in the same connected component. Singular typed cases in components with no other typed cases in the same component were excluded. Each (unordered) case pair was then considered individually to estimate the sensitivity and specificity of epidemiological linkage in predicting molecular linkage and vice versa.

## 5.4 Results

### 5.4.1 Overview of the static contact tracing network

During the study period there were 14,909 infection episodes in 14,590 individuals and 56,562 contact episodes in 53,344 individuals. Sixty-seven per cent (9703/14,590) of infected individuals were involved in contact tracing networks either as an index case naming contacts (42%), by being named as a contact (20%) or both (5%) (Figure 19). The vast majority of contacts (93%) were not diagnosed with infection.

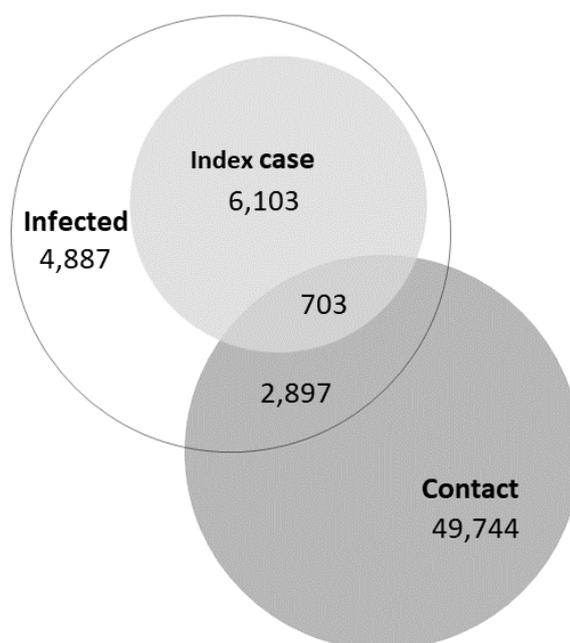


Figure 19. Status of 64,334 individuals in the case and contact database. Shaded areas denote those involved in contact tracing networks.

## **5.4.2 Components in the static contact tracing network**

### **5.4.2.1 Distance**

Individuals in the contact tracing network formed a total of 5729 separate connected components. Ninety-one per cent (5192/5729) of components were of star topology: a central infected node with outgoing edges to one or more contact nodes. These star components were of diameter 1 or 2 (Figure 20A and Figure 20B). Component topology changed if infected contacts named more contacts (Figure 20C, left-hand subgraph) or if separate star components had mutual contacts (Figure 20C, right-hand subgraph). Two per cent (118/5729) of components had a diameter of 5 or greater (Figure 20E). This included one largest component with diameter 23 containing 5% (3148/59534) of all nodes in the contact tracing network. This outlying largest component was excluded from further analysis in this chapter and is discussed in Chapter 6.

For diseased nodes in the same component the vast majority were adjacent i.e. connected by one path length (Figure 21). Iterative contact tracing (outgoing edges in the directed network) reached a maximum of five path lengths. However when both outgoing and incident edges were considered in an undirected network the maximum shortest path length between diseased nodes was 12. The path length of 2 in the undirected network is the artifact mode path length because it also reflects the common scenario of two infected contacts with an incident edge from the same index node.

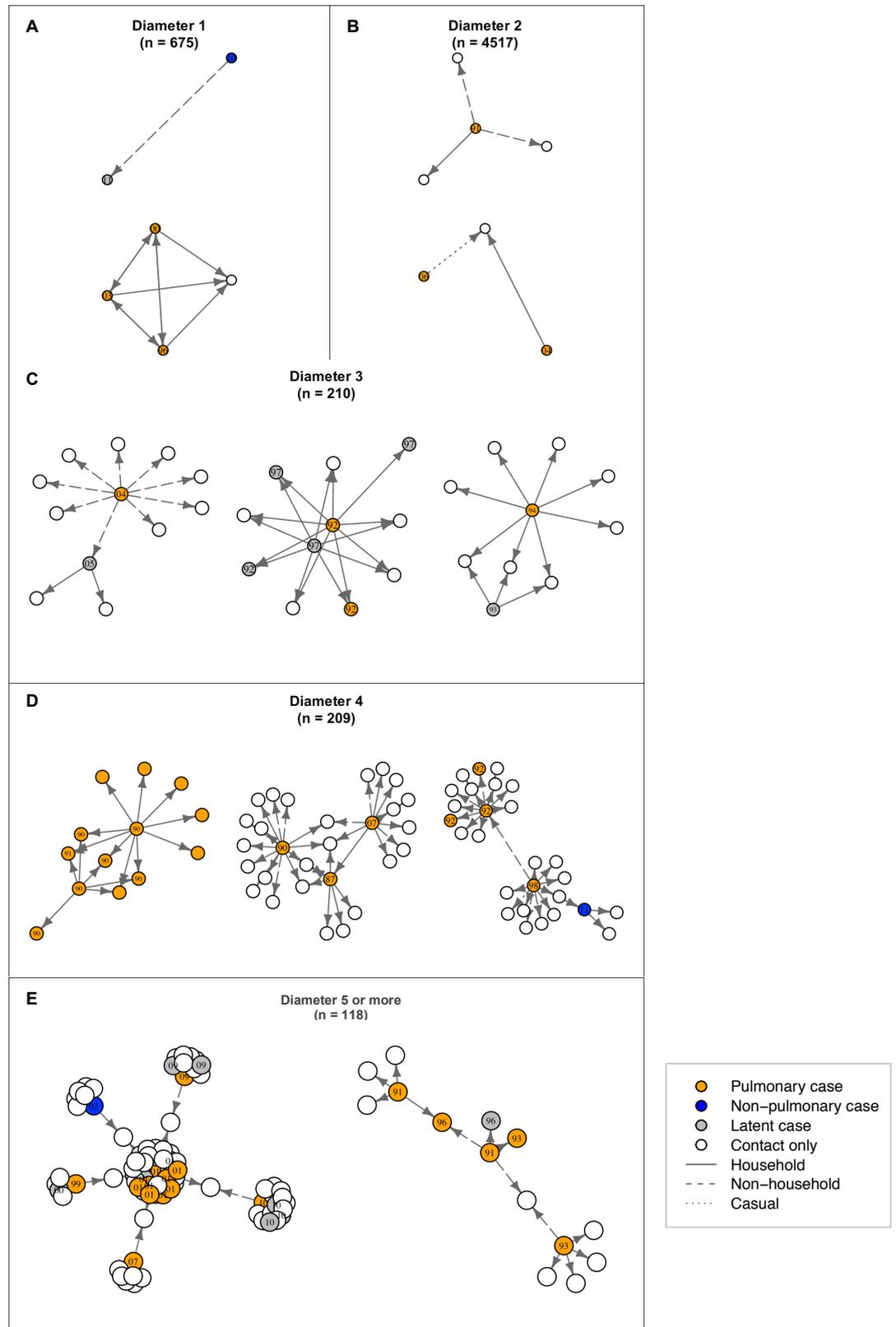


Figure 20. Undirected and unweighted diameters of 5,729 connected components in the static contact tracing network. The directed and weighted (different edge types) subgraphs are shown to illustrate evolution of the contact tracing network. Node labels denote the year of diagnosis.

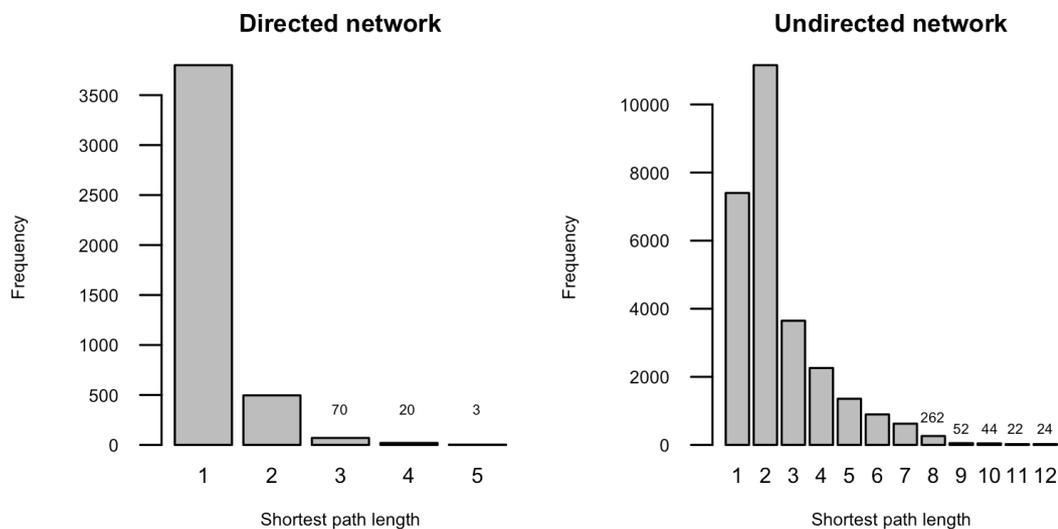


Figure 21. Pairwise distance distribution between 9,247 infected nodes in the directed and undirected contact tracing network. Numbers above bars denote actual number of pairs with the path length. The largest component was excluded.

#### 5.4.2.2 Cohesion

Table 13 summarises network characteristics of the contact tracing network by component size. Small components (i.e. components with diameter 1 to 2) had a median of 6 nodes (interquartile range, IQR 4 – 9). The median number of nodes in a component doubled with increasing diameter: 15 (IQR 11 – 25) in medium components (diameter 3 to 4) and 47 (IQR 31 – 72) in large components (diameter 5 or greater). The median number of infected nodes increased with increasing diameter (1, 3 and 7 for small, medium and large components respectively). Medium components had a marginally higher proportion of their nodes infected when compared to small and large components (median proportion infected 0.2 (IQR 0.03 – 0.29) versus median of 0.2 (IQR 0.01 – 0.25) and 0.17 (IQR 0.01 – 0.28) respectively,  $P < 0.01$ ).

Density reduced with increasing component size and differences were significant between components of all sizes (median density for small, medium and large were 0.35 (IQR 0.22 – 0.5), 0.17 (IQR 0.09 – 0.22) and 0.06 (IQR 0.03 – 0.06) respectively,  $P < 0.001$ ).

The overall clustering coefficient between diseased nodes in the contact tracing network was 0.01 (Table 13). Five per cent (270/5728) of components had at least one triplet of diseased nodes. In 93% (70/75) of small components, 69% (83/121) of medium components and 65% (48/74) of large components the probability that a triplet closed to form a clique was zero. The clustering coefficient distribution was similar between medium (median 0, IQR 0 – 0.38) and large components (median 0, IQR 0 – 0.32) ( $P = 0.74$ ) but different between small (median 0, IQR 0 – 0) and medium/large components ( $P < 0.001$ ).

When only diseased nodes were considered in the contact tracing network, the global reciprocity or probability of a mutual connection between two (diseased) nodes was 0.06. Only 11% (653/5728) of components had at least two diseased nodes with 95% (282/296) of small, 83% (214/258) of medium and 80% (78/98) of large components having a reciprocity of zero. Similar to the clustering coefficient distribution there was a difference in reciprocity distribution between small and medium/large components ( $P < 0.001$ ) but no difference between the large and medium components ( $P = 0.74$ ).

	Component size	Small	Medium	Large	
	Diameter	1 – 2 (N=4517)	3 – 4 (N=419)	≥ 5 (N=117)	
No. of nodes	Mean	8	20	60	
	SD	6.27	12.89	45.98	
	Minimum	3	4	10	
	Maximum	144	131	285	
	Percentiles	25	4	11	31
		50	6	15	47
		75	9	25	72
No. of infected nodes	Mean	1	4	10	
	SD	0.81	2.4	7.36	
	Minimum	1	2	3	
	Maximum	12	16	43	
	Percentiles	25	1	2	4
		50	1	3	7
		75	1	4	12
Proportion infected	Mean	0.21	0.23	0.2	
	SD	0.13	0.13	0.12	
	Minimum	0.007	0.032	0.014	
	Maximum	1	1	0.7	
	Percentiles	25	0.13	0.13	0.12
		50	0.2	0.2	0.17
		75	0.25	0.29	0.28
Density	Mean	0.35	0.17	0.06	
	SD	0.17	0.1	0.04	
	Minimum	0.01	0.02	0.0007	
	Maximum	1.33	0.5	0.2	
	Percentiles	25	0.22	0.09	0.03
		50	0.33	0.15	0.05
		75	0.5	0.22	0.08
Clustering coefficient of diseased nodes <sup>a</sup>		N=75	N=121	N=74	
	Mean	0.05	0.21	0.15	
	SD	0.2	0.35	0.22	
	Minimum	0	0	0	
	Maximum	1	1	1	
	Percentiles	25	0	0	0
		50	0	0	0
	75	0	0.38	0.32	
Reciprocity of diseased nodes <sup>b</sup>		N=296	N=258	N=98	
	Mean	0.04	0.09	0.05	
	SD	0.19	0.25	0.16	
	Minimum	0	0	0	
	Maximum	1	1	1	
	Percentiles	25	0	0	0
		50	0	0	0
	75	0	0	0	

SD, standard deviation

<sup>a</sup>Components with no triplets of diseased nodes were excluded.

<sup>b</sup>Components with no edge between diseased nodes were excluded.

Table 13. Descriptive statistics for 5,728 connected components in the static contact tracing network. The largest component with diameter 23 was excluded.

### 5.4.2.3 Degree distribution

Table 14 shows the number contacts named by infected nodes (outdegree). The outdegree varied by type of infection in the index case and type of edge (household, close or casual contact). The median number of household contacts (household outdegree) for nodes with pulmonary disease was two (IQR 0 – 5) versus three (IQR 1 – 5) for nodes with non-pulmonary disease ( $P < 0.001$ ).

The median number of close contacts (close outdegree) for nodes with pulmonary disease was also low - one (IQR 0 – 5) (Table 14). Thirty-nine per cent (2,075/5,293) of nodes with pulmonary disease did not name any close contacts. Nodes with non-pulmonary disease and latent infection did not routinely have contacts outside the household traced. Thus their close outdegree was lower compared to nodes with pulmonary disease.

Casual contacts in congregate settings were sought in only 16% (858/5,293) of nodes with pulmonary disease. The majority (85%, 551/858) that did have casual contacts sought had a maximum casual outdegree of two. Nodes with non-pulmonary disease and latent infection were not considered infectious but casual contacts were documented for 10% (223/2140) and 2% (55/2290) of non-pulmonary and latent nodes respectively.

	No. of household contacts			No. of close contacts			No. of casual contacts		
	P	NP	L <sup>b</sup>	P	NP	L <sup>b</sup>	P	NP	L <sup>b</sup>
Mean	3.25	3.18	0.56	3.82	1.53	0.23	0.45	0.14	0.03
SD	4.56	2.9	1.69	7.15	3.25	1.27	4.61	0.49	0.19
Minimum	0	0	0	0	0	0	0	0	0
Maximum	138 <sup>a</sup>	31	25	111	71	35	168	7	2
Percentiles									
25	0	1	0	0	0	0	0	0	0
50	2	3	0	1	0	0	0	0	0
75	5	5	0	5	2	0	0	0	0

P, pulmonary; NP, non-pulmonary; L, latent infection; SD, standard deviation

<sup>a</sup>This involved a large contact tracing exercise at a single hostel

<sup>b</sup>Contacts were not solicited but volunteered by these latent cases

Table 14. Outdegree distribution for nodes with pulmonary disease (n=5,293), non-pulmonary disease (N=2,140) and latent infection (N=2,290), by type of edge.

The distribution of indegrees (of all types) differed between node infection type ( $P < 0.001$  for each two-way comparison) (Figure 22). A higher proportion of nodes with pulmonary disease had at least one indegree when compared to nodes with non-pulmonary disease (24%, 1263/5293 versus 13%, 274/2140) while the majority of nodes with latent infection (85%, 1955/2290) had at least one indegree.

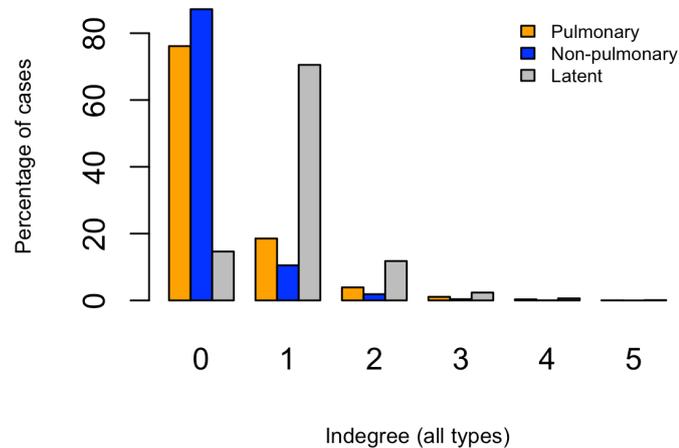


Figure 22. Indegree distribution (household, close and casual) for nodes in the contact tracing network, by node infection type.

#### **5.4.2.4 Mixing patterns**

Table 15 shows the assortativity coefficients for nodes linked by different edge types. In household case-contact pairs mixing was highly assortative by ethnic group and postal sector as expected (0.94 and 0.96 respectively). Among non-household case-contacts the assortativity coefficient for ethnic group remained high (0.88) but it was much lower among casual case-contacts pairs (0.27).

Edge type	N	Assortativity coefficient			
		Age	Sex	Ethnic group	Postal sector
Household	29,653	0.03	-0.01	0.94	0.96
Non-household	26,174	0.02	0.003	0.88	0.27
Casual	2,890	0.09	0.05	0.27	0.1
All	54,771	0.03	-0.02	0.89	0.6

Table 15. Mixing patterns between pairs of nodes linked by household, close (non-household) and casual edges.

#### **5.4.2.5 Network metrics and number of infected diagnosed**

Figure 23 shows the distribution of number of infected contacts by the number of contacts named (outdegree). Finding an infected contact was a rare event with the median number of infected being greater than one only for pulmonary cases with very high outdegrees - 15 or greater for household contacts (Figure 23A and Figure 23B) and 25 or greater for non-household outdegrees (Figure 23C and Figure 23D).

Centrality measures for index cases that had no infected or at least one infected contact were mostly similar (Figure 24). Of centrality scores only the eigenvector centrality differed between cases who had none or at least 1 infected contact diagnosed (mean eigenvector centrality 0.969, 95% CI 0.965 – 0.973 versus 0.945, 95% CI 0.933 – 0.957,  $P < 0.001$ ) (Figure 24A). Mean betweenness centrality score was 0.010 (95% CI 0.003 – 0.017) versus 0.018 (95% CI 0.001 – 0.034) ( $P = 0.2$ ) (Figure 24B), median degree centrality score was 0.071 (IQR 0.027 – 0.167) versus 0.067 (IQR 0.036 – 0.2) ( $P = 0.2$ ) (Figure 24C) and median closeness centrality score was 0.2 (IQR 0.174 – 0.24) versus 0.2 (9% CI 0.178 – 0.244) ( $P = 0.74$ ) (Figure 24D).

When clustering coefficient of the network once an index case had named his/her contacts were considered (but before any of the contacts were diagnosed if infected and named their secondary contacts) the mean score in index cases with at least one infected contact was 0.019 (95% CI 0.014 – 0.024) versus 0.01 (95% CI 0.008 – 0.012) ( $P = 0.002$ ).

Infected contacts were diagnosed both prior to and after infection diagnosis in the index case (Figure 25). More infected contacts of non-pulmonary index cases were diagnosed prior to infection diagnosis in the index case compared to contacts of pulmonary index cases. For infected contacts diagnosed after the index case diagnosis, the majority occurred within the following five years.

Compared to infection diagnosed de novo (i.e. in an individual not known to be a contact in the past), infected contacts had a shorter duration of symptoms by 11 – 28 days ( $P < 0.05$ ).

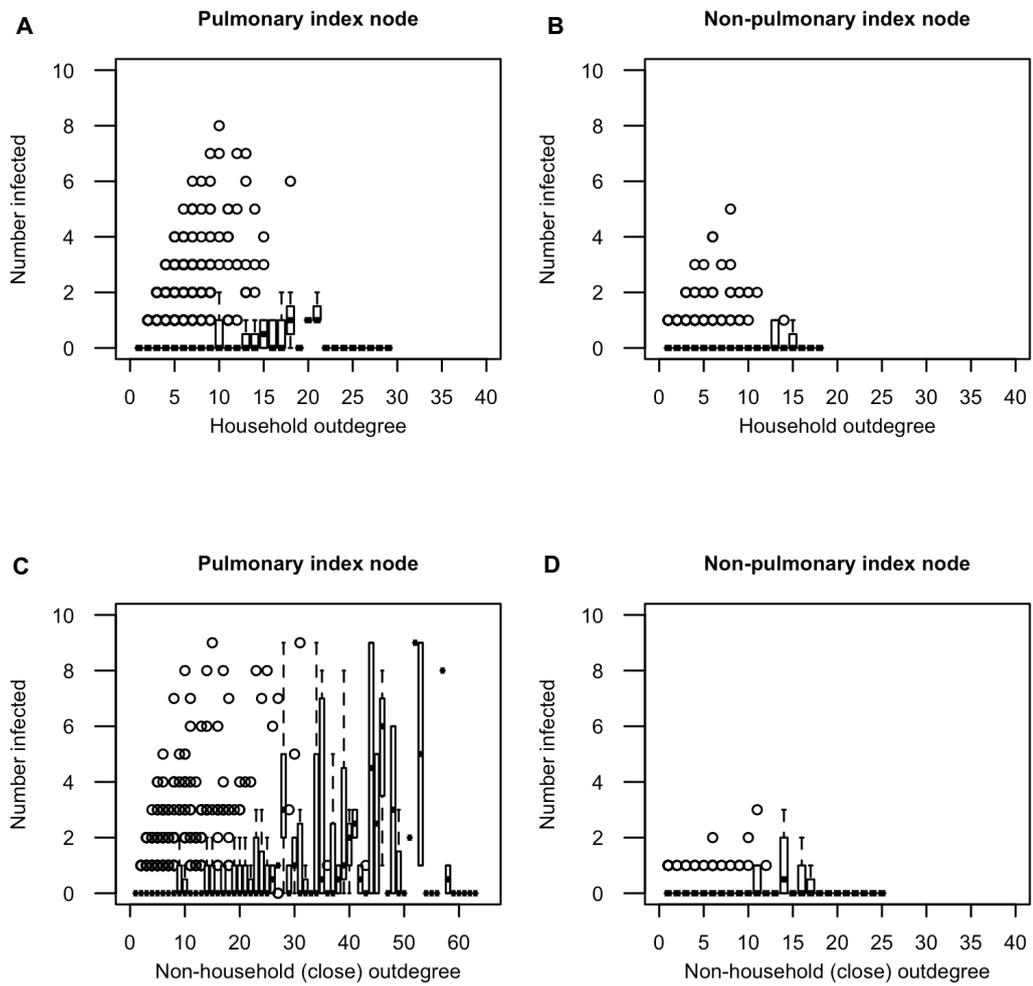


Figure 23. Box plots of the number of infected contacts for pulmonary (N=5,293) and non-pulmonary (N=2,140) index cases (with at least one outdegree), by edge type and number of outdegrees. Median number is denoted by band in the centre of boxes, first and third quartile by box edges, maximum number within 1.5 times the interquartile range by the uppermost whisker, outliers by circles.

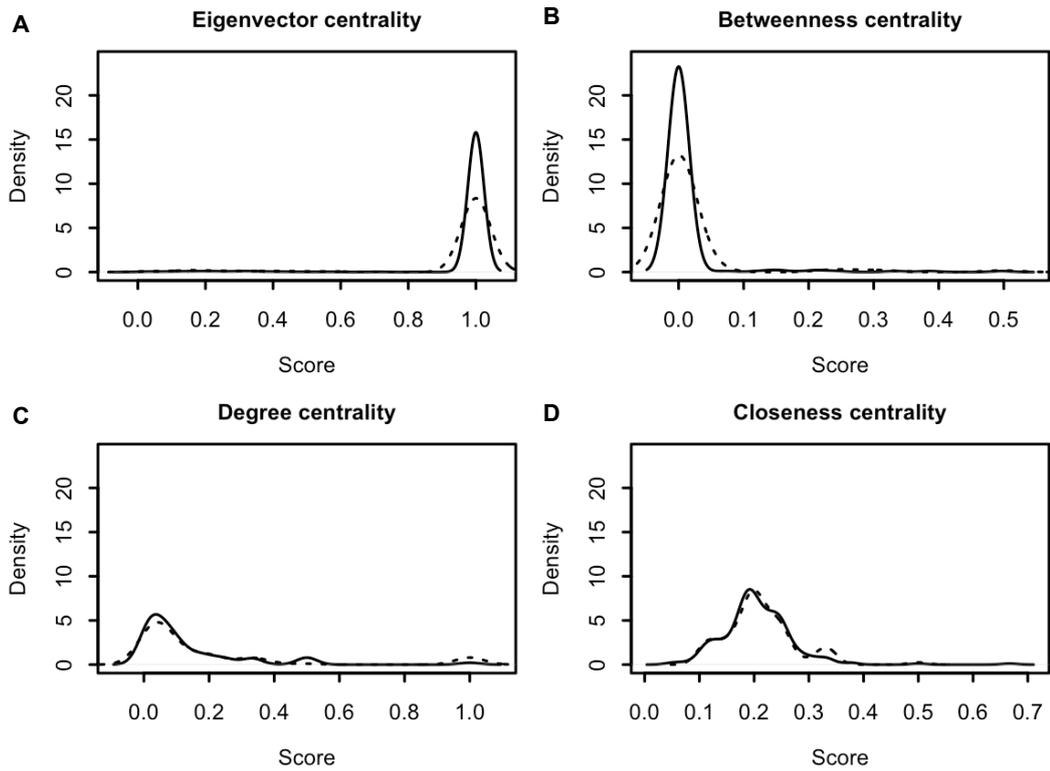


Figure 24. Distribution of centrality measures for nodes with no infected contacts (solid line) versus egos with at least one infected contact (dotted line).

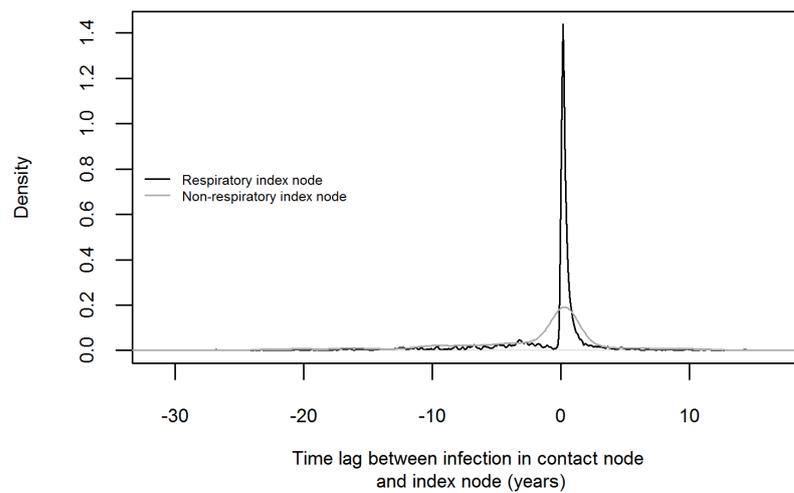


Figure 25. Distribution of time between infection diagnosis in an index case and contact by disease.

### 5.4.3 MIRU-VNTR typing

Mycobacterial interspersed repetitive unit-variable number tandem repeat (MIRU-VNTR) typing was universally available from year 2004 onwards in the study area. Although over 80% of culture-positive cases were typed, just over half (55%) of all diseased cases were culture-positive and available for typing (Figure 26).

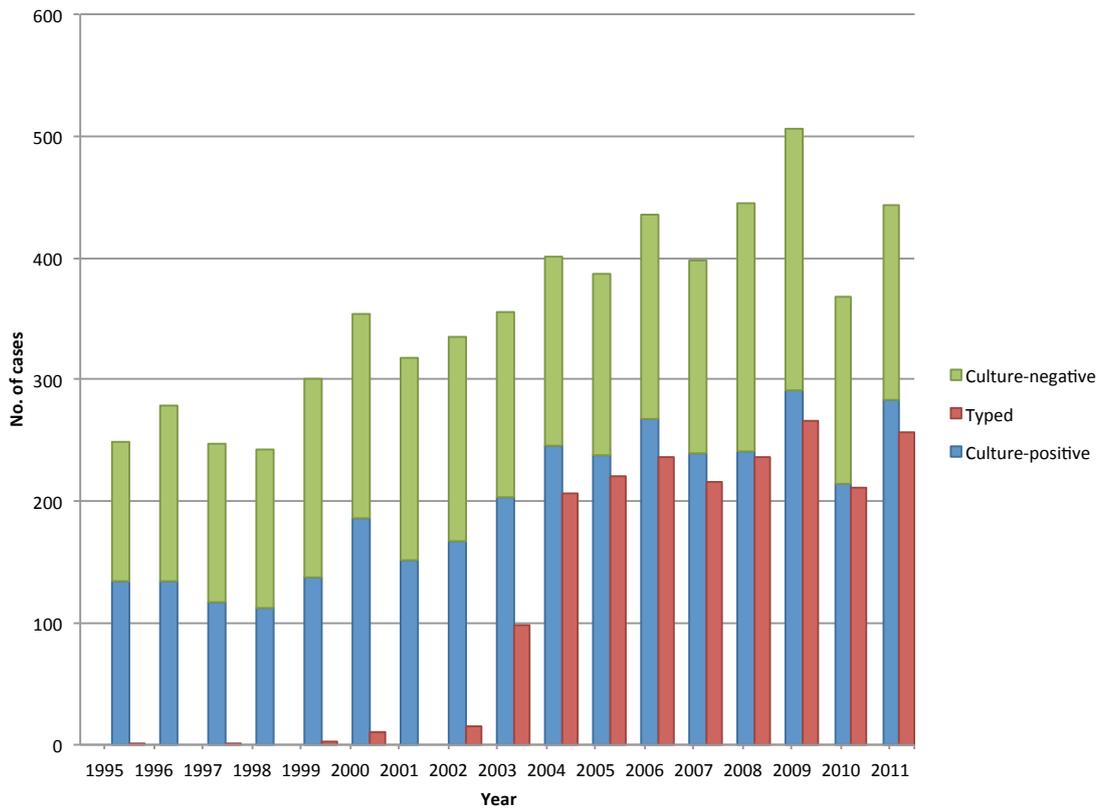


Figure 26. No. of typed isolates compared to the total burden of disease by year.

Sixty four per cent (1184/1848) of typed cases were in components with at least two typed cases. These cases were considered in unordered pairs and occurrence in the same connected component (i.e. presence of an epidemiological link) and match on MIRU-VNTR typing (i.e. presence of a genotypic link) was assessed. The overall sensitivity of epidemiological linkage in predicting molecular linkage was only 2% (48/2814, 95% CI 1 – 2%) and specificity was 100% (Table 16). Conversely, the sensitivity of molecular linkage in predicting epidemiological linkage was 39% (95% CI 30 – 48) (Table 17). Sensitivities did not alter significantly by demographic or clinical characteristics of case pairs (Table 17).

		Contact tracing network	
		Linked	Not linked
Typing by MIRU- VNTR	Linked	48	2766
	Not linked	76	697446

Table 16. Genotypic and epidemiological links for 1184 diseased nodes involved in contact tracing networks evolving from 2004 onwards. Numbers denote counts of (unordered) node pairs.

*Chapter 5. The contact tracing network: global structure, local properties and extent of tuberculosis transmission*

	No. linked epidemiologically and by MIRU-VNTR typing/No. linked epidemiologically	Sensitivity (%) (95% CI)	No. not linked epidemiologically and not linked by MIRU-VNTR typing/No. not linked epidemiologically	Specificity (95% CI)
All	48/124	39 (30 – 48)	697,446/700,212	100 (100 – 100)
One or both cases UK born	29/54	54 (40 – 67)	299,752/301,111	100 (100 – 100)
One or both non-UK born	35/97	36 (27 – 46)	655,509/657,753	100 (100 – 100)
One or both cases of White ethnicity	10/25	40 (21 – 61)	154,184/154,821	100 (100 – 100)
One or both cases of Black Caribbean ethnicity	3/6	50 (12 – 88)	78,673/79,275	99 (99 – 99)
One or both cases of non-White, non-Black Caribbean ethnicity	31/72	43 (31 – 55)	614,489/616,828	100 (100 – 100)
Both cases in the same age group	18/48	38 (24 – 53)	136,703/137,297	100 (100 – 100)
One or both cases with pulmonary disease	48/122	39 (31 – 49)	646,529/649,174	100 (100 – 100)
Both cases non-pulmonary disease	0/121	0 (0 – 91)	50,917/51,038	100 (100 – 100)

Table 17. Sensitivity and specificity of MIRU-VNTR typing in predicting epidemiological linkage, by ethnicity, birthplace, age and disease extent.

## 5.5 Discussion

In contrast to extensive examination of contact networks relevant to sexually transmitted diseases (Doherty, 2011) only a handful of studies have investigated TB contacts networks as outlined in the introduction. In TB network studies done in the context of transmission (known due to molecular epidemiology), the network approach provided information about the immediate, local contacts who were at higher risk of infection. A few studies investigated networks across their population irrespective of transmission (Cook *et al.*, 2007; Cozatt, 2016). These studies were small with less than 200 cases and spanned short time frames (up to 3 years). Thus description of the very large network in Birmingham over a long time frame may add new data about how contact tracing networks can be used to ascertain not only causality (i.e. transmission) but identify interventions (i.e. the community at risk).

The network that can be constructed from contact tracing data are subject to several complications. Because the data was collected over time both past and future individuals and/or their links are censored so only a partial view of the complete network is obtained. Individuals may have many interactions that come and go over their lifetime and this dynamic process is not taken into account as our network was analysed from a static viewpoint. Thus not all links observed may be relevant at a particular time. The other difficulty is the incompleteness of contacts as identified through contact tracing. Collection of contact data was biased because contacts were sought differently for infectious and non-infectious TB cases. In addition the collection method was not standardised (unstructured interview) and contact identification was voluntary.

Nevertheless the contact tracing network was fragmented with over 90% of components remaining small (up to ten individuals within 1-2 path lengths). Within clusters individuals were loosely connected with low density, clustering coefficient and reciprocity. The low measures of cohesiveness are likely a consequence of the way contact tracing is done rather than a reflection of the true network, for example for clusters of individuals within a household only edges from an index case to his/her household members are recorded. As contacts becoming a case was a rare event (just under 90% of all component contained just one case), edges between non-

infected household members are falsely absent. As the number of cases increased components grew in diameter and clustering coefficient and reciprocity also increased.

Larger components were rare and their significance are unclear. Is it possible to limit their growth? The distribution of cases in time may reflect the natural history of TB which is very slow branching, while contact tracing occurs at a much shorter time scale. Due the age limits of latent infection assessment in the past no intervention was possible for older contacts, but current guidelines now advocate testing for latent TB in all contacts under 65 (National Institute for Health and Clinical Excellence, 2016).

The median household degree of 2 reflects known UK data. Close degree was also very small (0-1). A minimum number of five contacts per infectious has been set as a quality metric for TB control boards (Anderson *et al.*, 2014). This may be difficult to achieve and may not increase case detection as degree centrality did not appear to influence whether at least one contact developed TB disease/infection. However eigenvector centrality, whether a node was connected to another highly connected node may be more useful.

There was low concordance between epidemiological links and MIRU-VNR relatedness as evidenced by the low sensitivity of each entity in predicting the other. For cases with matched genotypic data the probability that contact tracing detects an epidemiological link was only 2%. This is not surprising as MIRU-VNTR performs poorly in distinguishing relatedness for certain lineages (Sloot *et al.*, 2013) i.e. matched MIRU-VNTR isolates may in fact be unrelated. Additionally current forms of contact tracing are biased samples of the true contact network so not all epidemiological links are detected. These limitations may have contributed to findings in an evaluation of the national strain typing service in which it did not increase the contact tracing yield or reduce diagnostic delay (Mears *et al.*, 2014). Thus discontinuing routine cluster investigation (search for epidemiological links between cases in a MIRU-VNTR cluster) as the authors suggest is supported by our data.

For cases with observed connections in the contact tracing network the probability that their TB isolates were genetically similar was higher at 39%. As contact tracing information is usually available before genotypic information, it can be a powerful tool in identifying transmission early. However, name-based contact tracing as undertaken in our study has been described to have limited utility in finding cases (Asgar *et al.*, 2009). Changing the way contacts are sampled, for example repeat interviewing of cases, obtaining contacts of contacts ((Fitzpatrick *et al.*, 2001; McElroy *et al.*, 2003) and information about location (Gardy *et al.*, 2011) could reduce transmission events. None of these strategies have been evaluated in the UK. With wider experience of whole genome sequencing higher specificity in detecting probable transmission is expected and the number of true molecular clusters within a control programme is likely to reduce. This new molecular tool can only be fully exploited with corresponding advancements in the way contact tracing is programmatically done. In contrast to extensive contact tracing studies in sexually transmitted diseases, TB studies are lacking and more research into this area is needed.

## 6 The largest component

### 6.1 Introduction

Many social and biological networks in the real world typically consist of many small, disconnected components (groups of linked individuals or nodes), but also strikingly a single, largest component that contains the majority of the individuals/nodes (Marcotte *et al.*, 1999; Newman, 2001; Ripeanu *et al.*, 2002). This component is often termed a “giant” component, a phenomenon that has been calculated analytically to emerge when the number of links (edges) in a network approaches half the number of nodes in a network (Erdős and Rényi, 1959). In a random network model with a fixed number of nodes but no connections between them, if two nodes (not previously connected) are iteratively linked at random small chain-like components are initially formed. If a link is formed between nodes from two different components, these merge into one larger component. When the average number of connections per node is one, the network of fragmented small components undergoes a phase transition to form a network where most nodes are connected to the giant component.

In the static TB contact tracing network described in Chapter 5, cases and their contacts grouped into components ranging from single cases with few contacts through to larger structures. The largest component had 3148 individuals and 3461 edges formed through the years 1982 to 2011. This component contained 5% of the individuals in the contact tracing network, and was significantly bigger than the next largest component (285 individuals). While existence of this largest component was not surprising given previous knowledge of other large scale networks, it was clearly unique. In particular, did its formation reflect random processes or were there critical events that occurred for smaller components to merge into the largest component? The latter would signal opportunities for intervention.

This chapter explores the largest component in the contact tracing network in detail. Firstly, we describe the largest component and give an overview of its evolution with time. Secondly, we compare whether edge formation between nodes in the largest component and nodes in other components was different in terms of their number (per TB case and in the components overall) and cohesion (whether edges are widely distributed in the component or only routed through a few nodes). Thirdly, we detect subgroups of nodes that were more closely clustered together and identify the most central nodes within the largest component, and correlate this with their epidemiological features to understand events that contributed to creation of the largest component. Finally, we attempt to ascertain the extent of disease transmission within the largest component by analysing available molecular epidemiology and contrast this with position of the nodes in the component.

The largest component will be viewed as a static network only in this analysis. As introduced in Chapter 5, edges are considered fixed once formed although linkages between individuals are likely to vary with time.

## **6.2 Objectives**

- Ascertain if infection type (diseased or latent infected) and disease extent (pulmonary or non-pulmonary) contributed to formation of the largest component.
- Explore whether network measures (outdegree, transitivity, reciprocity and centrality) differed between the largest component and other components.
- Describe the epidemiological events that occurred within and between subgroups of the largest component to understand what drove its formation.
- Examine the extent to which the largest component acts to influence infection transmission or simply identifies a social network, by mapping available molecular genotyping to node position within the component.

### **6.3 Method**

The largest component is a subset of the larger contact tracing network described in Chapter 5. Here we note some particular aspects of the methods for readers coming to this chapter without detailed reading of previous chapters.

All data preparation and analyses were conducted in *R* version 3.3.3 (R Core Team, 2017).

#### **6.3.1 Terminology**

An infected individual refers to those with either stages of TB infection i.e. latent infection or active disease. Individuals were also referred to as an index case or contact. Contact type could be further categorised as household, close or casual. These terms are defined in alphabetical order in Table 18.

Table 19 defines the terms used to describe composition and structure of the largest component. These standard network terminologies were adapted from Hanneman and Riddle (2005).

Table 20 explains the network metrics used in this chapter, as calculated in the *R* package *igraph* (version 1.1.2) (Csardi and Nepusz, 2006).

Term	Definition
Casual contact	An individual not named directly by an index case but identified to be present at the same non-residential address at the time of the infectious period of the index case
Close contact	An individual named by an index case but does not reside at a mutual residential address
Household contact	An individual named by an index case and resides at a mutual residential address
Index case	An infected individual with at least one contact
Infected	Infected with either latent TB or (active) disease

Table 18. Definition of terms defining individuals in the contact tracing data.

Term	Explanation
Component	A set of nodes that are linked to each other by edges.
Degree	Number of edges associated with a node. Indegree is the number of edges the node receives. Outdegree is the number of edges the node sends.
Directed edge	Edges have a direction, <u>from</u> the case <u>to</u> the contact.
Edge	The relationship or link between nodes. Two nodes are adjacent if an edge exists between them.
Egocentric	Analysis on an individual node, with each node treated as a separate unit/case.
Network	The total collection of edges between nodes. Some nodes may not have edges to another node
Node	An individual
Sociocentric	Analysis on whole network, with each network being a separate unit/case
Undirected edge	Nodes are not classified as a sender or receiver of an edge

Table 19. Definition of network terminology adapted from Hanneman and Riddle (2005).

Term	Calculation
Component/ sociocentric level	
Assortativity coefficient	The degree to which nodes associated with those similar to themselves, calculated according to (Newman, 2002). This coefficient ranges from -1 to 1 with coefficients >1 denoting homophily and <1 denoting dissortative mixing
Density	Number of links expressed as a proportion of all possible links between nodes ( $N*(N-1)/2$ )
Diameter	The maximum shortest path length between any two nodes in a component, assuming an undirected and unweighted network.
Clustering coefficient	The number of closed triplets/the number of triplets, with edge direction and weights ignored. In addition components with no triplets of nodes were not included in the global average
Reciprocity	The number of (unordered) node pairs that had reciprocal edges divided by the number of (unordered) node pairs that had no edge between plus the number of node pairs that had non-reciprocal edges (a directed network was assumed)
Node/ Egocentric level	
Adjacent	Two nodes are adjacent if linked by an edge
Betweenness centrality	The sum of the number of paths that pass through it divided by the total number of shortest paths for each pair of nodes in the graph. This score was normalised by $(N-1)(N-2)/2$ (considering an undirected path) as the score scale with the number of pairs of nodes in the graph. Nodes with high betweenness centrality scores act as bridges between other nodes in the network.
Closeness centrality	The reciprocal of the average length of the shortest path between a node and all other nodes in the graph. Nodes with higher closeness centrality scores reach other nodes in the network quicker.
Degree centrality	The number of incident edges on a node, normalised by the maximum number of edges possible $(N-1)$ . Edge direction is ignored.
Eigenvector centrality	The weighted sum of the centrality of the nodes linked to the node in question.
Shortest path length	The number of edges, between a pair of nodes in a component, which traverses the minimal number of edges.

Table 20. Definition of component metrics adapted from the *R* package igraph (version 1.1.2) (Csardi and Nepusz, 2006).

### 6.3.2 Visualisation

The largest component was visualised in the *R* package *igraph* (version 1.1.2) (Csardi and Nepusz, 2006).

An individual with multiple infection events was coloured according to their earliest infection status only. An individual named by a case, but never infected (i.e. never diagnosed as a case) during the study period was coloured as a contact. Contacts who later became cases were coloured according to their infection status alone. For recurrent events, the year of the first infection diagnosis was used to label nodes, including if the node was a contact prior to first infection. For recurrent contacts, the year of the first contact only was used to label nodes.

Edges in the largest component were visualised without edge direction noted when describing the full component to simplify the complex diagram. Edge direction was included when visualising the most central nodes and their subgroups to enable easier interpretation of the time course of epidemiological events. Edge type was distinguished as household, close or casual according to relationship between the nodes, but was not weighted. Multiple edges connecting the same two nodes (6 node pairs in all) were merged to a single (weighted) edge for simplification.

For visualisation purposes 2,145 of 3,148 nodes whom were never cases (i.e. were only named as contacts) and whose removal did not disconnect the largest component were not displayed. This reduced representation of the largest component is designated as the simplified largest component.

### 6.3.3 Network metrics

Metrics were calculated assuming a static network in the *R* package *igraph* (version 1.1.2) (Csardi and Nepusz, 2006). Calculations were performed on the full, largest component i.e. with no nodes deleted, unlike the visualisation diagrams where a reduced network was shown (see section 6.3.2).

Metrics for the largest component was compared with components of diameter 3 or more in the contact tracing network, i.e. medium and large components only. Small components of diameter 1 to 2 were excluded from comparison because their network measures differed from the medium and large components, as highlighted in Chapter 5. I will refer to the comparator, medium and large components collectively as *other* components.

The Wilcoxon-rank-sum test was used to compare network metrics because values were not normally distributed.

### 6.3.4 Subgroup detection

The largest component was partitioned into groups to aid interpretation of its structure and evolution. The Clauset-Newman Moore algorithm was chosen as a first approximation due to its speed of calculation. This is a hierarchical algorithm that starts with each node belonging to a separate group. At each iteration of the algorithm two groups are merged if the modularity score improves i.e. if the proportion of edges within the merged group exceeds what would be expected if they occurred at random (Clauset *et al.*, 2004). A node cannot belong to more than one group but a group can contain subgroups. The “fastgreedy.group” command in the *R* package *igraph* version 1.1.2 (Csardi and Nepusz, 2006), was used, with edge weights (whether household, close or casual relationship) ignored. When the same partitioning algorithm was performed on the simplified largest component, the number of groups reduced from 62 to 24 while maintaining nodes from known large contact tracing exercises within different congregate settings in the correct groups. Therefore groups are discussed as partitioned on the simplified largest component. Groups were identified by sequential letters, A – X.

### 6.3.5 Genotypic data

As described in Chapter 5, genetic relationships between nodes were assessed using all 15- and 24- loci MIRU-VNTR data available. Tree-based analysis was done using the web application MIRU-VNTR*plus* (<http://www.miru-vntrplus.org>) which holds 186 reference strains from all major lineages of TB. Genetic distance was calculated using categorical distance i.e. the number of loci with a different allele divided by the total number of loci (24). The default distance cutoff was 0.17, which allowed a difference at up to four loci. Missing loci were ignored. The neighbour-joining algorithm (Saitou and Nei, 1987) was used to generate phylogenetic trees. A random sample of 65 24-loci MIRU-VNTR sequences, which represented approximately 20% of all available MIRU-VNTR sequences in the database, were included in the dendograms for reference. The unique MIRU-VNTR sequences were identified by their designated national cluster name whereby prefix letters indicate the phylogenetic lineage followed by a unique four digit number (B for Beijing, E for Euro-American, C for Delhi/ Central Asian, A for East-African-Indian, X for multiple lineages and U for none of the recognised lineages) (Gibson *et al.*, 2005).

## 6.4 Results

### 6.4.1 Overview of the (static) largest component

The largest component consisted of 5% (3,148/59,534) of all nodes in the study population. Figure 27 displays the simplified largest component. Of 3,468 edges in the largest component 43% (1502/3,468) were casual contact edges, 30% (1027/3,468) were close contact edges and 27% were household contact edges. The proportion of nodes that were cases (yellow and blue) was 15% (476/3,148). Eighty-five per cent (193/228) of diseased nodes had pulmonary disease compared to 54% (5,100/9,495) in the overall contact tracing network ( $P < 0.001$ ). A relatively high proportion of nodes (8%, 248/3148) had latent infection compared to 4% (2042/51623) across the contact tracing network ( $P < 0.001$ ). Median age of nodes in the largest component was 16 (interquartile range, IQR 13 – 26) versus 29 (IQR 16 – 49) in the whole network ( $P < 0.05$ ).

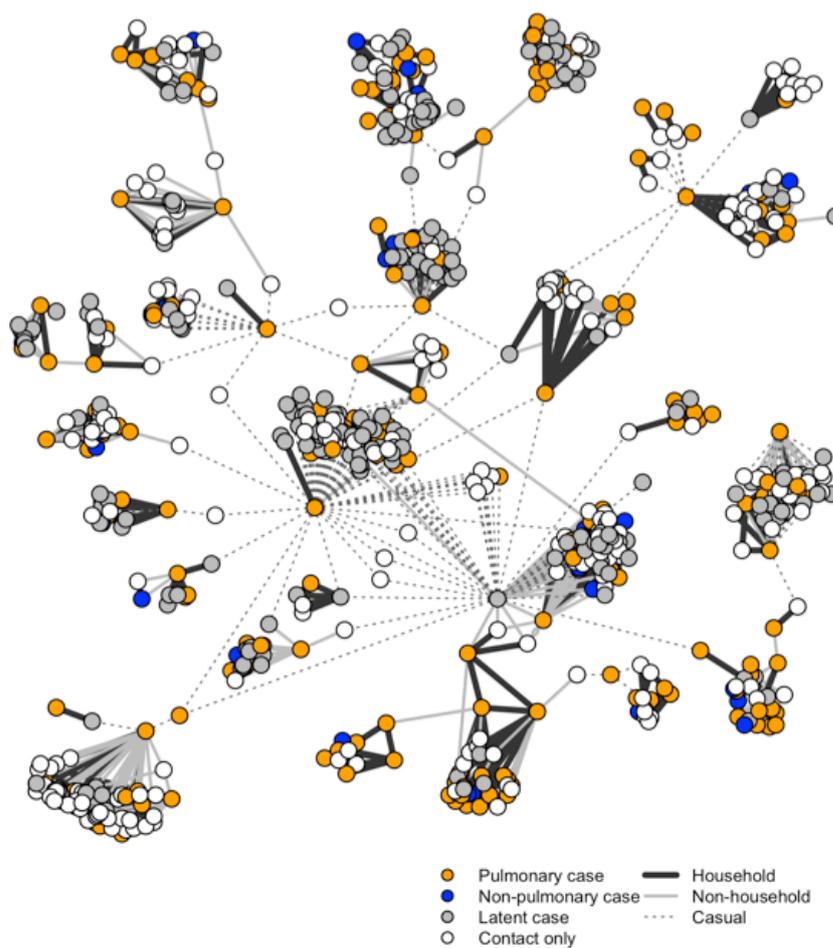


Figure 27. The simplified largest component, with 2,415 contact nodes with only one incident edge removed.

### 6.4.2 Evolution of the largest component with time

Figure 28 shows the growth curve for the largest component, showing the number of cases and contacts added each year. The timeline spanned 29 years (1982 to 2011). Peaks in the number of cases and contacts occurred in 2009 and 2011. There was also a peak in the number of contacts in 2004. I will now describe in detail the events that led to these larger and smaller attachments to the components.

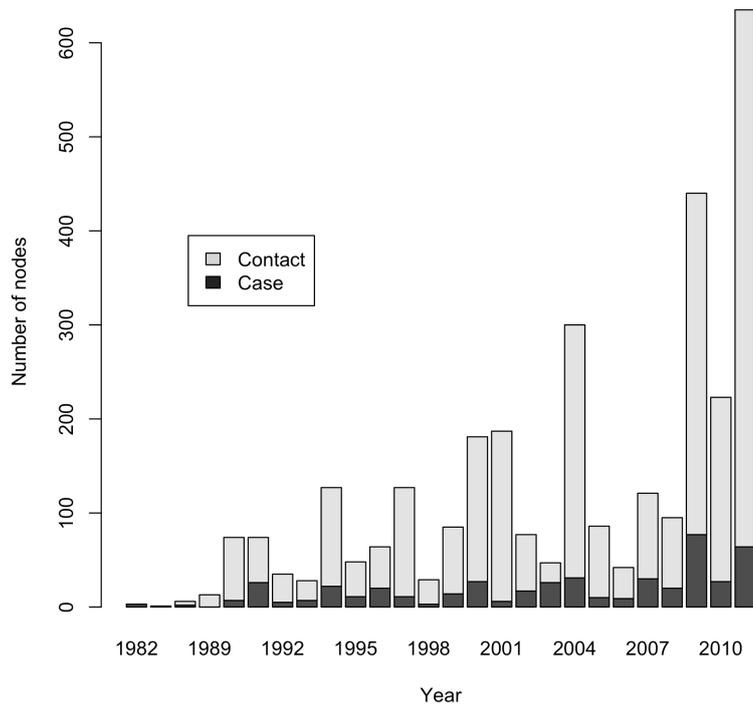


Figure 28. Number of cases and contacts added per year in the largest component.

Growth of this component occurred through two processes – the diagnosis (or re-diagnosis) of new cases amongst existing members of the network, leading to addition of new contacts and possibly secondary cases and contacts and the identification of new cases whose contacts (or secondary cases and contacts) are members of the existing network. We can see the evolution of the network by looking at the nodes and edges of the largest component which are in the dataset over time (Figure 29) Prior to 2005 the largest component consisted of disparate sub-components consisting of household/ close contact clusters. These sub-components continued to grow independently, with very little linkage of existing sub-components until 2005, where in the top right and lower right of the picture there are some larger linkage events. The peak of number of new nodes in 2009 and 2011 not only introduced a number of new nodes, but those nodes connected the smaller subgroups to form the largest component (Figure 29). Casual edges formed 43% (1502/3461) of the edge types in the component, followed by close edges (30%, 1027/3461) and household edges (27%, 932/3361).

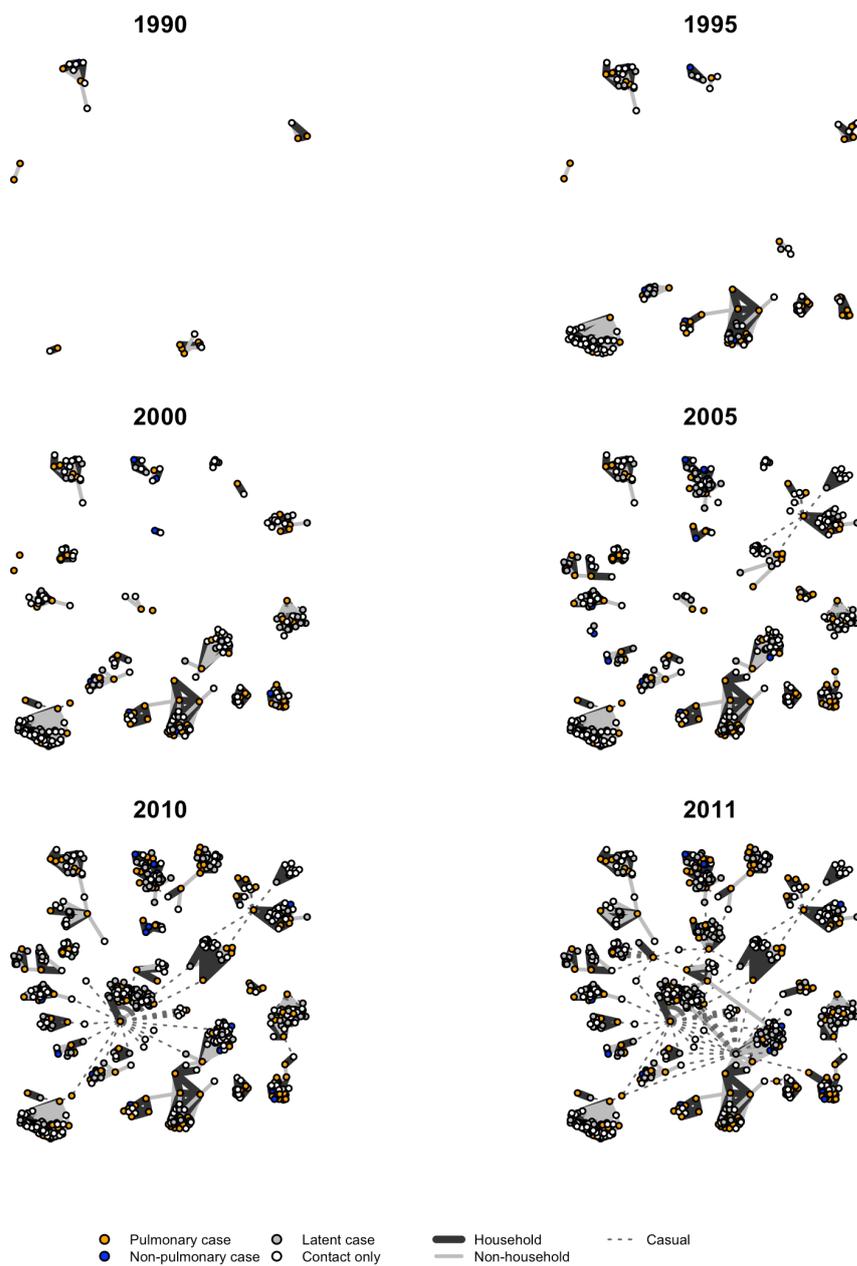


Figure 29. Evolution of the simplified largest component from year 1990 to 2011.

### 6.4.3 Connections in the largest component versus medium and large components

The outdegree distribution of case nodes in the largest component was similar to case nodes in other components (Figure 30).

Nodes with pulmonary disease in the largest component had a median household outdegree of 4 (IQR 0 – 7) versus 3 (IQR 0 – 6) in other components ( $P=0.12$ ) and a median close outdegree of 1 (IQR 0 – 8) versus 1 (IQR 0 – 7) ( $P=0.34$ ) (Figure 30A). Nodes with pulmonary disease in both the largest and other components had a median casual outdegree of zero (IQR 0 – 0) (Figure 30B). The casual outdegree for nodes with pulmonary disease in the largest component was more distributed compared to other components (Figure 30C) but this difference was not statistically significant ( $P=0.31$ ).

For nodes with non-pulmonary disease the median household outdegree in the largest component was 4 (IQR 3 – 7) versus 3 (IQR 0 – 6) in other components ( $P=0.13$ ) (Figure 30D). The median close and casual outdegree for nodes with non-pulmonary disease was zero in all components with no difference in distribution between those in the largest component and other components ( $P=0.12$  and  $P=0.61$  for close and casual outdegree respectively) (Figure 30E and Figure 30F). Very few nodes with latent infection in the largest and other components had any outdegrees (results not shown).

The median indegree for respiratory case nodes in the largest component was 1 (IQR 0 – 1) with a mean of 0.9 (SD 0.97) while those in the other components had a median of 1 (IQR 0 – 1) with a mean of 0.72 (SD 0.86) ( $P=0.01$ ) (Figure 31A). The indegree distribution for non-respiratory cases in the largest component approached statistically significant difference when compared to other components (median 0 (IQR 0 – 1), mean 0.37 (SD 0.65) versus median 0 (IQR 0 – 1), mean 0.62 (SD 0.77) in other components ( $P=0.05$ ) (Figure 31B). Latent cases had similar indegree distributions in both the largest and other components (median 1 (IQR 0 – 1), mean 1.3 (SD 0.51) and median 1 (IQR 0 – 2), mean 1.4 (SD 0.74);  $P=0.09$ ) (Figure 31C).

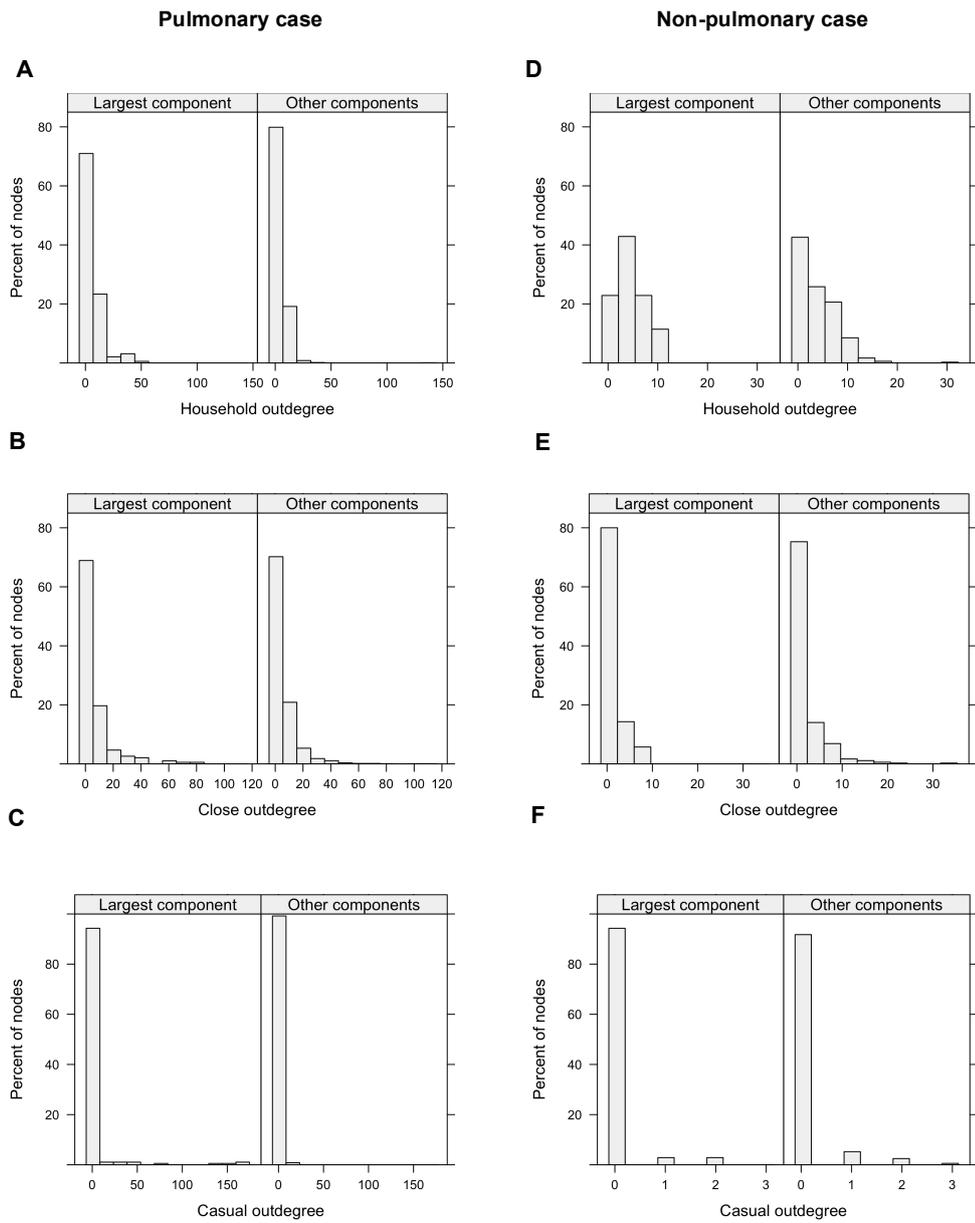


Figure 30. Outdegree distribution in pulmonary (A, B, C) and non-pulmonary (D, E, F) nodes in the largest component versus 536 medium and large (other) components in the contact tracing network, by type of edge.

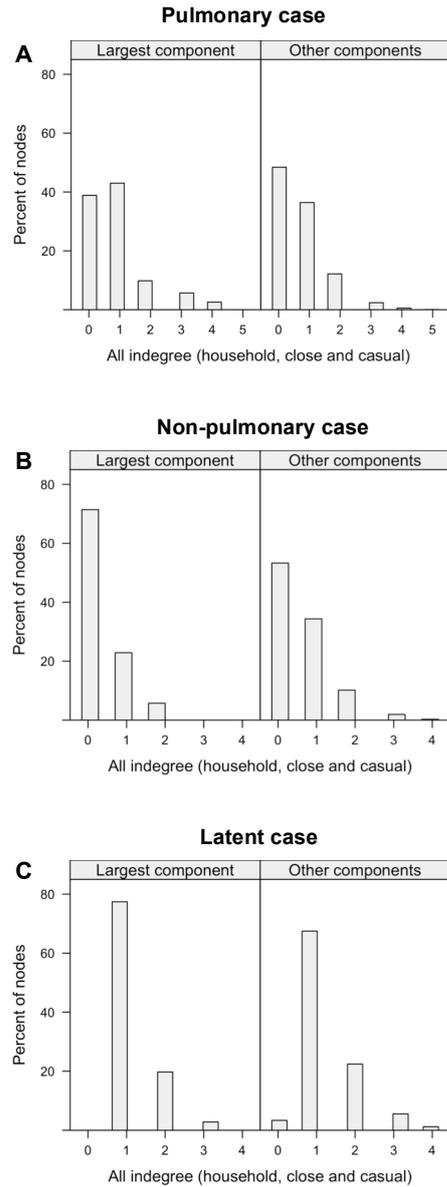


Figure 31. Indegree distribution in pulmonary (A), non-pulmonary (B) and latent case (C) nodes in the largest component versus 536 medium and large (other) components in the contact tracing network.

The overall density of the largest component was  $6.99 \times 10^{-4}$  versus a median density of 0.12 (IQR 0.67 – 0.19) in the other components ( $P=0.08$ ). Transitivity between case nodes in the largest component was 0.22 versus a median transitivity of 0 (IQR 0 – 0.37), mean of 0.19 (SD 0.31) in other components ( $P=0.45$ ). Reciprocity between case nodes was also no different in the largest component and other components (0.04 versus a median of 0 (IQR 0 – 0), mean 0.08 (SD 0.23);  $P=0.1$ ).

#### **6.4.4 Groups in the largest component**

Given the evolution of the component is driven by these independent clusters who are linked in later years, it is helpful to define ‘groups’ within the network, so that their characteristics can be defined.

Figure 32 shows the static, simplified largest component with member nodes coloured according to their group membership. Using the Clauset-Newman algorithm (as described in the methods), there were 24 groups in total. These groups are summarised in Figure 33 where each group has been represented by a node, its’ size scaled according to the number of nodes in each group. Median group size was 24 nodes (IQR 15 – 35). The largest group was group C (96 nodes) followed by groups H (82 nodes), B (62 nodes), D (51 nodes) and F (48 nodes).

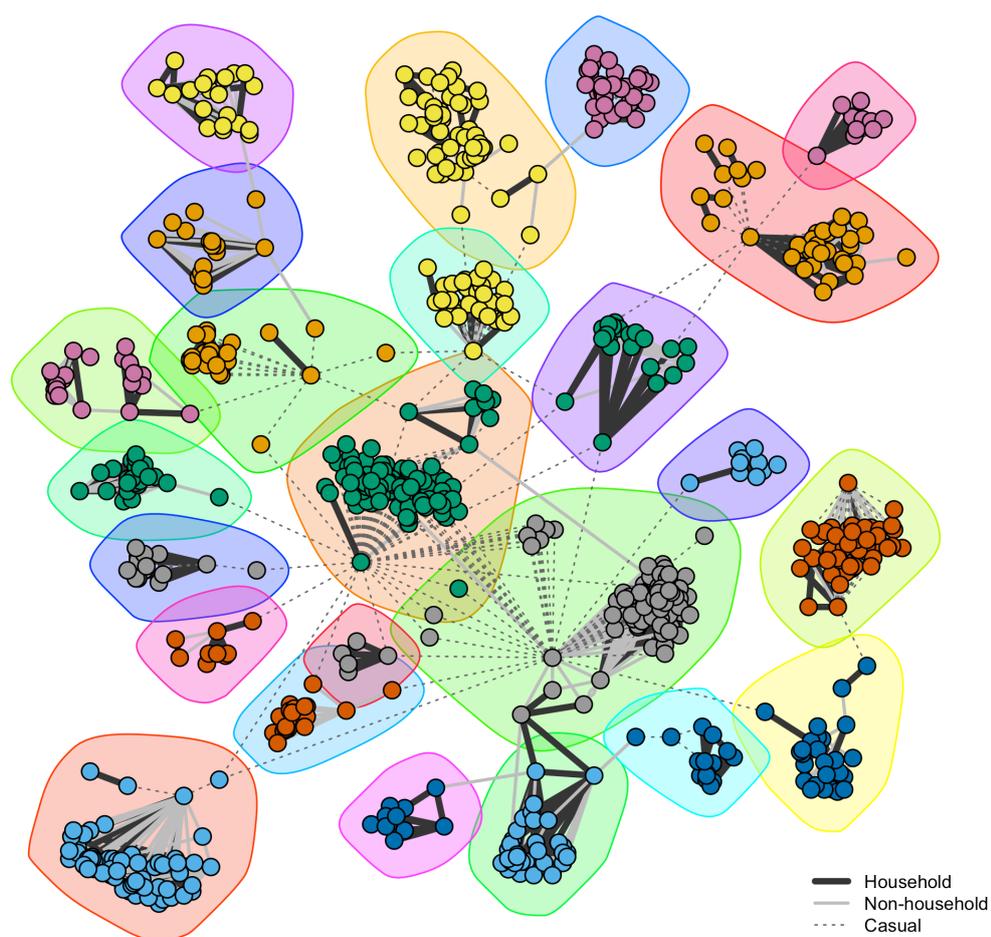


Figure 32. The simplified largest component grouped, in coloured polygons, as delineated by the Clauset-Newman algorithm. Nodes are not coloured to simplify the diagram.

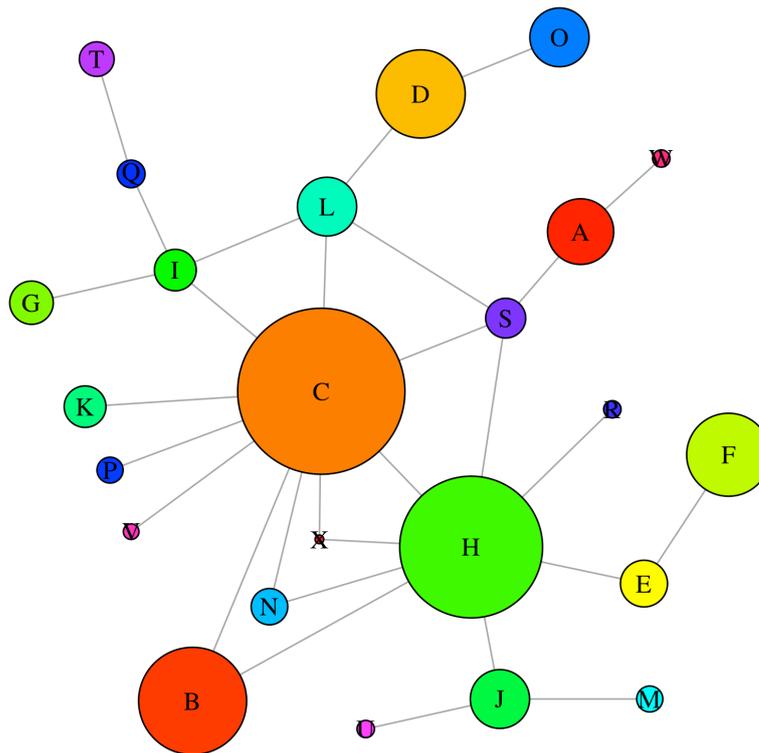


Figure 33. The simplified largest component contracted into groups. Groups are referenced by a letter (label). Size of each node is proportional to the sum of case and contact nodes with at least 2 edges in each group. Edges between groups were simplified and do not reflect the actual number of links between groups.

#### 6.4.5 Important nodes in the largest component: centrality measures

In order to understand which nodes acted as brokers or bridges for linkage, I first examined where particular nodes sit in the final static network. Figure 33 showed that nodes in groups C and H played a central role in connecting subcomponents of the largest component. This is affirmed in Table 21 which shows nodes with the highest centrality measures (closeness, betweenness, eigenvector and degree centralities) at the final timepoint. Node 12063 (group C) and node 122256 (group H) had very high scores for all. The node with the highest degree centrality (278) was node 141476 in group F (not shown in Table 19), but this node had low scores on other centrality measures (closeness centrality of 0.08, betweenness centrality of  $8.8 \times 10^5$  and eigenvector centrality  $1.3 \times 10^{-10}$ ).

Other nodes in groups C and H that ranked highly on closeness centrality also ranked highly on eigenvector centrality but not betweenness or degree centrality (Table 21). These nodes were mutual contacts of the most central nodes (node 12063 and 122256).

Similarly, two nodes in group S (node 124648 and 129127) were mutual contacts of node 12064 and node 122256.

A node in group X (148438) had high closeness and betweenness scores, but lower eigenvector scores. This node was named as a contact three times by nodes from communities C, I and L.

Node ID	Group	Closeness centrality	Rank	Betweenness centrality	Rank	Eigen-vector centrality	Rank	Degree centrality	Rank
12063	C	0.198	1	$2.68 \times 10^6$	2	1.000	1	270	2
122256	H	0.197	2	$2.90 \times 10^6$	1	0.803	2	246	3
148438	X	0.182	3	$9.81 \times 10^5$	11	0.064	30	9	82
133456	H	0.182	4	$6.93 \times 10^5$	17	0.111	4	20	29
124648	S	0.180	5	$5.49 \times 10^5$	23	0.105	7	6	128
133375	C	0.176	6	$5.84 \times 10^5$	22	0.059	45	5	144
129127	S	0.176	7	$2.51 \times 10^4$	151	0.107	5	9	82
148045	C	0.176	7	$2.51 \times 10^4$	152	0.104	8	3	177
149058	C	0.176	9	$1.89 \times 10^4$	181	0.105	6	8	91
149044	H	0.176	10	0	358	0.103	9	2	223
149068	C	0.176	10	0	358	0.103	10	2	223
148824	C	0.176	10	0	358	0.103	10	2	223
150010	H	0.176	10	0	358	0.103	12	2	223
148815	H	0.176	10	0	358	0.103	12	2	223
148821	C	0.176	10	0	358	0.103	12	2	223
149094	H	0.176	10	0	358	0.103	15	2	223
149054	C	0.176	10	0	358	0.103	15	2	223
148902	H	0.176	10	0	358	0.103	15	2	223
149001	H	0.176	10	0	358	0.103	15	2	223
149050	H	0.176	10	0	358	0.103	19	2	223

Table 21. Centrality scores for key nodes in the static largest component.

#### **6.4.6 Description of groups containing the most central nodes (node 12063, group C and node 122256, group H)**

Groups C and H play a crucial role in the development of the component, and so I have investigated them in more detail here. Figure 34 shows a directed subgraph for groups C and H, which contained the most central nodes (node 12063 circled in red and node 122256 circled in green). Nodes within one path length from these nodes and also the immediate neighbours of the first order nodes were included.

##### **6.4.6.1 Group C – a school-based super-spreading event.**

Node 12063 was a 15-year old who was symptomatic for two months while attending school before diagnosis of extensive cavitary pulmonary disease in October 2008. The case had no known prior exposure to TB, was UK-born, of Indian subcontinent ethnic group and had not travelled abroad in the five years prior to diagnosis. Contact tracing was done sequentially within the household and class groups of the index case but later expanded to students and school staff in the same year, the year above and year below the case (172 adults and 98 children aged 16 years old and under). This large contact tracing exercise was responsible for the peak of individuals joining the largest component in 2009 (Figure 28). The majority of contacts were UK-born (72%, 195/270) and of white ethnicity (65%, 175/270).

Eleven school children had evidence of incident pulmonary disease and 57 had latent infection (Figure 34, directed links from node 12063 circled in red).). All incident diseased cases were symptomatic after symptom onset in the index case. No cases were diagnosed in adult contacts.

Four additional school children had already been treated for latent infection in the past after contact tracing from separate index cases, while another child had been treated for active disease in 1997. Figure 34 shows that many of the contacts from this school outbreak had a second contact episode either before or after the school microepidemic.

Molecular typing (by 24-loci VNTR) for node 12063 matched to isolates from seven previous cases diagnosed in 2004 and 2005, all of whom were of Black African ethnicity and born in a single North African country. Four of these cases could not be epidemiologically linked to any other case, three of whom had recently arrived (within five years of their diagnosis) in the UK. The remaining three cases were a household (parent and two children). One of these children was 16 years old at the time of her diagnosis with pulmonary TB in 2004 and attended the 2008 outbreak school. Her attendance at school did overlap with the 2008 index case's attendance but no direct epidemiological links were identified. Contact tracing at the school was limited to 19 close friends of the 2004 case, 11 of whom completed screening with one latent infection case diagnosed. No further cases were diagnosed in any of these contacts up to the end the study period in 2011.

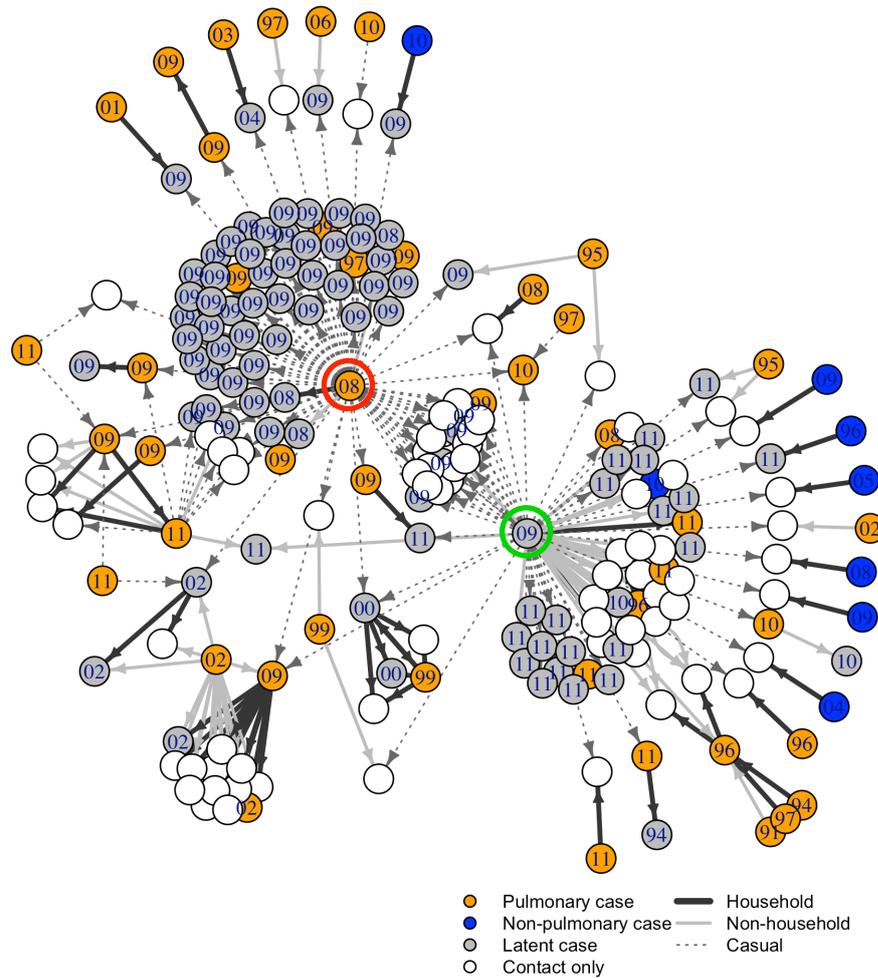


Figure 34. Subgraph of group C (containing node 12063 circled in red) and group H (containing node 122256 circled in green) in the largest component. Labels denote the year of diagnosis (09=2009, 99=1999). Nodes who did not become cases and had only one incident edge were removed. Only nodes up to 2 (undirected) path lengths away were included. A microepidemic of 11 secondary disease and 57 latent infection cases were detected on contact tracing.

#### **6.4.6.2 Group H**

Node 122256 in group H was a child directly linked to node 12063 in group C (Figure 34). Screened during the expanded school contact tracing exercise for node 12063 in 2008, node 122256 was diagnosed and treated for latent infection in 2009 (aged 14). In 2011 node 122256 was diagnosed with pulmonary TB following symptoms for four months; the TB isolate had identical by 24-loci VNTR to that of node 12063. This index case had previously also been exposed to TB and contact traced following disease in a grandparent in 1996, although the grandparent's isolate had no typing information.

Four household contacts were diagnosed with incident infection following contact tracing (disease in a 15-year old sibling and one parent, and latent infection in the other parent). Forty-two close, non-household contacts consisting of extended family and school friends were also contact traced with two further disease cases diagnosed (school friend and an adult) and four latent cases (three under 16s and one adult). A further 200 individuals (students and staff) in the same year at school were subsequently screened and 20 latent infection cases were diagnosis; all except one were in school children.

Among school contacts of node 122256, two children had previously been treated for pulmonary TB in 2008 and 2010 respectively (Figure 34). The 2008 case (node 154226) had self-presented with pleural TB two weeks before the diagnosis of group C's index case (node 12063). Her symptoms had started two months after symptom onset in node 12063. She had no known TB exposure in the past and her TB isolate was identical to the outbreak strain. Nineteen household and close contacts were identified for contact tracing but none completed screening. As contact tracing for node 12063 was undertaken at the school not long after her diagnosis, no further school contacts for her were sought.

The second previously treated case among node 122256's contacts occurred in 2010 (node 152450). This student also self-presented with pulmonary TB after two months of symptoms and the isolate matched the outbreak strain. This student was not identified as a contact to group H's index case (node 12063) in 2008. The reason for this is unclear as she did attend the same school in 2008.

Similar to the situation in group H, school contacts of group H's index case were commonly contacts again from TB cases before and after the school outbreak.

#### **6.4.7 Molecular epidemiology of the largest component**

Of the 223 diseased cases in the largest component, 63 (28%) had molecular typing information (Figure 35). Sixty-five per cent (41/63) of those with available MIRU-VNTR typing had 15-loci data only.

Genetic distances between typed nodes are shown in Figure 36. Nodes within group C and H (the school outbreak) were molecularly clustered together, and also were very close to group R (a household) and two second generation cases (L1 and I2). These nodes were distinct from other circulating strains in Birmingham, suggesting a true outbreak (Figure 37). Nodes L1 and I2 were cases that generated further large contact tracing exercises within a second school. Over 200 individuals were examined in total with 31 diagnosed with latent infection and two with disease.

Three further nodes in group C (C2, C3 and C5), linked via their household contacts who in turn were casual contacts to the central node in group C (C1) were genetically further away (Figure 36). Similarly other nodes in group H (H1, H2, H3) were more distant both molecularly and epidemiologically to the central node in this group (H5).

Beyond the (molecular) school cluster was a group of genetically close nodes related to the East-African lineage (root 4, Figure 36). These nodes were within households in groups F, G, I and V (Figure 35). Epidemiologically the households were linked by casual edges to the school outbreak. Cases occurred in different years with group

F in 2010, group G in 2003, group I in 2008, group V in 2003 and 2005. Another molecularly close node within the East-African lineage was I1, diagnosed in 2007. Although this node was epidemiologically linked to the central node in group H (H5) by a household who named both parties as close contacts (Figure 35), they were molecularly distant (Figure 38).

A further molecular cluster within the Delhi/ Central Asian lineage was evident among group O nodes (Figure 40). Epidemiologically these nodes were schoolmates and household contacts of node O7 diagnosed in 2007 at a third school (Figure 35). Node D4, another genetically similar node to node O7, was diagnosed in 2011 and was a teacher at the same school as O7 but not a contact to him/her. A second teacher and student at the same school as node O7, nodes D1 and D5 respectively, were diagnosed in 2004. Genetically nodes D1 and D5 were further away from node O7's molecular cluster (Figure 40) and therefore unlikely to have been part of the same transmission chain.

Nodes A5, A6 and B1 genetically shared a common ancestor with nodes D1 and D5 (Figure 40) and were epidemiologically distant (Figure 35). Nodes A5 and A6 shared a household. Node A6 attended a fourth school that required contact tracing at school, with several school contacts themselves in households affected by TB (Figure 35). Finally node B1 was a teacher at the outbreak school involving groups C and H, diagnosed in 2009, but clearly was not part of the school outbreak given the molecular distance between their isolates.

The remainder of nodes with typing data available were less genetically clustered (Figure 39 and Figure 40). A minority of these nodes were in households that were epidemiologically close (within 2, close edge, path lengths) but genetically distant to nodes in other household (e.g. G2 and G1, U1 and J1, R2 and R4) (Figure 35). Within the same household there was one instance of genetically distant nodes (D9 and D4). The remainder nodes were epidemiologically linked casually to other nodes as expected from their genetic distance.

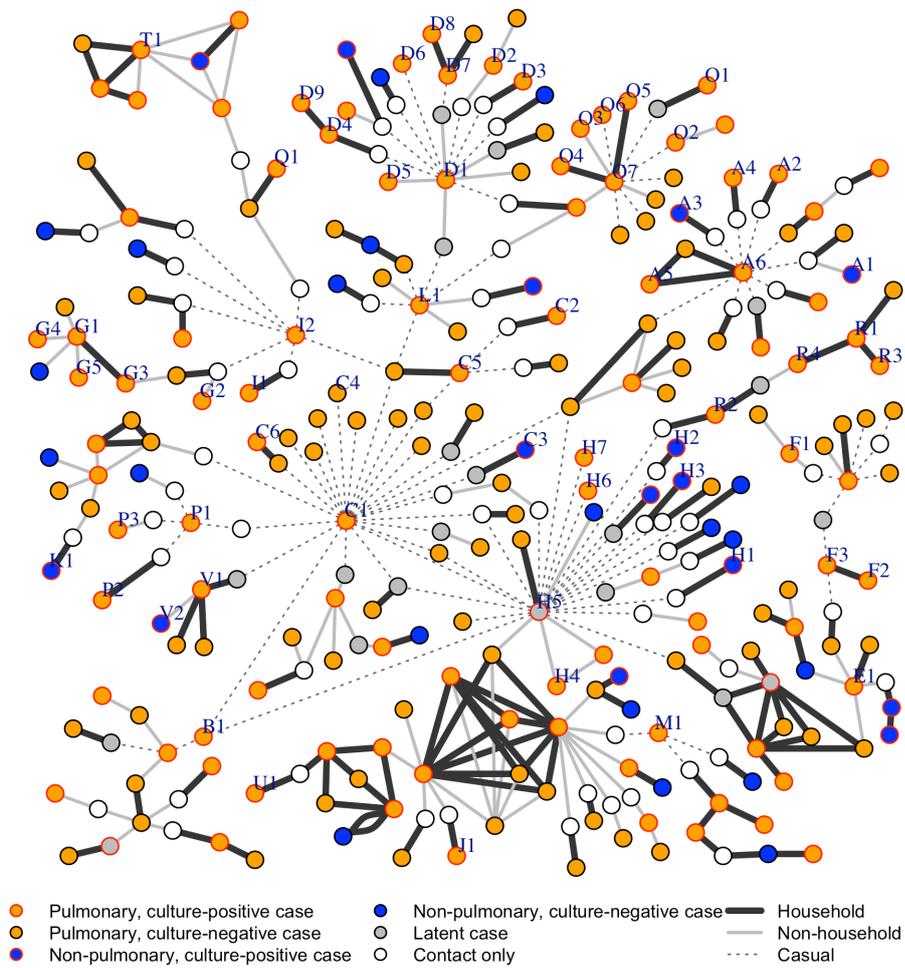


Figure 35. Reduced graph of the largest component showing culture-positive and culture-negative nodes and their MIRU-VNTR typing availability. All case nodes were retained in this graph. Latent and contact only nodes were excluded if they did not result in an increase of the number of connected components. Labels are depicted for nodes with MIRU-VNTR typing information with the alphabet referencing group membership (see Figure 33) and unique number in each group referencing an individual node.

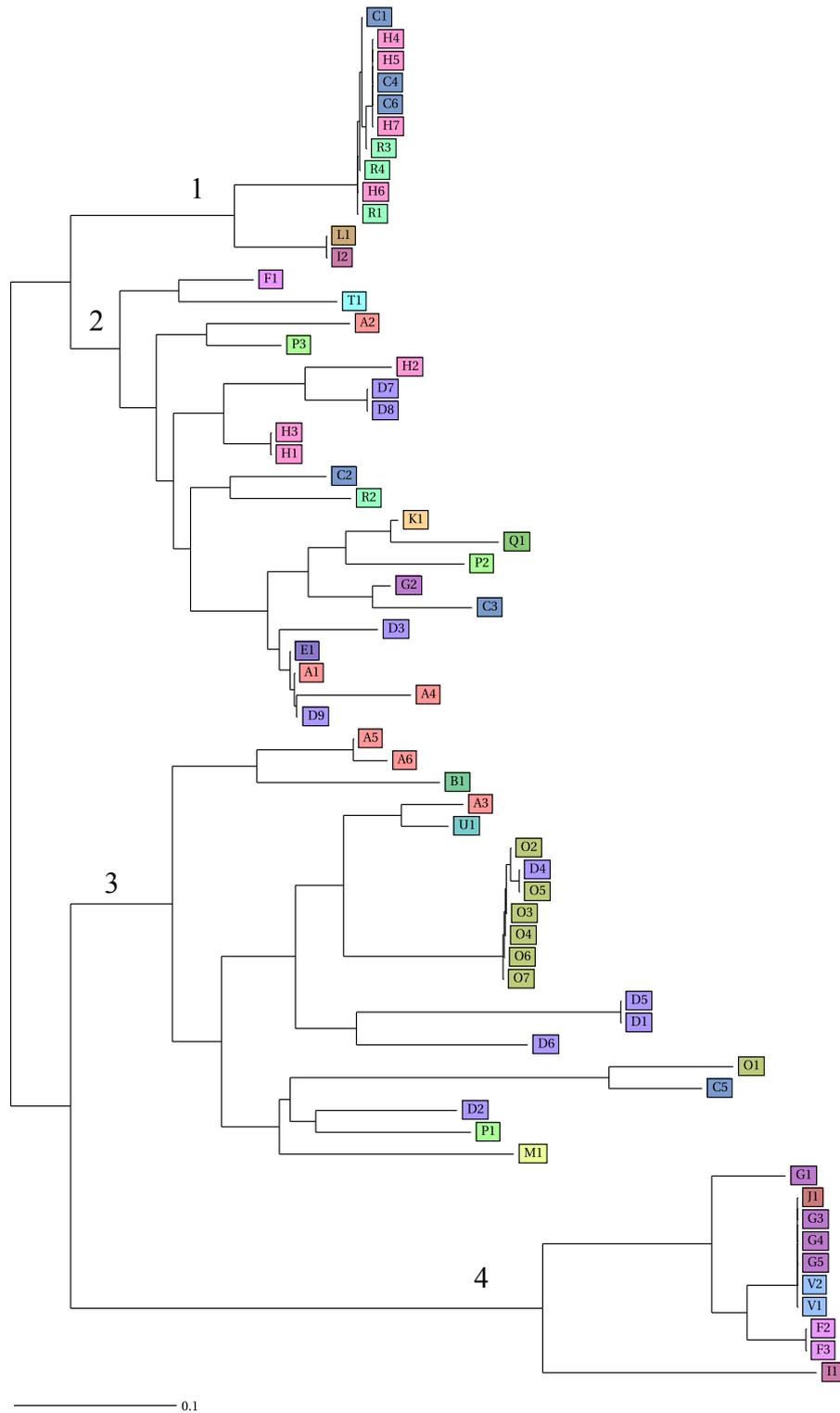


Figure 36. Phylogenetic distribution of 63 nodes with 15- or 24-loci MIRU-VNTR sequences in the largest component. Numbers highlight the root at which more detailed tree analyses was done with a sample of unique MIRU-VNTR sequences in the wider database included for reference (Figures 37 – 40). Nodes were labelled with an alphabet referencing epidemiological group membership (see Figure 33) and unique number in each group referencing an individual node, and were coloured according to their epidemiological group.

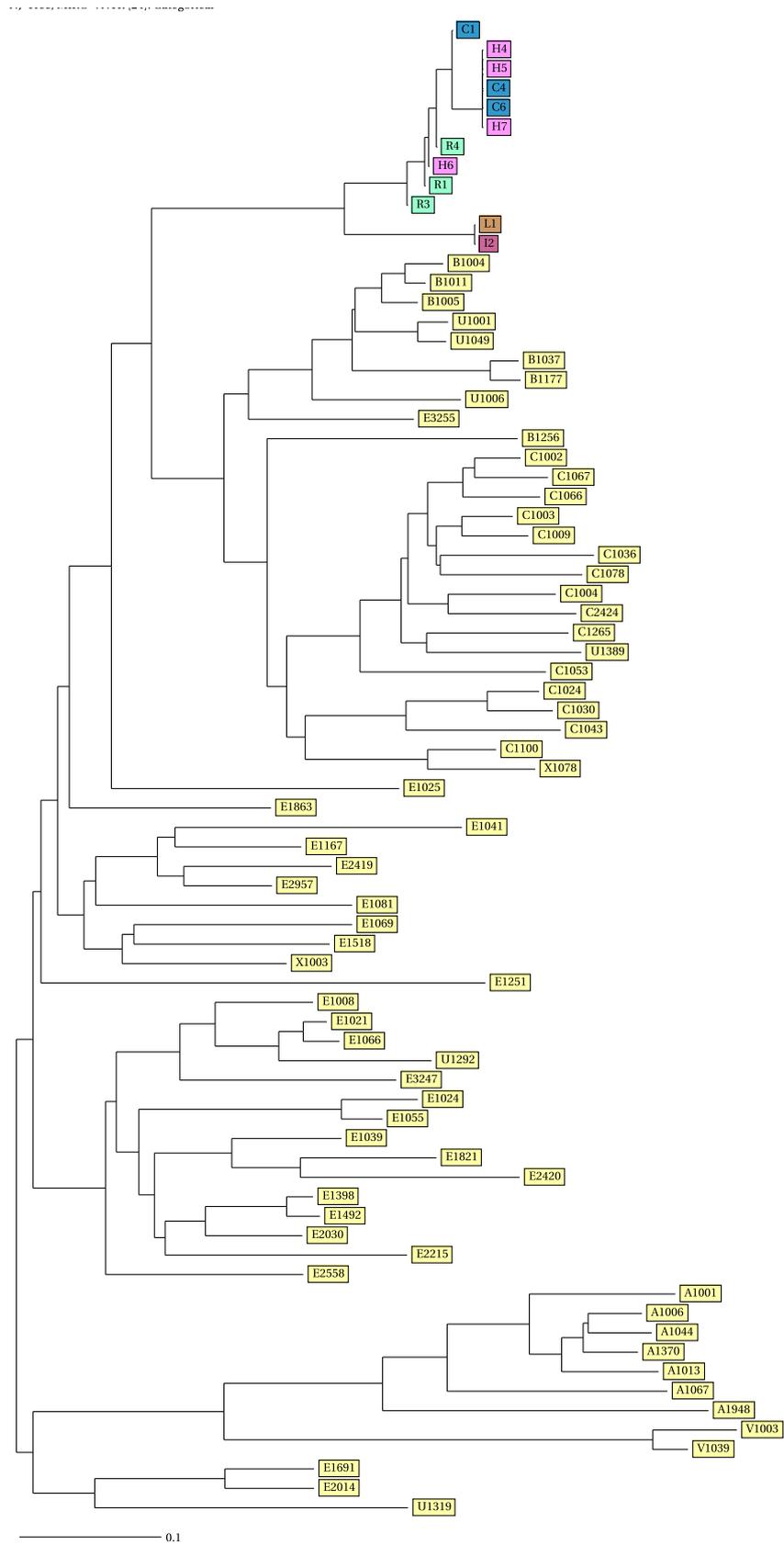


Figure 37. Phylogenetic distribution of nodes within root 1 (Figure 36) in comparison to the wider MIRU-VNTR sequences in Birmingham.

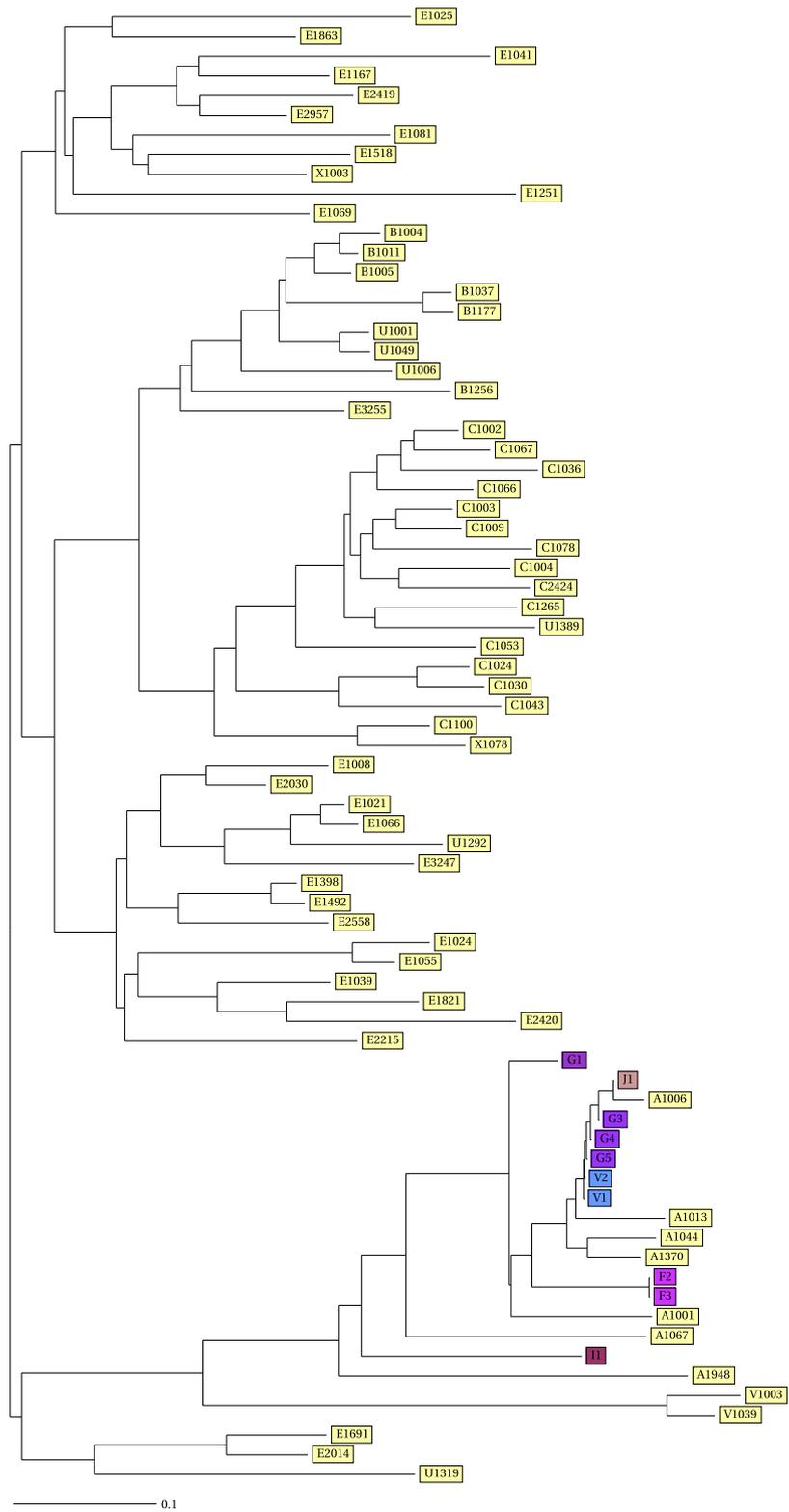


Figure 38. Phylogenetic distribution nodes within root 4 (Figure 36) in comparison to the wider MIRU-VNTR sequences in Birmingham.

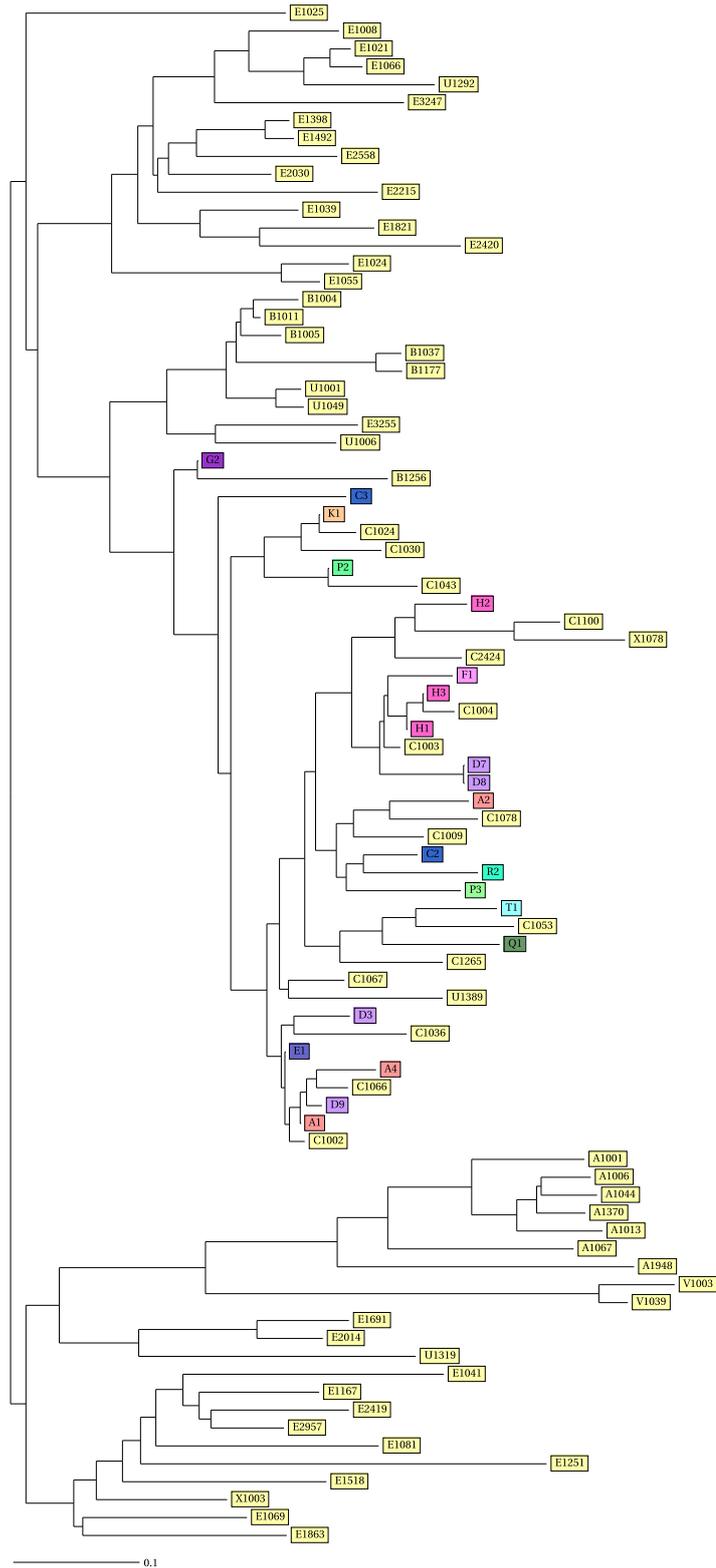


Figure 39. Phylogenetic distribution of nodes within root 2 (Figure 36) in comparison to the wider MIRU-VNTR sequences in Birmingham.

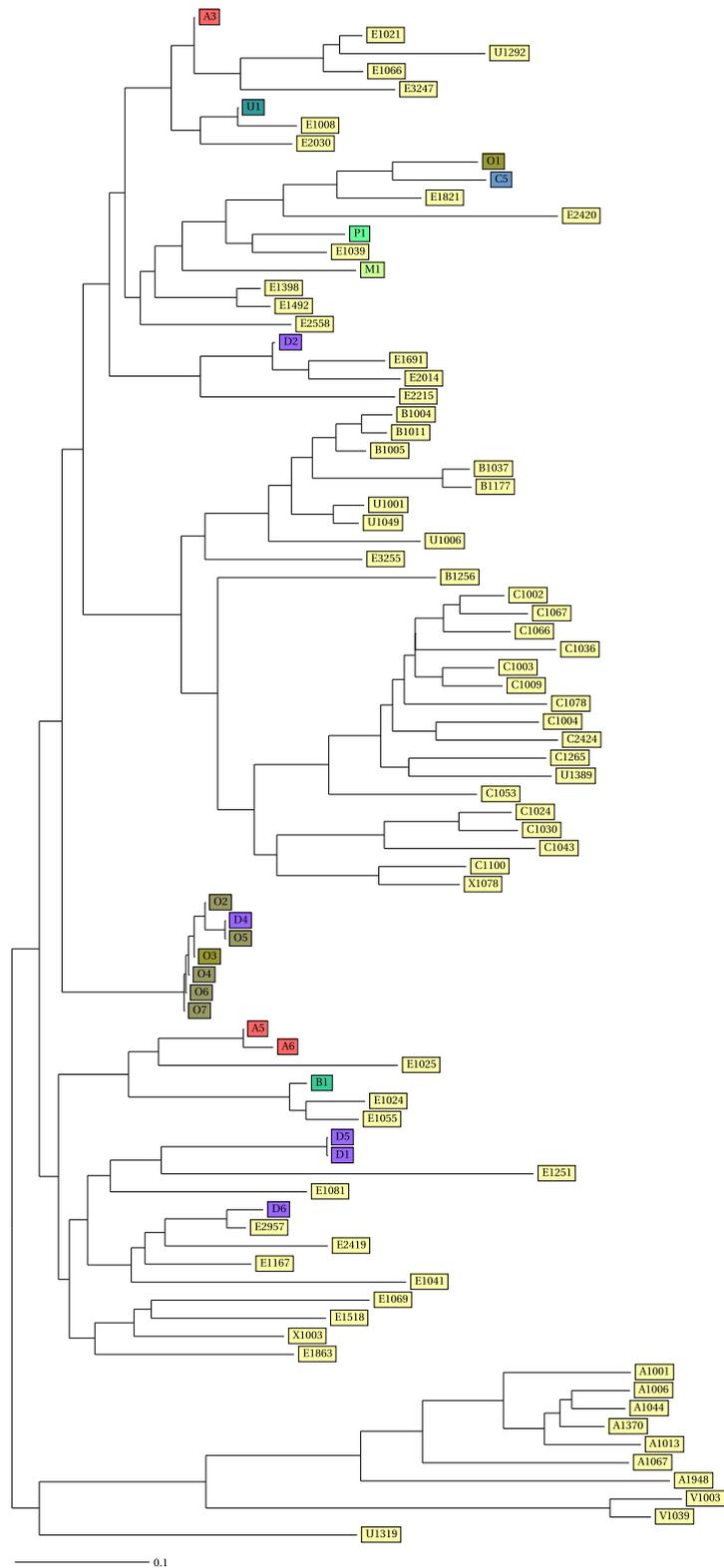


Figure 40. Phylogenetic distribution of nodes within root 3 (Figure 36) in comparison to the wider MIRU-VNTR sequences in Birmingham.

## 6.5 Discussion

The largest component differed from other components by a higher proportion of case nodes with pulmonary TB. Thus the largest component may have formed as a result of more contact tracing done overall (as non-pulmonary cases usually only have household contacts traced). However the proportion of case nodes with latent infection was also double that in the overall network. This suggests that the host population in the largest component was different – half of the individuals involved were adolescents with a median age of 16. Unlike younger children, adolescents can present with features of adult-type disease (most often pulmonary) as well as intrathoracic lymph node or extrapulmonary disease that is often seen in children (Cruz *et al.*, 2013; Snow *et al.*, 2018). They may be more susceptible to developing infection and disease due to altered metabolic processes associated with adolescent growth ((Wilcox and Laufer, 1994). In a cross-sectional survey of tuberculin skin test response by age groups in a high TB burden setting, the rate of acquiring latent TB (i.e. the force of infection) increased through childhood and reached a maximum at age 15 years (Wood *et al.*, 2010). In this analysis nodes that were in a vulnerable age group within a congregate setting were also factors likely to have contributed to the size of the largest component.

At the same time, the significant number of infectious nodes could enable contagion to spread wider and more rapidly resulting in a bigger connected component. However, case nodes in the largest component had similar social networks to cases outside the largest component. Cohesion, or how tightly linked the component was as measured by density, clustering coefficient and transitivity was no different. Although there was some suggestion of larger households in the largest component, the difference was not statistically significant. Pulmonary case nodes however had slightly higher indegree, i.e. they were more likely to have been named as a contact before or after their diagnosis. This is likely a reflection of the school outbreak in which subsequent secondary cases resulted in repeat contact tracing of the same individuals within the school as there was no significant difference in the outdegree distribution between the largest and other components.

The community or group-finding algorithm was able to readily identify subgroups in the largest component that mapped correctly to epidemiological clusters of index cases and their concentric circle of traced contacts. When the most central nodes were identified it was evident that these nodes were a probable source for spread of infection among a susceptible school-age population, compounded by prolonged infectious periods.

The genetic distance between nodes was underestimated in this analysis. Missing loci were ignored so apparent genetically close nodes may still be distant to each other. Nevertheless overlap of the available molecular data and epidemiological links in the most central nodes was highly suggestive of a superspreading event at a school spanning four years. At the same time molecular data enabled distinguishing several individuals in the school outbreak who were not part of the outbreak strain but involved in concurrent outbreaks in the wider community. This suggests that contact tracing may be identifying risk groups rather than transmission groups and the number of secondary cases due to the outbreak strain may have been overestimated. Additionally molecular data was able to highlight potential gaps in contact tracing in the case where molecular data matched the outbreak strain but no epidemiological link was confirmed. Thus mismatches of molecular and epidemiological data can reflect the quality or comprehensiveness of contact tracing within a TB control programme. The interpretation of epidemiological and genomic data was resource intensive but retrospective information about exact chains of transmission or missed opportunities for contact tracing could inform more intelligent contact tracing in the future.

TB in children are important markers of transmission. The burden of TB in adolescents and young adults in particular has only recently become a focus for surveillance and control (Snow *et al.*, 2018). School outbreaks involving adolescents are well reported (Watson and Moss, 2001), including the school outbreak described here (Caley *et al.*, 2010). Previous cross-sectional analyses implied that the wider community surrounding children with TB are involved. The descriptive analysis of the largest component shows the extent to which this can occur: multiple, separate components (developing at different times over a three decades) were bridged by a clustering of social contacts in a school. The large

numbers of individuals screened at a single location was able to link in to this high risk network in the community more deeply than contact tracing around a single active case. These features may make school-based intervention where there is TB case efficient. At present there is no prioritisation of control activities for adolescents with TB in the UK. Wider epidemiological evaluation of TB in adolescents in UK cities is necessary to confirm whether enhanced focus on this group may be useful and what strategies are feasible.

## 7 Measuring growth in the contact tracing network

### 7.1 Introduction

The goal of contact tracing is early detection of individuals in close proximity to an infectious case, among whom prevalence of disease is likely to be high. Knowledge of these individuals allows local control measures to limit disease spread, e.g. vaccination (Fenner *et al.*, 1988), isolation (Donnelly *et al.*, 2003) or treatment to prevent secondary disease or reduce the infectiousness of those already infected (Althaus *et al.*, 2014). Because the formation of contact tracing networks relies on disease diagnosis events, i.e. each new case is in turn asked for their contacts that are then examined for infection, continued branching of the network suggests wider contagion.

In chapters 5 and 6 we examined the properties of the observed static contact tracing network, in which the vast majority (over 90%) of linked individuals remained in small radial or star-like components (clusters) over time. Predicting which component will gain more nodes (individuals) and edges (links) may enable control programmes to target case finding and intervention efforts better. If component growth can be detected but not prevented, e.g. the epidemic of concern was already in a phase where multiple individuals were already infected, ability to predict component growth may also be useful as a monitoring tool to distinguish failure of contact tracing from expected discovery of infected individuals during the process.

In network theory, the best-known example of how networks grow is the “preferential attachment” model. In this model, nodes added sequentially to a network attach to nodes already in the network with a probability that is proportional to its degree (Barabasi and Albert, 1999; Price, 1976). This mechanism of growth has been shown to explain real networks that have power-law degree distributions, ie

networks where most nodes have very few connections with only a few nodes, acting as hubs, being highly connected. Preferential attachment based on degree distribution does not intuitively explain growth of the contact tracing network as observed thus far – higher outdegree was not associated with increased number of diagnoses in contacts (Chapter 5). However the contact tracing process may not reveal all links and the branching topology of the network it produces may not be the true structure of contact between individuals.

A method of explaining growth that simulates implementation of the contact tracing process has been explored using the Galton-Watson process. In a transmission tree of infected individuals where the primary infected is termed an infector and secondary cases arising from the infectors are termed infectees, infector/infectees can be discovered via contact tracing with a certain probability  $p_c$ . Both infectors and infectees can be traced from the first generation onwards, but the primary infector can only be backward traced. The algorithm runs recursively over the transmission tree and thus discovers the size of the connected component (Muller *et al.*, 2000). The minimum tracing probability required to limit the eventual component size, or the effective reproduction number, can then be estimated. The model relies on individuals being in the same chain of transmission and does not take into account infection from another source. However, previous investigations in Chapter 5 and Chapter 6 suggest that in our setting linked individuals often have different molecular isolates and do not always share a transmission tree.

Change in longitudinal data can be empirically studied using the well-documented methods of survival analysis (Cox and Oakes, 1984), which focuses on if and when an event of interest occurs in time and how its occurrence varies with predictors; and individual growth modeling (Rogosa *et al.*, 1982; Willett, 1988) or multilevel modeling (Goldstein, 1995), which focuses on how a dependent variable, in our example, size of a component, changes with time. Neither of these methods has been applied to explore the processes that generate networks. Nevertheless their familiarity makes them applicable to our problem of measuring growth in the contact tracing network.

When the contact tracing network was visualised in Chapter 5, components grew in diameter if infected contacts named more contacts or if separate radial/ star components had mutual contacts. Applying a survival approach, events causing component growth are a contact 1) naming a further contact, because the contact was diagnosed with infection (secondary contact tracing) or 2) is named as a contact again by a different case (regardless of disease/infection status in the contact) (Figure 41). The advantage of a survival approach is that partially observed (censored) events are taken into account because they still contribute to the total time at risk until the end of the study period or individuals are lost to follow-up. This is particularly important because larger components were more likely to have had more time to evolve. Conducting time to event analysis on individual nodes clustered within the same component will involve repetition of analysis linked to the earliest, primary case (zeroth generation). Thus failure times will not be independent. However, marginal or frailty models can still provide consistent estimates of covariate effects. Node level attributes of the first generation contact and his/her adjacent nodes on event occurrence/timing, e.g. composition of relationship (within household or outside household) or demographic similarity between them can be examined.

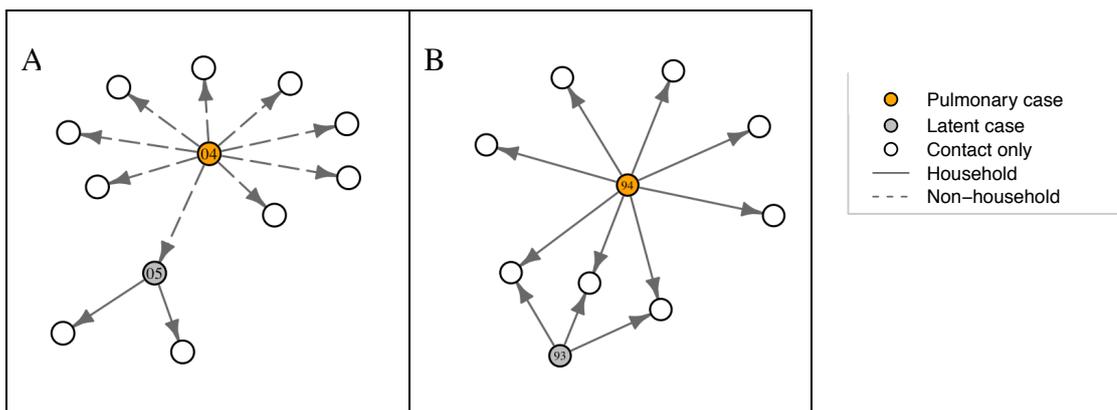


Figure 41. Examples of events causing component growth. A contact is diagnosed with TB and secondary contact tracing occurs (A) or a contact is named twice by different cases (B). Node labels denote the year of diagnosis.

Individual growth models can be applied to whole networks (socio-centric approach). This alternative method to study change in component sizes can answer different questions such as identifying fast- or slow-growing components and whether cluster-level characteristics (such as presence of recurrent cases, proportion of nodes born in the UK, median time delay between connected cases) determine growth trajectories. The main disadvantage of this approach is obtaining biased estimates of (individual) cluster growth due to right censoring. Furthermore application of growth models to contact tracing networks has not been attempted previously and there are no well-defined cluster-level variables.

The feasibility of both approaches will be explored in this chapter. Methods and results will be presented separately for the node-level survival analysis followed by individual growth/ multilevel modeling at the component level.

## **7.2 Objectives**

- Estimate the risks and factors associated with node-level events that cause component growth, i.e. secondary contact tracing and (outdegree from ego) and being named as a contact a second time (indegree from alter).
- Quantify the risk of a TB diagnosis in a second generation contact
- Describe the growth trajectories of components in the contact tracing network and explore the extent of systematic variation between components.

### 7.3 Survival analysis approach

Data for this analysis are all cases and contacts in the Birmingham TB register (described in Chapter 2) with at least one edge. Growth was defined to occur when a contact, in a case-contact pair (the ego), developed a link to a further individual (the alter), by (1) naming the individual (sending an ‘outdegree’) or (2) being named by the individual (receiving an ‘indegree’) (Figure 42).

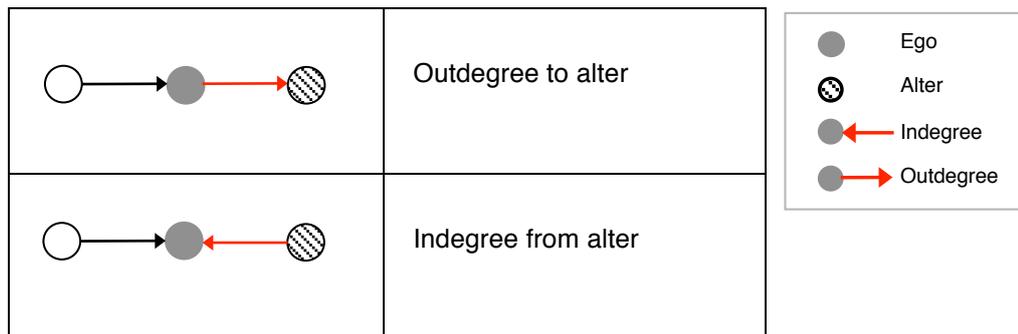


Figure 42. Schematic depicting growth events in a network. Circles are individuals in the contact tracing network.

Case-contact pairs were at risk for multiple events, as depicted in Figure 43. In particular, both ego contacts and alter contacts could be diagnosed after (forward traced) or prior to (backward traced) receiving an indegree or outdegree. As only future cases can subject to intervention, only alter events occurring forward in time were analysed (boxed events, Figure 43). In addition because different event types can occur to the same ego and were not mutually exclusive, separate one-state survival models were constructed for each event type.

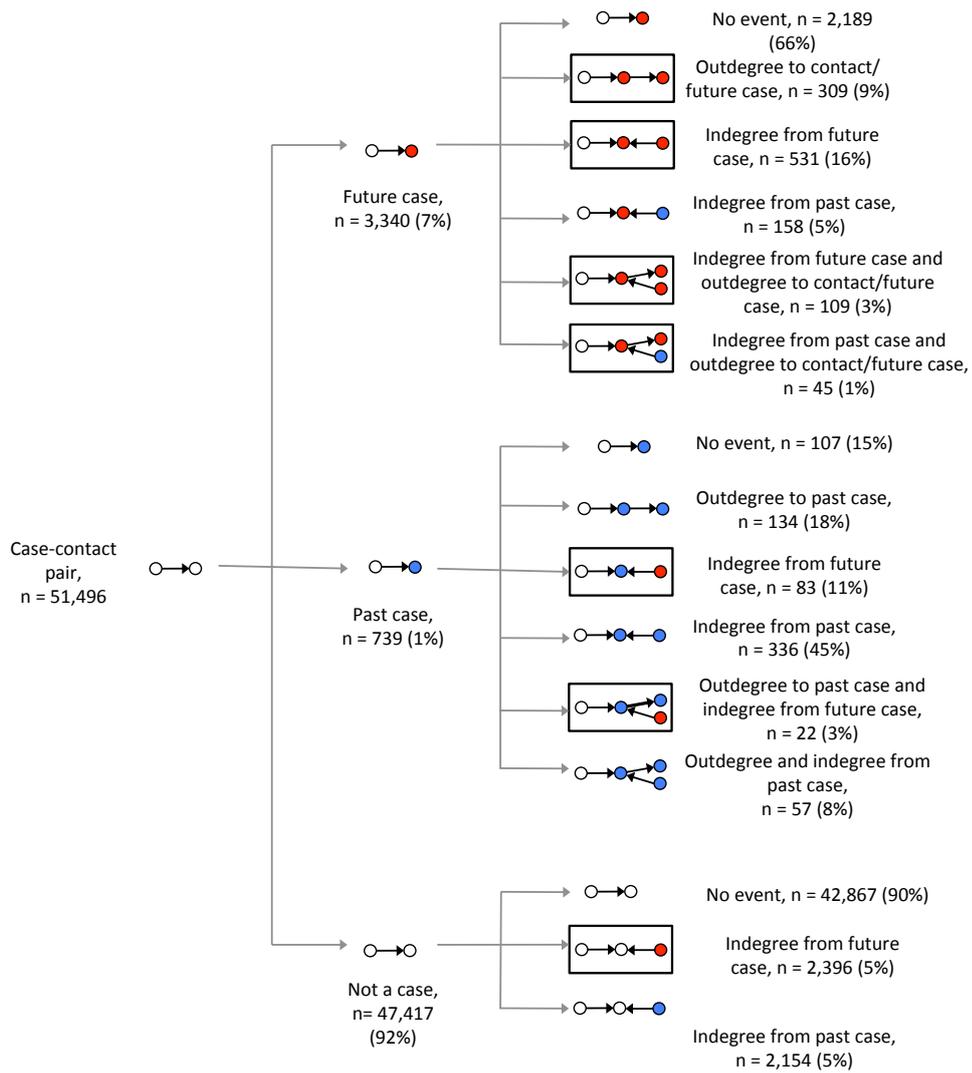


Figure 43. Event histories for (directed) case-contact pairs. Blue node denotes backward traced case, red node denotes forward traced case.

### 7.3.1 Method

All data for the survival approach was prepared and formatted using the *R* package Biograph (version 2.0.6) (Willekens, 2016). The package mvna (version 2.0.1) (Allignol *et al.*, 2008) was used to estimate the cumulative hazard and etm (version 0.6-2) (Arthur Allignol *et al.*, 2011) was used to estimate the cumulative incidence function. The coxph function from the survival package (version 2.42-3) was used to investigate the association between time to event of interest and predictor variables. The Efron method was used to handle ties and the robust method was used to adjust for egos linked in multiple case-contact pairs. Proportionality of hazards was examined graphically from log-log hazard plots and from Schonfield residuals.

#### 7.3.1.1 Terminology

The terminology used in this chapter is consistent with previous chapters. A summary of terms used in this section is provided in Table 22 for reference.

Term	Definition
Alter	A node that sends or receives links or edges to an ego.
Case/infection	Infected with either latent TB or (active) disease.
Casual edge	An edge between individuals not named directly by an index case but identified to be present at the same non-residential address at the time of the infectious period of the index case
Close edge	An edge between an individual named by an index case who does not reside at a mutual residential address
Degree	The number of edges associated with a node. Indegree is the number of edges the node receives. Outdegree is the number of edges the node sends.
Edge	The relationship or link between nodes.
Ego	The contact in case-contact pair. Growth of the contact tracing network is measured by addition of links from this node. The ego is also the first generation contact.
Household edge	An edge between an individual named by an index case who resides at a mutual residential address.
Index case	An infected individual with at least one contact.
Node	An individual in the contact tracing network.
Network	The total collection of edges between a group of nodes. Some nodes may not have edges to another node.

Table 22. Definition of common terms used in the survival approach to measuring growth in the contact tracing network.

### 7.3.1.2 Outdegree to alter

The purpose of this analysis was to quantify the rate of second generation contact tracing, explore whether second generation contact tracing was associated with certain characteristics and quantify the risk of infection in second generation contacts.

The initial status in this model was a case-contact pair in which the contact/ego was a forward-traced case. Two sequential events were subsequently of interest (1) outdegree to an alter (second generation contact tracing) (2) the alter being diagnosed as a TB case (latent infection or active disease) in the future (risk of infection in second generation contact) (Figure 44). If multiple second-generation cases were observed, only the first (earliest in time) was considered.

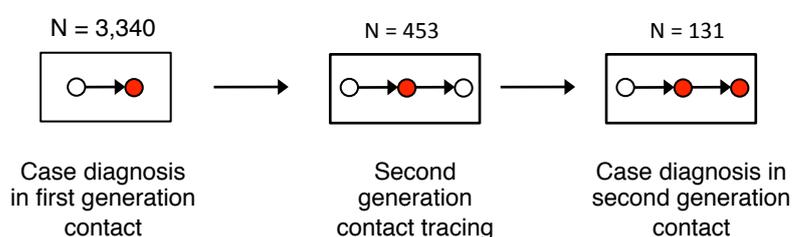


Figure 44. Transition states in a survival model to evaluate second generation contact tracing and its yield in case diagnosis.

Time to an outdegree to alter was calculated from the case notification date of the ego. Time to case diagnosis in an alter was calculated from the time from entry into the first event, i.e. time from diagnosis in the ego. Follow-up was censored at the end of the study period (31 December 2011).

Edge type between the initial case-contact/ego pair, ego disease status, age, gender, ethnic group and birthplace and cohort (year of diagnosis in the index case) as risk factors for an outdegree to alter was assessed in a univariable Cox proportional hazards models. A multivariable model was not constructed as it was known that secondary contact tracing was done depending on the experience of the TB nurse managing the case (Catherine Brown, 2017) and a full model would not be clinically useful. A Cox model was also not constructed for time to case diagnosis in an alter, because determinants of infection following TB contact were explored in Chapter 4.

### **7.3.1.3 Indegree from alter**

The initial state was the case-contact pair. All contacts, regardless of case status (forward traced case, backward traced or not a case) were considered at risk. The start of the follow-up time was the date the ego was registered to the case in the register. This time is generally after diagnosis in the case but before diagnosis in the ego. Follow-up was censored at the end of the study period (31 December 2011). Only indegrees received forward in time were considered, i.e. receiving an indegree after the start time. For egos receiving multiple indegrees, only the first indegree closest in time to the start time was included.

The main variable of interest in predicting whether an indegree was received from an alter was the case status of the ego, as this is a clinical entity that is easily identified for more intensive contact tracing. Other variables assessed in a Cox model were the ego's age, gender, ethnic group, place of birth and their outdegree, if they were a case. The type of edge (household, close or casual) between the case-contact/ego pair was termed the primary edge and the year of this edge formation was also considered.

The multivariable Cox model was stratified according to first edge type between the case-contact pair because the hazard for this variable was not proportional over time. Ego's age, gender, ethnicity and birthplace were included *de novo* because these factors were associated with the number of outdegrees (i.e. the size of their social circle).

## **7.3.2 Results**

### **7.3.2.1 Outdegree to alter**

Of the first generation contacts who were forward traced as a case, only 14% (453/3340) had second generation contact tracing (Figure 43). Thus the cumulative hazard for this event was low but gradually increased with time since diagnosis in the ego (Figure 45) corresponding to a probability of 5% at year 1 (95% CI 4.1%, 5.6%) and 10% (95% 9.1%, 11.7%) by year 10 (Table 23).

From 453 cases that had at least one second generation contact named, 131 (29%) had a contact that was subsequently diagnosed as case. The cumulative hazard for future infection diagnosis in second generation contacts was high and greatest soon after the exposure (Figure 45). This risk continued to increase steadily over time since infection in the first generation contact. Table 24 shows that if second generation contacts are traced there is a 28% chance (95% CI 18.7%, 37.4%) of a case diagnosis within two years of infection in the first generation case. This risk is not directly comparable to previous work that quantified the risk of infection in all generation contacts per TB exposure.

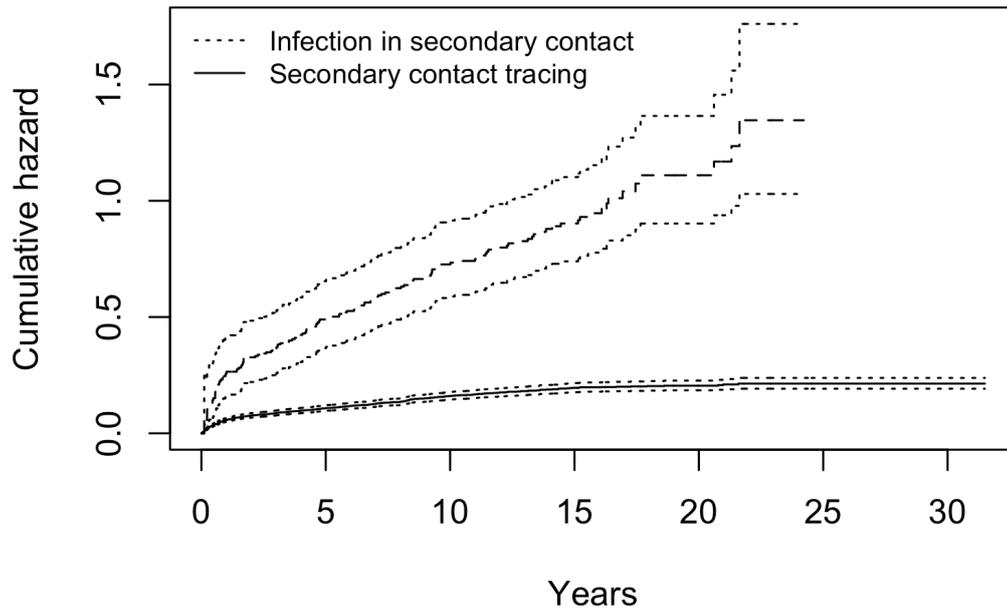


Figure 45. Nelson-Aalen estimate of the cumulative hazard for a case-case pair to have second generation contact tracing followed by infection in a secondary contact.

Years since zeroth generation diagnosis	Probability	Variance	Lower	Upper
1	0.05	1.41E-05	0.04	0.06
2	0.06	1.84E-05	0.05	0.07
5	0.08	2.59E-05	0.07	0.09
10	0.10	4.32E-05	0.09	0.12
15	0.11	4.32E-05	0.10	0.13
20	0.10	9.80E-05	0.08	0.12

Table 23. Probability (Aalen-Johansen estimator) for a first generation infected to have contacts traced.

Years since first generation diagnosis	Probability	Variance	Lower	Upper
1	0.23	2.37E-03	0.13	0.32
2	0.28	2.27E-03	0.19	0.37
5	0.39	2.00E-03	0.30	0.48
10	0.52	1.60E-03	0.44	0.60
15	0.60	1.40E-03	0.52	0.67
20	0.67	1.50E-03	0.59	0.75

Table 24. Probability (Aalen-Johansen estimator) for infection in a second

generation contact.

Table 25 shows factors associated with second generation contact tracing. In the univariable Cox model this was more likely to occur if the ego was not a household contact of the index case in the zeroth generation, had disease rather than infection, was female, was non-UK born and was diagnosed as case later in the cohort.

Variable	Hazard ratio	95% CI	P
Edge type in initial case-contact pair (categorical)			
Casual	3.31	2.45, 4.46	<0.001
Close	1		
Household	0.64	0.53, 0.79	<0.001
Ego case status			
Latent infection	1		
Respiratory	9.82	7.46, 12.93	<0.001
Non-respiratory	9.1	6.39, 12.94	<0.001
Ego age (years)	1.03	1.03, 1.04	<0.001
Ego gender			
Male	1		
Female	1.3	1.08, 1.57	<0.01
Ego ethnic group			
White	1		
Black Caribbean	0.8	0.53, 1.2	0.27
Pakistani	1.03	2.35, 3.14	0.98
Bangladeshi	0.74	0.4, 1.37	0.34
Indian	1.05	0.76, 1.43	0.77
Black African	0.88	0.56, 1.38	0.58
Other	0.44	0.22, 0.88	0.02
Unknown	2.125e-06	0, inf	0.98
Ego birthplace			
UK	1		
Non-UK	1.27	1.05, 1.53	0.02
Unknown	2.424e-06	0, inf	0.98
Year of diagnosis in ego	1	1, 1.001	<0.001

Table 25. Univariable Cox regression for time to second generation contact tracing.

### 7.3.2.2 Indegree from alter

The overall proportion of case-contact pairs extending their network by an indegree from an alter was 6% (3113/51,496). Figure 46 shows that indegree from an alter was an infrequent event with 91.5% (95% confidence interval, CI 91.1%, 91.9%) remaining as case-contact pairs only. Most of the risk occurred in the first five years after diagnosis in the zeroth generation case (five-year survival probability 96.4% (95% CI 96.2%, 96.5%).

Table 26 shows the univariable and multivariable cause-specific hazard ratios for time to indegree from an alter. Case-contact pairs with a household edge were at increased risk of an indegree from an alter on univariable analysis. However, hazard estimates for the case-contact pair edge type could not be evaluated in the multivariable model because its hazard was not proportional over time. Of note the number of outdegrees from an ego was not statistically associated with the risk of an indegree from an alter.

In the multivariable Cox model egos who became respiratory or latent cases were at higher risk of an indegree from a second alter compared to egos who were never a case (hazard ratio 1.52 and 1.16 respectively, 95% CI 1.16-1.98 and 1.16-1.93 respectively) (Table 26). Egos who were cases in the past were not at increased risk. Egos' ethnic group and birthplace remained independently statistically associated with risk of an indegree from an alter. Pakistani, Indian and Black African ethnicity had a hazard ratio of 2.96, 2.32 and 1.82 respectively (95% CI 2.21-3.95, 1.7-3.16, 1.16-2.84 respectively) when compared to egos of White ethnicity while UK-born egos had 1.24 times the risk of non-UK born egos (95% CI 1.07, 1.44).

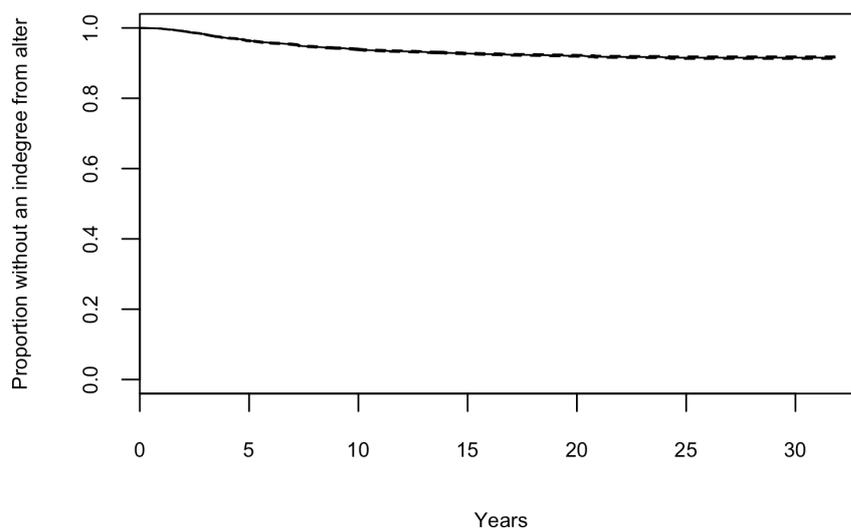


Figure 46. Kaplan-Meier estimator of duration remaining without an indegree from an alter for 51,496 case-contact pairs. Dotted lines denote 95% confidence intervals.

Variable	Univariable			Multivariable		
	Hazard ratio	95% CI	P	Hazard ratio	95% CI	P
Edge type in initial case-contact pair (categorical)				-		
Casual	1.02	0.82, 1.28	0.849			
Close	1					
Household	1.18	1.1, 1.27	<0.001			
Year initial edge formed	1.02	1.01, 1.02	<0.001	0.99	0.98, 1.01	0.27
Ego status						
Contact only	1					
Forward-traced infection						
Respiratory	4.25	3.72, 4.85	<0.001	1.52	1.16, 1.98	0.002
Non-respiratory	3.58	2.65, 4.84	<0.001	1.14	0.81, 1.62	0.45
Latent	4.41	3.95, 4.93	<0.001	1.49	1.16, 1.93	0.002
Backward-traced infection						
Respiratory	3.98	3.07, 5.17	<0.001	1.36	0.98, 1.91	0.072
Non-respiratory	1.96	0.93, 4.11	0.076	0.63	0.3, 1.3	0.21
Latent	4.15	2.97, 5.79	<0.001	1.35	0.86, 2.11	0.2
Ego outdegree	1	0.99, 1	0.237	-		
Ego age (years)	0.99	0.99, 1	<0.001	0.99		0.49
Ego gender						
Male	1			1	1	
Female	1.04	0.97, 1.12	0.258	0.05	0.96, 1.19	0.46
Unknown	0.03	0.004, 0.21	<0.001	1.02	0.007, 0.36	0.003
Ego ethnic group						
White	1			1		
Black Caribbean	1.31	1.03, 1.67	<0.001		0.66, 2.05	0.59
Pakistani	2.72	2.35, 3.14	<0.001	2.96	2.21, 3.95	<0.001
Bangladeshi	1.15	0.88, 1.5	0.321	1.36	0.84, 2.2	0.21
Indian	2.04	1.74, 2.39	0.031	2.32	1.7, 3.16	<0.001
Black African	1.94	1.49, 2.52	<0.001	1.82	1.16, 2.84	0.009
Other	1.24	0.94, 1.65	0.133	1.34	0.83, 2.18	0.24
Unknown	0.24	0.19, 0.3	<0.001	0.29	0.19, 0.46	<0.001
Ego birthplace						
UK	1			1		
Non-UK	0.76	0.67, 0.87	<0.001	0.81	0.69, 0.94	0.005
Unknown	0.22	0.2, 0.25	<0.001	0.4	0.31, 0.53	<0.001

NOTE: The hazard from edge type was not proportional with time and therefore the multivariable model was stratified by this variable.

Table 26. Cox regression for variables associated with time to indegree from alter.

## 7.4 Growth curve model

### 7.4.1 Method

A multi-level framework was applied to explore how connected components grew over time. Growth was defined as the addition of an infected (diseased or latently infected) case to a component. The level-1 submodel explored the cumulative number of infected cases as a function of time in years since the first infected case in each component, represented by

$$Y_{ij} = \pi_{0i} + \pi_{1i}(TIME_{ij}) + e_{ij}$$

where

$Y_{ij}$  the cumulative number of cases in component  $i$  at time  $j$

$\pi_{0i}$  the initial number of cumulative cases in component  $i$

$\pi_{1i}$  component  $i$ 's rate of change during the study period

$e_{ij}$  the portion of component  $i$ 's outcome that is unpredicted on occasion  $j$

The level-2 submodel to investigate predictors of inter-component differences in growth was represented by

$$\pi_{0i} = \gamma_{00} + \gamma_{01}PRED_i + u_{0i}$$

$$\pi_{1i} = \gamma_{10} + \gamma_{11}PRED_i + u_{1i}$$

where the level-2 intercepts  $\gamma_{00}$  and  $\gamma_{10}$  are the population average initial number of cases and growth rate respectively, for a component with the reference predictor ( $PRED = 0$ ), the level-2 slopes  $\gamma_{01}$  and  $\gamma_{11}$  are the effect of  $PRED$  on initial number of cases and growth rate for components with the reference predictor, and the level-2 residuals  $u_{0i}$  and  $u_{1i}$  are the portions of initial number of cases and growth rate that are unexplained at level-2.

The following easily observed diagnosis events in an individual (node) were proposed as predictors of component growth:

- Delayed infection in a contact (i.e. more than 2 years after contact event) (**delayed node**)
- Recurrent infection (or known previous infection) (**recurrent case**)
- Recurrent contact episode (**link node**)

Note that these events were not mutually exclusive in a node.

A demographic predictor, the median age of a component was also evaluated as a control predictor by default because of expected association between age and the natural history of TB and social contact patterns.

As a preliminary analysis these predictors were simplified and evaluated as non-time varying i.e. taken from the final static network at the end of the observation period.

Models were fitted using the *R* package lme 4 (version 1.1-14) (Bates *et al.*, 2015).

Components that had only one infected case were excluded. Components were observed at different times (i.e. time between cases in a component was variable) and for different time periods and this structure was maintained in the analysis.

Smooth non-parametric trajectories from empirical growth plots were examined.

Because 70% of components had up to three infected cases only and had linear growth forms, ordinary least squares regression (OLS) was used to fit the level-1 submodel. All other parameters were estimated using maximum likelihood.

Association between intercepts and slopes were assessed by the Pearson correlation coefficient.

## 7.4.2 Results

### 7.4.2.1 Empirical and parametric component growth trajectories over time

Of 5729 separate connected components, 1417 (25%) had more than one infected case. Figure 48 shows the individual growth over time for these connected components (excluding the largest component). Seventy-one percent (1004/1416) had up to three infected cases and therefore had a linear growth change. However 29% (412/1416) of the components had more than three infected cases. The growth trajectories for these components were more complex with periods of accelerated and slow growth at variable time points suggesting logistic, exponential or logarithmic growth forms. Overall there was wide variation in the slopes of the different components, which suggested that a random slope model would be appropriate. The average, smooth non-parametric trajectory imposed on the empirical growth plots suggested that on average growth rate was highest within the first five years.

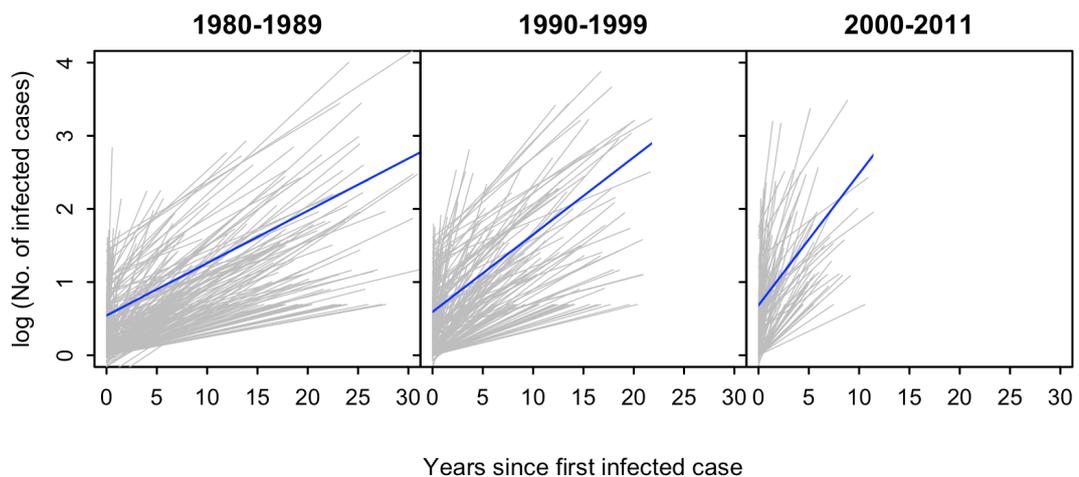


Figure 47. Ordinary least squares fitted growth trajectories for components, by calendar year of first initial infected case. The average change change trajectory in for all components within each time period is shown in blue.

Figure 48 shows the OLS trajectories fitted to individual components. Quality of fit of the linear change model varied substantially between different components with 5% (74/1416) having a low  $R^2$  statistic of 50% or less. The average initial number of cases in a component was 1 (intercept median -0.006, IQR -0.04, 0.1; mean 0.1, SD 0.33). This suggests that a model that varies by intercept may be unnecessary. The median estimated slope was 0.77 (IQR 0.13, 3.47) with a mean of 3.32 (SD 13.44). Thus on average a connected component grew by an estimated 2 cases per year. The correlation coefficient between the intercept and slope was -0.45 suggesting that components with a lower number of initial cases grew faster.

When OLS trajectories were examined by calendar year of first initial case, slopes were flatter for those components with a longer period of observation (Figure 49). This agrees with the earlier observation that there is a rapid increase of component cases in the initial first five years followed by a period of much slower growth.

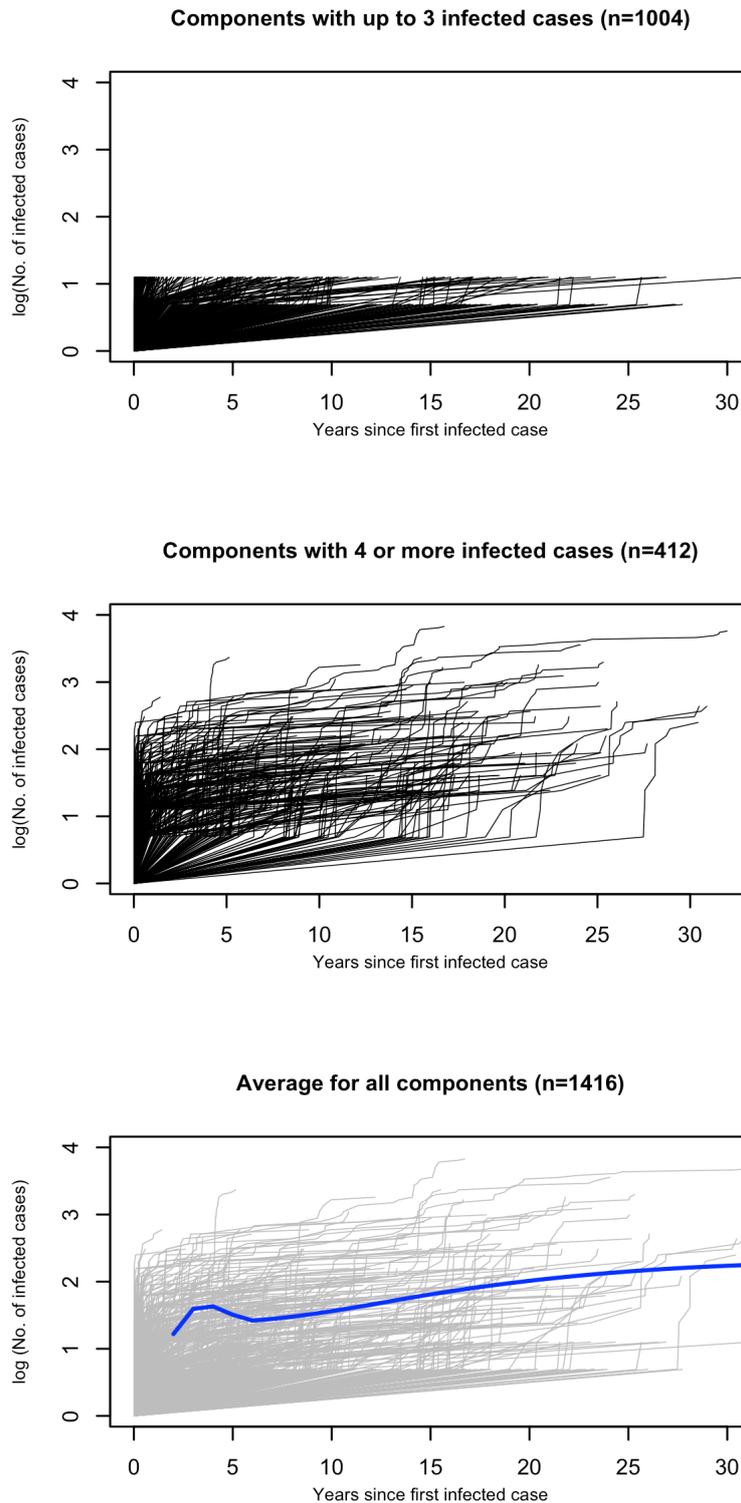


Figure 48. Empirical and average growth trajectories of 1,416 connected components with more than 1 infected case in the contact tracing network. The largest component was excluded. The smooth, non-parametric average change trajectory for all connected components is shown in blue.

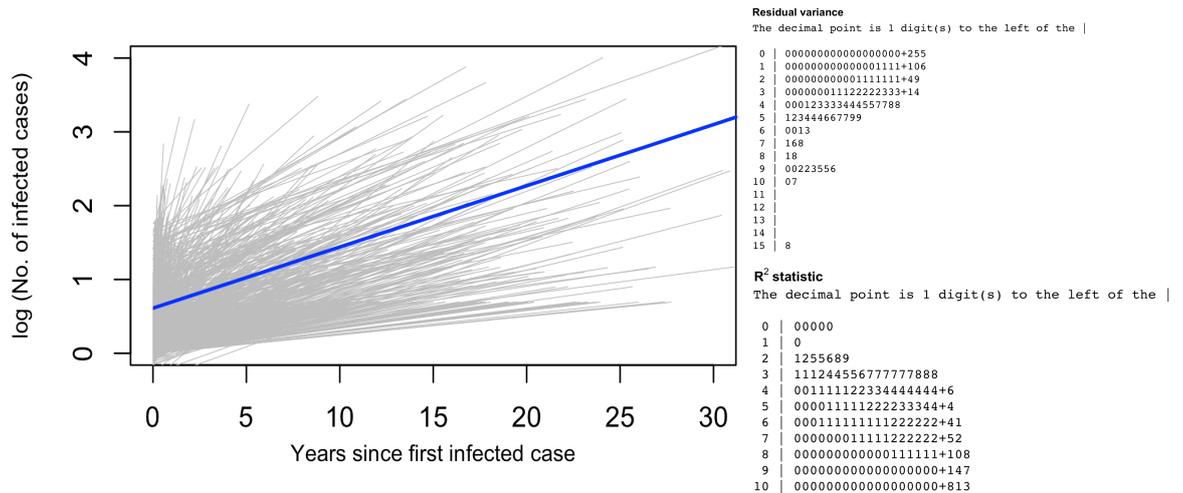


Figure 49. Linear change trajectories, residual variance and  $R^2$  statistic resulting from fitting separate ordinary least squares regression models for 1,416 connected components with more than one infected case in the contact tracing network. The average change trajectory for all components is shown in blue.

#### 7.4.2.2 Exploratory analysis of relationship between component growth and static component predictors

Median age, presence of a delayed diagnosis case, recurrent case and recurrent contact (link node) all showed an apparent effect on component growth rate (Figure 50). Components with a younger median age appeared to grow faster (Figure 50A). Slopes for components with a delayed diagnosis and recurrent case were less different to components without such cases but appeared worthy of further exploration (Figure 50B and Figure 50C). Components with link nodes or recurrent contacts also grew faster compared to those without link nodes (Figure 50D).

#### 7.4.2.3 Unconditional means model for component growth

Model A in Table 27 shows the estimates from the unconditional means linear growth model. The grand mean or average (log) number of cases at the start of component growth was 0.7 (i.e. 2 cases). The proportion of residual variance due to difference between components was estimated as  $0.149/(0.149 + 0.39) = 0.28$ . Thus most of the variance was within each component from event (the addition of a case) to event (i.e. at level 1). The intraclass correlation coefficient interpretation of this figure is that correlation between any two events in a component was low.

#### 7.4.2.4 Unconditional growth model to evaluate the fixed and random effects of time

Model B in Table 27 shows the effect of time in the level-1 submodel. The initial status in this model was 0.491 (1.6 cases) and the rate of change was 0.09 (1 case per event). The level-1 residual variance reduced from 0.39 to 0.215 and the between-component intercept variance reduced from 0.149 to 0.069 with the addition of fixed and random effect of time in this model. Linear time explained approximately 70% of the total variability of cumulative number of cases (conditional  $R^2$  of 0.703). A smaller proportion of the variance accounted for by the fixed effect of time, 43% (marginal  $R^2$  of 0.427), was due to linear time.

Given the unexplained within-component residual variance that still exists further predictors at level-1 should be explored. The between-component slope had minimal variance of 0.001 after controlling for time since the first case in a component. The correlation coefficient between the level 2 (between component) intercept and slope was high at 0.77, suggesting that the higher the initial number of cases the steeper the slope.

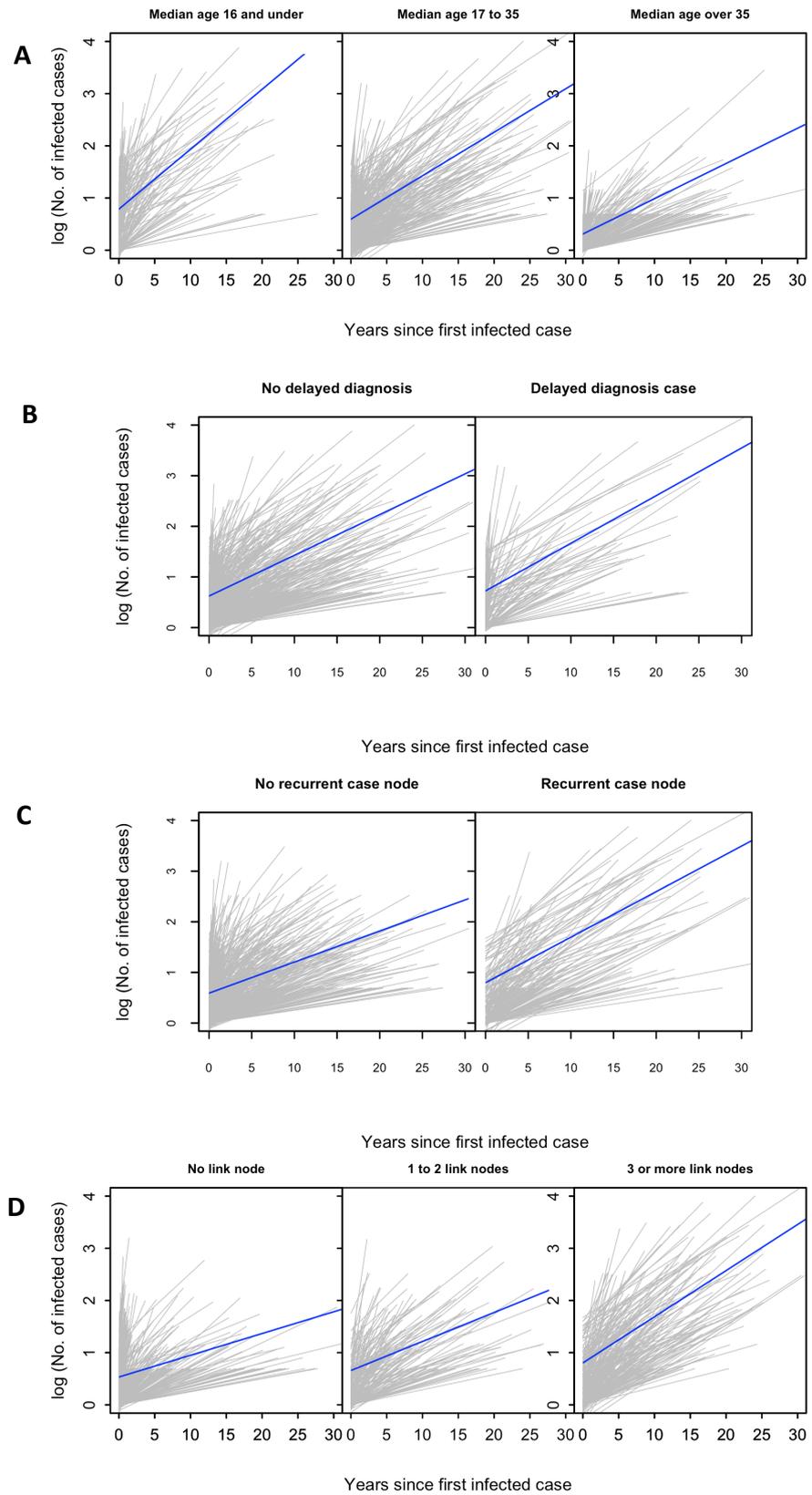


Figure 50. Ordinary least squares fitted trajectories for component growth, by levels of selected predictors. The average change trajectory for all components is shown in blue.

	Model A	Model B	Model C	Model D
Fixed effects				
Initial status				
Intercept	0.711	0.491	0.7590	0.724
Age			-0.01	-0.011
Recurrent case				0.071
Delayed diagnosis				0.104
Link node				0.077
Rate of change				
Intercept		0.0899	0.113	0.105
Age			-0.001	-0.001
Recurrent case				-0.007
Delayed diagnosis				-0.005
Link node				0.012
Variance components				
Level-1 within	0.39	0.215	0.212	0.215
Level-2 between				
Initial status	0.149	0.069	0.05	0.045
Rate of change		0.001	0.001	0.001
Correlation coefficient		0.77	0.72	0.87
Goodness of fit				
Conditional $R^2$		0.703	0.711	0.706
Marginal $R^2$		0.427	0.487	0.507
Deviance	10994.6	8263.7	8011.6	7967.5
AIC	11000.6	8275.7	8027.6	7995.5
BIC	11020.3	8315.1	8080.0	8087.3

Table 27. Results of fitting preliminary multi-level models of component growth.

## **7.5 Discussion**

This chapter has repeated some of the analyses regarding the risk of infection/disease following a contact but is included here due to its context for second generation contacts, for whom contact tracing is routinely undertaken. Developing a model for component growth is likely to need inclusion of time-dependent variables which is beyond the scope of this work.

## 8 Conclusions

This work presents a systematic exploration of routine tuberculosis (TB) register data collected at the Birmingham Chest Clinic, UK from 1980 to 2011. Birmingham is an urban TB hotspot with heterogenous distribution of the disease in certain risk groups, an epidemiological situation typical of most low incidence countries.

Because TB elimination is on the horizon for settings similar to the UK and Birmingham, there is an urgent need to find multi-modal strategies to identify cases and eliminate local transmission of TB.

The dataset was unique in that it contained individual level information of not only active TB cases but those with latent infection, recorded contacts details with preservation of their connection to a case and spanned a long time period. The data were more of a census than a sample, which resulted in more descriptive than inferential analysis. Coding of the clinical dataset into the research dataset required considerable effort, with validation against clinical notes required in many cases. We took a conservative approach to minimize the possibility of incorrect linkage, but at the cost of missing links (i.e. the same individual might have many unique identifiers within the dataset).

Identification of repeat individuals by record linkage across time was a prominent finding during preliminary examination of the data. The same individuals or their contacts appearing more than once to the clinic will be a familiar situation to healthcare workers, but creation of the research dataset from the Birmingham TB register has made the extent of this clustering and its epidemiological significance clearer. Once the data were created, an interdisciplinary approach combining traditional longitudinal research methods such as survival analysis and multilevel modelling with newer tools based on molecular epidemiology and network theory was applied.

Use of routine data for research such as presented here has inherent limitations. Although the research dataset comprised near complete recording of all treated individuals affected by TB in the population of interest, several sources of bias exist. TB cases outside the geographic area were not included so repeat treatment episodes elsewhere and contacts out of area could not be included. Detection bias was also present in both case and contact data – contacts were selectively assessed for latent infection based on age, detected latent infection that were not treated were omitted and the voluntary nature of contact tracing meant that not all contacts relevant to exposure were named. In addition, programmatic changes to data collection and contact tracing have inevitably evolved over the prolonged time frame and have not always been documented fully. Omissions and inaccuracies in data entry further contributed to incomplete ascertainment of unique individuals and missing data may have biased the estimated effects of predictor variables. Nevertheless, evaluation of this real-life dataset has discovered several findings that may inform public health action.

Three main themes have emerged from this work. Firstly, the persistence of TB with repeated diagnosis in individuals already treated with effective drug therapy. This is a recognised challenge for TB elimination. Repeated diagnosis, whether due to intrinsic host/pathogens factors or behavioral factors such as adherence to chemotherapy, have been a primary motivator for increased investment to develop shorter and better drug regimens (Tiberi *et al.*, 2018). Until this is a reality, repeated TB episodes will contribute to the ongoing case burden and transmission. Here we have confirmed consistent estimates of low risk of repeated treatment episodes following active TB treatment in low incidence settings (Rosser *et al.*, 2018). We found that latent infection treatment, additionally, contributed to retreatment cases at the same rate as active TB treatment. Most analyses have considered relapse from latent infection and active disease separately. However when projecting future case burden, both must be taken into account. It is unclear to what extent recording of latent TB infection is systematically done throughout the UK but this work supports its inclusion in surveillance. Recently a national initiative for testing new entrants to the UK for latent TB in primary care was launched in high burden areas where clinical commissioning allows and these cases are recorded centrally (Public Health

England, 2015). This is a step in the right direction but latent cases detected via other routes should not be missed.

The second theme concerns the persistence of TB manifested through contact patterns that are joined into networks. TB contacts were often identified as contacts again due to retreatment episodes in a source case or exposure to a new case. Networks of linked individuals affected by TB grew through second generation contact tracing and separate clusters merged through a mutual case/contact due to new or retreatment case diagnoses. These events could be observed both near and distant in time to the first case diagnosis, reflecting the long natural history of TB.

The first finding from examination of contact patterns was that recurrent contacts had twice the risk of being diagnosed as a case. Few studies combining empirical data and theoretical modelling have evaluated the risks of contact repetition (Al-Mouaiad Al-Azem, 2006; Read *et al.*, 2008; Smieszek *et al.*, 2009) and we have added to the literature in this regard using a multi-state model. Recurrent contacts are an easily recognised clinical group who should be prioritised but their identification within a TB control programme requires consistent recording of unique patient identity. The long-term nature of TB natural history suggests that individual identification on the basis of address and name (given that names change and transcription from non-English into English is not reliable) is insufficient for this purpose. Health providers in England often utilise both a local identifier as well as the NHS number so it would be important to develop clinical systems that can reliably distinguish the same individual locally and nationally.

We provided a description of the largest population-level TB contact network to date. We find that most clusters of linked individuals in the network were of small size and had very low density, clustering coefficient and reciprocity. This may not be a reflection of true social structure due to the way contact networks are formed. Thus no clear network measure for a case embedded in a cluster, except the eigenvector centrality i.e. how connected an individual was to other highly connected individuals, was indicative of higher number of case detection in contacts. While a network approach has demonstrated utility in isolated outbreaks, routine visualisation and measurement of network features may not be useful in identifying

high risk contacts or clusters in settings similar to Birmingham due to the sporadic nature of true outbreaks. However changing the way contacts are identified may produce different network structure and yield different conclusions.

Within the contact patterns we found an example of a superspreading event where network metrics clearly identified the key players in transmission. This outbreak in a school setting was the catalyst in linking up over 3000 individuals who were previously in separate clusters within the contact tracing network. A narrative of this event highlighted potential gaps in its management and underlines the vulnerability of a susceptible population. Here the outbreak reflected a community with persistent burden of TB. In a declining epidemic these pockets of TB are expected to shape the trend of cases for the overall population (Dowdy *et al.*, 2012). Thus the wider community setting in which outbreaks occur needs to be considered when making public health decisions because strategies in addition to traditional contact tracing may be required.

In the third and final theme, we did not find that TB persisted when viewed from a molecular point of view. Although we had very limited mycobacterial repetitive unit-variable number tandem repeats (MIRU-VNTR) typing information for the data, the overall sensitivity of epidemiological links in predicting molecular links was only 2%. It was difficult to infer transmission due to the low resolution of the typing method and only partially seen epidemiological links. A recent analysis found that both MIRU-VNTR typing and epidemiological links based on shared household were poor at delineating true transmission as defined by whole-genome sequencing (Wyllie *et al.*, 2018). Thus routine study of who infected whom using MIRU-VNTR data appears impractical outside the research setting (Munang *et al.*, 2016).

Much scope exists for future work. Although contact tracing is a major strategy little evidence base exists regarding its practical deployment (Erkens *et al.*, 2010).

Qualitative studies about interview technique such as the range and number of questions that encourage optimal identification of individuals at risk, in what setting and over what time frame remain to be completed. Given that with our conservative linkage approach we have found continued appearance of the same people as contacts and cases over a period of 30 years, the identification of individuals is critical, as is the creation of a national system, as people frequently move between local health systems.

Application of a multilevel model incorporating time-dependent variables would be worth developing to gain further understanding of the dynamics of component growth and know if this is preventable or if growth is simply a consequence of prolonged serial intervals in TB. Ultimately recommendations suggested for action here may benefit an individual, but which can have population level impact?

Recurrent individuals affected by TB and larger than average groups of linked individuals constitute only a small fraction of the cohort and therefore interventions in these groups may have little consequence on the course of the epidemic. Such questions can be explored by modelling and the findings from this thesis could aid model parameterisation appropriate to our setting.

In summary, analysis of a routinely collected case and contact data from Birmingham was able to demonstrate some useful results that have applicability to disease control. The findings echo known albeit recently developed core principles from infectious diseases modelling but translation from theory to the clinic has been remote. It is hoped that this work will open the door to bridging this gap.

## 9 Appendix

### 9.1 Case table fields in the Birmingham TB Register, 1980 - 2011

Field	Type	Comment	Percentage completeness
Case no.	Numeric	Primary key for each episode, not the unique individual	100
Year of notification	Numeric		100
Surname	Text		100
Forename	Text		100
Title	Text		Not assessed
Address 1-4	Text		100 (at level 1)
Postcode	Text		94
Telephone number	Numeric		41
Sex	Dropdown		100
Date of birth	Date		94
Age	Numeric		100
Ethnic group	Text		96
Place of birth	Dropdown	UK or non-UK	91
Year of UK entry	Numeric		89
Country of birth	Text		30
Previous TB	Yes/No		64
Date of previous TB	Numeric	New since 2011	3
Previous BCG	Yes/No		75
BCG scar present	Yes/No		32
Date of BCG	Date		4
Occupation	Text		34
NHS no.	Numeric		15
Hospital no.	Alphanumeric		67
Other hospital no.	Alphanumeric	New since 2011	25
Consultant	Text		Not assessed
GP	Text		80
Health visitor	Text		78
Risk group	Text Yes/No from 2011	List is problem drug use, currently homeless, been in prison, drug abuse, alcohol abuse	1 prior to 2011, 100 in 2011
Date of notification	Date		100
Extent	Pulmonary/Non-pulmonary/ Prophylaxis		99

Organs	Dropdown	e.g. lungs, bone and joint, gastrointestinal	78
Smear positive from which site?	Dropdown	As above	29
Culture positive from which site?	Dropdown	As above	43
Culture negative from which site?	Dropdown	As above	8
Organism	Dropdown	<i>M. tuberculosis, africanum, avium etc</i>	100
Sensitivity	Dropdown	List of drugs	40
Resistance	Dropdown	List of drugs	4
Histology	Text		4
Post-mortem diagnosis	Yes/No		59
No. of contacts examined	Numeric	Calculated from contacts table	100
No. of contacts treated for active TB	Numeric	Calculated from contacts table	100
No. of contacts treated for latent TB	Numeric	Calculated from contacts table	100
Co-morbidity 1	Text	e.g. diabetes, HIV, but also risk factors e.g. alcohol, drug abuse	8
Co-morbidity 2	Text	As above	0
HIV offered	Yes/No	Since 2011, introduced as quality metric that requires reporting	89
Reason not offered	Text		Not assessed
Date of HIV test	Date		Not assessed
HIV	Positive/Negative		55
Patient notes	Text		
Year and case no.	Auto	Concatenated from year of notification and primary case no.	
Denotified	Yes/No		27
Symptoms present	Yes/No		57
Case status	Dropdown	List is denotified, probable, or confirmed	69

First year status	Dropdown	List is died of TB, died not of TB, transferred out, not known/lost to follow up, denotified, treatment complete, other problem, culture positive	40
IGRA result	Text	Since 2011	0
Mantoux test done?	Yes/No		58
Mantoux size	Numeric		57
Treatment start date	Date		39
Treatment end date	Date		11
Date of symptom onset	Date		18
Workplace	Text		6
Enhanced tuberculosis surveillance (ETS) no	Numeric	Since 2011	100
Date of death	Numeric		2
MIRU	12 digits	VNTR typing	13
MIRU_PLUS	9 digits	VNTR typing	2
VNTR	5 digits		11
ETR	5 digits		2

## 9.2 Contact table fields in the Birmingham TB Register, 1980 - 2011

Field	Type	Comment	Percentage completeness
Case no.	Numeric	Primary key for each episode, not the unique individual	100
Year of notification	Numeric		100
Surname	Text		100
Forename	Text		100
Title	Text		Not assessed
Address 1-4	Text		100 (at level 1)
Postcode	Text		94
Telephone number	Numeric		47
Sex	Dropdown		98
Date of birth	Date		65
Age	Numeric	Calculated by program	99
Ethnic group	Text		73
Place of birth	Dropdown	UK or non-UK	24
Year of UK entry	Numeric		15
Country of birth	Text		4
Previous TB	Yes/No		6
Date of previous TB	Numeric	New since 2011	<1
Previous BCG	Yes/No		76
BCG scar present	Yes/No		40
Date of BCG	Date		18
Occupation	Text		12
Relationship to index	Text	e.g. cousin, sister	24
NHS no.	Numeric		12
Hospital no.	Alphanumeric		67
Consultant	Text		Not assessed
GP	Text		67
Health visitor	Text		8
Comorbidity 1 and 2	Text		0
Index case surname	Text		98
Index case forename	Text		96
Index case year and case no.	Numeric		88

Index case year	Numeric		97
Index case pulmonary?	Yes/No		98
Index case sputum smear positive?	Yes/No		97
Index case notification	Date		94
Index case close contact?	Yes/No		90
Outcome	Dropdown	List is: OK, fail, prophylaxis, BCG	4
Workplace	Text		<1

## 10 Bibliography

- Al-Mouaiad Al-Azem, A.** (2006). *Social network analysis in tuberculosis control among the aboriginal population of Manitoba*. University of Manitoba. Retrieved from [https://mspace.lib.umanitoba.ca/bitstream/handle/1993/29657/Al-Mouaiad\\_Al-Azem\\_Social\\_network.pdf?sequence=1](https://mspace.lib.umanitoba.ca/bitstream/handle/1993/29657/Al-Mouaiad_Al-Azem_Social_network.pdf?sequence=1)
- Allignol, A, Beyersmann, J., & Schumacher, M.** (2008). mvna, a R-package for the Multivariate Nelson-Aalen Estimator in Multistate Models. *R News*, **8**.
- Allignol, Arthur, Beyersmann, J., & Schumacher, M.** (2011). Empirical Transition Matrix of Multi-State Models: The (etm) Package. *R News*, **38**(2), 1–15.
- Althaus, C. L., Turner, K. M. E., Mercer, C. H., Auguste, P., Roberts, T. E., Bell, G., ... Low, N.** (2014). Effectiveness and cost-effectiveness of traditional and new partner notification technologies for curable sexually transmitted infections: observational study, systematic reviews and mathematical modelling. *Health Technol. Assess.*, **18**(2), 1–100, vii–viii.
- American Thoracic Society.** (2000). Targeted tuberculin testing and treatment of latent tuberculosis infection. *MMWR. Recomm. Reports Morb. Mortal. Wkly. Report. Recomm. Reports*, **49**(RR-6), 1–51.
- Anderson, C., White, J., Abubakar, I., Lipman, M., Tamne, S., Anderson, S. R., ... Dart, S.** (2014). Raising standards in UK TB control: introducing cohort review. *Thorax*, **69**(2), 187–189.
- Anderson, R. M., Gupta, S., & Ng, W.** (1990). The significance of sexual partner contact networks for the transmission dynamics of HIV. *J. Acquir. Immune Defic. Syndr.*, **3**(4), 417–429.
- Andre, M., Ijaz, K., Tillinghast, J. D., Krebs, V. E., Diem, L. A., Metchock, B., ... McElroy, P. D.** (2007). Transmission network analysis to complement routine tuberculosis contact investigations. *Am. J. Public Health*, **97**(3), 470–477.
- Antonucci, G., Girardi, E., Raviglione, M. C., & Ippolito, G.** (1995). Risk factors for tuberculosis in HIV-infected persons. A prospective cohort study. The Gruppo Italiano di Studio Tubercolosi e AIDS (GISTA). *JAMA*, **274**(2), 143–148.
- Asghar, R. J., Patlan, D. E., Miner, M. C., Rhodes, H. D., Solages, A., Katz, D. J., ... Oeltmann, J. E.** (2009, September). Limited Utility of Name-Based Tuberculosis Contact Investigations among Persons Using Illicit Drugs: Results of an Outbreak Investigation. *J. Urban Health*. Boston.
- Ayres, J. G., Ellis, C. J., & Portsmouth, O. H. D.** (1995). *East Birmingham Hospitals 1895 - 1995: from City Hospital, Little Bromwich to Birmingham Heartlands Hospital*. Aldridge: Norman A. Tector Limited.
- Bailey, W. C., Gerald, L. B., Kimerling, M. E., Redden, D., Brook, N., Bruce, F., ... Dunlap, N. E.** (2002). Predictive model to identify positive tuberculosis skin test results during contact investigations. *JAMA*, **287**(8), 996–1002.
- Bang, D., Andersen, A. B., Thomsen, V. O., & Lillebaek, T.** (2010). Recurrent tuberculosis in Denmark: relapse vs. re-infection. *Int. J. Tuberc. Lung Dis.*, **14**(4), 447–453.

- Barabasi, & Albert.** (1999). Emergence of scaling in random networks. *Science*, **286**(5439), 509–512.
- Barnes, P. F., & Cave, M. D.** (2003). Molecular epidemiology of tuberculosis. *N. Engl. J. Med.*, **349**(12), 1149–1156.
- Bates, D., Maechler, M., Bolker, B., & Walker, S.** (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.*, **67**(1), 1–48.
- Bhatti, N., Law, M. R., Morris, J. K., Halliday, R., & Moore-Gillon, J.** (1995). Increasing incidence of tuberculosis in England and Wales: a study of the likely causes. *BMJ*, **310**(6985), 967–969.
- Bifani, P. J., Plikaytis, B. B., Kapur, V., Stockbauer, K., Pan, X., Lutfey, M. L., ... Kreiswirth, B. N.** (1996). Origin and interstate spread of a New York City multidrug-resistant Mycobacterium tuberculosis clone family. *JAMA*, **275**(6), 452–457.
- Bolotin S Guthrie JL, Drews SJ, Jamieson F, A. D. C.** (2010). The Ontario Universal Typing of Tuberculosis (OUT-TB) Surveillance Program –what it means to you. *Can Resp J*, **17**(3), e51-4.
- Bothamley, G. H., Kruijshaar, M. E., Kunst, H., Woltmann, G., Cotton, M., Saralaya, D., ... Chapman, A. L. N.** (2011). Tuberculosis in UK cities: workload and effectiveness of tuberculosis control programmes. *BMC Public Health*.
- Bryant, J. M., Harris, S. R., Parkhill, J., Dawson, R., Diacon, A. H., van Helden, P., ... Bentley, S. D.** (2013). Whole-genome sequencing to establish relapse or re-infection with Mycobacterium tuberculosis: a retrospective observational study. *Lancet. Respir. Med.*, **1**(10), 786–792.
- Cain, K. P., Haley, C. A., Armstrong, L. R., Garman, K. N., Wells, C. D., Iademarco, M. F., ... Laserson, K. F.** (2007). Tuberculosis among foreign-born persons in the United States: achieving tuberculosis elimination. *Am. J. Respir. Crit. Care Med.*, **175**(1), 75–79.
- Caley, M., Fowler, T., Welch, S., & Wood, A.** (2010). Risk of developing tuberculosis from a school contact: retrospective cohort study, United Kingdom, 2009. *Euro Surveill. Bull. Eur. Sur Les Mal. Transm. = Eur. Commun. Dis. Bull.*, **15**(11).
- Cauchemez, S., Bhattarai, A., Marchbanks, T. L., Fagan, R. P., Ostroff, S., Ferguson, N. M., & Swerdlow, D.** (2011). Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. *Proc. Natl. Acad. Sci. U. S. A.*, **108**(7), 2825–2830.
- Cavany, S. M., Sumner, T., Vynnycky, E., Flach, C., White, R. G., Thomas, H. L., ... Anderson, C.** (2017). An evaluation of tuberculosis contact investigations against national standards. *Thorax*, **72**(8), 736–745.
- Cave, A. J. E.** (1939). The evidence for the incidence of tuberculosis in ancient Egypt. *Br. J. Tuberc.*, **33**(3), 142–152.
- Clauset, A., Newman, M. E. J., & Moore, C.** (2004). Finding community structure in very large networks. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.*, **70**(6 Pt 2), 66111.
- Comstock, G. W., Livesay, V. T., & Woolpert, S. F.** (1974). The prognosis of a positive tuberculin reaction in childhood and adolescence. *Am. J. Epidemiol.*, **99**(2), 131–138.
- Cook, V. J., Shah, L., Gardy, J., & Bourgeois, A. C.** (2012). Recommendations on modern contact investigation methods for enhancing tuberculosis control. *Int J Tuberc Lung Dis*, **16**(3), 297–305.

- Cook, V. J., Sun, S. J., Tapia, J., Muth, S. Q., Arguello, D. F., Lewis, B. L., ... McElroy, P. D.** (2007). Transmission network analysis in tuberculosis contact investigations. *J Infect Dis*, **196**(10), 1517–1527.
- Cox, D. R., & Oakes, D.** (1984). *Analysis of survival data*. New York: Chapman and Hall.
- Cozatt, D. M.** (2016). *Social network analysis (SNA) of Clark county, Nevada tuberculosis (TB) case and contact investigation data to determine pediatric risk factors for disease transmission. Dissertation Abstracts International: Section B: The Sciences and Engineering*.
- Cruz, A. T., Hwang, K. M., Birnbaum, G. D., & Starke, J. R.** (2013). Adolescents with tuberculosis: a review of 145 cases. *Pediatr. Infect. Dis. J.*, **32**(9), 937–941.
- Csardi, G., & Nepusz, T.** (2006). The igraph software package for complex network research. *InterJournal, Complex Sy*, 2006.
- Daly, R., Khatib, N., Larkins, A., & Dediccoat, M.** (2016). Testing for latent tuberculosis infection using interferon gamma release assays in commercial sex workers at an outreach clinic in Birmingham. *Int. J. STD AIDS*, **27**(8), 676–679.
- de Vries, G., Aldridge, R. W., Cayla, J. A., Haas, W. H., Sandgren, A., van Hest, N. A., & Abubakar, I.** (2014). Epidemiology of tuberculosis in big cities of the European Union and European Economic Area countries. *Euro Surveill. Bull. Eur. Sur Les Mal. Transm. = Eur. Commun. Dis. Bull.*, **19**(9).
- de Wreede, L., Fiocco, M., & Putter, H.** (2011). mstate: An R Package for the Analysis of Competing Risks and Multi-State Models. *J. Stat. Softw.*, **7**.
- Dillon, E. S., Boucot, K. R., Cooper, D. A., Meier, P., & Richardson, R.** (1952). A survey of tuberculosis among diabetics. *Diabetes*, **1**(4), 283–289.
- Doherty, I. A.** (2011). Sexual networks and sexually transmitted infections: innovations and findings. *Curr. Opin. Infect. Dis.*, **24**(1), 70–77.
- Donnelly, C. A., Ghani, A. C., Leung, G. M., Hedley, A. J., Fraser, C., Riley, S., ... Anderson, R. M.** (2003). Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong. *Lancet (London, England)*, **361**(9371), 1761–1766.
- Dowdy, D. W., Golub, J. E., Chaisson, R. E., & Saraceni, V.** (2012). Heterogeneity in tuberculosis transmission and the role of geographic hotspots in propagating epidemics. *Proc Natl Acad Sci U S A*, **109**(24), 9557–9562.
- Driver, C. R., Munsiff, S. S., Li, J., Kundamal, N., & Osahan, S. S.** (2001). Relapse in persons treated for drug-susceptible tuberculosis in a population with high coinfection with human immunodeficiency virus in New York City. *Clin. Infect. Dis.*, **33**(10), 1762–1769.
- Driver, C. R., Valway, S. E., Morgan, W. M., Onorato, I. M., & Castro, K. G.** (1994). Transmission of Mycobacterium tuberculosis associated with air travel. *JAMA*, **272**(13), 1031–1035.
- Drobniewski, F., Rusch-Gerdes, S., & Hoffner, S.** (2007). Antimicrobial susceptibility testing of Mycobacterium tuberculosis (EUCAST document E.DEF 8.1)--report of the Subcommittee on Antimicrobial Susceptibility Testing of Mycobacterium tuberculosis of the European Committee for Antimicrobial Susceptibility Tes. *Clin. Microbiol. Infect.*, **13**(12), 1144–1156.
- Dubos, R. J., & Dubos, J.** (1952). *The White Plague, Tuberculosis; Man and Society*. New Jersey: Rutgers University Press.
- Dye, C., Glaziou, P., Floyd, K., & Raviglione, M.** (2013). Prospects for tuberculosis elimination. *Annu. Rev. Public Health*, **34**, 271–286.

- Ena, J., & Valls, V.** (2005). Short-course therapy with rifampin plus isoniazid, compared with standard therapy with isoniazid, for latent tuberculosis infection: a meta-analysis. *Clin. Infect. Dis.*, **40**(5), 670–676.
- Erdős, P., & Rényi, A.** (1959). On random graphs I. *Publ. Math. Debrecen*, **6**, 290–297.
- Erkens, C. G. M., Kamphorst, M., Abubakar, I., Bothamley, G. H., Chemtob, D., Haas, W., ... Lange, C.** (2010). Tuberculosis contact investigation in low prevalence countries: a European consensus. *Eur. Respir. J.*, **36**(4), 925–949.
- European Centre for Disease Prevention and Control/WHO Regional Office for Europe, M.** (2018). *Tuberculosis surveillance and monitoring in Europe, 2018 - 2016 data*. Stockholm. Retrieved from <https://ecdc.europa.eu/en/publications-data/tuberculosis-surveillance-and-monitoring-europe-2018>
- Fenner, F., Henderson, D. A., Arita, I., Jezek, Z., & Ladnyi, I. D.** (1988). *Smallpox and its eradication*. Geneva.
- Ferebee, S. H., & Mount, F. W.** (1962). Tuberculosis morbidity in a controlled trial of the prophylactic use of isoniazid among household contacts. *Am. Rev. Respir. Dis.*, **85**, 490–510.
- Fitzpatrick, L. K., Hardacker, J. A., Heirendt, W., Agerton, T., Streicher, A., Melnyk, H., ... Onorato, I.** (2001). A preventable outbreak of tuberculosis investigated through an intricate social network. *Clin. Infect. Dis.*, **33**(11), 1801–1806.
- Fox, G. J., Barry, S. E., Britton, W. J., & Marks, G. B.** (2013). Contact investigation for tuberculosis: a systematic review and meta-analysis. *Eur Respir J*, **41**(1), 140–156.
- Frieden, T. R., Woodley, C. L., Crawford, J. T., Lew, D., & Dooley, S. M.** (1996). The molecular epidemiology of tuberculosis in New York City: the importance of nosocomial transmission and laboratory error. *Tuber. Lung Dis.*, **77**(5), 407–413.
- Fujikawa, A., Fujii, T., Mimura, S., Takahashi, R., Sakai, M., Suzuki, S., ... Mori, T.** (2014). Tuberculosis contact investigation using interferon-gamma release assay with chest x-ray and computed tomography. *PLoS One*, **9**(1), e85612.
- Gardam, M. A., Keystone, E. C., Menzies, R., Manners, S., Skamene, E., Long, R., & Vinh, D. C.** (2003). Anti-tumour necrosis factor agents and tuberculosis risk: mechanisms of action and clinical management. *Lancet. Infect. Dis.*, **3**(3), 148–155.
- Gardy, J. L., Johnston, J. C., Ho Sui, S. J., Cook, V. J., Shah, L., Brodtkin, E., ... Tang, P.** (2011). Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N. Engl. J. Med.*, **364**(8), 730–739.
- Gaudette, L. A., & Ellis, E.** (1993). Tuberculosis in Canada: a focal disease requiring distinct control strategies for different risk groups. *Tuber. Lung Dis.*, **74**(4), 244–253.
- Genewein, A., Telenti, A., Bernasconi, C., Mordasini, C., Weiss, S., Maurer, A. M., ... Bodmer, T.** (1993). Molecular approach to identifying route of transmission of tuberculosis in the community. *Lancet (London, England)*, **342**(8875), 841–844.
- Ghani, A. C., Swinton, J., & Garnett, G. P.** (1997). The role of sexual partnership networks in the epidemiology of gonorrhoea. *Sex. Transm. Dis.*, **24**(1), 45–56.

- Gilbert, R. L., Antoine, D., French, C. E., Abubakar, I., Watson, J. M., & Jones, J. A.** (2009). The impact of immigration on tuberculosis rates in the United Kingdom compared with other European countries. *Int. J. Tuberc. Lung Dis.*, **13**(5), 645–651.
- Goldstein, H.** (1995). *Multilevel statistical models* (Halstead P). New York.
- Grange, J. M., & Zumla, A.** (2002). The global emergency of tuberculosis: what is the cause? *J. R. Soc. Promot. Health*, **122**(2), 78–81.
- Grinsdale, J. A., Ho, C. S., Banouvong, H., & Kawamura, L. M.** (2011). Programmatic impact of using QuantiFERON(R)-TB Gold in routine contact investigation activities. *Int J Tuberc Lung Dis*, **15**(12), 1614–1620.
- Haddad, M. B., Wilson, T. W., Ijaz, K., Marks, S. M., & Moore, M.** (2005). Tuberculosis and homelessness in the United States, 1994–2003. *JAMA*, **293**(22), 2762–2766.
- Hanneman, R. A., & Riddle, M.** (2005). *Introduction to social network methods*. Riverside, CA: University of California.
- Harries, A. D.** (1990). Tuberculosis and human immunodeficiency virus infection in developing countries. *Lancet (London, England)*, **335**(8686), 387–390.
- Hawker, J. I., Bakhshi, S. S., Ali, S., & Farrington, C. P.** (1999). Ecological analysis of ethnic differences in relation between tuberculosis and poverty. *BMJ*, **319**(7216), 1031–1034.
- Hayward, A. C., Darton, T., Van-Tam, J. N., Watson, J. M., Coker, R., & Schwoebel, V.** (2003). Epidemiology and control of tuberculosis in Western European cities. *Int. J. Tuberc. Lung Dis.*, **7**(8), 751–757.
- Hong Kong Chest Service, & British Medical Research Council.** (1987). Five-year follow-up of a controlled trial of five 6-month regimens of chemotherapy for pulmonary tuberculosis. *Am. Rev. Respir. Dis.*, **136**(6), 1339–1342.
- Houk, V. H., Kent, D. C., Baker, J. H., Sorensen, K., & Hanzel, G. D.** (1968). The Byrd study. In-depth analysis of a micro-outbreak of tuberculosis in a closed environment. *Arch. Environ. Health*, **16**(1), 4–6.
- Hsu, K. H.** (1963). CONTACT INVESTIGATION: A PRACTICAL APPROACH TO TUBERCULOSIS ERADICATION. *Am. J. Public Health Nations. Health*, **53**, 1761–1769.
- Jajou, R., Neeling, A. de, Hunen, R. van, Vries, G. de, Schimmel, H., Mulder, A., ... Soolingen, D. van.** (2018). Epidemiological links between tuberculosis cases identified twice as efficiently by whole genome sequencing than conventional molecular typing: A population-based study. *PLoS One*, **13**(4), e0195413.
- Jit, M., Stagg, H. R., Aldridge, R. W., White, P. J., & Abubakar, I.** (2011). Dedicated outreach service for hard to reach patients with tuberculosis in London: observational study and economic evaluation. *BMJ*, **343**, d5376.
- Joint Tuberculosis Committee of the British Thoracic and Tuberculosis Association.** (1973). Chemoprophylaxis against tuberculosis in Britain. *Tubercle*, **54**(4), 309–316.
- Kasaie, P., Andrews, J. R., Kelton, W. D., & Dowdy, D. W.** (2014). Timing of tuberculosis transmission and the impact of household contact tracing. An agent-based simulation model. *Am. J. Respir. Crit. Care Med.*, **189**(7), 845–852.

- Kawatsu, L., Izumi, K., Uchimura, K., Urakawa, M., Ohkado, A., & Takahashi, I.** (2015). Can social network analysis assist in the prioritisation of contacts in a tuberculosis contact investigation? *Int. J. Tuberc. Lung Dis.*, **19**(11), 1293–1299.
- Keane, J., Gershon, S., Wise, R. P., Mirabile-Levens, E., Kasznica, J., Schwieterman, W. D., ... Braun, M. M.** (2001). Tuberculosis associated with infliximab, a tumor necrosis factor alpha-neutralizing agent. *N. Engl. J. Med.*, **345**(15), 1098–1104.
- Kenyon, T. A., Valway, S. E., Ihle, W. W., Onorato, I. M., & Castro, K. G.** (1996). Transmission of multidrug-resistant Mycobacterium tuberculosis during a long airplane flight. *N. Engl. J. Med.*, **334**(15), 933–938.
- Kim, L., Moonan, P. K., Yelk Woodruff, R. S., Kammerer, J. S., & Haddad, M. B.** (2013). Epidemiology of recurrent tuberculosis in the United States, 1993–2010. *Int. J. Tuberc. Lung Dis.*, **17**(3), 357–360.
- Klovdahl, A. S., Graviss, E. A., Yaganehdooost, A., Ross, M. W., Wanger, A., Adams, G. J., & Musser, J. M.** (2001). Networks and tuberculosis: an undetected community outbreak involving public places. *Soc. Sci. Med.*, **52**(5), 681–694.
- Kremer, K., Glynn, J. R., Lillebaek, T., Niemann, S., Kurepina, N. E., Kreiswirth, B. N., ... van Soolingen, D.** (2004). Definition of the Beijing/W lineage of Mycobacterium tuberculosis on the basis of genetic markers. *J. Clin. Microbiol.*, **42**(9), 4040–4049.
- Kruijshaar, M. E., Abubakar, I., Dedicat, M., Bothamley, G. H., Maguire, H., Moore, J., ... Lipman, M.** (2012). Evidence for a national problem: continued rise in tuberculosis case numbers in urban areas outside London. *Thorax*, **67**(3), 275–277.
- Lambregts-van Weezenbeek CSB van Gerven PJHJ, de Vries G, Verver S, Kalisvaart NA, van Soolingen D, S. M.** (2003). Tuberculosis contact investigation and DNA fingerprint surveillance in The Netherlands: 6 years' experience with nation-wide cluster feedback and cluster monitoring. *Int J Tuberc Lung Dis*, **7**(12), s463–s470.
- Long, R., Njoo, H., & Hershfield, E.** (1999). Tuberculosis: 3. Epidemiology of the disease in Canada. *CMAJ*, **160**(8), 1185–1190.
- Lonroth, K., Migliori, G. B., Abubakar, I., D'Ambrosio, L., de Vries, G., Diel, R., ... Raviglione, M. C.** (2015). Towards tuberculosis elimination: an action framework for low-incidence countries. *Eur. Respir. J.*, **45**(4), 928–952.
- Loudon, R. G., & Roberts, R. M.** (1967). Droplet expulsion from the respiratory tract. *Am. Rev. Respir. Dis.*, **95**(3), 435–442.
- Maguire, H., Dale, J. W., McHugh, T. D., Butcher, P. D., Gillespie, S. H., Costetsos, A., ... Banerjee, D. K.** (2002). Molecular epidemiology of tuberculosis in London 1995–7 showing low rate of active transmission. *Thorax*, **57**(7), 617–622.
- Mak, A., Thomas, A., Del Granado, M., Zaleskis, R., Mouzafarova, N., & Menzies, D.** (2008). Influence of multidrug resistance on tuberculosis treatment outcomes with standardized regimens. *Am. J. Respir. Crit. Care Med.*, **178**(3), 306–312.
- Mandal, P., Craxton, R., Chalmers, J. D., Gilhooley, S., Laurenson, I. F., McSparron, C., ... Hill, A. T.** (2012). Contact tracing in pulmonary and non-pulmonary tuberculosis. *QJM*, **105**(8), 741–747.

- Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O., & Eisenberg, D.** (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**(5428), 751–753.
- Marks, S. M., Taylor, Z., Qualls, N. L., Shrestha-Kuwahara, R. J., Wilce, M. A., & Nguyen, C. H.** (2000). Outcomes of contact investigations of infectious tuberculosis patients. *Am. J. Respir. Crit. Care Med.*, **162**(6), 2033–2038.
- McElroy, P. D., Rothenberg, R. B., Varghese, R., Woodruff, R., Minns, G. O., Muth, S. Q., ... Ridzon, R.** (2003). A network-informed approach to investigating a tuberculosis outbreak: implications for enhancing contact investigations. *Int. J. Tuberc. Lung Dis.*, **7**(12 Suppl 3), S486-93.
- Mears, J., Abubakar, I., Crisp, D., Maguire, H., Innes, J. A., Lilley, M., ... Sonnenberg, P.** (2014). Prospective evaluation of a complex public health intervention: lessons from an initial and follow-up cross-sectional survey of the tuberculosis strain typing service in England. *BMC Public Health*, **14**(1), 1023.
- Melegaro, A., Jit, M., Gay, N., Zagheni, E., & Edmunds, W. J.** (2011). What types of contacts are important for the spread of infections?: using contact survey data to explore European mixing patterns. *Epidemics*, **3**(3–4), 143–151.
- Mistry, R., Cliff, J. M., Clayton, C. L., Beyers, N., Mohamed, Y. S., Wilson, P. A., ... Lukey, P. T.** (2007). Gene-expression patterns in whole blood identify subjects at risk for recurrent tuberculosis. *J. Infect. Dis.*, **195**(3), 357–365.
- Moonan, P. K., Ghosh, S., Oeltmann, J. E., Kammerer, J. S., Cowan, L. S., & Navin, T. R.** (2012). Using genotyping and geospatial scanning to estimate recent mycobacterium tuberculosis transmission, United States. *Emerg Infect Dis*, **18**(3), 458–465.
- Morán-Mendoza, O., Marion, S. A., Elwood, K., Patrick, D. M., & Fitzgerald, J. M.** (2007). Tuberculin skin test size and risk of tuberculosis development: a large population-based study in contacts. *Int J Tuberc Lung Dis*, **11**(9), 1014–1020.
- Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., ... Edmunds, W. J.** (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med.*, **5**(3), e74.
- MRC Tuberculosis and Chest Diseases Unit.** (1980). National survey of tuberculosis notifications in England and Wales 1978-9. *Br. Med. J.*, **281**(6245), 895–898.
- Muller, J., Kretzschmar, M., & Dietz, K.** (2000). Contact tracing in stochastic and deterministic epidemic models. *Math. Biosci.*, **164**(1), 39–64.
- Munang, M. L., Browne, C., Evans, J. T., Smith, E. G., Hawkey, P. M., Welch, S. B., ... Dedicoat, M. J.** (2016). Programmatic utility of tuberculosis cluster investigation using a social network approach in Birmingham, United Kingdom. *Int. J. Tuberc. Lung Dis.*, **20**(10), 1300–1305.
- National Institute for Health and Clinical Excellence.** Tuberculosis, Pub. L. No. NG33 (2016). London. Retrieved from <https://www.nice.org.uk/guidance/ng33>
- Nerlich, A. G., Haas, C. J., Zink, A., Szeimies, U., & Hagedorn, H. G.** (1997, November). Molecular evidence for tuberculosis in an ancient Egyptian mummy. *Lancet (London, England)*. England.
- Newman, M. E.** (2001). The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. U. S. A.*, **98**(2), 404–409.
- Newman, M. E. J.** (2002). Assortative mixing in networks. *Phys. Rev. Lett.*, **89**(20), 208701.

- Newton, S. M., Smith, R. J., Wilkinson, K. A., Nicol, M. P., Garton, N. J., Staples, K. J., ... Wilkinson, R. J.** (2006). A deletion defining a common Asian lineage of Mycobacterium tuberculosis associates with immune subversion. *Proc. Natl. Acad. Sci. U. S. A.*, **103**(42), 15594–15598.
- Nikolayevskyy, V., Trovato, A., Broda, A., Borroni, E., Cirillo, D., & Drobniewski, F.** (2016). MIRU-VNTR Genotyping of Mycobacterium tuberculosis Strains Using QIAxcel Technology: A Multicentre Evaluation Study. *PLoS One*, **11**(3), e0149435.
- O'Shea, M. K., Koh, G. C. K. W., Munang, M., Smith, G., Banerjee, A., & Dedicoat, M.** (2014). Time-to-detection in culture predicts risk of Mycobacterium tuberculosis transmission: a cohort study. *Clin. Infect. Dis.*, **59**(2), 177–185.
- Office of Population Censuses and Surveys ; General Register Office for Scotland ; Registrar General for Northern Ireland.** (1997). *1991 Census aggregate data*. UK Data Service (Edition: 1997). DOI: <http://dx.doi.org/10.5257/census/aggregate-1991-1>
- Office of Population Censuses and Surveys ; Registrar General for Scotland.** (2000). *1981 Census aggregate data* (Edition: 2000). UK Data Service. DOI: <http://dx.doi.org/10.5257/census/aggregate-1981-1>
- Office for National Statistics ; General Register Office for Scotland ; Northern Ireland Statistics and Research Agency.** (2005). *2001 Census aggregate data* (Edition: 2005). UK Data Service. DOI: <http://dx.doi.org/10.5257/census/aggregate-2001-1>
- Office for National Statistics ; National Records of Scotland ; Northern Ireland Statistics and Research Agency.** (2016). *2011 Census aggregate data*. UK Data Service (Edition: June 2016). DOI: <http://dx.doi.org/10.5257/census/aggregate-2011-1>
- Oren, E., Winston, C. A., Pratt, R., Robison, V. A., & Narita, M.** (2011). Epidemiology of urban tuberculosis in the United States, 2000-2007. *Am. J. Public Health*, **101**(7), 1256–1263.
- Osoba, O. A.** (2004). Microbiology of tuberculosis. In M. M. Madkour (Ed.), *Tuberculosis* (pp. 115–132). Berlin: Springer.
- Pai, M., Zwerling, A., & Menzies, D.** (2008). Systematic review: T-cell-based assays for the diagnosis of latent tuberculosis infection: an update. *Ann. Intern. Med.*, **149**(3), 177–184.
- Pareek, M., Abubakar, I., White, P. J., Garnett, G. P., & Lalvani, A.** (2011). Tuberculosis screening of migrants to low-burden nations: insights from evaluation of UK practice. *Eur. Respir. J.*, **37**(5), 1175–1182.
- Pascopella, L., Deriemer, K., Watt, J. P., & Flood, J. M.** (2011). When tuberculosis comes back: who develops recurrent tuberculosis in California? *PLoS One*, **6**(11), e26541.
- Price, D. D. S.** (1976). A general theory of bibliometric and other cumulative advantage processes. *J Amer Soc Inf. Sci.*, **27**, 292–306.
- Public Health (Tuberculosis) Regulations (1911). Retrieved May 19, 2018 from <http://www.jstor.org/stable/25295148>
- Public Health Act (1875). Retrieved Feb 19, 2018 from <http://www.legislation.gov.uk/ukpga/Vict/38-39/55/contents>

- Public Health England.** (2013). Enhanced TB surveillance. Retrieved May 19, 2018, from <http://webarchive.nationalarchives.gov.uk/20140714073824/http://www.hpa.org.uk/Topics/InfectiousDiseases/InfectionsAZ/Tuberculosis/TBUKSurveillanceData/EnhancedTuberculosisSurveillance/>
- Public Health England.** (2014). *TB strain typing and cluster investigation handbook: 3rd edition*. London.
- Public Health England.** (2015). *Collaborative tuberculosis strategy for England: 2015 to 2020*.
- Public Health England.** (2017a). *Three-year average annual number of TB case notifications and rates by PHE Centre, upper tier local authority and local authority district, England, 2014-2016*. Retrieved from <https://www.gov.uk/government/publications/tuberculosis-tb-in-england-surveillance-data>
- Public Health England.** (2017b). *Tuberculosis in England 2017 report (presenting data to end of 2016)*.
- Public Health England.** (2017c). *UK pre-entry tuberculosis screening report 2016*.
- Public Health England, & Department of Health.** (2013). Tuberculosis: chapter 32. In *Immunisation against infectious disease* (pp. 391–409).
- R Core Team.** (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.r-project.org/>
- Raviglione, M. C., Sudre, P., Rieder, H. L., Spinaci, S., & Kochi, A.** (1993). Secular trends of tuberculosis in western Europe. *Bull. World Health Organ.*, **71**(3–4), 297–306.
- Read, J. M., Eames, K. T. D., & Edmunds, W. J.** (2008). Dynamic social networks and the implications for the spread of infectious disease. *J. R. Soc. Interface*, **5**(26), 1001–1007.
- Rieder, H. L., Snider, D. E. J., & Cauthen, G. M.** (1990). Extrapulmonary tuberculosis in the United States. *Am. Rev. Respir. Dis.*, **141**(2), 347–351.
- Riley, R. L., Mills, C. C., Nyka, W., Weinstock, N., Storey, P. B., Sultan, L. U., ... Wells, W. F.** (1995). Aerial dissemination of pulmonary tuberculosis. A two-year study of contagion in a tuberculosis ward. 1959. *Am. J. Epidemiol.*, **142**(1), 3–14.
- Ripeanu, M., Iamnitchi, A., & Foster, I.** (2002). Mapping the Gnutella network: properties of large-scale peer-to-peer systems and implications for system design. *IEEE Internet Comput.*, **6**(1), 50–57.
- Rivers, C. M., Lofgren, E. T., Marathe, M., Eubank, S., & Lewis, B. L.** (2014). Modeling the impact of interventions on an epidemic of ebola in sierra leone and liberia. *PLoS Curr.*, **6**.
- Roetzer, A., Diel, R., Kohl, T. A., Rückert, C., Nübel, U., Blom, J., ... Niemann, S.** (2013). Whole Genome Sequencing versus Traditional Genotyping for Investigation of a Mycobacterium tuberculosis Outbreak: A Longitudinal Molecular Epidemiological Study. *PLOS Med.*, **10**(2), e1001387. Retrieved from <https://doi.org/10.1371/journal.pmed.1001387>
- Rogosa, D., Brandt, D., & Zimowski, M.** (1982). A growth curve approach to measurement of change. *Psychol. Bull.*, **92**(3), 726–748.

- Rose, A. M., Watson, J. M., Graham, C., Nunn, A. J., Drobniewski, F., Ormerod, L. P., ... Leese, J.** (2001). Tuberculosis at the end of the 20th century in England and Wales: results of a national survey in 1998. *Thorax*, **56**(3), 173–179.
- Rosser, A, Marx, F. M., & Pareek, M.** (2018). Recurrent tuberculosis in the pre-elimination era. *Int. J. Tuberc. Lung Dis.*, **22**(2), 139–150.
- Rosser, Andrew, Richardson, M., Wiselka, M. J., Free, R. C., Woltmann, G., Mukamolova, G. V., & Pareek, M.** (2018). A nested case-control study of predictors for tuberculosis recurrence in a large UK Centre. *BMC Infect. Dis.*, **18**(1), 94.
- Saitou, N., & Nei, M.** (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**(4), 406–425.
- Saunders, M. J., Koh, G. C. K. W., Small, A. D., & Dedicoat, M.** (2014). Predictors of contact tracing completion and outcomes in tuberculosis: a 21-year retrospective cohort study. *Int. J. Tuberc. Lung Dis.*, **18**(6), 640–646.
- Schluger, N. W., & Rom, W. N.** (1998). The host immune response to tuberculosis. *Am. J. Respir. Crit. Care Med.*, **157**(3 Pt 1), 679–691.
- Schneider, E., & Castro, K. G.** (2003). Tuberculosis trends in the United States, 1992–2001. *Tuberculosis (Edinb.)*, **83**(1–3), 21–29.
- Seddon, J. A., Paton, J., Nademi, Z., Keane, D., Williams, B., Williams, A., ... Kampmann, B.** (2016). The impact of BCG vaccination on tuberculin skin test responses in children is age dependent: evidence to be considered when screening children for tuberculosis infection. *Thorax*, **71**(10), 932 LP – 939.
- Slot, R., Borgdorff, M. W., de Beer, J. L., van Ingen, J., Supply, P., & van Soolingen, D.** (2013). Clustering of tuberculosis cases based on variable-number tandem-repeat typing in relation to the population structure of *Mycobacterium tuberculosis* in the Netherlands. *J. Clin. Microbiol.*, **51**(7), 2427–2431.
- Slot, R., Schim van der Loeff, M. F., Kouw, P. M., & Borgdorff, M. W.** (2014). Risk of tuberculosis after recent exposure. A 10-year follow-up study of contacts in Amsterdam. *Am. J. Respir. Crit. Care Med.*, **190**(9), 1044–1052.
- Smieja, M. J., Marchetti, C. A., Cook, D. J., & Smaill, F. M.** (2000). Isoniazid for preventing tuberculosis in non-HIV infected persons. *Cochrane Database Syst. Rev.*, (2), CD001363.
- Smieszek, T., Fiebig, L., & Scholz, R. W.** (2009). Models of epidemics: when contact repetition and clustering should be included. *Theor. Biol. Med. Model.*, **6**, 11.
- Smith, C. M., Maguire, H., Anderson, C., Macdonald, N., & Hayward, A. C.** (2017). Multiple large clusters of tuberculosis in London: a cross-sectional analysis of molecular and spatial data. *ERJ Open Res.*, **3**(1).
- Snider, G. L.** (1997). Tuberculosis then and now: a personal perspective on the last 50 years. *Ann. Intern. Med.*, **126**(3), 237–243.
- Snow, K. J., Sismanidis, C., Denholm, J., Sawyer, S. M., & Graham, S. M.** (2018). The incidence of tuberculosis among adolescents and young adults: a global estimate. *Eur. Respir. J.*, **51**(2).
- Springett, V. H.** (1972, February). Tuberculosis--epidemiology in England and Wales. *Br. Med. J.*
- Stead, W. W., & Lofgren, J. P.** (1983). Does the risk of tuberculosis increase in old age? *J. Infect. Dis.*, **147**(5), 951–955.

- Stead, W. W., To, T., Harrison, R. W., & Abraham, J. H. 3rd.** (1987). Benefit-risk considerations in preventive treatment for tuberculosis in elderly persons. *Ann. Intern. Med.*, **107**(6), 843–845.
- Sterling, T. R., Villarino, M. E., Borisov, A. S., Shang, N., Gordin, F., Bliven-Sizemore, E., ... Chaisson, R. E.** (2011). Three months of rifapentine and isoniazid for latent tuberculosis infection. *N. Engl. J. Med.*, **365**(23), 2155–2166.
- Story, A., Murad, S., Roberts, W., Verheyen, M., & Hayward, A. C.** (2007). Tuberculosis in London: the importance of homelessness, problem drug use and prison. *Thorax*, **62**(8), 667–671.
- Subcommittee of the Joint Tuberculosis Committee of the British Thoracic Society.** (1990). Control and prevention of tuberculosis in Britain: an updated code of practice. *BMJ*, **300**(6730), 995–999.
- Supply, P., Lesjean, S., Savine, E., Kremer, K., van Soolingen, D., & Locht, C.** (2001). Automated high-throughput genotyping for study of global epidemiology of Mycobacterium tuberculosis based on mycobacterial interspersed repetitive units. *J. Clin. Microbiol.*, **39**(10), 3563–3571.
- Supply, Philip, Allix, C., Lesjean, S., Cardoso-Oelemann, M., Rusch-Gerdes, S., Willery, E., ... van Soolingen, D.** (2006). Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of Mycobacterium tuberculosis. *J. Clin. Microbiol.*, **44**(12), 4498–4510.
- Taylor, J. G., Yates, T. A., Mthethwa, M., Tanser, F., Abubakar, I., & Altamirano, H.** (2016). Measuring ventilation and modelling M. tuberculosis transmission in indoor congregate settings, rural KwaZulu-Natal. *Int. J. Tuberc. Lung Dis.*, **20**(9), 1155–1161.
- Therneau, T.** (2015). A Package for Survival Analysis in S.
- Thomas, H. L., Harris, R. J., Muzyamba, M. C., Davidson, J. A., Lalor, M. K., Campbell, C. N. J., ... Zenner, D.** (2018). Reduction in tuberculosis incidence in the UK from 2011 to 2015: a population-based study. *Thorax*. Retrieved from <http://thorax.bmj.com/content/early/2018/04/19/thoraxjnl-2017-211074.abstract>
- Tiberi, S., du Plessis, N., Walzl, G., Vjecha, M. J., Rao, M., Ntoumi, F., ... Zumla, A.** (2018). Tuberculosis: progress and advances in development of new drugs, treatment regimens, and host-directed therapies. *Lancet. Infect. Dis.*
- Trauer, J. M., Moyo, N., Tay, E.-L., Dale, K., Ragonnet, R., McBryde, E. S., & Denholm, J. T.** (2016). Risk of Active Tuberculosis in the Five Years Following Infection . . . 15%? *Chest*, **149**(2), 516–525.
- Tufariello, J. M., Chan, J., & Flynn, J. L.** (2003). Latent tuberculosis: mechanisms of host and bacillus that contribute to persistent infection. *Lancet. Infect. Dis.*, **3**(9), 578–590.
- Underwood, B. R., White, V. L. C., Baker, T., Law, M., & Moore-Gillon, J. C.** (2003). Contact tracing and population screening for tuberculosis - who should be assessed? *J. Public Health (Bangkok)*, **25**(1), 59–61.
- Van Geuns, H. A., Meijer, J., & Styblo, K.** (1950). Results of the examination of subjects in contact with TB patients in Rotterdam, 1967–1969. *Bull Int Union Tuberc Lung Dis*, 105–119.
- Veen, J.** (1992). Microepidemics of tuberculosis: the stone-in-the-pond principle. *Tuber. Lung Dis.*, **73**(2), 73–76.

- Vynnycky, E., & Fine, P. E.** (1997). The natural history of tuberculosis: the implications of age-dependent risks of disease and the role of reinfection. *Epidemiol Infect*, **119**, 183–201.
- Vynnycky, E., & Fine, P. E.** (1999). Interpreting the decline in tuberculosis: the role of secular trends in effective contact. *Int. J. Epidemiol.*, **28**(2), 327–334.
- Walker, T. M., Ip, C. L. C., Harrell, R. H., Evans, J. T., Kapatai, G., Dediccoat, M. J., ... Peto, T. E. A.** (2013). Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet. Infect. Dis.*, **13**(2), 137–146.
- Watson, J. M., & Moss, F.** (2001, May). TB in Leicester: out of control, or just one of those things? *BMJ*.
- WHO Global Tuberculosis Programme.** (1994). *TB : a global emergency, WHO report on the TB epidemic*. Geneva. Retrieved from <http://apps.who.int/iris/handle/10665/58749>
- Wigfield, A. S.** (1972). 27 years of uninterrupted contact tracing. The “Tyneside Scheme”. *Br. J. Vener. Dis.*, **48**(1), 37–50.
- Wilcox, W. D., & Laufer, S.** (1994). Tuberculosis in adolescents. A case commentary. *Clin. Pediatr. (Phila.)*, **33**(5), 258–262.
- Willekens, F.** (2016). Biograph: Explore Life Histories.
- Willett, J. B.** (1988). Questions and answers in the measurement of change. In E. Rothkopf (Ed.), *Review of research in education (1988-89)* (pp. 345–422). Washington, DC: Americal Educational Research Association.
- Wood, R., Liang, H., Wu, H., Middelkoop, K., Oni, T., Rangaka, M. X., ... Lawn, S. D.** (2010). Changing prevalence of tuberculosis infection with increasing age in high-burden townships in South Africa. *Int. J. Tuberc. Lung Dis.*, **14**(4), 406–412.
- World Health Organization.** (2014). *Contact tracing during an outbreak of Ebola virus disease: Disease surveillance and response programme area disease prevention and control cluster*. Geneva.
- World Health Organization.** (2017). *Global tuberculosis report 2017*. Geneva. Retrieved from [http://www.who.int/tb/publications/global\\_report/en/](http://www.who.int/tb/publications/global_report/en/)
- World Health Organization & Stop TB Partnership.** (2006). *The Stop TB Strategy: building on and enhancing DOTS to meet TB-related Millenium Development Goals*. Geneva.
- World Health Organization & Stop TB Partnership.** (2010). *The global plan to stop TB 2011-2015: transforming the fight towards elimination of tuberculosis*. Geneva.
- Wyllie, D., Davidson, J., Walker, T., Rathod, P., Peto, T., Robinson, E., ... Campbell, C.** (2018). A quantitative evaluation of MIRU-VNTR typing against whole-genome sequencing for identifying *Mycobacterium tuberculosis* transmission: A prospective observational cohort study. *BioRxiv*. Retrieved from <http://biorxiv.org/content/early/2018/01/24/252734.abstract>
- Yang, C., Luo, T., Sun, G., Qiao, K., Sun, G., DeRiemer, K., ... Gao, Q.** (2012). *Mycobacterium tuberculosis* Beijing strains favor transmission but not drug resistance in China. *Clin. Infect. Dis.*, **55**(9), 1179–1187.