INVESTIGATING TUBERCULOSIS TRANSMISSION USING SPATIAL METHODS

Catherine Mary Smith

UCL

Institute of Health Informatics

PhD thesis

I, Catherine Mary Smith, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed ____

ABSTRACT

Background: Tuberculosis remains a leading infectious cause of death worldwide. Reducing transmission requires an increased focus on local control measures informed by spatial data. Effective use of spatial methods will improve understanding of tuberculosis transmission and support outbreak investigations.

Methods: I conducted a systematic literature review to describe spatial methods that have been used in previous outbreak investigations (Chapter 2). I developed and evaluated a novel interactive mapping tool, written using the R programming language (Chapter 3). Using multinomial logistic regression and spatial scan statistics, I investigated molecular and spatial clustering of tuberculosis in London (Chapter 4). I described the evolution of a large outbreak of drug-resistant tuberculosis in London in space and time (Chapter 5). Through three case studies, I assessed the utility of a novel spatial tool, geographic profiling, which aims to identify the locations of sources of infectious disease using locations of linked cases (Chapter 6). I analysed the spatial accessibility of tuberculosis services in London using travel time data (Chapter 7).

Key findings:

- Spatial methods provide an important complementary tool to epidemiological analyses, but are currently under-used (less than half a percent of published outbreak investigations used spatial methods).
- Large numbers of tuberculosis cases in London have resulted from local transmission, with more than one in ten cases part of large clusters.
- Social complexity and area-level deprivation are associated with transmission of tuberculosis in large clusters.
- Geographic profiling may assist with epidemiological investigations of infectious diseases in some circumstances by prioritising areas for investigation.
- Pan-London commissioning could improve tuberculosis services by enhancing spatial accessibility.

Conclusions: Spatial methods provide many valuable contributions to investigations of tuberculosis. Development of new tools and wider use of existing methods could limit the public health impacts of infectious disease outbreaks.

ACKNOWLEDGEMENTS

I would like to thank my supervisors, Andrew Hayward, Hannah Fry and Steve Le Comber, for their insight, support and enthusiasm throughout this work. I am also grateful to all at the Public Health England South East and London Field Epidemiology Service, in particular Helen Maguire and Charlotte Anderson, for supporting my placement with the team. Many thanks to Suzan Trienekens, PHE, for her help with the analysis of the isoniazid-resistant tuberculosis outbreak, and to Maeve Lalor and Tim Brown, PHE, for providing data. Thanks also to Neil Macdonald, PHE, for assistance with analysis of molecular clustering data; Al Story, UCLH, for providing access to Find and Treat data, and Sara Downs and Andy Mitchell at the Animal and Plant Health Agency for supporting the bovine tuberculosis project. I would also like to acknowledge the many online contributors to the open-source community who donate their time to share programming expertise, without which I could not have developed the mapping application. Finally, thanks to my lovely friends, housemates and family.

LIST OF CONTENTS

Abstract 2

Acknowledgements 3

List of contents 4

Publications and presentations arising from this thesis 7

List of figures 9

List of tables 12

List of boxes 14

List of electronic content 14

Abbreviations 15

1 Introduction to tuberculosis transmission and spatial analysis 16

1.1 Description of chapter contents 16

1.2 Tuberculosis natural history, epidemiology and control 16

1.3 Spatial analysis in infectious disease investigations 24

1.4 Thesis rationale and aims 32

2 Systematic literature review of use of spatial methods for infectious disease outbreak investigations 35

2.1 Description of chapter contents 35

2.2 Study rationale and introduction 35

2.3 Aims and objectives 35

2.4 Methods 36

2.5 Results 38

2.6 Discussion 48

3 Development of an open-source tool for visualising disease point locations 56

3.1 Description of chapter contents 56

3.2 Study rationale and introduction 56

3.3 Aims and objectives 60

3.4 Methods 60

3.5 Results 65

3.6 Discussion 69

4 Molecular and spatial epidemiology of tuberculosis in London 75

4.1 Description of chapter contents 75

4.2 Study rationale and introduction 75

4.3 Aims and objectives 77

4.4 Methods 77

4.5 Results 82
4.6 Discussion 97
5 Epidemiology and spatial analysis of a large isoniazid-resistant tuberculosis outbreak 105
5.1 Description of chapter contents 105
5.2 Study rationale and introduction 105
5.3 Aims and objectives 107
5.4 Methods 108
5.5 Results 111
5.6 Discussion 121
6 Geographic profiling as a tool for targeting tuberculosis control measures ... 129
6.1 Description of chapter contents 129
6.2 Study rationale and introduction 129
6.3 Aims and objectives 132
6.4 Case study 1: Isoniazid-resistant tuberculosis outbreak in London 132
6.5 Case study 2: Molecular cluster of tuberculosis in a rural town in England 137
6.6 Case study 3: Spatial targeting of badger culling for bovine tuberculosis control 140
6.7 Discussion 159
7 Spatial accessibility of tuberculosis services in London 163
7.1 Description of chapter contents 163
7.2 Study rationale and introduction 163
7.3 Methods 167
7.4 Results 170
7.5 Discussion 179
8 Summary of research and main findings, recommendations and final conclusions 184
8.1 Description of chapter contents 184
8.2 Summary of research undertaken 184
8.3 Key findings 186
8.4 Implications and recommendations for policy and practice 194
8.5 Future directions 199
8.6 Conclusions 201
9 References 202
10 Appendices 229

Appendix 10.1: Details of studies included in systematic literature review on spatial methods for infectious disease outbreak investigations (Chapter 2). 229

Appendix 10.2: Questionnaire used in evaluation of interactive mapping application (Chapter 3). 246

Appendix 10.3: Box plots of distribution of number of cases in molecular tuberculosis cluster by risk factor, London, 2010-2014 (Chapter 4). 249

Appendix 10.4: Screen shots from interactive mapping application showing key characteristics of molecular clusters of tuberculosis with significant spatial clustering in London, 2010-2014 (Chapter 4). 250

Appendix 10.5: Sensitivity analysis using different implementations of geographic profile model for analysis of Randomised Badger Culling Trial data (Chapter 6). 254

Appendix 10.6: Ethical approval application exemption letter. 259

PUBLICATIONS AND PRESENTATIONS ARISING FROM THIS THESIS

PUBLICATIONS

1. Smith CM, Downs SH, Mitchell A, Hayward AC, Fry H & Le Comber SC. Spatial targeting for bovine tuberculosis control: Can the locations of infected cattle be used to find infected badgers? *PLoS ONE* 2015 10(11): e0142710. (Based on work in Chapter 6).
2. Smith CM, Le Comber SC, Fry H, Bull M, Leach S & Hayward AC. Spatial methods for infectious disease outbreak investigations: Systematic literature review. *Euro Surveill.* 2015;20(39):pii=30026. (Based on work in Chapter 2).
3. Smith CM & Hayward AC. DotMapper: an open source tool for creating interactive disease point maps. *BMC Infect Dis* 2016 Apr 12;16:145. (Based on work in Chapter 3).
4. Smith CM & Emmett L. Navigating an outbreak: geospatial methods for STI outbreak investigations. *Sex Transm Infect.* 2016;92:327-328. (Based on work in Chapter 2).
5. Smith CM, Trienekens SCM, Anderson C, Lalor MK, Brown T, Fry H, Story A, Hayward AC & Maguire H. Twenty years and counting: Epidemiology of an outbreak of isoniazid-resistant tuberculosis in England and Wales, 1995-2014. *Accepted for publication in Euro Surveill., September 2016.* (Based on work in Chapter 5).
6. Smith CM, Maguire H, Anderson C, Macdonald N & Hayward AC. Multiple large clusters of tuberculosis in London: a cross-sectional analysis of molecular and spatial data. *Accepted for publication in ERJ Open Research, November 2016.* (Based on work in Chapter 4).

MAJOR PRESENTATIONS

1. Smith CM et al. Approaching twenty years and counting: Epidemiology of an outbreak of isoniazid-resistant tuberculosis in England and Wales, 1995-2013. Oral presentation at Public Health England Applied Epidemiology Scientific Meeting, Warwick, March 2015. (Based on work in Chapter 5).

2. Smith CM et al. Spatial methods for infectious disease outbreak investigations: Systematic literature review. Poster presentation at Farr Institute International Conference, St Andrews, August 2015. (Based on work in Chapter 2).
3. Smith CM et al. Development of an open source tool for mapping disease clusters. Oral presentation at Public Health Science conference, London, November 2015. (Based on work in Chapter 3).
4. Smith CM et al. Development of an open source tool for mapping disease clusters. Oral presentation at Public Health England Applied Epidemiology Scientific Conference, Warwick, March 2016. (Based on work in Chapter 3).

LIST OF FIGURES

Figure 1.1: Estimated tuberculosis incidence rates, 2014. 17

Figure 1.2: Tuberculosis case notifications and rates, England, 2000-2014. 18

Figure 1.3: Three-year average tuberculosis rates by local authority district, England and London, 2012-2014. 19

Figure 1.4: Tuberculosis rates in England by risk group, 2014. 20

Figure 1.5: Dot map of John Snow's cholera outbreak investigation in London in 1854. 26

Figure 1.6: Smoothed intensity map and Voronoi diagram of John Snow's cholera outbreak investigation in London in 1854. 29

Figure 1.7: Geographic profile of John Snow's cholera investigation in London in 1854. 32

Figure 2.1: Study selection, systematic literature review on spatial methods in infectious disease outbreak investigations. 39

Figure 2.2: Reports of outbreak investigations using spatial methods by year. 39

Figure 2.3: Locations of outbreak investigations using spatial methods by country and continent. 42

Figure 3.1: Workflow for running interactive dot mapping application on a local machine. 63

Figure 3.2: Structure of data that can be imported into the interactive dot mapping application. 63

Figure 4.1: Definition of small and large molecular clusters of tuberculosis used for prediction of large clusters. 81

Figure 4.2: Cases included in analysis of molecular clusters of tuberculosis in London, 2010-2014. 82

Figure 4.3: Median interval between successive tuberculosis cases in molecular clusters in London, 2010-2014, by cluster size. 83

Figure 4.4: Box plots of distribution of number of cases in molecular tuberculosis cluster by number of social risk factors, London, 2010-2014. 84

Figure 4.5: Forest plot of adjusted odds ratios and 95% confidence intervals from multivariable multinomial logistic regression analysis for risk factors associated with tuberculosis cases in molecular clusters of different sizes in London, 2010-2014 (Table 4.4). 92

Figure 4.6: Locations of significant spatial clusters of cases within eight molecular clusters (denoted a-h) of tuberculosis in London (2010-2014), overlaid on smoothed incidence maps. 93

Figure 5.1: Case ascertainment, sampling and strain typing methods for defining cases in the isoniazid-resistant outbreak, 1999-2014. 109

Figure 5.2: Number of cases in the isoniazid-resistant tuberculosis outbreak by year, England and Wales, 1995-2014. 112

Figure 5.3: Percentage of cases in the isoniazid-resistant tuberculosis outbreak by year and social risk factor, 1999-2014. 115

Figure 5.4: Cases included in analyses of treatment outcomes and contribution of Find and Treat to control of the isoniazid-resistant tuberculosis outbreak. 116

Figure 5.5: Numbers of cases in the isoniazid-resistant outbreak in England (by PHE Region) and Wales, 1995-2014. 118

Figure 5.6: Smoothed incidence maps of cases in the isoniazid-resistant tuberculosis outbreak in London, by five-year time period, 1995-2014. 119

Figure 5.7: K-function analysis of spatial clustering in isoniazid-resistant tuberculosis outbreak in London, 1995-2014.. 120

Figure 5.8: Study selection, systematic literature review on large outbreaks of tuberculosis. 121

Figure 6.1: Model results derived from locations of cases in isoniazid-resistant tuberculosis outbreak in London, 1995-2014.. 135

Figure 6.2: Geographic profile derived from residential locations of tuberculosis cases and latent infections in molecular cluster of tuberculosis in a rural town in England. 138

Figure 6.3: Example search areas using ring cull approach. 146

Figure 6.4: Locations of proactive trial areas of Randomised Badger Culling Trial included in this analysis. 149

Figure 6.5: Distributions of hit scores around cattle herd breakdowns, designed using ring cull and geographic profiling search strategies. 150

Figure 6.6: Kaplan-Meier curves comparing sizes of search areas for setts housing tuberculosis-infected and uninfected badgers, by ring cull and geographic profile methods. 151

Figure 6.7: Kaplan-Meier curves comparing search areas for setts housing TB-infected badgers by ring cull and geographic profile methods for all trial regions; region B2 and region E3.. 154

Figure 7.1: Locations of tuberculosis clinics in London by Health Protection Team, Clinical Commissioning Group, Sector, and cohort review Group. Small areas are CCG boundaries. 164
Figure 7.2: Cases included in analysis of tuberculosis service accessibility in London, 2010-2013. 171
Figure 7.3: Tuberculosis patients and clinics in London, 2010-2013. Cases are colour coded according to the tuberculosis clinic that they used (*Used* panel) and to their nearest clinic based on travel time (*Catchment* panel). 172
Figure 7.4: Distribution of travel times for tuberculosis patients to their used and catchment clinics, London, 2010-2013. 173
Figure 7.5: Distribution of tuberculosis patient travel times by clinic, London, 2010-2013. 174
Figure 7.6: Tuberculosis patients who attended University College London Hospital clinic, 2010-2013. 175
Figure 7.7: Impact of removing each tuberculosis clinic on average patient travel times, London, 2010-2013. 176
Figure 7.8: Optimum configurations of tuberculosis clinics to minimise patient travel time, by number of clinics included, London, 2010-2013. 177
Figure 7.9: Patient travel times to tuberculosis clinics for different numbers of optimum clinic configurations, London, 2010-2013. Patients assigned to clinics based on minimum travel times. 178
Figure 7.10: Average annual number of tuberculosis patients per clinic for different numbers of optimum clinic configurations, London, 2010-2013. Patients assigned to clinics based on minimum travel times. 179

LIST OF TABLES

Table 2.1: Infectious diseases investigated by category. 40

Table 2.2: Contexts of outbreak investigations of human and animal diseases...... 41

Table 2.3: Spatial methods used in outbreak investigations. 42

Table 3.1: R packages used in the interactive mapping application........................ 62

Table 3.2: Features suggested at user demonstration for interactive mapping tool, and how they were implemented. 68

Table 4.1: Distribution of tuberculosis cases and molecular clusters by cluster size in London, 2010-2014. 83

Table 4.2: Baseline characteristics of tuberculosis cases in molecular clusters of different sizes in London, 2010-2014. 85

Table 4.3: Single variable multinomial logistic regression analysis for risk factors associated with tuberculosis cases in molecular clusters of different sizes in London, 2010-2014 87

Table 4.4: Multivariable multinomial logistic regression analysis for risk factors associated with tuberculosis cases in molecular clusters of different sizes in London, 2010-2014, adjusted for random effects of Lower Super Output Area. 91

Table 4.5: Description of molecular clusters of tuberculosis with significant spatial clustering in London, 2010-2014.. 95

Table 4.6: Characteristics of large (N ≥ 5 cases) and small (N < 5 cases) molecular clusters of tuberculosis and results of single variable logistic regression analysis for cluster-level risk factors associated with cluster size in London, 2010-2014. 96

Table 5.1: Demographic characteristics of cases in the isoniazid-resistant tuberculosis outbreak, England and Wales, 1995-2014. 113

Table 5.2: Treatment outcomes of cases in the isoniazid-resistant tuberculosis outbreak, England and Wales... 117

Table 6.1: Application of geographic profiling to criminology, biology and epidemiology.. 130

Table 6.2: Tuberculosis cluster investigation questionnaire and geographic profile results (for venues visited by two or more individuals) in molecular cluster of tuberculosis in rural town in England.. 139

Table 6.3: Badger culling strategies used in the current bovine tuberculosis policy in England, the Randomised Badger Culling Trial, and hypothetical spatially targeted culling strategies used in this study. 142

Table 6.4: Characteristics of included areas from proactive trial regions of first cull of the Randomised Badger Culling Trial. 148

Table 6.5: Infected and uninfected setts included in searches with different thresholds according to search strategy, aggregated across all trial areas. 152

Table 6.6: Cox regression spatial survival analysis comparing search strategies. 153

Table 8.1: Application of spatial methods to steps in outbreak investigation. 196

LIST OF BOXES

Box 1.1: Factors determining the probability of transmission of tuberculosis. 22
Box 1.2: Summary of Chapter 1. 34
Box 2.1: Steps in an outbreak investigation. 38
Box 2.2: Summary of Chapter 2. 55
Box 3.1: Functional and technical specifications for interactive mapping application. 61
Box 3.2: Feedback from user demonstration during development of interactive mapping application (27 July 2015). 67
Box 3.3: Summary of Chapter 3. 74
Box 4.1: Summary of Chapter 4. 104
Box 5.1: Summary of Chapter 5. 128
Box 6.1: Summary of Chapter 6. 162
Box 7.1: Summary of Chapter 7. 183

LIST OF ELECTRONIC CONTENT

Provided with attached CD-ROM

E 3.1: DotMapper demonstration movie 1.
E 3.2: DotMapper demonstration movie 2.
E 3.3 DotMappter demonstration movie 3.
E 5.1 Animation of spatial locations of cases in isoniazid-resistant outbreak.

Abbreviations

API	Application Programming Interface
CGT	Criminal Geographic Targeting
CI	Confidence Interval
CSS	Cascading Style Sheets
DOT	Directly Observed Therapy
DPM	Dirichlet Process Mixture
ECDC	European Centre for Disease Control
ETS	Enhanced Tuberculosis Surveillance
FMD	Foot-and-Mouth Disease
GIS	Geographic Information Systems
HIV	Human Immunodeficiency Virus
IMD	Index of Multiple Deprivation
LSOA	Lower-Layer Super Output Area
LTBR	London Tuberculosis Register
MCMC	Markov Chain Monte Carlo
MDR	Multidrug-Resistant
MeSH	Medical Subject Headings
NCL	North Central London
OR	Odds Ratio
PHE	Public Health England
RBCT	Randomised Badger Culling Trial
RFLP	Restriction Fragment Length Polymorphism
TfL	Transport for London
UCLH	University College London Hospital
WHO	World Health Organization
XDR	Extremely Drug-Resistant

1 INTRODUCTION TO TUBERCULOSIS TRANSMISSION AND SPATIAL ANALYSIS

1.1 DESCRIPTION OF CHAPTER CONTENTS

Tuberculosis is a major global health problem that requires an increasingly local approach to control. Although England is a low incidence country, high rates of disease are still found in London. Spatial methods have long been used to investigate infectious diseases, but the extent to which they could contribute to understanding of tuberculosis transmission has not been explored in depth.

In this chapter I introduce the background to tuberculosis and the use of spatial methods for investigating infectious disease transmission. This includes a summary of the main epidemiological characteristics and risk factors for tuberculosis; its natural history, detection and treatment, and a description of control measures used to interrupt its transmission. I then provide a historical perspective of the use of spatial data in infectious disease epidemiology, before describing some of the most common methods for visualising, describing, identifying clusters and modelling spatial data that are used today. I also introduce a new tool, geographic profiling, which has recently been applied to infectious disease epidemiology as a means identifying likely locations of sources of infectious diseases. Finally, I summarise the rationale for this thesis and describe its overall aims and objectives.

1.2 TUBERCULOSIS NATURAL HISTORY, EPIDEMIOLOGY AND CONTROL

1.2.1 Epidemiology of tuberculosis worldwide

Tuberculosis remains a major global health problem, and is a leading infectious cause of death worldwide.[1] In 2014, 9.6 million people are estimated to have fallen ill with the disease, equivalent to 133 per 100,000 population, and 1.5 million died.[2] Since 2000, the absolute number of cases worldwide has been slowly declining, at a rate of 1.5% per year.[2]

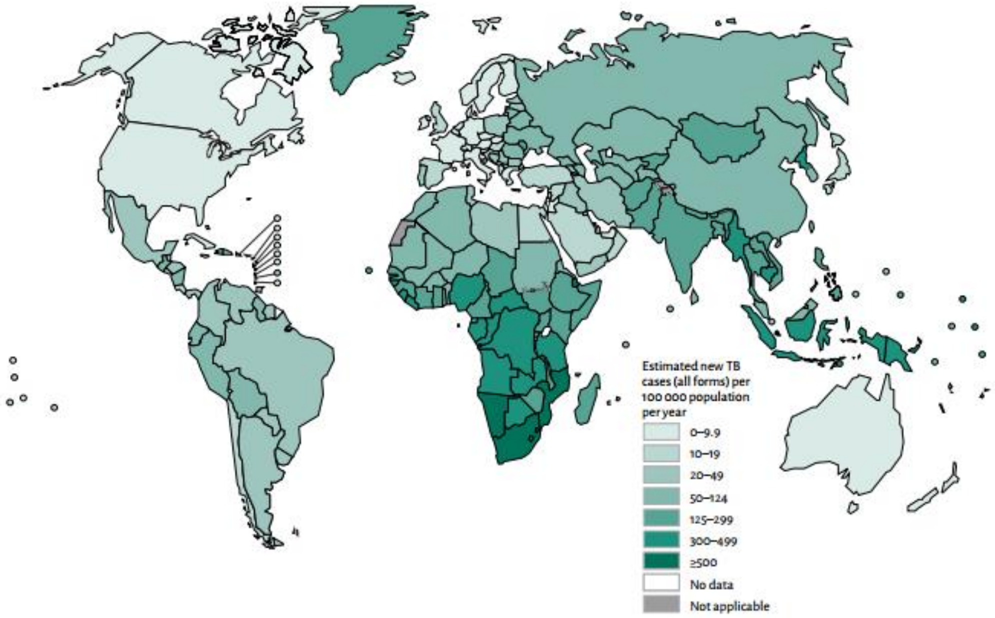
Globally, the incidence of tuberculosis is estimated to be 1.7 times higher in men than women, and the majority of cases are in adults.[2] In the World Health

Organization (WHO) regions of the Americas, Europe and Africa, rates are highest in younger adults, whereas in the Eastern Mediterranean, Western Pacific and South-East Asian regions notification rates increase with age.[2] In 2014, multidrug-resistant (MDR) tuberculosis contributed an estimated 3.3% of new cases and 20% of previously treated cases worldwide. Highest levels of MDR disease are found in Eastern European and central Asian countries.[2]

The incidence rate of tuberculosis tends to be lowest in high income countries, including most countries in Western Europe, Canada, the United States of America, Australia and New Zealand (Figure 1.1).[2] These countries generally had rates of 10 per 100,000 or lower in 2014. The highest rates are found in Africa and Asia, and the highest disease burdens in terms of absolute numbers of cases are in India, Indonesia, China, Nigeria and Pakistan. In South Africa, for example, the 2014 incidence rate was estimated at 834 per 100,000 population, a total of 450,000 new cases.[2]

Figure 1.1: Estimated tuberculosis incidence rates, 2014.

Reprinted from WHO Global tuberculosis report 2015, Chapter 2: Disease burden and 2015 targets assessment, page 18.[2]


Estimated new TB cases (all forms) per 100 000 population per year
0–9.9
10–19
20–49
50–124
125–299
300–499
≥500
No data
Not applicable

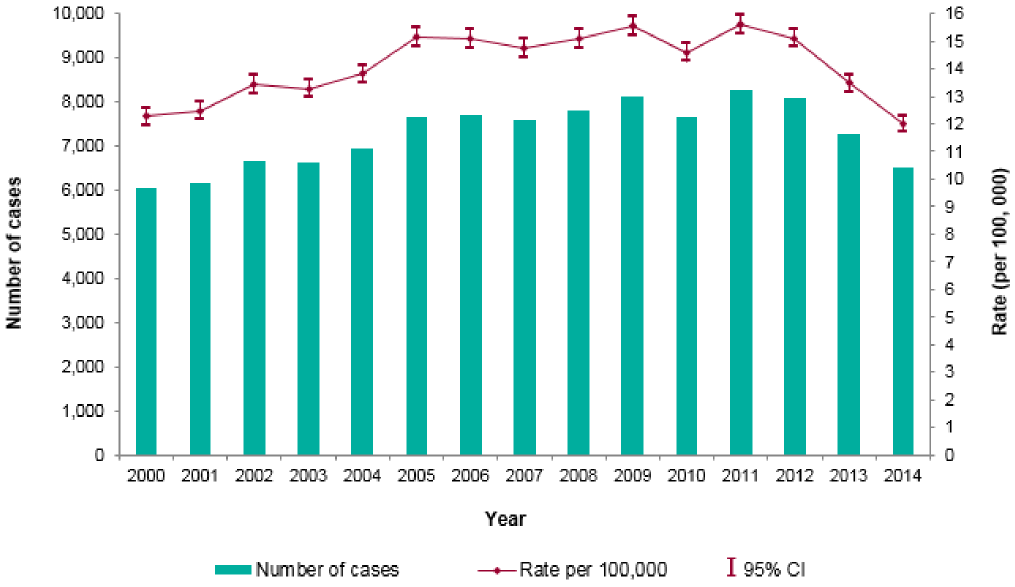
Human Immunodeficiency Virus (HIV) is an important risk factor for tuberculosis. This is because it impairs the immune system, leading to increased susceptibility to infection with tuberculosis bacteria and probability of progression to active disease.[3,4] An estimated 1.2 million (12%) of the incident cases of tuberculosis in 2014 occurred amongst HIV positive people.[2] This problem is particularly concentrated in the African WHO region, in which 74% of these people resided. HIV co-infected people are also more likely to die from tuberculosis infection, with a quarter of all deaths from tuberculosis in 2014 amongst those who were HIV positive.[2]

1.2.2 Epidemiology of tuberculosis in England

Tuberculosis is a notifiable disease in England, and all reported cases are recorded using the Enhanced Tuberculosis Surveillance (ETS) system, maintained by Public Health England (PHE). There were 6,520 cases of tuberculosis notified in 2014 in England, a rate of 12.0 per 100,000 population.[5] This represented a decrease compared to the 2013 rate, but no overall change since 2000 (Figure 1.2).[5] In the UK, rates of tuberculosis declined in the 20th century, then increased from the early 1980s until mid-2000s, and have since remained relatively stable.[6]

Figure 1.2: Tuberculosis case notifications and rates, England, 2000-2014.

Source: Public Health England. Tuberculosis in England 2015 report.[5]

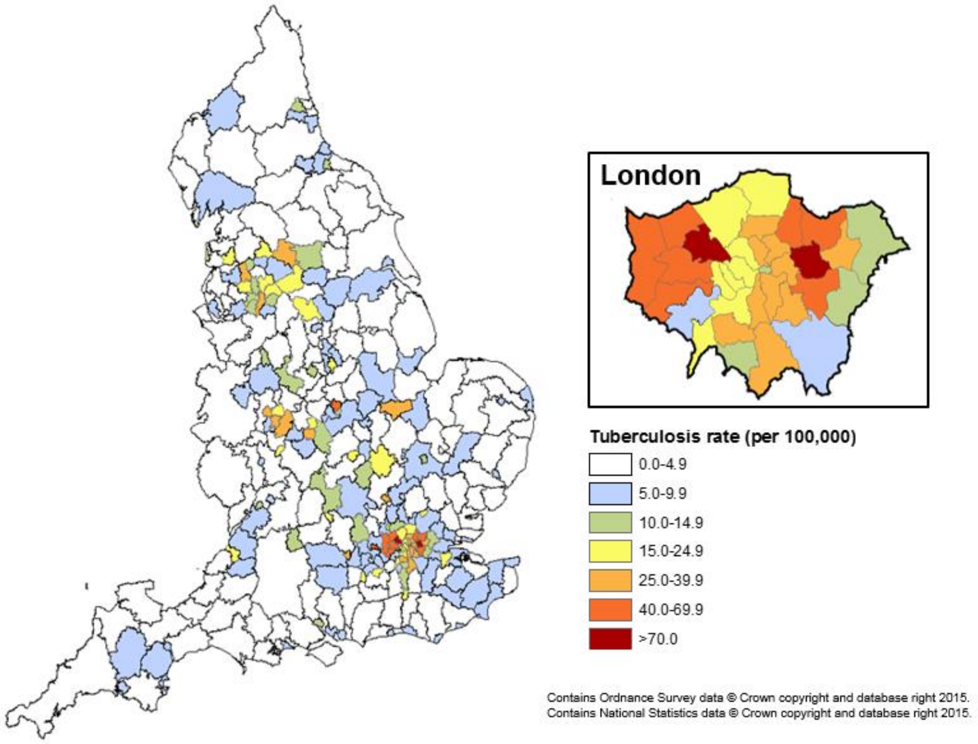


More than half (59%) of the cases in England in 2014 were male, and the age group with the highest rate was 30 to 34 years (23.4 per 100,000 population). The majority of cases (72%) occurred in individuals who were born outside of the UK, and the rate in this population was fifteen times higher than in the UK born population. Amongst those born outside the UK, the most frequent countries of birth were India (21%), Pakistan (13%), and Somalia (4%). The proportion of new cases in England in 2014 who had MDR disease was 1.4% (56 cases), a decrease since 2011, when this proportion peaked at 1.8% (88 cases).

In England, like other European countries, highest rates of tuberculosis are found in large cities (Figure 1.3).[7] The rate of tuberculosis in London in 2014 was 30.1 per 100,000 population (95% CI 29.0-31.3), and this accounted for 39.4% of all cases in England.[8]

Figure 1.3: Three-year average tuberculosis rates by local authority district, England and London, 2012-2014.

Source: Public Health England. Tuberculosis in England 2015 report.[5]

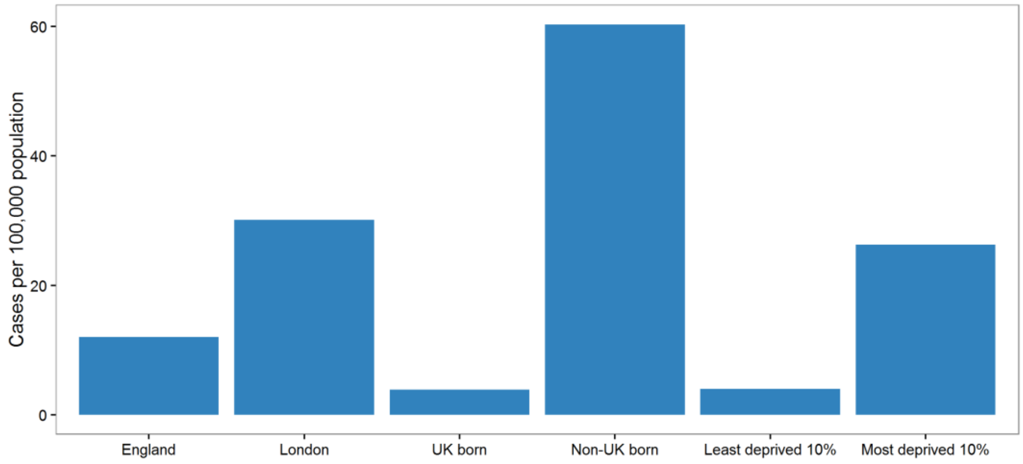


Contains Ordnance Survey data © Crown copyright and database right 2015.
Contains National Statistics data © Crown copyright and database right 2015.

Tuberculosis in England is also associated with social deprivation. In 2014, the rate of tuberculosis was 26.3 per 100,000 in the 10% of the population living in the most deprived areas of England, compared with 4.0 per 100,000 in the 10% of the population living in the least deprived areas.[5] Social risk factors (current or history of homelessness, imprisonment, drug or alcohol misuse) are particularly important in the UK born population.[9] Of the cases in England in 2014 who were born in the UK, 15% had at least one of these risk factors, compared to 7% born outside the UK. Rates of tuberculosis in different populations in England are summarised in Figure 1.4.

Figure 1.4: Tuberculosis rates in England by risk group, 2014.

Data source: Public Health England. Tuberculosis in England 2015 report.[5]



1.2.3 Natural history, detection and treatment

Tuberculosis is a bacterial disease that has persisted throughout human history.[10] It usually results from infection with *Mycobacterium tuberculosis*, but can also be caused by other *Mycobacteria* including *M. africanum* and *M. bovis*. The disease is most often associated with the lungs (pulmonary tuberculosis), but can affect any part of the body.[11]

Most people who are infected with tuberculosis do not develop symptoms but have latent infection. This is defined as a state of persistent immune stimulation by *M. tuberculosis* antigens without evidence of clinical disease.[12] These individuals are not infectious because the *Mycobacteria* in their body are being controlled by their immune system. Properties of the immune response can be used to screen for latent infection using, for example, the tuberculin skin test or the Interferon Gamma Release Assay.[13] Effective treatments for latent tuberculosis are available, but they are not always administered, partly because the drugs can cause harmful side effects.[14] However, modelling estimates suggest that if widespread treatment for latent tuberculosis infection could be attained it would greatly increase the chance of eliminating the disease.[15]

Approximately one in ten individuals who are infected with tuberculosis will progress to active disease.[16] This may occur after a period of latency (termed reactivation) or soon after infection (recent transmission). The main clinical symptoms of pulmonary tuberculosis include persistent cough, fever and weight loss, and individuals with these symptoms are infectious.[17] Active disease can be detected by inspection of chest radiographs followed by sputum smear microscopy and microbial culture testing, the current reference standard.[18] A molecular assay, Gene Xpert MTB/RIF is also available, which additionally provides rapid detection of drug resistance.[19,20] This test has improved sensitivity when compared to microscopy or radiography but is expensive and limited by throughput capacity.[21]

Tuberculosis can be treated effectively with antibiotics. The current standard regimen for drug-sensitive disease is six months of isoniazid and rifampicin, supplemented in the first two months by pyrazinamide and ethambutol.[12,18] Treatment of tuberculosis can be complicated by drug resistance, which emerged shortly after the introduction of anti-tuberculosis treatments.[22] Disease that is resistant to at least isoniazid and rifampicin is known as MDR tuberculosis, and requires an extended treatment regimen using additional second line drugs.[12] More recently, extremely drug-resistant (XDR) tuberculosis strains have emerged.[23] XDR tuberculosis provides further challenges to treatment because it is resistant to drugs in at least two of the additional classes of second-line antibiotics usually used in treatment of MDR disease.

1.2.4 Molecular strain typing

Strain typing methods can be used to characterise tuberculosis isolates at the molecular level. Early methods developed for molecular strain typing included spoligotyping and restriction fragment length polymorphism (RFLP) analysis. These are both based on polymerase chain reaction (PCR)-based amplification of regions of the mycobacterial genome which vary in size in different strains.[24] The Mycobacterial Interspersed Repetitive Unit-Variable Number Tandem Repeats (MIRU-VNTR) method has been introduced more recently.[5] This method distinguishes strains of *M. tuberculosis* by the number of copies of tandem repeats at specific regions in the genome.[24] The discriminatory power of MIRU-VNTR typing depends on the number of loci evaluated, typically either 12 or 24.

There are various applications of molecular strain typing to epidemiological investigations of tuberculosis. For example, it can be used to distinguish between cases of reactivation and reinfection, to confirm instances of laboratory cross-contamination, and to identify clusters of cases that may be linked through transmission.[25,26] The National Tuberculosis Strain Typing Service in England has been typing isolates prospectively using 24 loci MIRU-VNTR since 2010. In the first five years of prospective typing, 82% of culture confirmed cases had at least 23 loci typed.[5] Of these, 57% were in 2,245 molecular clusters, indicating potential transmission links.[5] A detailed analysis of MIRU-VNTR cluster data from London is presented in Chapter 4.

1.2.5 Transmission

Transmission of tuberculosis usually occurs when an individual with active disease expels droplet nuclei containing *Mycobacterium* bacilli by coughing, sneezing, shouting or singing.[27] These droplet nuclei are then inhaled through nasal passages into the lungs of a susceptible individual. Transmission can also occur through ingestion, for example by drinking unpasteurised milk.

The probability that tuberculosis is transmitted from one person to another is determined by four main factors (Box 1.1).[27]

Box 1.1: Factors determining the probability of transmission of tuberculosis.

Adapted from: Centers for Disease Control and Prevention. Core Curriculum on Tuberculosis: What the Clinician Should Know.[27]

1. The infectiousness of the transmitting individual.
2. The susceptibility of the exposed individual.
3. The proximity, frequency and duration of contact between infectious and exposed individuals.
4. The environment in which the disease is being transmitted.

Infectiousness is determined by the number of bacilli that the individual expels into the environment. Increased infectiousness is associated with persistent cough, failure to cover the mouth when coughing, and a positive sputum smear test. It is also related to age, as young children are less likely to produce sputum when they cough.[27] Susceptibility depends on the immune status of the exposed individual. It can be influenced by prior exposure to tubercle bacilli, for example through vaccination, and by other factors which compromise the immune system such as infection with HIV.[4]

Proximity of infectious and exposed individuals relates to the frequency and duration of contact between them, as well as their physical nearness. It is important because *Mycobacterium* bacilli can only remain airborne for a limited period of time, and can therefore only travel over a limited distance.[28] The environment in which the disease is being transmitted affects the concentration of infectious droplet nuclei in the air. Factors such as the size of the space, ventilation, air pressure, and handling of specimens (for example in laboratories) determine the suitability of an environment for tuberculosis transmission.[27]

1.2.6 Control

Control measures to reduce transmission of tuberculosis can be directed at any of the above four factors. Prior to the discovery of effective anti-tuberculosis drugs in the mid-twentieth century, reducing proximity between tuberculosis patients and healthy individuals was one of the only approaches available. Patients were therefore referred to 'sanatoria', long term care facilities set up, in part, to isolate tuberculosis sufferers from those who they might infect.[10]

Today, reducing infectiousness by treatment of active cases is the main focus of control. Ensuring that individuals with disease receive effective treatment not only benefits the patients themselves but also has public health benefit as it reduces the length of time that they are able to infect others. The WHO therefore recommends monitoring of treatment outcomes for all bacteriologically confirmed and clinically diagnosed cases of tuberculosis.[29] In England, the proportion of drug-sensitive cases that were notified in 2013 who completed treatment within one year was 85%.[5] Directly Observed Therapy (DOT) is a strategy which is recommended for patients thought to be at risk of not completing treatment.[18] It aims to improve adherence to antibiotic regimens through active monitoring and recording of drug consumption by an observer acceptable to the patient and health system.[30]

Another method to reduce the length of the infectious period (or avert it completely) is screening of high risk individuals for active or latent infection.[31] For example, the UK operates a pre-entry screening programme for people migrating from high incidence countries.[18] Contact tracing is also used to identify high risk individuals for screening, and involves systematic evaluation of the contacts of known tuberculosis patients.[32] Screening can additionally be focused on groups in the community who are known to be at risk of tuberculosis, including those with social risk factors. For example, *Find and Treat* is a service in London which uses a mobile digital x-ray unit to identify cases in these hard-to-reach populations.[31] Early diagnosis of tuberculosis relies on recognition of symptoms and access to appropriate services. Raising and maintaining awareness of tuberculosis amongst populations most affected and professionals working with them is therefore also important.[18]

Reducing susceptibility of exposed individuals to infection with *M. tuberculosis* requires effective vaccines. This has the potential to contribute major improvements in the control of the disease, but the only currently available vaccine is Bacille Calmette-Guérin (BCG), which has variable protective efficacy.[33] In the UK, BCG vaccination is currently recommended only for individuals thought to have an increased risk of coming into contact with tuberculosis, for example children from high incidence countries.[18]

Altering environments in which tuberculosis may be transmitted can reduce the concentration of tubercle bacilli in the air. To date, these control measures have generally been focused on improving ventilation in health care settings, however, they are increasingly being recognised as important measures in public settings.[34,35]

The interventions outlined here have contributed to great advances in the control of tuberculosis over the last 20 years, including a reduction in associated mortality of 47% since 1990.[2] However, there is now a changing emphasis in global tuberculosis programmes from disease control to elimination. The WHO's *End TB Strategy*, published in 2015, aims for a 90% reduction in tuberculosis incidence by 2035, and has a long term vision of eliminating the disease as a public health problem.[36] Achieving this is likely to require an increased focus on localised strategies, which have been important in elimination of other diseases such as smallpox and polio.[37-39] A key component of local control measures is effective collection and analysis of spatial data, for example to identify transmission hotspots and direct community based interventions.[37]

1.3 Spatial analysis in infectious disease investigations

1.3.1 History

Mapping locations of individuals affected by disease has long formed an important part of epidemiological investigations. In early examples, maps were used to plan rudimentary public health interventions and to inform debates about the aetiology of diseases.

Throughout the Middle Ages and Renaissance periods, for example, plague epidemics were a recurrent threat across Europe, Asia and Africa. Maps of the progression of the disease showed that it followed trade networks, and a theory was developed that it was a portable by-product of trade.[40] As an early public health intervention, merchant ships were therefore quarantined in ports in an attempt to prevent the disease spreading into cities such as Venice and London. In the province of Bari, Italy, maps were used to plan the quarantine measures, directing troops to isolate plague-affected towns from those in which the disease had not yet appeared.[40]

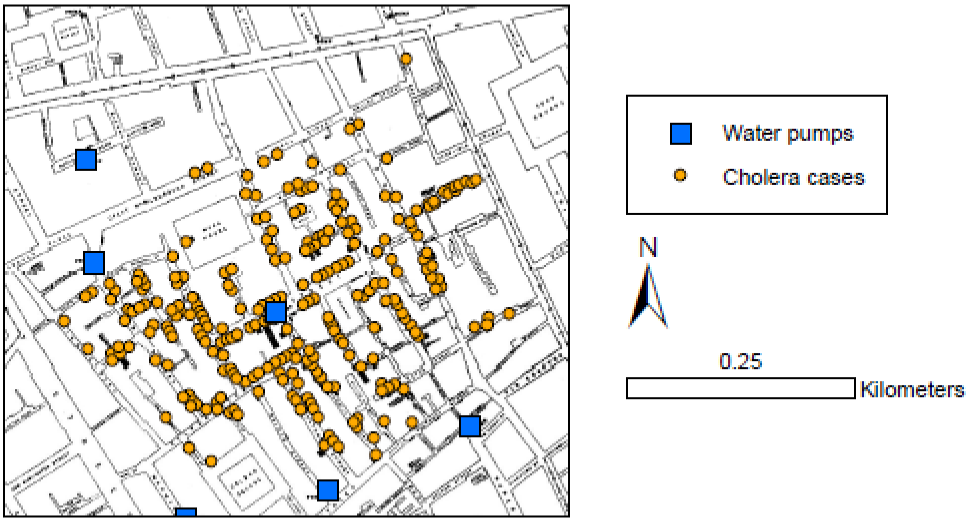
Maps also played an important role in understanding of cholera transmission. In 1831, the Lancet published a map which described the progress of the disease across the world and was used to correctly predict its imminent arrival in England.[41] At this time, many cholera epidemics were occurring worldwide, but its origin and mode of transmission were not known. The most pervasive theory was that it was transmitted through miasma, or 'bad air'. In support of this, a map of cases in the Baltic city of Dantzick appeared to show that the disease emerged from dirty alleys and then spread irregularly.[40]

John Snow first published the hypothesis that cholera had a waterborne source in 1849.[42] His theory was based on the logic that diseases of the gut are more commonly caused by something that has been swallowed rather than inhaled. In the report, he identified potential sources of contaminated water that may have caused outbreaks in Dumfries (river Nith), Glasgow (river Clyde) and London (river Thames).[40] However, his theory was not widely accepted at first, and critics pointed out that the disease was not universal among people who drank polluted water, and people in some areas were ill even though water they drank was assumed to be clean. In 1852, William Farr published an alternative theory of cholera transmission.[43] His report included maps of the disease at a national scale, and showed that it appeared to progress from port cities inland. He also observed a seasonality in the data, and this led him to hypothesise that cholera was generated in the warm airs that evaporated from polluted water in summer.

The now famous outbreak of cholera which would eventually provide more compelling evidence for John Snow's theory of waterborne transmission occurred in London in 1854.[44] This outbreak led to the death of more than 500 people in the Soho district of the city. Snow observed that the majority of those who died lived close to a particular water pump, on Broad Street, whilst few cases lived nearer to a different pump.[44] This led him to propose that the Broad Street pump was the source of the outbreak, and to produce his *Diagram of the topography of the* *outbreak* (reproduced in Figure 1.5). The map, on which cases were plotted at their residential locations, serves as a visual test of Snow's theory on the origin of disease. The theory was given further weight when the likely index patient was identified, a baby whose soiled clothes had been washed in a cesspool located a few feet from the Broad Street pump. Snow petitioned for the removal of the handle from the Broad Street water pump, and it was subsequently removed, although by this time the cases had already started to abate.

Figure 1.5: Dot map of John Snow's cholera outbreak investigation in London in 1854.

Adapted from: Snow. On the mode and transmission of cholera.[44]



These historical examples serve to highlight some of the opportunities afforded by spatial thinking in infectious disease investigations, as well as some limitations which remain pertinent to contemporary epidemiological analyses. Presenting data on a map is a powerful means of linking cases, transforming data from rows of a table into equal, shared events whose elements are placed in context.[40] It forces consideration of these events in their geographic, economic, and social landscapes which may promote or inhibit development of the disease. This can be used to generate or test hypotheses about the transmission of infectious diseases, but can also lead to incorrect conclusions if false assumptions are held. As demonstrated early on through quarantining, and later by Snow, maps can provide an objective means of proposing intervention measures. Used in isolation, however, spatial analyses are insufficient to prove hypotheses. Snow's theory of cholera transmission was not widely accepted until 1883 when Robert Koch definitively identified *Vibrio cholerae*, the waterborne agent of disease.[40] This emphasises the need to combine spatial analyses with other tools such as microbiological and epidemiological investigations to formulate robust conclusions.

1.3.2 Spatial methods and applications

Since Snow produced his dot map, spatial analysis has become a standard component of epidemiological outbreak investigations. Today, guidelines for investigating outbreaks of infectious diseases including the European Centre for Disease Prevention and Control (ECDC) toolbox invariably recommend consideration of case locations.[45-48] Spatial analyses have also been applied to a broad range of other epidemiological research themes, including surveillance, production of disease 'atlases', and health service analysis.

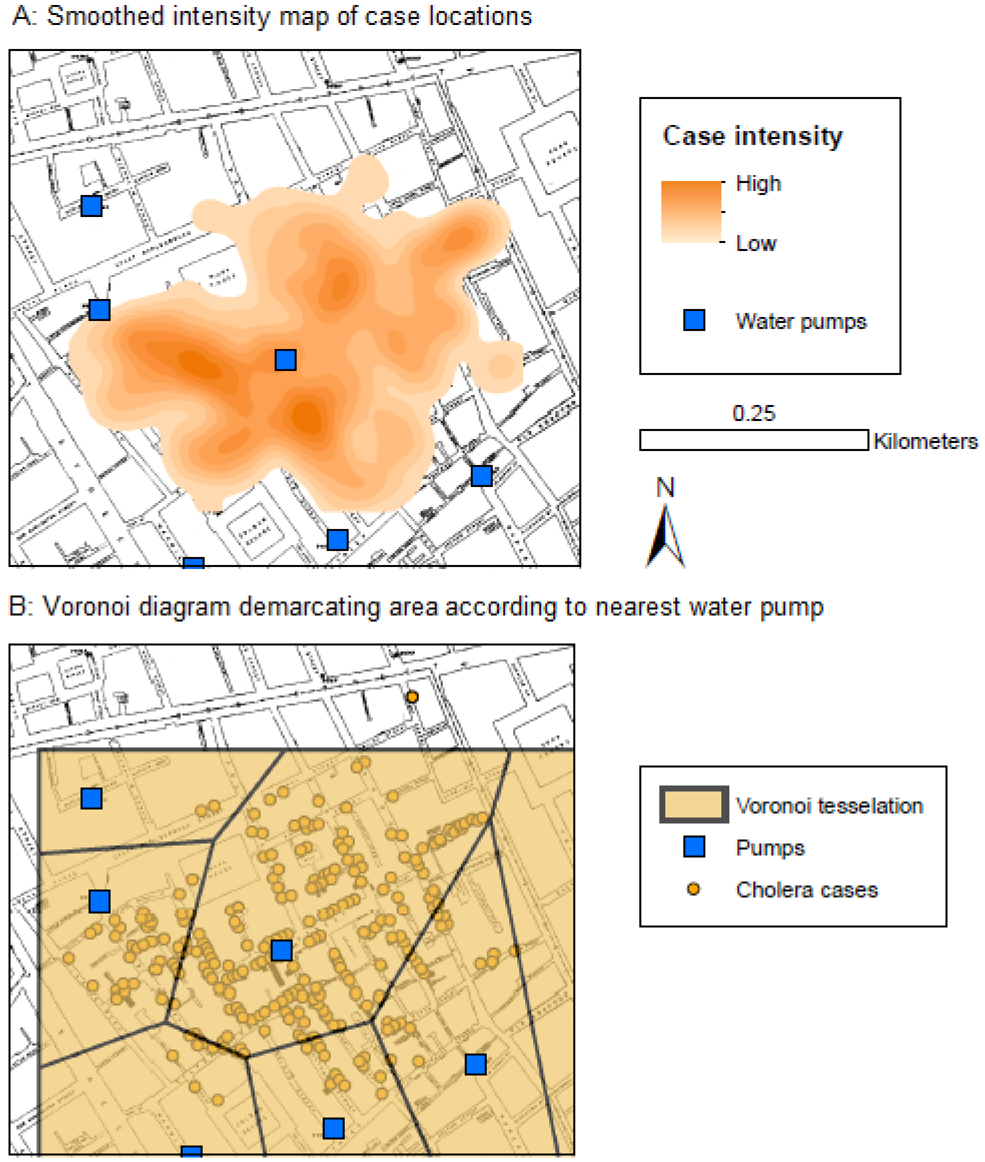
Geographic information systems (GIS) have increased the availability and range of tools that can be used to analyse spatial data. A GIS is a database designed to handle geographically-referenced information complemented by software tools for the input, management, analysis and display of data.[49] Methods available in GIS include those for analysis of point patterns, such as patient residential locations, as well as for data aggregated into aerial units, such as administrative areas. Broadly, spatial methods can be categorised into those designed for data visualisation, descriptive analyses, cluster analyses, and modelling. Here, I summarise some of the key methods commonly used in epidemiological investigations, and describe a new method, geographic profiling.

1.3.2.1 Visualisation

Visualisation of data on maps provides an easy-to-understand means of presenting information about disease in context. Cases can be plotted as point locations on dot maps or aggregated into administrative areas and displayed as cumulative counts or rates. These maps can be used to describe patterns, identify outliers, and communicate findings. Smoothed incidence maps are an alternative means of visualising point locations as continuous distributions of disease risk, generated by adjusting the density at each point according to the number of cases in adjacent areas.[49] Figure 1.6A shows an example of this using the data from Snow's 1854 cholera outbreak investigation, and they are also used in Chapters 4 and 5. Creation of a series of dot maps or smoothed incidence maps for different time periods can be a useful means of visualising changes in disease distributions over time. A tool enabling interactive dot mapping is developed in Chapter 3.

Areas on maps can also be demarcated according to other locations of interest such as potential sources of infection or service access points. For example, contour lines can be added to maps that join locations of equal distance (isodistances) or travel time (isochrones) from a certain point. Voronoi diagrams (also known as Dirichlet regions or Thiessen polygons) are a means of partitioning points on a two dimensional plane into convex polygons such that each polygon contains exactly one generating point and every point in a given polygon is closer to its generating point than any other.[50] An example of this is demonstrated in Figure 1.6B, in which the area included in Snow's cholera investigation has been divided according to its closest water pump, and shows that majority of cases lie within the area closest to the pump on Broad Street. Further examples of spatial visualisations include interactive maps, animations, schematic maps, and origin-destination plots.

Figure 1.6: Smoothed intensity map and Voronoi diagram of John Snow's cholera outbreak investigation in London in 1854.



A: Smoothed intensity map of case locations

B: Voronoi diagram demarcating area according to nearest water pump

1.3.2.2 *Descriptive analyses*

GIS can also be used to describe and explore spatial disease data. Measuring distances, for example between cases and potential sources of infection, can be informative if an infection is suspected to derive from an environmental point source. In outbreaks of Legionnaires' disease, this method has been applied to identify cooling towers or other aerosol-producing devices proximal to cases, and therefore to generate hypotheses about the likely source.[51] Similarly, distances or travel times can be used to determine accessibility to healthcare services. A study in rural South Africa, for example, demonstrated that there was a significant decline in usage of clinics with increasing travel time.[52] Analysis of accessibility of tuberculosis clinics in London based on travel times is presented in Chapter 7.

1.3.2.3 Cluster analyses

Clusters, in the context of spatial analysis of disease data, can be defined as areas with higher than expected levels of disease risk. Numerous statistical methods have been developed to detect clusters, including methods for point and aggregated data.[49] 'Global' tests evaluate the entire area for any evidence of clustering without pinpointing specific clusters, whilst 'local' (or 'cluster detection') tests identify the positions of specific clusters.

Cuzick and Edwards' k-nearest neighbour test, for example, is a global method for assessing clustering in case-control point data.[53] It involves counting the number of nearest neighbours of cases that are also cases, and comparing this to the number that would be expected under the null hypothesis that cases and controls were randomly distributed. Ripley's k-function is another global method for assessing clustering in point data which, unlike nearest-neighbour methods, explores spatial patterns across a range of spatial scales.[54,55] It is defined as the expected number of cases within the given range of distances from an arbitrary case location, and is compared with simulated distributions that are completely spatially random. K-function analysis is used to investigate spatial clustering of a large tuberculosis outbreak in London in Chapter 5. Moran's I is a measure of global clustering often used for area data.[56] Similar to Pearson's correlation coefficient, it measures the correlation between a variable (such as the rate of disease) in areas and the spatial distance or 'lag' between them. The resulting statistic ranges from +1 (spatial clustering of like values) to -1 (spatial dispersion of like values), where 0 indicates spatial randomness.[49]

Kulldorff's spatial scan statistic is a commonly used method used to identify local clustering, usually in point data.[57] Observed numbers of cases within windows of various sizes are compared with numbers that would be expected under a random distribution, and circular or elliptical regions of elevated risk of disease are then located. This method is used in Chapter 4 to identify areas of spatial clustering within molecular clusters of tuberculosis in London. Scan statistics and the k-nearest neighbour test have also been adapted to identify spatiotemporal clustering, testing the null hypothesis that cases geographically close to each other occur at random times.[58,59] The Knox test is another method used to identify spatiotemporal clustering.[60] It involves specification of distance and time thresholds which are considered 'near' in space and time. The number of case pairs that are within these limits is compared with an expected number based on a random distribution.[61] Clustering can also be detected in non-geographic spatial units, such as hospital beds, using the Grimson test.[62]

1.3.2.4 Modelling

Spatial relationships can also be analysed though modelling. A range of techniques can be used which, broadly, aim to create informative representations of features, events and processes in geographical space. Environmental risk mapping, for example, uses statistical methods to define relationships between spatially-referenced variables and disease risk.[49] These maps can be particularly useful in the context of emerging infections, such as the recent epidemics of ebola and zika virus disease.[63,64] By determining the ecological niches of these viruses and their vectors, areas that are at-risk of future outbreaks can be identified, and prevention measures directed appropriately.

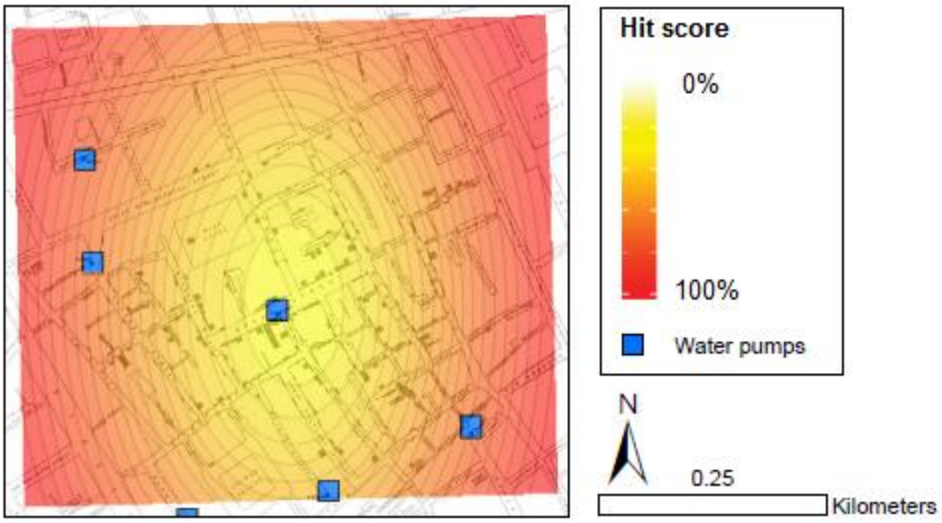
1.3.2.5 Geographic profiling

Geographic profiling is a novel statistical method which aims to identify the locations of sources, for example of infectious disease outbreaks, using the locations of linked cases. Conceptually, this works by first applying a clustering algorithm to the case locations to assign them into groups that may have been derived from the same source. Then, potential locations of sources are assessed, with locations closer to the centre of groups of cases given a higher likelihood of being a source than those further from the cases.

The unique approach of geographic profiling compared to other cluster detection methods is to frame the problem of spatial targeting in terms of creating an optimal *search strategy*. This means that, rather than solely locating clusters, the entire study area is ordered according to the likelihood that each point is a source of the observed cases. The likelihood is then expressed as a *hit score*, the percentage of the area that would have to be searched before reaching the given point, when starting the search with the most likely areas. This could theoretically be used to make evidence-based decisions about the distribution of resources for targeting interventions. In infectious disease control, this is an attractive prospect because improved targeting leads to improved efficiency and cost effectiveness of interventions.[65]

An example geographic profile generated using data from John Snow's cholera investigation is shown in Figure 1.7. The model assumes that areas in the centres of groups of cases (in this example around the Broad Street water pump) are most likely to contain a source, and lowest hit scores are therefore found in these locations. Areas that are isolated from cases (in this example around other water pumps) are less likely to contain a source and have higher hit scores.

Figure 1.7: Geographic profile of John Snow's cholera investigation in London in 1854.



Applied to investigations of tuberculosis outbreaks, geographic profiling could be used to identify areas in which the disease had been transmitted. This could potentially provide a focus for locally targeted control measures such as screening for latent or active disease and health promotion activities. The utility of the tool for tuberculosis investigations is investigated in Chapter 6 using three case studies.

1.4 THESIS RATIONALE AND AIMS

Tuberculosis is an important global health problem which requires an increasingly local approach to improve its control. In countries with low overall incidence, high rates are still found in large cities including London, England. Transmission of the infection requires that infected and susceptible individuals are in close proximity, often for protracted periods, and is thus an inherently spatial event. Spatial methods can be used to visualise, describe, detect clusters, and model infectious disease transmission. Use of these tools therefore offers an opportunity to gain insights into tuberculosis transmission dynamics and design locally-targeted interventions.

1.4.1 Aims and objectives

The overall aim of this thesis is to investigate the use of spatial methods to support local tuberculosis investigation and control. With a specific focus on London, the highest incidence area in England, results will be used to inform policy and practice. Specific objectives are to:

1. Identify the methods of spatial visualisation and analysis that have been used in previous infectious disease outbreak investigations worldwide through a systematic literature review (Chapter 2).
2. Develop an interactive mapping tool to support investigation of disease outbreaks and clusters (Chapter 3).
3. Investigate molecular and spatial clustering of tuberculosis in London (Chapter 4).
4. Describe the epidemiological and spatial characteristics of a large outbreak of isoniazid-resistant tuberculosis in England and Wales (Chapter 5).
5. Assess the utility of geographic profiling as a means of targeting tuberculosis control measures (Chapter 6).
6. Investigate geographic accessibility to tuberculosis services in London using travel time data (Chapter 7).
7. Discuss the main findings, strengths and limitations of the research; implications for future policy and practice, and opportunities for future research (Chapter 8).

Box 1.2: Summary of Chapter 1.

- Tuberculosis is a major global health problem with over 9 million new cases reported per year, and is a significant issue in England, which has a rate of 12 cases per 100,000 population.
- Improved local understanding of disease transmission and targeting of control measures is needed to progress towards elimination of the disease as a public health problem.
- Spatial methods for visualising, describing, identifying clusters and modelling infectious diseases can be used to assist epidemiological investigations.
- This thesis will explore the use of spatial methods for investigating local tuberculosis transmission, with a specific focus on transmission and control in London, the highest incidence area in England.

2 Systematic literature review of use of spatial methods for infectious disease outbreak investigations

2.1 Description of chapter contents

In this chapter, I present a systematic literature review of the use of spatial methods in published reports of infectious disease outbreak investigations. To identify relevant papers, I design and conduct a systematic search of electronic databases and review titles, abstracts and full texts against a set of inclusion criteria. I estimate the proportion of outbreak investigation reports that use spatial tools and extract key details of the included articles relating to the nature of the outbreak investigation and the spatial methods used. The ways in which spatial methods contributed to the outcomes of the investigation are described. I use the results to identify the types of investigations in which spatial methods have been most useful, and to highlight situations in which they have not yet been used to their full potential. Advantages and limitations of spatial methods are identified.

2.2 Study rationale and introduction

As outlined in Chapter 1, the importance of considering spatial locations of individuals affected in infectious disease outbreaks has long been acknowledged. Whilst a wide range of tools are available to present and analyse these locations, the extent to which they are implemented and the utility of the results they provide has not been assessed in detail. A systematic review of published reports of infectious disease outbreak investigations that have used spatial methods provides a means of establishing this, and of identifying potential areas for future development.

2.3 Aims and objectives

Aim: To examine the methods of spatial visualisation and analysis that have been used in previous infectious disease outbreak investigations worldwide through a systematic literature review.

The objectives of this study were to:

1. Design and perform a systematic search to identify published reports of infectious disease outbreak investigations that included spatial methods.
2. Estimate the proportion of published outbreak investigations that used spatial methods.
3. Extract descriptive, methodological and outcomes data relating to spatial methods from the reports that met the inclusion criteria.
4. Summarise reports according to the location of the outbreak investigated, date of publication, type of infection, and context or suspected source.
5. Classify the spatial methods used into classes of visualisation, description, cluster analysis, and modelling analyses.
6. Identify advantages and limitations of spatial methods.

2.4 METHODS

2.4.1 Search strategy

To minimise the risk of bias, I employed a broad search strategy of multiple electronic databases with few restrictions. I conducted searches of Embase (including articles from 1980 onwards), Medline (1946 onwards) and Web of Science (1900 onwards). These databases were selected in order to include outbreak investigations of both human and animal infections.

Search terms included keywords relating to outbreaks (“disease outbreak”, “outbreak”, “epidemic”) combined with those for spatial analysis (“spatial”, “cluster”, “geographic information systems”, “GIS”, “mapping”). MeSH terms used were “Geography, Medical” OR “Geographic Information Systems” OR “Spatial Analysis” AND “Disease Outbreaks”. I ran the search on 28 November 2013 and also included additional relevant articles identified from bibliographies of key studies.

2.4.2 Inclusion and exclusion criteria

After deduplication, I reviewed titles and abstracts to identify articles that met the inclusion criteria: Articles had to relate to an infectious disease; had to describe investigation of an outbreak (as defined above), and to involve application of spatial analysis or mapping. For the purposes of this review, I defined an outbreak as the occurrence of a series of cases of disease in excess of the number expected in a given time and place. I included only outbreaks with local or regional impacts, and excluded large national or multinational-scale studies of epidemics or pandemics, such as pandemic influenza.

Studies describing retrospective analyses of outbreaks that used spatial methods which could theoretically be applied in real-time investigations were included. No exclusions were made on the basis of language or outbreak location. Abstracts that did not include clear information on the inclusion criteria were brought forward for full text review. Full texts of articles were assessed with the same inclusion criteria to determine the final list of studies to be included.

2.4.3 Estimation of proportion of outbreak investigations using spatial methods

To obtain a crude estimate of the overall number of published reports of outbreak investigations, regardless of whether they used spatial methods, I repeated the search of the same databases using only the outbreak investigation terms. I then simulated the deduplication and screening process by assuming that the same proportion of studies would be excluded at each step as in the original search. From this total, I calculated the approximate proportion of published reports of outbreak investigations that used spatial methods.

2.4.4 Data extraction and synthesis

I reviewed each included study and extracted information about the spatial methods and outcomes using a bespoke data extraction spreadsheet. Descriptive details obtained were the location of the outbreak, date of publication, type of infection, context or suspected source, and whether the study was prospective or retrospective. Methodological details were the type of spatial methods used and the tools employed. Outcomes were results of the investigations that related specifically to the use of spatial methods and any comments on their advantages or limitations.

I summarised the reports according to location of outbreak investigated, date of publication, type of infection, and context or suspected source. Four broad classes were also used to categorise the spatial methods used: visualisation, description, cluster analysis and modelling.

To demonstrate the utility of spatial methods during outbreak investigations, I identified the stage(s) of the investigation to which they were applied. Outbreak investigations can be delineated into steps in various ways, and for the purpose of this review I adapted steps from the ECDC *Field Epidemiology Manual* (Box 2.1).[66]

Box 2.1: Steps in an outbreak investigation.

Adapted from: European Centre for Disease Control. Field Epidemiology Manual.[66]

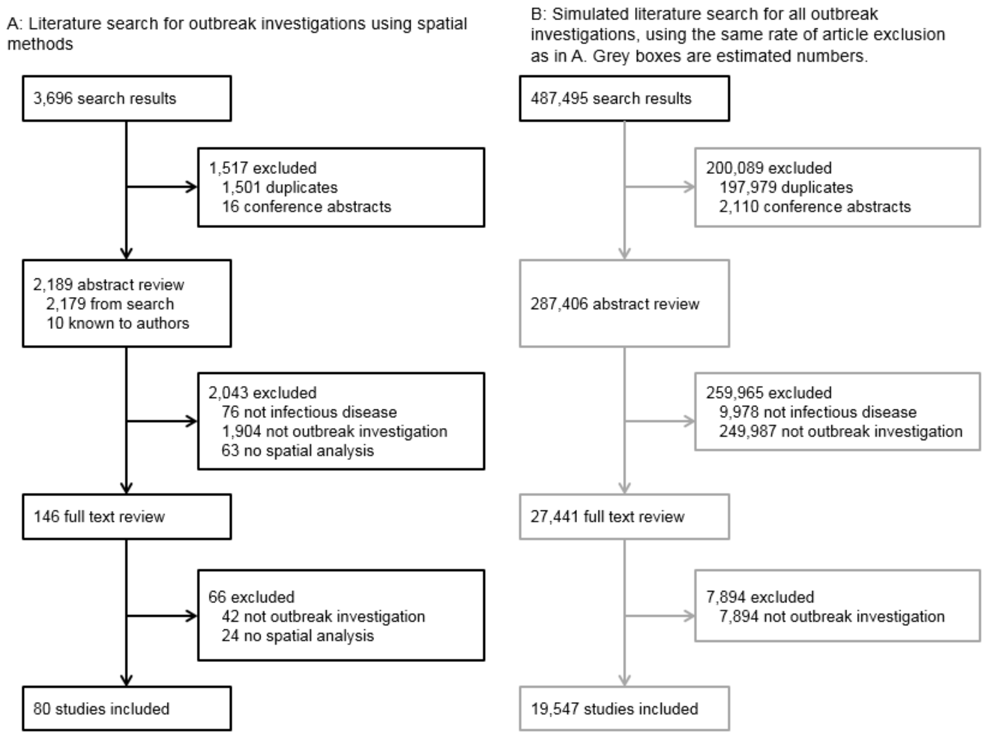
1. Establishing existence of an outbreak.
2. Confirming diagnosis.
3. Defining and identifying outbreak cases.
4. Describing cases and developing hypotheses.
5. Evaluating hypotheses and drawing conclusions.
6. Comparing with established facts.
7. Executing prevention measures.
8. Communicating findings.

2.5 RESULTS

2.5.1 Article screening and estimation of proportion using spatial methods

After excluding duplicates, the search yielded a total of 2,189 articles for abstract screening. Of these, 146 were selected for full text review and 80 were included in the analysis. Reasons for article exclusion are summarised in Figure 2.1A. Repeating the search without any terms specific to spatial analysis identified 487,495 articles. Assuming the same rate of article exclusion at each step in the review process, the total number of published articles relating to outbreak investigations of infectious diseases can therefore be estimated as approximately 20,000 (Figure 2.1B). The overall proportion of published outbreak investigation reports that explicitly described spatial methods was therefore around 0.4%. Key details of all 80 included articles[18,67-145] are shown in Appendix 10.1.

Figure 2.1: Study selection, systematic literature review on spatial methods in infectious disease outbreak investigations.


A: Literature search for outbreak investigations using spatial methods
3,696 search results
1,517 excluded
1,501 duplicates
16 conference abstracts
2,189 abstract review
2,179 from search
10 known to authors
2,043 excluded
76 not infectious disease
1,904 not outbreak investigation
63 no spatial analysis
146 full text review
66 excluded
42 not outbreak investigation
24 no spatial analysis
80 studies included
B: Simulated literature search for all outbreak investigations, using the same rate of article exclusion as in A. Grey boxes are estimated numbers.
487,495 search results
200,089 excluded
197,979 duplicates
2,110 conference abstracts
287,406 abstract review
259,965 excluded
9,978 not infectious disease
249,987 not outbreak investigation
27,441 full text review
7,894 excluded
7,894 not outbreak investigation
19,547 studies included

2.5.2 Characteristics of studies included

Publication of outbreak investigations using spatial methods has increased markedly since 2000, with over half (42) of the studies published since 2008 (Figure 2.2). Most articles (66, 83%) concerned infections in human populations, of which the most frequently investigated infections were Legionnaires' disease (12), cholera (7) and influenza (7) (Table 2.1). Correspondingly, the most common transmission contexts for human infections were water/ sanitation (20), followed by environmental (14) and community (10) (Table 2.2).

Figure 2.2: Reports of outbreak investigations using spatial methods by year.

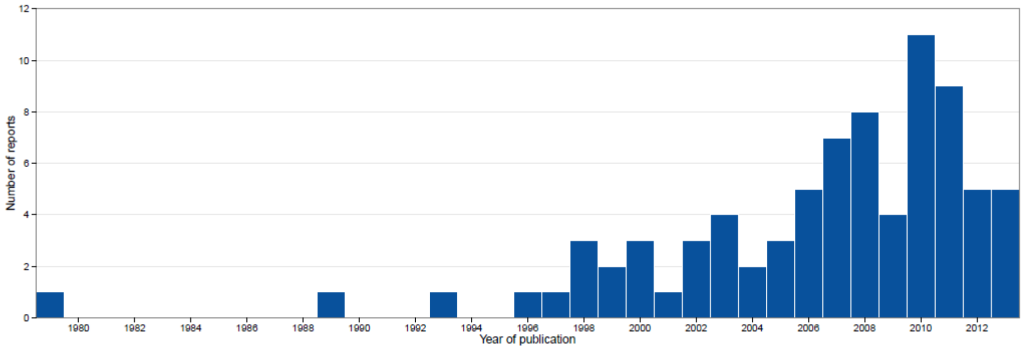


Table 2.1: Infectious diseases investigated by category.

Infection	N*	References
Respiratory		
Legionnaires' Disease	12	18,67,80,81,92,95-98,112,117,143
Influenza	7	87,90,101,114,120,123,144
SARS	3	100,104,145
Acute respiratory disease	1	115
Intestinal		
Cholera	7	75,88,102,105,125,131,133
Cryptosporidiosis	2	111,141
Diarrhoea†	2	76,130
Salmonellosis	2	78,99
Shigellosis	2	107,128
Campylobacteriosis	1	93
Giardiasis	1	116
Necrotizing enterocolitis	1	137
Viral haemorrhagic fever		
Dengue	5	70,83,110,113,126
Ebola	1	108
Porcine high fever disease	1	103
West Nile Virus	1	84
Viral skin infections		
Measles	3	91,118,138
Foot and mouth disease	2	124,129
Varicella	1	135
Variola minor	1	71
Protozoal		
Toxoplasmosis	2	79,85
Leishmaniasis	1	140
Malaria	1	102
Trypanosomiasis	1	89
Rickettsioses		
Q-fever	5	94,106,132,139,142
Bacterial zoonotic		
Anthrax	3	86,109,119
Leptospirosis	1	73
Mycoses		
Blastomycosis	3	77,122,127
Viral CNS		
Rabies	2	121,136
Viral hepatitis		
Hepatitis A	1	134
Hepatitis E	1	72
Helminthiases		
Schistosomiasis	1	74

Other bacterial			
	Meningococcal meningitis	1	82
STI			
	Syphilis	1	68
Tuberculosis		1	69

*Total 81 because one study reported two investigations.
†Infectious agent not known or not specified.
CNS: central nervous system; SARS: severe acute respiratory syndrome.

Table 2.2: Contexts of outbreak investigations of human and animal diseases.

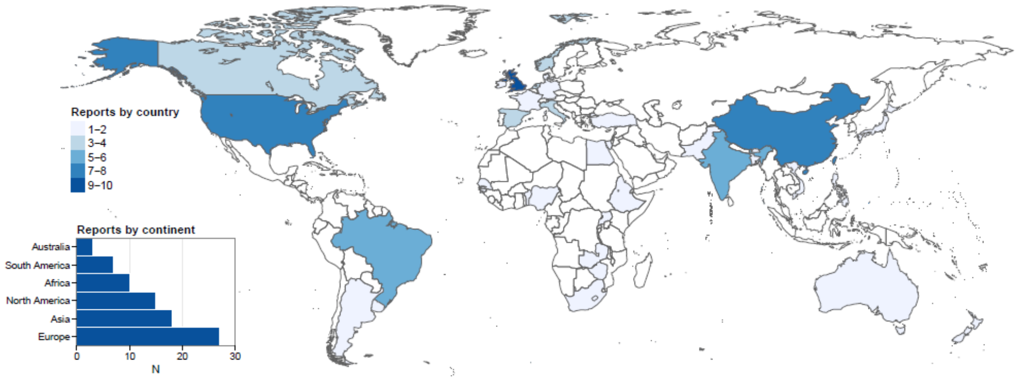
Context	Human†		Animal	
	N	References	N	References
Water/ sanitation	20	72,73,75,76,79,85,88,93,102,105,107,111,116,125,128,130,131,133,134,141	0	
Environmental	14	18,67,80,81,92,95-98,112,117,122,127,143	1	77
Community	10	69,71,82,87,91,100,101,104,118,138	2	121,136
Vector-borne	10	70,74,83,84,89,102,110,113,126,140	0	
Farm/ breeding facility	5	94,106,132,139,142	12	86,90,103,108,109,114,115,119,120,123,124,129
Healthcare-associated	5	99,135,137,144,145	0	
Food	1	78	0	
STI	1	68	0	
Total*	**66**		**15**	

* Total 81 because one article reported two investigations.
†Includes outbreaks affecting humans that had animal origin.

Healthcare-associated infections were reported in five of the articles whilst food-borne and sexually transmitted infections were reported once apiece. Veterinary infections were almost exclusively linked to farms or other breeding facilities (12) and influenza was the most frequently investigated infection affecting animals (4). Prospective outbreak investigations comprised around half (39, 49%) of the articles included, with the remainder describing retrospective analyses of outbreak data.

Figure 2.3 displays the outbreaks by country, with the most reports in the United Kingdom (10) or United States (8); and by continent, with a third of reports in Europe (27) and fewer in Africa (10 total).

Figure 2.3: Locations of outbreak investigations using spatial methods by country and continent.



2.5.3 Spatial methods

Spatial methods used are listed and classified according to type in Table 2.3.

Table 2.3: Spatial methods used in outbreak investigations.

Method	N	References
Visualisation, 80 studies (39 prospective, 41 retrospective)		
Dot map	68	18,67-85,87-91,93-98,100-103,105,107,109-119,121-127,129-134,136,138-143
Thematic map	25	70,73,74,77,79,80,86-88,91,93,99,106,109,111,116,119,120,123,124,130,132,138,139,143
Rate map	14	68,84,85,88,92,100,104-106,112,125,128,131,139
Smoothed incidence map	13	67,70,71,86,90,94,100,101,103,105,115,126,139
Case movement map	7	69,81,95,97,98,112,117
Schematic map	6	99,108,135,137,144,145
Standard deviation ellipse	4	100,101,120,136
Origin-destination plot	1	100
Velocity vector map	1	86
Voronoi diagram	1	131
Spatial description, 47 studies (28 prospective, 19 retrospective)		
Spatial case definition	32	18,72,78,80,81,86,89,92-99,111,112,115-117,122,123,127-130,132,133,137,142,144,145
Source proximity	16	18,70,77,78,80,89,92,94,98,117,124,129,130,132,139,144
Spatial case finding	8	70,72,76,105,106,112,113,130

Spatial average	5	102,120,122,132,136
Case-case distance	3	82,123,129
Risk factor proximity	2	73,123
Spatial social network analysis	1	90
Cluster, 24 studies (8 prospective, 16 retrospective)		
Kulldorff's spatial/ spatiotemporal scan statistic	13	86,89,90,103,105,113-115,120,127,130,134,143
Cuzick Edwards k-nearest neighbour test/ Jacquez's k-nearest neighbours for space time interaction	7	83,86,103,107,115,120,133
Knox test	5	103,110,115,120,126
K-function/ space-time K-function	5	67,86,103,105,110
Moran's I	4	100,101,131,143
Nearest neighbour analysis	3	100,120,131
Getis Ord G*i*(*d*) statistic	2	84,101
Barton & David's test	1	110
Grimson test	1	137
Oden's Ipop	1	86
Mantel's test	1	120
Spatial modelling, 13 studies (3 prospective, 10 retrospective)		
Air dispersion modelling	7	112,117,129,135,142,144,145
Environmental risk prediction model	2	87,126
Kriging	2	70,90
Empirical Beyes smoothing	1	88
Geographic profiling	1	102

All articles presented or referred to at least one method of visualising case distributions to describe outbreaks in space. Plotting cases as dots on a map is the simplest form of visualisation and was used in the majority (68, 85%) of studies. Dot maps were either presented using case locations only, or were enhanced with further information such as their vaccination status,[138] migratory status,[74] or date of disease onset.[125] Thematic maps provide context to case locations by displaying the spatial distributions of other variables. Such maps were used in 25 studies and variables plotted included socioeconomic status,[73] soil type,[119] and road density.[123]

Maps of disease rates were used in 14 studies, with data usually aggregated according to administrative boundaries. Smoothed incidence maps were used in 13 studies. Other methods for visualising outbreaks that were used in fewer studies included standard deviation ellipses and velocity vector maps. Both of these methods use the locations of cases to describe the direction of spread of outbreaks.

Cluster analyses were used in 24 studies (30%), and spatial scan statistics were the most frequently used (13 studies). k-nearest neighbour tests, K-function analyses, and the Knox test were also used relatively frequently (7, 5 and 5 studies respectively). Modelling approaches were used in 13 studies, including seven which used air dispersion models to identify areas that may have been exposed to air from suspected contaminated environmental sources.

A range of other spatial methods based on geographic attributes of cases were also identified. These included methods for defining (31 studies) and identifying (8 studies) cases, summarising the average locations of cases (5 studies), and assessing proximity to potential sources (16 studies).

Analytic methods were used less frequently in prospective than retrospective articles: Cluster methods were used in 16 (39%) retrospective compared with eight (21%) prospective studies, and modelling in 10 (24%) and 3 (8%) of retrospective and prospective analyses respectively.

The most frequently cited GIS software was ArcGIS/ ArcView, used in 30 studies, with MapInfo the other commonly used programme (7 studies). Various other programmes including R, ClusterSeer, GeoDa and SaTScan were used for specific analyses.

2.5.4 Application of spatial methods to outbreak investigations

Applications of spatial methods to different stages during outbreak investigations are described below.

2.5.4.1 1. Establishing existence of an outbreak

Few studies (4) used spatial methods to assist with establishing the existence of an outbreak. Methods that were used aimed to identify unusual patterns of cases, either visually or through formal statistical tests of clustering.

For example, Affolabi and colleagues described complementary use of molecular and geographic methods to identify an outbreak of tuberculosis in Benin.[69]

Amongst a series of 194 *M. tuberculosis* isolates, 17 belonged to the Beijing genotype and exhibited an identical 12-loci subtype. Mapping of patients' residences, workplaces and movements revealed a corresponding spatial cluster, confirming that the cases were likely to be linked. In another study, Roy and colleagues plotted the locations of cases of blastomycosis in Wisconsin after noting an increase in the number of reports.[127] They visually identified clustering within five neighbourhoods and used the spatiotemporal scan statistic to confirm that this was statistically significant.

2.5.4.2 2. Confirming diagnosis

Although knowledge of the endemicity of diseases in the geographic regions in which outbreaks arise is useful in developing plausible preliminary diagnostic hypotheses, spatial methods alone are not able to confirm a diagnosis and were therefore not used for this purpose in any of the studies.

2.5.4.3 3. Defining and identifying outbreak cases

Geographic boundaries in which outbreak cases were defined were stated explicitly in over a third (31) of the studies. For instance, Keramarou and colleagues' investigation of an outbreak of Legionnaires' disease included only cases that lived or worked in the outbreak area, defined as a 12 km corridor on either side of a major road.[97]

Spatial methods were also used to assist with active case finding in eight studies. Bali and colleagues describe a search for cases of hepatitis E prompted by identification of three cases in a small town in northern India.[72] A house-to-house survey in this region identified 3,170 cases of jaundice with an attack rate of 5.2%.

2.5.4.4 4. Describing outbreak cases and developing hypotheses

Use of dot mapping to support an outbreak in real time is described by Fitzpatrick and colleagues, who investigated a rise in measles cases in Dublin, Ireland.[91] By continuously updating their maps throughout the outbreak, they were able to identify clustering of cases as soon as it developed, which ultimately assisted with targeting of control interventions.

Simple maps were also used to develop hypotheses about the origins of outbreaks. For example, Kistemann and colleagues plotted cases by date of onset in an investigation of a nosocomial *Salmonella* outbreak.[99] Their schematic map revealed the central kitchen in the hospital as the only functional relationship linking the cases, which they therefore hypothesised to be the source of the infection.

Sasaki and colleagues created a Voronoi diagram to demarcate their study area using locations of water taps.[131] Plotting incidence rates in different areas defined by these water tap boundaries helped to visualise clear spatial clustering of cholera cases associated with poor water and sanitation facilities. Smoothed incidence maps were used in an investigation by Norström and colleagues into acute respiratory disease in Norwegian cattle herds. They used smoothing based on kernel density estimation to describe the progression of the outbreak, which was shown to spread locally before jumping to new areas.[115]

A common method to develop hypotheses about sources of infections was to construct concentric circles of varying radii around potential sources and compare the attack rates in each. Nygard and colleagues used this technique in an investigation of Legionnaires' disease in Norway.[117] They calculated attack rates in five rings of increasing distance around eight potential sources and observed a trend of decreasing rate ratios with increasing distance from an air scrubber. Other metrics used to describe cases included calculating their average location and proximity to risk factors.

Possible airborne spread of Q fever from farms near Cheltenham, UK was investigated by Wallensten and colleagues using the Numerical Atmospheric-dispersion Modelling Environment model.[142] Plotting the modelled distribution showed that air from each of the suspected farms may have exposed the town. Le Comber and colleagues used the geographic profiling method to identify most likely locations of mosquito breeding sites using residential locations of a series of cases of malaria in Cairo, Egypt.[102]

2.5.4.5 5. Evaluating hypotheses and drawing conclusions

More than half of the studies (42) used statistical tests, such as cluster and regression analyses, to conduct formal evaluations of hypotheses arising from observations of case distributions. Fevre and colleagues, for example, assessed clustering of cases of trypanosomiasis under the hypothesis that a cattle market was the source of the outbreak.[89] A significant cluster encompassing the location of the market was detected using the spatio-temporal scan statistic, supporting this theory.

In an investigation on a military installation in North Carolina, McKee and colleagues used the k-nearest neighbour method to identify significant spatio-temporal clustering of shigellosis.[107] They used dot maps to locate the area with intense transmission, and targeted it with educational interventions to bring the outbreak under control.

Combinations of multiple tests for clustering were used in some studies, such as Norström and colleagues' investigation of acute respiratory disease in Norwegian cattle herds.[115] They combined the Knox test,[60] a global test for space-time clustering, with the k-nearest neighbour test and space-time scan statistic. These tests allowed them, respectively, to define the smallest distance and time frame in which the events had been clustered; to determine whether cases tended to be close to other cases, and to locate the most significant clusters. All methods indicated presence of space-time clustering, adding weight to the conclusion that a common contagious source was responsible for the outbreak.

Regression analysis was used in several studies to test the hypothesis that the risk of infection decreased with increasing distance from a suspected source. Hackert and colleagues, for example, used linear regression of log-transformed attack rates to assess a cluster of human cases of Q fever in the Netherlands.[94] Incidence increased by a statistically significant exposure-response gradient with proximity to a dairy goat farm, which they concluded was likely to be the primary and sole source of the infection.

2.5.4.6 6. Comparing results with established facts

Results from spatial analyses provided updates to knowledge about the dynamics of infectious agents, such as their minimum infective dose and mode of transmission. For example, in an outbreak of Legionnaires' disease in Christchurch, New Zealand, cases were identified at a distance of 12 km from the implicated cooling tower.[143] White and colleagues therefore proposed updates to WHO guidelines which at the time placed the area at risk from such sources at only 3.2 km.

Wong and colleagues used a computational fluid dynamics model to study the spread of an influenza outbreak in a hospital setting.[144] Concentrations of hypothetical virus-laden particles from modelled air distributions correlated closely with locations of infected patients. This suggested a possible role for aerosol transmission of influenza, which is predominantly associated with transmission by droplets and direct contact.

2.5.4.7 7. Executing prevention measures

Spatially targeted interventions to control the outbreak, or prevent future cases, were described in many studies. Measures that aimed to control outbreaks included cleaning implicated cooling towers;[98] issuing water boiling orders to areas served by contaminated supplies;[93,116] vaccination catch-up campaigns;[91] removal of breeding sites for mosquito larvae,[113] and targeted information campaigns.[107] For example, Acheson and colleagues placed post code-targeted information on social networks during an outbreak of heterosexually acquired syphilis in Teesside, UK.[68]

Attempts to prevent future outbreaks included improvement of infrastructure;[88,130] change of policy,[95,99,117] and generation of risk maps.[126] Luquero and colleagues, for instance, used results of their analysis to recommend specific regions in which to focus preparedness activities to avoid future cholera outbreaks in Guinea-Bissau.[105]

2.5.4.8 8. Communicating findings

All studies in this review had, by definition, used their spatial analyses in communication of findings through reports in peer-reviewed publications. Several studies also highlighted the usefulness of maps in reports or presentations to communicate results to health officials, policy makers and the public. Sarkar and colleagues, for example, presented dot maps of cases of acute diarrhoeal disease in a village in southern India to the local community and health authorities.[130] Their maps visualised the proximity of cases to a contaminated water supply, and the presentation resulted in release of funds to improve sanitation in the area.

2.6 DISCUSSION

2.6.1 Summary of findings

A range of spatial tools are available for outbreak investigations and can provide useful insights that lead to public health actions, but these methods have generally been under-used. Although the simple dot map was the most commonly used method, a wide range of techniques were applied, including more sophisticated data visualisations and analytic tools. Outbreak reports using spatial tools constitute less than half a percent of the overall total number of published reports of outbreak investigations, and there are discrepancies in the extent to which they are used for similar problems in different contexts.

Across the range of studies, there were examples of spatial tools being usefully applied throughout the course of an outbreak investigation; from initial confirmation of the outbreak to describing and analysing cases and communicating findings. Spatial techniques often provided valuable insights that supplemented traditional epidemiological analyses of person and time and led to public health actions.

2.6.2 Interpretation of results – use of spatial methods

Outbreak investigations of infectious diseases occurring in any context were included in this study. Thus, it extended the scope of two previous reviews that focused, respectively, on use of spatial methods in outbreaks of Legionnaires' disease,[51] and on spatiotemporal methods to investigate transmission of infections in hospital settings.[146] In doing so, it has highlighted imbalances in application of spatial methods in investigations depending on the country in which the outbreak has occurred, the context, and the infectious disease being investigated.

For example, ten studies described investigations of outbreaks in the UK. This was the largest number of any country and the same number as in the whole of Africa, which clearly does not correlate with the distribution of the global burden of infectious diseases. It was also notable that although there was a large number of reports from Europe compared with other parts of the world, reports derived from only ten different counties in the continent. These were predominantly in Western Europe, with one report from Turkey the only investigation from eastern Europe. This highlights the importance of sharing expertise internationally to expand the use of these tools in under-represented areas.

Spatial tools were used most often for outbreaks that were thought to have an environmental point source, such as Legionnaires' disease and those associated with water-borne illness. In these contexts, spatial methods have clear applications to the stages of the investigation concerned with analysing results and drawing conclusions about the likely source of the infection. However, this study has demonstrated that these tools can also have useful applications in other stages of the investigation. There is therefore scope for expansion of the use of spatial methods even in these contexts, for example in initial confirmation of the outbreak, describing cases, and communication of findings.

Outbreak investigations in any context follow similar steps, and could therefore potentially incorporate spatial methods. It was notable, however, that only one study reported an outbreak of foodborne illness. Annual summary statistics from 2013 report a total of 5,196 food- and waterborne outbreaks in the European Union[147] and 831 reports of foodborne outbreaks in 2012 in the USA.[148] Although only a small proportion of these are likely to have been published in academic journals, this still indicates a substantial shortfall in use of spatial methods in this context.

2.6.3 Interpretation of results – limitations of spatial methods

There are several limitations of spatial methods, and barriers to their use, which may account for the unequal and under-use of these tools.

First, reliable spatial analyses can only be conducted with accurate location data. This can be a particular challenge in developing countries in which good quality maps of residential areas are often not available.[149] Several investigations of outbreaks in such settings conducted field surveys and used Global Positioning Systems (GPS) to accurately record patient residence or risk factor locations.[83,89,113,130,131,134] However, this is a time and cost intensive approach and will not always be feasible. In settings in which good quality maps of residential areas are available, quality of location data is also not assured. Errors can arise from incomplete or mistranscribed addresses, out of date GIS databases, or incomplete information on potential source locations. During outbreaks of Legionnaires' disease, for example, some investigators had to conduct visual searches or make public enquiries to ascertain the locations of aerosol-producing devices because there was no central registry.[80,81,92,96-98,117]

Simplification of case locations to static points, usually residential locations, also impacts the utility of location data. In reality, individuals can become exposed to infectious agents at any place where they spend time, and, similarly, traditional census population denominators that record night time populations are not necessarily reflective of population distributions during the day.[150,151] Although a number of studies made attempts to record case movements,[69,81,95,97,98,112,117] none accounted for diurnal fluctuations in populations. Ideally, this spatial uncertainty should be accounted for in the data collection, analysis and visualisation stages to improve reliability of estimates of spatial risk, and new analytic methods may be required to achieve this.

Second, even if reliable location data are available, presentation of information on maps can be open to misinterpretation. Dot maps, for instance, were used widely but do not take into account the geographic distribution of the underlying population and can therefore mask important trends. Similarly, patterns in aggregated data are subject to the ecological fallacy, that the characteristics of a group can be used to deduce the characteristics of individuals in the group. They are also sensitive to changes in the boundaries into which they are grouped, a phenomenon known as the modifiable aerial unit problem.[49] Presentation of data on maps fails to highlight these limitations, and relatively few prospective investigations used statistical methods to formally confirm observations identified from visual displays of data.

Third, researchers may be deterred from using spatial analytic methods because they involve selection of parameter values, often with an element of subjectivity. Methods that display or identify clustering require specification of the degree to which distant points may be considered part of the same neighbourhood. For the spatial scan statistic, the user must define the maximum spatial extent of clusters in terms of the percentage of the population that can be included; in k-nearest neighbour analysis, the number of neighbours included must be specified, and equivalent parameters must be selected for other spatial cluster and modelling analyses.[49] Altering these parameters can have a profound influence on the results, and a trial and error approach is often required to arrive at an appropriate value. This can raise issues of multiple hypothesis testing, although some methods, including the spatial scan statistic and Tango's maximised excess events test,[152] are able to adjust for this whilst evaluating clustering at multiple scales. Results of spatial analyses can also suffer from lack of specificity. For example, in several studies of Legionnaires' disease, spatial methods identified areas most likely to be the source of the infection, but could not discriminate between potential sources that were close together.[92,96,98]

Another barrier to the effective use of spatial methods that is often cited is the expense of specialised GIS software and the need for trained personnel to operate it. Although some GIS programmes are available free of charge, the most commonly used was a commercial package, ArcGIS.

2.6.4 Implications for policy and practice

The results of this study point to a number of recommendations for improved practice and opportunities for further development of spatial methods. Given the potential utility of existing tools demonstrated here, under-use of these methods has doubtless resulted in missed opportunities for more effective real-time outbreak investigations. Public health officials must be supported to address this issue, and a useful first step would be development of protocols describing the application of appropriate analyses. Provision of training, for example through short courses, and interdisciplinary working with specialists in geographic analysis, would also be beneficial to improve the skills base of the workforce.

The majority of studies identified in this review that used analytic methods described retrospective analysis of data collected during outbreaks. These reports demonstrated the potential utility of analytic methods, but would be of greater public health benefit when used in real-time. Assembly of GIS databases in advance is essential to allow spatial analyses during prospective outbreak investigations. Improving data accessibility will save time during investigations, improve accuracy of analyses and prevent duplication of effort.

Reports of analyses using spatial methods would also benefit from some degree of standardisation. For example, reporting of the sources and level of precision of spatial data would enable more accurate interpretation of the results by researchers not familiar with the study site. This could be achieved, for example, through extension of the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement with items specific to spatial data.[153]

2.6.5 Study strengths

In this review, I used a systematic approach with a transparent search strategy to identify published reports of outbreak investigations that incorporated spatial methods. I used a robust methodology, following the relevant sections of the PRISMA guidelines (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) to ensure that the results were as accurate and potentially reproducible as possible.[154]

The review had a broad set of inclusion criteria, encompassing reports of outbreaks of any infectious disease in any context. It included infections of both humans and animals, and did not have any restrictions on the basis of language or publication date. This breadth allowed me to identify areas in which the use of spatial methods is most prevalent, and those in which it could be improved.

A further strength of this review is that it has a practical focus: I categorised the spatial methods used in each study according to the stage of the outbreak investigation during which they were applied, assisting implementation of similar approaches in future investigations.

2.6.6 Study limitations

The primary limitation of this study was the challenge of designing the database search strategy. Although I employed a broad search which identified a large number of abstracts for screening, the number of studies identified here will inevitably be an underestimate of the outbreak investigations that used spatial methods. My search will not have captured studies that used spatial methods but did not refer to them explicitly in the title, abstract, subject headings or MeSH terms. Restricting the search to articles published in academic journals also excluded reports in the 'grey' (unpublished) literature. Inclusion of such reports would increase the number of investigations using spatial methods, but would be unlikely to reveal novel approaches or tools not identified here.

There was also a possible publication bias in this study. Spatial analyses may have been more likely to be presented in published reports if they were found to be useful. Concerns of breaching patient confidentiality could have further limited the number of studies that published maps. Nevertheless, the proportion of studies using spatial methods was very small, and even if the estimate is an order of magnitude too low, it would still represent less than 5% of the estimated total number of investigations published.

Articles published since the database search was run at the end of 2013 are also not included in this study. This study showed that the number reports using spatial methods has increased in recent years, probably due to increased availability of GIS software. This trend is likely to have continued, and recent publications will focus on current public health issues, for example the recent Ebola outbreak in West Africa.

Another limitation intrinsic to the nature of the study question was that I was unable to conduct a formal quality assessment of the included studies. Quality assessment tools, such as the STROBE statement, are typically used to aid interpretation of results of studies by assessing the likelihood of bias in estimates of effect.[153] In this review, I did not have a quantifiable effect estimate that could be assessed using these methods. However, I classified methods used into categories and identified the ways in which they contributed to outcomes of investigations. This provided an indication of the sophistication of the analysis undertaken and their utility.

2.6.7 Future directions

As well as increased adoption of existing methods, there is scope for development of new tools for analysis and visualisation of spatial data. A move towards web-based applications with user-friendly interfaces would be a natural progression, provided that these platforms included adequate training materials and data governance infrastructure.

It is notable that spatial scan statistics were the most frequently adopted analytic methods. Scan statistics can be implemented with relatively little training through SaTScan,[155] a programme free to download from the internet. This suggests a possible model for wider adoption of other more advanced techniques and could potentially make spatial analyses more accessible to non-experts.

The quantity and detail of geo-located data available to researchers is also increasing. GPS-enabled mobile devices and applications for self-reported or crowd-sourced information (for example *sickweather*, based on reports on social networks[156]), have the potential to provide near real-time data including information on individuals' movements. Development of new analytic techniques will be needed to ensure that these data are effectively exploited and potential benefits are met. In the context of outbreak investigations, possible applications include contact tracing and improved estimation of exposure to environmental risk factors.

Box 2.2: Summary of Chapter 2.

- A systematic review of published reports of infectious disease outbreak investigations using spatial methods was conducted.
- A total of 80 studies using spatial methods were identified, less than half a percent of the estimated overall number of published reports of outbreak investigations.
- A range of techniques including sophisticated means of visualisation, cluster and analytic tools have been used; but the simple dot map was the most common.
- There is a disparity in the use of spatial tools according to the country in which the outbreak has occurred, the context, and the infectious disease being investigated.
- Spatial tools have been usefully applied throughout the course of an outbreak investigation, from initial confirmation of the outbreak to describing and analysing cases and communicating findings.
- Spatial techniques often provided valuable insights that supplemented traditional epidemiological analyses of person and time and led to public health actions.
- There is scope for much wider implementation of existing methods and development of new tools.

3 DEVELOPMENT OF AN OPEN-SOURCE TOOL FOR VISUALISING DISEASE POINT LOCATIONS

3.1 DESCRIPTION OF CHAPTER CONTENTS

In this chapter I develop an interactive tool for visualising disease point locations. The tool is written using Shiny, a web application framework for the statistical software, R. It allows plotting of geographically referenced point locations on an interactive map displayed in a web browser window. Its features include colour coding of cases according to categorical variables; filtering of cases plotted by combinations of variables; and additional displays of data as a summary table and epidemic curve. I demonstrate how this tool can be used in the context of investigations of clusters of tuberculosis linked through molecular strain typing. It can be used to improve understanding of tuberculosis transmission by identifying groups of cases which are close to each other in space and also share epidemiological characteristics. The application is not restricted to investigation of tuberculosis clusters and could be used to plot locations with associated characteristics using any geographically referenced data. I evaluate the application through a demonstration to potential users and an online survey.

3.2 STUDY RATIONALE AND INTRODUCTION

3.2.1 Spatial data visualisation

In Chapter 2, I conducted a systematic literature review which described the use of spatial data in infectious disease outbreak investigations. This review demonstrated that careful consideration of the spatial locations of cases of disease on a map can prompt infectious disease outbreak investigations; highlight important relationships between cases; generate hypotheses about transmission, and guide control measures. Visualisation of data on maps also provides an easily-understood means of presenting information in context that is not as readily derived from tables of data or written reports, and can be a powerful tool for advocacy. Provided that accurate spatial data are available, these tools can be used for a range of different disease areas and in high- and low-income settings.

Dot maps were the most widely used methods of visualising spatial data. They display the locations of cases and can be colour coded to convey additional information such as categorical variables regarding patient demographics and risk factors. Production of dot maps for different time periods can describe the progression of the cluster in space and time; and including data on contextual locations such as potential transmission venues can aid hypothesis generation.

In spite of these broad applications, I found that spatial data visualisation tools were used in only approximately half of the real-time outbreak investigations included in the review. One of the barriers to use is lack of flexibility in existing tools, i.e. the ability to make rapid changes to the numbers of cases being displayed and the characteristics highlighted. The applications also often require specialised software which can be expensive and may require trained personnel to operate. Confidentiality issues can also arise when sharing data or maps between organisations, particularly if information must be uploaded to the internet.

Another recent systematic review explored the use of visualisation tools for infectious disease epidemiology.[157] As well as describing features offered by different tools for specific types of visualisations, the authors aimed to identify the needs and preferences of public health professionals using these tools; the software architecture in which they were designed, and discussed implementation and evaluations of their use in practice.

Studies of user needs emphasised a preference for dynamic, interactive graphics for data exploration. Recent developments in visualisation technologies such as scalable vector graphics, dynamic HTML, and R Shiny have the potential to facilitate delivery of these interactive features. However, the majority of existing tools identified in the review used static visualisations. The BioSense surveillance system developed by the Centers for Disease Control was one exception to this,[158] which uses R and RStudio to embed interactive data analysis and visualisation within its web application.

Despite studies recognising the importance and potential of effective visualisation tools, the review found minimal success in widespread implementation and adoption of such tools. The authors identified similar reasons for this as those which emerged from my systematic review, including user-level barriers such as lack of adequate training, misconceptions about the use of the tools and lack of trust in new systems, and system-level barriers such as access issues and lack of organisational support.

To mitigate these barriers, the authors recommend that future projects should avoid developments in 'silos', and instead aim to build tools that are interoperable and compatible with existing data formats and standards. Further, the authors suggest using a participatory design process, which involves potential users in the development of the tools from an early stage and incorporates their feedback. Finally, they note the importance of collaboration between agencies and organisations to integrate novel tools into workflows.

Together, the results of these systematic reviews highlight a clear scope for development of new interactive tools for analysis and visualisation of spatial data, for example production of dot maps, in infectious disease epidemiology.

3.2.2 Current tools available

Tools currently available for visualisation of epidemiological data range from those designed to produce 'flat' (non-interactive) images to bespoke interactive data visualisation platforms. Analysis and visualisation of spatial data, meanwhile, has traditionally been undertaken using a GIS.

Flat images are the most commonly used form of visualisation in infectious disease epidemiology,[157] and can be produced using numerous tools. Spreadsheet packages such as Microsoft Excel and Apple Numbers offer many types of graphs generated through a graphical user interface, whilst statistical software packages such as Stata, SAS, Matlab and R enable creation of similar graphics using code. These packages allow outputs to be reproduced more easily but require a greater degree of technical knowledge.

In recent years, tools enabling users to produce interactive graphs have become increasingly prevalent. These include 'off-the-shelf' products, for example Google Charts, which are able to generate visualisations with interactive elements that can be customised to some degree. Bespoke interactive tools that offer complete control over the features displayed can also be produced, for example using the JavaScript library D3. These 'data-driven-documents' are flexible, fast and support large data sets, but need to be produced by web developers with knowledge of the JavaScript programming language. Shiny is a web application framework for the statistical software R that offers an alternative to programming using a tool such as D3.[159] It allows interactive tools to be produced using the R framework, and does not require knowledge of programming in languages such as JavaScript.

One of the most commonly used GIS in epidemiology is ArcGis/ ArcView, a commercial package with many features ranging from production of simple dot maps to sophisticated analyses using spatial statistics.[160] QGIS is an open-source alternative to ArcGIS which shares some of its features and is free to download.[161] However, both of these programs require a degree of technical expertise to operate, even if the desired outputs are relatively simple. Lightweight applications which are designed for a single function can provide an attractive alternative to broad GIS packages in some circumstances. SaTScan, a program used for performing spatial scan statistics to identify significant clustering in data,[155] and the ECDC Map Maker (EMMa), an online tool used to create maps of area-level data,[162] are examples of such programmes. Both are free to use, have simple data requirements, and avoid the need to manually process geospatial data.

3.2.3 Investigation of molecular clusters of tuberculosis

Molecular clusters of tuberculosis are groups of cases that share indistinguishable strain types and may therefore be linked through transmission. In the United Kingdom, routine molecular strain typing by MIRU-VNTR was introduced in 2010. From January 2010 to December 2013, 81% (16,602) of isolates for culture-confirmed cases were strain typed for at least 23 loci.[163] Over half (8,890) of these cases shared a strain type with at least one other case, and were therefore classified as being part of a molecular cluster. A total of 1,854 distinct molecular clusters were identified, and initial guidelines required prospective investigation of all clusters that met certain thresholds.[164]

Tuberculosis cluster investigations aim to reduce public health impacts by detecting and diagnosing previously unidentified latently infected and active tuberculosis cases.[165] An evaluation of the strain typing service in 2013 found that routine cluster investigation based on thresholds was neither effective nor cost effective, and it was therefore discontinued.[166] Current recommendations state that local cluster investigations should be conducted when deemed appropriate by public health professionals.[164]

Interactive dot mapping could be a useful means of interrogating data about molecular clusters of tuberculosis by highlighting cases that share links in person, time and place. This could be implemented relatively easily where national case registers that include geographic data are maintained. The WHO recommends collection of address-level geographic information in electronic case registers for tuberculosis.[167] Many countries including the United States and at least 23 in Europe also collect strain typing data on a routine basis.[168,169] In the United Kingdom, for example, the ETS system includes post code-level information for all cases, in addition to the routine molecular strain typing data. Cases could therefore be plotted to a high degree of precision, and linked to other cases with the same molecular strain type.

A tool that makes use of the interactive features of web browsers but does not require a stand-alone GIS could therefore be useful for creating dot maps to support disease outbreak and cluster investigations.

3.3 Aims and objectives

Aim: To develop an interactive, open-source tool to create dot maps to support investigation of disease outbreaks and clusters.

The objectives of this study were to:

1. Define the functional and technical specifications of an interactive dot mapping tool.
2. Write the code for the tool.
3. Demonstrate the utility of the tool through interrogation of an example data set based on molecular clusters of tuberculosis in London and the South East of England.
4. Evaluate the tool through a survey of potential users.
5. Make the code available through an online repository.

3.4 Methods

3.4.1 Specification

I defined the functional and technical specification for this tool, based on the research described above, as shown in Box 3.1.

Box 3.1: Functional and technical specifications for interactive mapping application.

Functional:

i. Displays locations of cases of disease on an interactive map.
ii. Group of cases being displayed can be changed according to user selection, with ability to display two or more groups for comparison.
iii. Colour codes cases according to epidemiological characteristics such as demographics and risk factors.
iv. Cases can be filtered by combinations of one or more epidemiological characteristics (such as demographics and risk factors).
v. Additional locations of interest can be plotted for context (for example locations of clinics).
vi. No specialist training in GIS or advanced computer programming skills required to operate.

Technical:

i. Runs on a standalone computer.
ii. Flexible data input (not restricted, for example, to data extracted from one surveillance system).
iii. Does not require upload of data to the internet (maintaining patient confidentiality).
iv. Uses free, open-source software.
v. Code freely available to download from a repository for further development or customisation.

3.4.2 Implementation

I chose to develop the interactive mapping tool using *Shiny*, a package for developing interactive web applications using the statistical software, R.[159,170] This is a useful framework for interrogation of sensitive data such as locations of cases of disease because it can be run locally and therefore does not require upload of information to the internet. I used R version 3.1.3 implemented in RStudio version 0.98.1103 to write the application.

Shiny applications are composed of two R scripts: A user-interface script (frontend) which defines the layout and appearance of the application by converting R code into HTML, and a server script (backend) which contains the R code to process data and produce the outputs that are displayed in the user interface.

The application that I have developed also contains two additional files: An R script which converts data containing information about cases of disease in one or more clusters, for example an extract from the ETS surveillance system, into the format that can be read by the Shiny application; and a cascading style sheet (CSS) document that enhances the appearance of the application.

In addition to the Shiny package itself, the application also uses a number of other R packages for data processing and plotting. Details of packages used, versions and their functions are shown in Table 3.1. I used the *leaflet* package for the JavaScript library of the same name to enable interactive mapping.[171] Base map tiles in the application are provided by OpenStreetMap, a free, editable world map, enabling visualisations produced to be shared without copyright restrictions.[172] Points on the map are automatically colour coded according to levels of categorical variables using the package *RColorBrewer* to select colour palettes appropriate for cartography.[173]

Table 3.1: R packages used in the interactive mapping application.

Package name	Version	Purpose
shiny	0.12.2	Producing interactive web application.
ggmap	2.3	Geocoding points.
plyr	1.8.1	Aggregating data to produce summary tables.
RColorBrewer	1.0.5	Producing colour scales.
leaflet	1.0.0.9999	Generating interactive maps.
zoo	1.7.12	Formatting dates.
epitools	0.5.7	Formatting dates for epidemic curve.
ggplot2	1.0.0	Plotting epidemic curve.

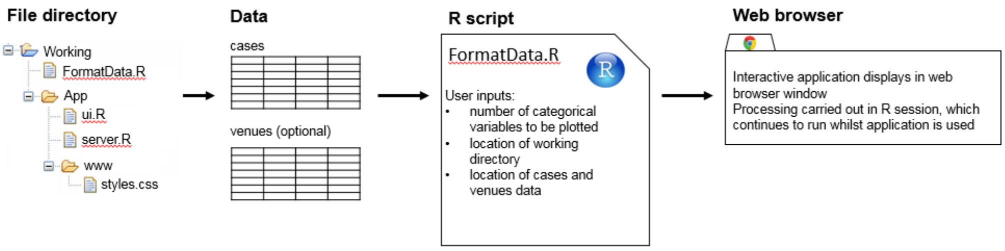
Additional features of the application include optional geocoding of postcodes or named geographic locations using the R package *ggmap*;[174] construction of epidemic curves using the packages *ggplot2* and *epitools*,[175,176] and a summary data table. Design of the application was inspired by the *SuperZip* interactive visualisation by RStudio.[177]

The code for the application, *DotMapper*, is free to download from its GitHub repository (https://github.com/cathsmith57/DotMapper), which includes the R scripts for running the application and formatting the data. It also has links to a working online version of the application; a video demonstration of its features; a short user guide, and example data sets.

3.4.3 User workflow

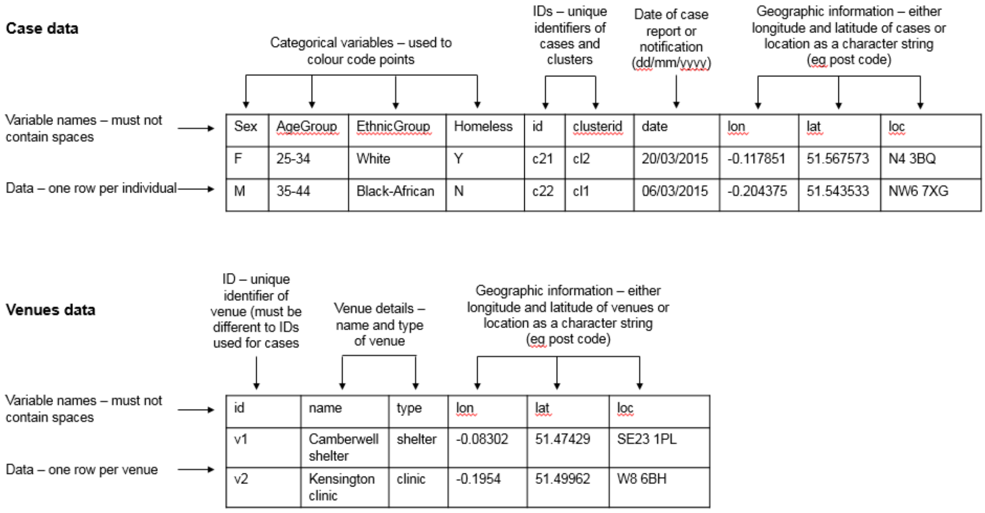
The workflow through which this application can be run on a local machine is shown in Figure 3.1. First, the user must save the four files that comprise the application using the file structure shown.

Figure 3.1: Workflow for running interactive dot mapping application on a local machine.


File directory
Working
FormatData.R
App
ui.R
server.R
www
styles.css
Data
cases
venues (optional)
R script
FormatData.R
R
User inputs:
• number of categorical variables to be plotted
• location of working directory
• location of cases and venues data
Web browser
Interactive application displays in web browser window
Processing carried out in R session, which continues to run whilst application is used

Second, the user compiles the data sets that they wish to display. The application plots data of two types: cases (i.e. patient locations and associated characteristics) and, optionally, venues (i.e. other locations of interest such as clinics or potential sources of infection). I designed the application to be as flexible as possible to enable rapid plotting of data collected from different surveillance systems or surveys, although there are some requirements about the structure of the data (Figure 3.2).

Figure 3.2: Structure of data that can be imported into the interactive dot mapping application.


Case data
Categorical variables – used to colour code points
IDs – unique identifiers of cases and clusters
Date of case report or notification (dd/mm/yyyy)
Geographic information – either longitude and latitude of cases or location as a character string (eg post code)
Variable names – must not contain spaces
Data – one row per individual
Venues data
ID – unique identifier of venue (must be different to IDs used for cases
Venue details – name and type of venue
Geographic information – either longitude and latitude of venues or location as a character string (eg post code)
Variable names – must not contain spaces
Data – one row per venue

Sex	AgeGroup	EthnicGroup	Homeless	id	clusterid	date	lon	lat	loc
F	25-34	White	Y	c21	cl2	20/03/2015	-0.117851	51.567573	N4 3BQ
M	35-44	Black-African	N	c22	cl1	06/03/2015	-0.204375	51.543533	NW6 7XG

id	name	type	lon	lat	loc
v1	Camberwell shelter	shelter	-0.08302	51.47429	SE23 1PL
v2	Kensington clinic	clinic	-0.1954	51.49962	W8 6BH

Case data must be in ‘wide’ format, with one row per individual; categorical variables used for colour coding points should be the first columns of the data; there must be a unique identifier for each case and cluster; the date of the case report or notification must be included and formatted dd/mm/yyyy (e.g.

'25/05/2005'), and geographic information must be included either as a longitude and latitude or as a character string (e.g. post code). Venues data must also be formatted with one row per location; have a unique identifier (that differs from the case IDs); a name and type of venue, and geographic information as for cases.

Third, the user opens the data formatting R script in an R console. Within the R script, the user must set a single parameter, which is the number of categorical variables in their data set. The user then runs the R script, and will be prompted to select the locations of their working directory (in which the files were saved), the case data, and (if used) venues data. Running the R script also requires that the additional packages that the application uses have been installed and are loaded.

The final command in the R script runs the application, which will open in a web browser window. Whilst the application is being used, the data are processed in the R console, which must therefore not be closed during the session.

3.4.4 Case study

As a case study, I have generated data for three example molecular clusters of tuberculosis (denoted c1, c2 and c3). These molecular clusters use data of the same structure as that in the ETS system, and mimic the broad trends seen in real molecular clusters. I have anonymised the data by altering the exact characteristics, including demographics, risk factors, and spatial locations of cases.

3.4.5 Evaluation

The tool was evaluated in two stages. First, at an early stage of its development, I demonstrated the tool to a small group of potential users. This provided an opportunity for me to gain feedback about the potential for implementation and barriers to use, and additional features that would be of interest.

Second, when I had completed development of the tool to its current form, I ran an online survey to gain further feedback about the application. The questions in the survey are provided in Appendix 10.2. I gave presentations about the application and requested that attendees complete my survey. These presentations included the UK Public Health Science conference (London, November 2015); the Farr Institute Festival for Digital Health (London, March 2016); the PHE Field Epidemiology Conference (Wawrick, March 2016), and an internal meeting at the South East and London Field Epidemiology Service (July 2016).

3.5 RESULTS

3.5.1 Features

The primary output of the application is an interactive map. By default, the map displays the locations of the cases, colour coded according to the first categorical variable in the data set. Locations of contextual venues can be toggled on or off. The map can be panned and zoomed in and out to explore the data, and clicking on cases or venues produces a popup displaying further information. The application can be used to plot just one group of cases, for example in an outbreak situation, or to load multiple groups and compare their characteristics. Drop-down menus are used to select the group to display; to filter groups by size (number of cases in cluster) if necessary, and to change the variable being plotted.

Cases plotted can also be filtered by interactively selecting subsets of data: A date range slider is provided to select cases in any time period according to their notification date, and selecting *Subset* facilitates display of cases which satisfy selected combinations of characteristics of categorical variables. The *Reset groups* button returns the display to showing all cases.

Other tabs in the application display a summary data table and an epidemic curve. The data table presents the number and percentage of cases according to each categorical variable. If multiple groups are included in the data it also displays totals and percentages for all cases, which could be used to assess whether patterns in the selected group reflect the overall epidemiology of the disease. Cases are plotted as a function of time using an epidemic curve. The time periods into which the cases are grouped in the epidemic curve can be switched between days, weeks, months, quarters and years, as appropriate for the specific disease being investigated.

3.5.2 Case study

This case study presents data from three example molecular clusters of tuberculosis (denoted c1, c2 and c3) using altered and anonymised data of the same structure as that in ETS system. I have produced three short screencast movies which demonstrate the interactive analyses possible using the application. These movies are provided in the CD-ROM attached with this thesis.

Movie 1 shows all cases in molecular cluster c1 displayed by ethnic group. There is a notable group of cases in the north east of the city which are of Pakistani ethnicity, whilst cases of other ethnicities appear to be more dispersed. Displaying only the cases in the first month of the molecular cluster reveals that the initial cases were all in this spatially-constrained group of Pakistani ethnicity, and the molecular cluster became more dispersed and affected different ethnic groups in later months. This visualisation can be used to generate hypotheses about transmission and potentially highlight a missed opportunity for early control in a targeted population: The strain appears to have been transmitted amongst a distinct population group before being spread more widely in the community.

Assessment of molecular cluster c2 demonstrates how this tool could be used to target interventions for specific risk populations (Movie 2). Locations of cases in this molecular cluster are displayed according to whether they have a history of homelessness, and locations of sheltered accommodation services are also shown. There appears to be an association with homelessness in the south central areas of London. The shelter in the south of the city may therefore be a suitable focus for interventions, such as screening by the mobile digital screening unit, Find and Treat.[31]

Movie 3 displays molecular cluster c3, comprised of individuals in a tight geographic group in the north central areas of the city, about half of whom were born in the United Kingdom and aged between 20 and 40. The epidemic curve shows that these cases all occurred within five calendar quarters with numbers increasing recently, indicating a possible opportunity for control.

3.5.3 Evaluation

3.5.3.1 User demonstration

The user demonstration was carried out on 27th July 2015, and attended by a Tuberculosis Cluster Investigator (PHE), a Senior Epidemiology Scientist (PHE), and a Professor of Infectious Disease Epidemiology (UCL). Feedback from this session included potential ways in which the tool could be used and barriers to its use (Box 3.2), and suggestions of new features.

Box 3.2: Feedback from user demonstration during development of interactive mapping application (27 July 2015).

Potential uses:

i. Mapping of drug-resistant tuberculosis cases, for example the group of multidrug-resistant cases in London to inform need for further contact tracing.
ii. Description of major tuberculosis clusters in London.
iii. Investigation of geographic distribution of extended-spectrum beta-lactamases (ESBLs) to highlight areas of potential community transmission.
iv. Investigation of outbreaks, for example of foodborne or sexually-transmitted infections.

Barriers to use:

i. Technical skills of potential users.
ii. Ease of adaptability to other infections.
iii. Availability of accurate geographic information.

To address this feedback, I implemented several new features, which are in the current version of the application. The main change was the addition of the option to display only a sub-set of cases which had certain characteristics. I also altered the mechanism for importing data so that the application could be used with any number of categorical variables, as it was previously designed specifically for tuberculosis molecular cluster investigation and therefore required data to be imported to be in the structure of the ETS system. Table 3.2 summarises the new features that were suggested at the user demonstration and how I implemented them.

Table 3.2: Features suggested at user demonstration for interactive mapping tool, and how they were implemented.

Feature suggestion	Implementation
Display of contextual locations such as clinics or potential venues of transmission.	Option to plot additional locations of interest using markers of a different shape (pins) to clearly distinguish them from case locations (dots).
Include additional categorical variables such as smear positivity, time since entry to the UK, treatment outcomes.	Import of data flexible to allow plotting of any number of categorical variable. User has to specify only the number of categorical variables being plotted.
Enable grouping by variables other than group ID, for example by drug resistance or treatment clinic.	Multiple groups can be displayed at one time, and cases can be filtered by any combination of categorical variables.
Include button for capturing screen shots of maps to include in cluster summaries.	Not currently implemented within tool, but screen shots can be captured using standard methods.

3.5.3.2 Survey

There were seven respondents to the online survey about the current version of the application. Of these, four were epidemiologists or information officers working in public health, two were researchers and one did not state their role. All respondents answered ‘strongly agree’ or ‘agree’ when asked if *DotMapper* was useful; easy to use; well designed and offers a new way to explore data. They were also confident that they could set up *DotMapper* using R software (six ‘agreed’, one ‘strongly agreed’), and all except one person indicated that they planned to try using the application.

Most people said that they envisaged using *DotMapper* in communicable disease surveillance or outbreak investigations (five respondents each). Investigations of a range of infections were suggested as potential uses, including sexually-transmitted infections, influenza and other respiratory infections. Two respondents said that they would use the application for non-communicable disease surveillance. Four respondents said that they may use the tool for healthcare service evaluation for example by mapping the locations of cases and clinics.

The most important barriers that would stop people using the tool were lack of familiarity with R software and concerns about patient confidentiality. Popular feature requests were the ability to map aggregated aerial data (rather than point locations); templates for loading data from specific surveillance systems, and custom reports based on selected maps. Potential features ranked as less important were development of a mobile version of the application and detailed online tutorials about how to use it.

3.6 Discussion

3.6.1 Summary of findings

The application that I have developed in this study enables rapid, interactive dot mapping of cases of disease. It is intended to be used as a means for public health officials and researchers to visualise and interrogate geographically-referenced data. In the context of investigation of tuberculosis molecular clusters, I have demonstrated how the data presented on a map can be used to identify patterns in both space and time; be used to generate hypotheses about disease transmission pathways; identify cases of interest for future investigation, and guide potential control measures.

3.6.2 Interpretation of results – strengths of the application

The application developed in this study meets the functional and technical specifications for the interactive, open-source dot mapping tool outlined in the methods. In its current form, it could be useful in tuberculosis molecular cluster investigations, and there was some evidence from the evaluation that potential users would be interested in implementing it. However, this version of the tool has not undergone rigorous user testing and should be considered as a proof-of-concept rather than a completed product.

The main advantage of the tool developed in this study over existing applications is that it uses an interactive, user-friendly web interface to display spatial data without the need for specialised GIS software. The design of the tool incorporates many interactive features which are valued by users.[157] These features, including the ability to zoom, pan, and click on points to display additional information, are familiar from commonly-used applications (such as Google Maps), and will therefore require little training to operate. This could allow interrogation of geographically-referenced data in molecular cluster investigations by users without GIS expertise.

Use of the statistical software R, as opposed to a bespoke GIS package or web-based visualisation tool, is another benefit of this application. This could help to streamline workflows by allowing mapping and epidemiological analyses to be conducted within the same software environment, eliminating the need to transfer data between software packages. This could be particularly advantageous in an outbreak situation, in which data are being updated on a regular basis.

This application is flexible, allowing plotting of any geo-coded point data and associated categorical information. The code can be freely downloaded from an online repository, and can therefore be adapted and improved by other users. This avoids the issue of development in 'silos' which has limited the adoption of some visualisation tools in the past,[157] and means that the capabilities can be expanded to suit user needs. For example, it could be adapted to import data from a specific surveillance system, or to plot area data.

3.6.3 Interpretation of results – limitations of the application

The main limitation of this application is that, as is common with open-source projects, it is not supported and therefore requires a degree of technical expertise for users to install and de-bug as necessary. However, the application benefits from using the R framework, which has a large user-base and online community that will be able to provide assistance in many situations.

Another potential limitation is the ease of sharing visualisations within and between agencies. Although screenshots, and screencast movies, as presented here, can be produced easily, these methods clearly detract from the interactive utility of the application. Avoiding the requirement to upload data to the internet gives this application the advantage of maintaining patient confidentiality. However, maps can be de-anonymising if displayed at a high zoom level, and incorporation of systems for sharing data may therefore be beneficial for cross-agency exercises.

3.6.4 Implications for policy and practice

This application has been designed principally for use on local machines in a single-user context. Public health officials and surveillance staff can use it to interrogate datasets without having to remove personal identifiable information from within secure systems. The application can also be used by researchers with appropriate data access rights. Installation of the software and its dependencies may be a challenge for some users who are less familiar with R, and it may therefore be best disseminated via existing networks, for example of local software 'champions'.

The utility of this application in practice will depend on the availability of reliable, well recorded location data in a timely manner. The ETS system in the United Kingdom records postcodes systematically, but this is not always the case for surveillance systems for other infections. For example, the Genitourinary Medicine Clinic Activity Data-set version 2 is the primary surveillance system for STIs in England. Post codes are collected in this system but converted to small area of residence when made available to public health surveillance staff. Use of this tool for STI cluster investigation would therefore rely on data collected by clinics and in laboratory systems, which are only available to public health staff when a specific concern has been identified, and completeness of location data is variable.

There are also limitations to the information which can be gleaned from interrogation of data in dot maps. Representation of case locations as static points is a simplification of their true geographic distribution and the maps do not take into account the distribution of the underlying population. The ability to produce multiple maps of different groups of cases and combinations of risk factors is a clear advantage of this application. However, a potential drawback is 'cognitive overload', when the user is presented with more information than they are able to process successfully.

3.6.5 Study strengths

One of the strengths of this study is that the tool was evaluated in collaboration with potential end users both at an early stage in the design process and on the final version of the application. This allowed the specifications for the tool to be refined to match the requirements and preferences of the users, and identified priorities for future developments. The completed application met all of these functional and technical requirements and, since the code is free to download, it could be adapted to suit specific user needs.

Another strength is that the application was tested with data similar to that from the ETS system, demonstrating specific scenarios in which it could be useful. The feedback from the online survey suggested that potential users thought that the tool could have a range of other uses, and that it was generally easy to use and set up.

3.6.6 Study limitations

The main limitation of this study was that the tool has not been thoroughly user-tested. Gathering feedback about this tool through an online survey proved challenging, and the evaluation was therefore limited by the number of respondents. Although the feedback was generally positive, it is possible that only staff and researchers with an interest in the topic were motivated to complete the survey. This may have biased the results, making the tool appear more useful and acceptable than it would be if a more general audience had completed the evaluation.

3.6.7 Future directions

The flexible nature of this application permits a large scope for future work, both using the tool in its current form, and by extending it with additional features. In this study, I have demonstrated the utility of the tool for investigations of molecular clusters of tuberculosis linked through molecular strain typing. Integrating this tool into routine practice would be a logical next step to assist tuberculosis molecular cluster investigators to interrogate large, complex data sets and identify opportunities for intervention.

Another potential use of the tool is for investigation of outbreaks of other infectious diseases. The systematic review in Chapter 2 showed that spatial methods including dot mapping were used relatively rarely for investigations of outbreaks of many infectious diseases including those resulting from foodborne or sexual transmission. During such investigations, PHE often uses online surveys to collect information from patients. Extracts from these surveys could be imported into the tool with minimal data processing, allowing cases to be displayed on the map and filtered, for example according to different food exposures or sexual behaviours. Display of contextual locations such as food outlets or potential venues of transmission may also be informative for hypothesis generation in these investigations.

Although the focus of the development of this tool has been for use on clusters of infectious diseases, there are also many potential uses outside this field. For example, commissioners of services for non-communicable diseases may find the application of use in identifying areas in greatest need of services by mapping locations of disease cases or events such as accidents. Researchers in other areas of science such as ecology could use the tool for mapping locations of field study sites;

and in the commercial sector businesses could use this framework, for example to assess opportunities for growth of services. Expanding the tool into these different fields would also require further evaluation to ensure that the design and features are appropriate for the user groups.

Further extensions to this tool could be implemented by users with an understanding of R and Shiny. This may include creating a bespoke version of the tool with adaptations to specific data sets. For example, users may wish to customise categories into which continuous variables are divided, add additional information to popups, or change the information displayed in the table. Such developments could be used locally or shared with professional networks through an online repository. Additional features of the leaflet JavaScript library could also be used to extend the tool, such as adding area data with polygons, overlaying raster images, enabling marker dragging, or using alternative base map tiles. I have created an example of one such extension by overlying a geographic profile risk surface onto the map. The code for this is available from https://github.com/cathsmith57/geoprofileShiny.

Finally, this study highlights an opportunity for expanding the use of interactive visualisations in infectious disease epidemiology and other areas of health informatics. With increasing volumes of healthcare data, there will be need to ensure that tools for visualising information are developed in tandem with new analytic methods. There are many potential applications that could use some of the same features employed in the tool developed in this study. For example, a similar application could be developed for interrogating data in non-geographic contexts, such as hospital-based outbreaks and social networks. Platforms such as R Shiny allow these applications to be created without the need to invest in bespoke software or employ specialist web developers. A likely challenge for the development of new tools, as was found in this study, will be engagement of potential users in the design and evaluation process, and dissemination of the final product. Further issues may include integration of visualisations into systems providing real-time data, and design of tools that are suitable for a range of different user groups, from healthcare professionals to patients or individuals collecting data to monitor their personal health.

Box 3.3: Summary of Chapter 3.

- An interactive mapping tool for plotting disease case locations was developed using R Shiny.
- The application is novel in providing rapid geographic displays of epidemiological characteristics of cases in a user-friendly way without the need for specialised GIS software.
- It has broad applicability to investigations of disease clusters in any setting worldwide.
- In the context of tuberculosis control, the tool can be used to generate hypotheses about disease transmission, potentially facilitating more appropriate targeting of services to diagnose and treat patients.
- User evaluation provided mostly positive feedback, suggesting that the application may be useful in practice.
- However, user feedback was limited as there were few respondents to an online survey.
- There are many opportunities for further developments of this tool, and to create other similar tools to improve visualisation of health informatics data more broadly.

4 Molecular and spatial epidemiology of tuberculosis in London

4.1 Description of chapter contents

In this chapter, I use results from the first five years of routine molecular strain typing of tuberculosis in London to investigate the evidence for local transmission. I describe the distribution of molecular clusters based on their size (number of cases in clusters) and the length of time between cases. I demonstrate that the characteristics of cases are different in clusters of different sizes, first by visual means using box plots, and then through multinomial regression analysis. I use spatial scan statistics to test for significant spatial clustering within the largest molecular clusters compared to the background of all tuberculosis cases. I present the locations of these clusters on smoothed incidence maps generated through kernel density estimation. I then use the interactive mapping tool, developed in Chapter 3, to visualise each of the large clusters and generate hypotheses about routes of transmission. Finally, I investigate the possibility of using the characteristics of the first two cases in a molecular cluster to predict how likely it is to develop into a large cluster.

4.2 Study rationale and introduction

In countries with low incidence of tuberculosis such as the United Kingdom, highest rates are often found in large cities.[7] The rate of tuberculosis in London in 2014, for example, was 30 per 100,000 population compared to 12 per 100,000 in the whole of England.[5] This high incidence has led to the city being described as the ‘tuberculosis capital of Western Europe’.[178] Large outbreaks of tuberculosis represent a particular threat to control because they reflect multiple instances of active transmission. Such large outbreaks have occurred previously in London and other large cities.[179-184] However, identification of outbreaks of tuberculosis is not straightforward, as it requires cases resulting from active transmission to be distinguished from those arising from reactivation of latent disease with absent or limited onward transmission. The extent to which they contribute to the overall burden of disease is therefore not known.

Since 2010, PHE has been undertaking routine molecular strain typing of all tuberculosis isolates by 24 locus MIRU-VNTR. Cases that share a molecular strain type may be linked through transmission and therefore form part of large outbreaks. PHE conducts investigations of molecular clusters which aim to identify previously unknown epidemiological links between cases through systematic review of patient records and, where indicated, re-interviewing of patients.[164] This can prompt public health interventions such as extended contact tracing and screening. However, prioritising clusters for investigation and identification of epidemiological links is challenging.[164,185]

The systematic review in Chapter 2 demonstrated that spatial analyses can also be useful to investigate links between cases of infectious diseases. For example, cluster detection tests can be used to identify outbreaks and to pinpoint areas where transmission may be occurring. Smoothed incidence maps provide an effective means of visualising distributions. The interactive mapping tool developed in Chapter 3 can be used to explore potential epidemiological links between cases and may therefore assist with identification of specific groups that could benefit from interventions.

An analysis of the first three years of molecular strain typing data in London showed that 46% of cases were part of a molecular cluster (i.e. shared a strain type with at least one other case).[186] Cluster size ranged from two to 55 cases, and over half of the clusters had only two cases. However, the study did not determine whether risk factors varied by cluster size; or assess spatial clustering.

The study also attempted to identify factors that could be used to prioritise molecular clusters warranting detailed investigation. It compared the characteristics of the initial two cases in small (fewer than five cases) molecular clusters with those in large (five or more cases) molecular clusters. Reduced time between notifications and a history of imprisonment were found to predict cluster growth, and the authors recommended that the analysis be repeated using five years of data.

Data from five years of molecular strain typing in London are now available (2010-2014). Combinations of epidemiological and spatial analyses of these data could be used to estimate the importance of long chains of transmission in the city, to identify risk factors for cases in large clusters, and to predict growth of clusters based on characteristics of initial cases.

4.3 Aims and objectives

Aim: To investigate the size and distribution of molecular clusters of tuberculosis in London between 2010 and 2014 and examine the evidence for local transmission.

The objectives of this study were to:

1. Describe the distribution of molecular clusters in London by cluster size (number of cases in the cluster).
2. Identify factors associated with molecular clusters of different sizes.
3. Examine the spatial distribution of large molecular clusters and test for significant spatial clustering.
4. Determine the extent to which early cases in molecular clusters can be used to predict development of large clusters.
5. Discuss implications for tuberculosis control and cluster investigation.

4.4 Methods

4.4.1 Data sources

This was a cross-sectional analysis of patients notified with tuberculosis between 1 January 2010 and 31 December 2014 resident in London. I extracted data from the ETS system, a national online register for real-time reporting of tuberculosis cases that is maintained by PHE. This system includes patient demographics (age, sex, ethnic group, country of birth, time since entry to the UK, occupation) and clinical characteristics (site of disease, sputum smear status, history of tuberculosis disease and treatment, drug sensitivity, whether the case spent time as a hospital inpatient). It also includes information on social risk factors for tuberculosis (whether the patient has a history of homelessness or problems with illicit drug or alcohol use). Surveillance data from ETS is routinely matched to the National Tuberculosis Strain Typing Service to provide molecular clustering information.

An estimate of level of social deprivation (the index of multiple deprivation, IMD) is also included in ETS through matching of residential postcodes to lower-layer super output areas (LSOAs). LSOAs are a geographic hierarchy used in England and Wales each encompassing a mean population of 1,500. The IMD is a measure of relative deprivation at the LSOA level in England and is based on seven domains of deprivation: income; employment; crime; living environment; barriers to housing and services; health and disability; and education, skills and training.[187] Low ranks indicate higher levels of deprivation and I converted ranks into London-level deprivation quintiles, with the lowest quintile representing the most deprived areas.

Ethical approval was not required for this study because PHE has Health Research Authority approval to hold and analyse national surveillance data for public health purposes.

4.4.2 Molecular clustering analysis

I categorised cases as unique or part of a molecular cluster, and by the number of cases in the molecular cluster. I used PHE convention for assigning cluster status: Cases with fewer than 23 loci successfully typed by MIRU-VNTR were excluded from the analysis, and each cluster must have had at least one case typed with 24 loci.[5]

Molecular clusters were therefore defined as groups of two or more cases, each typed with at least 23 loci, which shared an identical MIRU-VNTR strain type with another case notified in the study region during the study period, with the stipulation that at least one of the cases in the cluster had been typed with 24 loci.

Unique cases were individuals whose strain was typed with at least 23 loci but whose type did not cluster with another case. I excluded cases whose molecular strain type was unique within the study area but shared a molecular strain type with another case in England.

I described the distribution of molecular clusters by size (number of cases in the cluster) and calculated the proportion of cases that were part of clusters. I identified successive cases reported in a molecular cluster using case notification dates and calculated the median and interquartile range number of days between successive cases in molecular clusters by cluster size.

To compare the distribution of case characteristics by cluster size, I initially used box plots. I then categorised cases according to the size of the molecular cluster: not clustered (unique) cases; two cases; three to 20 cases, and more than 20 cases. In situations such as this, in which the outcome of interest is a categorical variable, a multinomial logistic regression model can be used.[188] This is an extension of the simple logistic regression model that is used for dichotomous outcomes. Coefficients resulting from the multinomial model are interpreted in a similar way to the odds ratios (ORs) derived from a logistic regression model.

I used this approach as opposed to a linear regression model with cluster size as a continuous outcome because the box plots showed highly skewed distributions of cluster sizes, and residuals of resulting linear regression models therefore deviated substantially from normality.

I investigated associations at single variable analysis and included variables with an association of $p<0.2$ in the initial multivariable model. I then used a backwards stepwise approach to eliminate variables which did not contribute significantly to produce a final model, with model fitting based on the Akaike information criterion (AIC). LSOA of residence was included as a random effect in models which included IMD to account for the hierarchical level at which this variable was measured. Social risk factors were considered separately and as a cumulative count of these risk factors at single variable analysis, and as a count at multivariable analysis.

4.4.3 Spatial clustering analysis

I used spatial scan statistics to assess spatial clustering within molecular clusters, implemented using the R package *rsatscan*, an interface for the SaTScan software.[155,189] I tested the hypothesis that cases in large molecular clusters (of more than 20 cases) were closer together in space than the underlying spatial distribution of tuberculosis cases. For each large molecular cluster, I therefore performed a spatial scan under the Bernoulli (case/control) model, using the locations of all other tuberculosis cases as controls.

Spatial scan statistics work by comparing observed and expected numbers of cases occurring within spatial windows of various sizes. Expected numbers of cases and controls are generated under the Bernoulli distribution using Monte Carlo simulations. A likelihood ratio is then calculated for each window of different size and used to determine a p value. This is the probability that the observed number of cases within the window would have arisen if the cases and controls had a random Bernoulli distribution.

I aimed to identify areas with evidence of local transmission and therefore set the maximum radius of the spatial window at 5 km, and identified clusters with a p value of less than 0.05 which encompassed at least ten cases.

To visualise the distributions of these clusters, I plotted the locations of significant spatial clusters for each molecular cluster overlaid on a smoothed incidence map. These maps were produced using kernel density estimation with a Gaussian kernel of bandwidth 5 km, and showed the relative distribution of the given molecular cluster compared with all other tuberculosis cases in London.

4.4.4 Spatial distribution of cases in large molecular clusters

I used the interactive mapping application developed in Chapter 3 to explore the spatial distribution of tuberculosis cases in the largest molecular clusters. I generated interactive dot maps of the locations of cases in each cluster, and explored the distribution of the risk factors included in the surveillance data. I used the interactive sub-setting feature of the application to highlight groups that were particularly affected for each cluster to generate hypotheses about transmission. I summarised key characteristics of each cluster and took screen shots of the maps.

4.4.5 Prediction of large clusters

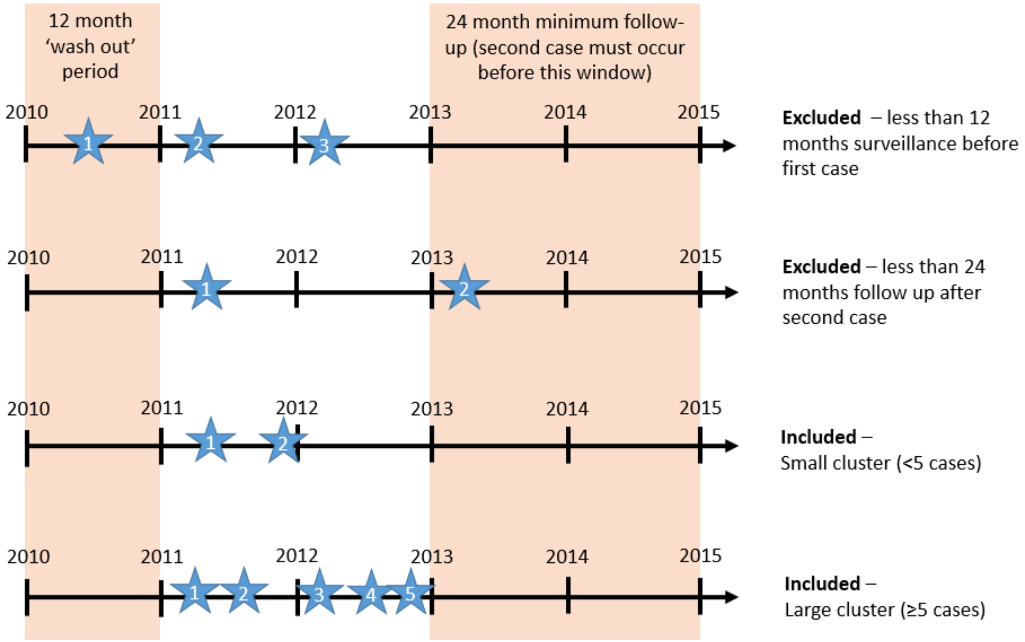
Having established factors that are associated with cases in largest clusters, I investigated whether the characteristics of initial two cases in a molecular cluster could be used to predict whether the cluster would be likely to develop into a large cluster.

The start of the study period for previous analyses in this chapter (1 January 2010) was determined by the start of routine molecular strain typing of tuberculosis isolates in London. Some of the clusters included in the analyses will therefore inevitably have been linked to cases that occurred before 2010, but for which strain typing data were unavailable. However, in this part of the analysis I aimed to identify initial cases in clusters, and thus to exclude clusters with linked cases that occurred prior to the start of routine strain typing. I therefore excluded molecular clusters which had any cases in London in 2010, allowing a period of one year to ‘wash out’ ongoing clusters. Similarly, to ensure that all clusters had an equal opportunity to develop into a large cluster, I included only those clusters which had at least 24 months follow up after the second case.

This analysis therefore included clusters with no cases reported in 2010 and at least two cases reported between 1 January 2011 and 31 December 2012 (Figure 4.1). Large clusters were defined as clusters which had progressed from two to five or more cases during the 24 month follow-up period; small clusters were those with fewer than five cases.

Figure 4.1: Definition of small and large molecular clusters of tuberculosis used for prediction of large clusters.

Timelines represent four different example molecular clusters. Stars represent cases in a molecular cluster.

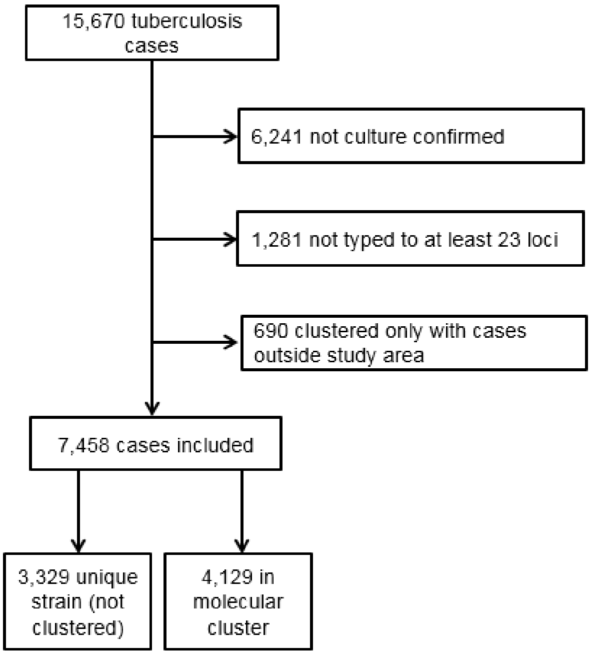

12 month 'wash out' period
24 month minimum follow-up (second case must occur before this window)
2010
2011
2012
2013
2014
2015
1
2
3
Excluded – less than 12 months surveillance before first case
1
2
Excluded – less than 24 months follow up after second case
1
2
Included – Small cluster (<5 cases)
1
2
3
4
5
Included – Large cluster (≥5 cases)

I conducted this analysis at the level of the cluster, and therefore calculated summary variables which defined clusters as 'exposed' (if at least one of the initial two cases had the given characteristic) or 'unexposed' (if neither case had the characteristic). To simplify the analysis, I considered only cluster characteristics that I had identified in the previous analysis to be significantly associated with large clusters. As a measure of the closeness of cases in time and space, I also calculated the number of days between the notification dates of the initial cases in the cluster and the distance in kilometres between their reported residences. I calculated numbers and proportions of small and large clusters with each characteristic. I then used logistic regression to test for significant associations.

4.5 RESULTS

Between 2010 and 2014, a total of 15,670 cases of tuberculosis were notified in London. Of these, 8,148 (52%) cases were successfully typed by MIRU-VNTR with at least 23 loci defined, whilst 6,241 (40%) were not culture confirmed and 1,281 (8%) were not typed and therefore excluded from this analysis (Figure 4.2). A further 690 cases were also excluded because they clustered only with cases that were not resident within the study area. This study therefore included 7,458 tuberculosis cases with a molecular strain type, of which 4,129 (55%) were part of 996 molecular clusters and 3,329 (45%) had a unique strain.

Figure 4.2: Cases included in analysis of molecular clusters of tuberculosis in London, 2010-2014.


15,670 tuberculosis cases
6,241 not culture confirmed
1,281 not typed to at least 23 loci
690 clustered only with cases outside study area
7,458 cases included
3,329 unique strain (not clustered)
4,129 in molecular cluster

4.5.1 Cluster size and time between cases

Cluster size ranged from two to 102 cases, with a median of two cases. There were 20 clusters with more than 20 cases, including 795 (11%) of all cases (Table 4.1). More than half of the clusters (522, 53%) comprised pairs of cases, but a larger proportion of cases were in the 454 clusters of three to 20 cases (2,290, 31% cases).

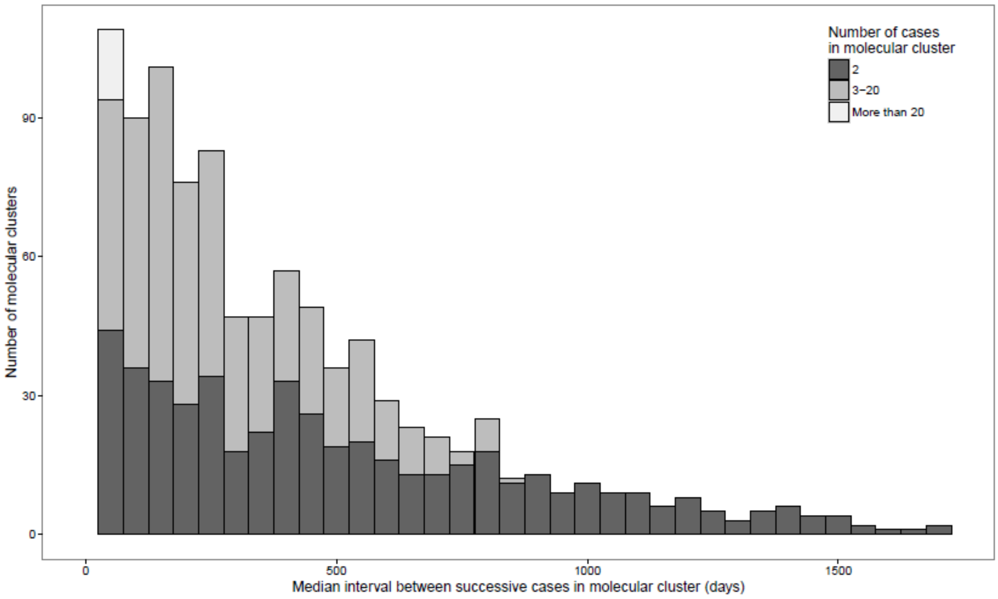
Table 4.1: Distribution of tuberculosis cases and molecular clusters by cluster size in London, 2010-2014.

Cluster size (number of cases)	Cases		Clusters	
	N	%	N	%
1 (Not clustered)	3329	44.6	NA	NA
2	1044	14.0	522	52.4
3-20	2290	30.7	454	45.6
More than 20	795	10.7	20	2.0
Total	**7458**		**996**	

Successive cases in clusters were defined using notification dates. The maximum possible time between cases was 1,825 days (five years, the length of the study period), and the maximum time observed between cases was 1,687 days. Of the 3,133 case-case intervals, 712 (23%) were longer than one year. Figure 4.3 displays the distribution of median intervals between successive cases in each cluster by cluster size. Overall, the median time between successive cases in a cluster was 114 days (IQR 32-323 days). For cases in clusters of more than 20 cases the median time was 23 days (IQR 8-54 days); for cases in clusters of 3-20 cases it was 149 days (54-335 days); and for clusters of two cases it was 406 days (IQR 162-752 days).

Figure 4.3: Median interval between successive tuberculosis cases in molecular clusters in London, 2010-2014, by cluster size.

Analysis based on clusters, n=996.

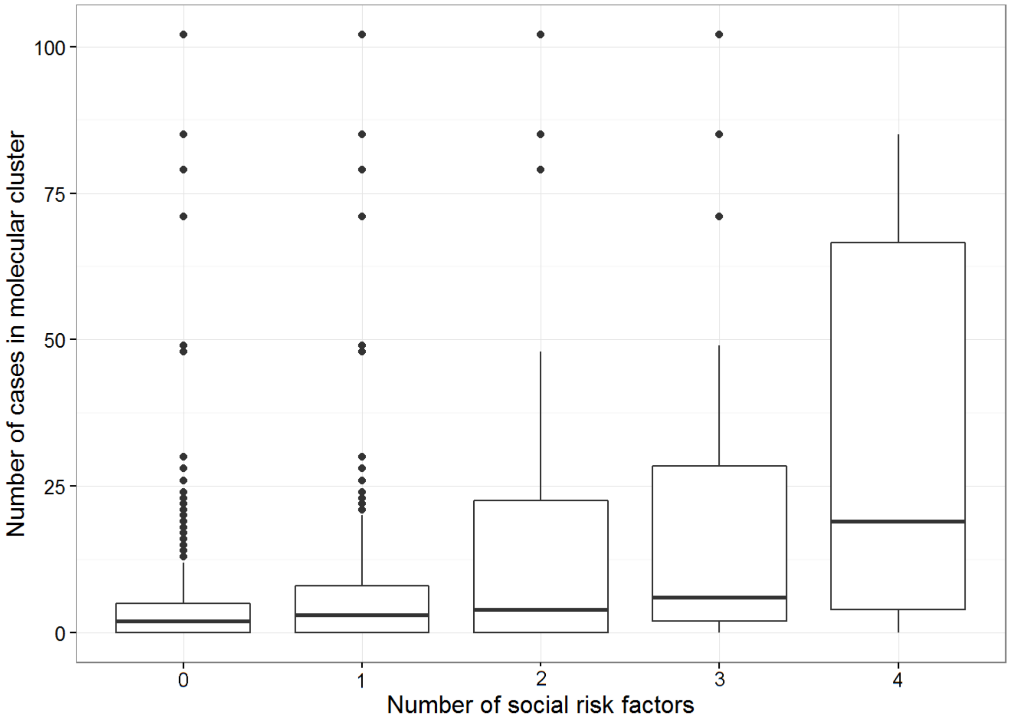


4.5.2 Factors associated with large molecular clusters

Box plots of the distribution of case characteristics by number of cases in the molecular cluster were highly skewed towards zero, owing to the large proportion of cases that were not clustered (cluster size one) or in small clusters (Appendix 10.3). Wider distributions with larger interquartile ranges (larger boxes) therefore indicated that a relatively large proportion of the cases with the given characteristic were in larger clusters. There was a stepped increase in the width of the distribution for cases with an increased count of social risk factors (Figure 4.4). Wider distributions were also present for cases who were younger (aged 0-14 years); of black-Caribbean ethnicity, and born in the UK. Differences in distributions were not evident for cases plotted by sex, time since entry to the UK, occupation, IMD, or any of the clinical characteristics (Appendix 10.3).

Figure 4.4: Box plots of distribution of number of cases in molecular tuberculosis cluster by number of social risk factors, London, 2010-2014.

Analysis based on cases, n=7,458.



Baseline characteristics of tuberculosis cases according to the number of cases in the cluster are shown in Table 4.2, and results of the single variable multinomial logistic regression analysis are presented in Table 4.3.

For each exposure, an OR was calculated for each of the three cluster size outcomes (two cases, three to 20 cases, more than 20 cases), with cases not in a cluster representing the comparison group. For example, the unadjusted ORs for being born in the UK were 4.11 (for cases in clusters of more than 20 cases); 2.55 (for cases in clusters of 3-20 cases), and 1.77 (for cases in clusters of two cases). This means that the odds of cases being in the largest clusters versus not being in a cluster for those born in the UK were 4.11 times that of those not born in the UK. Similarly, the odds of cases being in a cluster of 2-20 cases compared to not being in a cluster for those born in the UK were 2.55 times that of those not born in the UK; and the odds of cases being in a cluster of two cases compared to not being in a cluster for those born in the UK were 1.77 times those not born in the UK.

Table 4.2: Baseline characteristics of tuberculosis cases in molecular clusters of different sizes in London, 2010-2014.

				Number of cases in cluster		
		All cases	Not clustered	2	3-20	More than 20
Variable		**Total**	**N (% row)**	**N (% row)**	**N (% row)**	**N (% row)**
Total		7458	3329 (44.6)	1044 (14.0)	2290 (30.7)	795 (10.7)
Sex						
	Female	2911	1320 (45.3)	430 (14.8)	874 (30.0)	287 (9.9)
	Male	4546	2008 (44.2)	614(13.5)	1416 (31.1)	508 (11.2)
Age group (years)						
	0-14	158	43 (27.2)	24 (15.2)	60 (38.0)	31 (19.6)
	15-44	5284	2346 (44.4)	731 (13.8)	1628 (30.8)	579 (11.0)
	45-64	1369	585 (42.7)	207 (15.1)	433 (31.6)	144 (10.5)
	65+	647	355 (54.9)	82 (12.7)	169 (26.1)	41 (6.3)
Ethnic group						
	White	829	308 (37.2)	119 (14.4)	277 (33.4)	125 (15.1)
	Black-Caribbean	258	49 (19.0)	38 (14.7)	113 (43.8)	58 (22.5)
	Black-African	1722	609 (35.4)	242 (14.1)	614 (35.7)	257 (14.9)
	Black-Other	97	28 (28.9)	15 (15.5)	37 (38.1)	17 (17.5)
	Indian	2185	1126 (51.5)	303 (13.9)	615 (28.1)	141 (6.5)
	Pakistani	673	313 (46.5)	96 (14.3)	181 (26.9)	83 (12.3)
	Bangladeshi	365	246 (67.4)	38 (10.4)	69 (18.9)	12 (3.3)
	Chinese	86	49 (57.0)	11 (12.8)	21 (24.4)	5 (5.8)
	Mixed /Other	1177	561 (47.7)	176 (15.0)	347 (29.5)	93 (7.9)
Place of birth						
	Non-UK	6223	2990 (48.0)	876 (14.1)	1801 (28.9)	556 (8.9)
	UK	1150	301 (26.2)	156 (13.6)	463 (40.3)	230 (20.0)
Time since entry to UK (years)						
	0-1	1095	540 (49.3)	159 (14.5)	298 (27.2)	98 (9.0)
	2-4	1391	740 (53.2)	202 (14.5)	347 (24.9)	102 (7.3)
	5-9	1156	538 (46.5)	162 (14.0)	339 (29.3)	117 (10.1)
	10+	1803	785 (43.5)	261 (14.5)	583 (32.3)	174 (9.7)

Occupation					
Other	2465	1146 (46.5)	342 (13.9)	759 (30.8)	218 (8.8)
None	2615	1129 (43.2)	367 (14.0)	788 (30.1)	331 (12.7)
Education	1074	452 (42.1)	148 (13.8)	342 (31.8)	132 (12.3)
Health care	275	137 (49.8)	44 (16.0)	81 (29.5)	13 (4.7)
Pulmonary disease					
No	3006	1553 (51.7)	389 (12.9)	815 (27.1)	249 (8.3)
Yes	4452	1776 (39.9)	655 (14.7)	1475 (33.1)	546 (12.3)
Sputum smear					
Negative	2371	1066 (45.0)	338 (14.3)	734 (31.0)	233 (9.8)
Positive	2062	745 (36.1)	314 (15.2)	718 (34.8)	285 (13.8)
Previous diagnosis					
No	6821	3088 (45.3)	941 (13.8)	2078 (30.5)	714 (10.5)
Yes	354	124 (35.0)	57 (16.1)	125 (35.3)	48 (13.6)
Previous treatment					
No	13	6 (46.2)	1 (7.7)	6 (46.2)	0 (0)
Yes	261	82 (31.4)	41 (15.7)	98 (37.5)	40 (15.3)
Drug resistance*					
No	6738	3033 (45.0)	921 (13.7)	2114 (31.4)	670 (9.9)
Yes	667	276 (41.4)	111 (16.6)	161 (24.1)	119 (17.8)
Inpatient					
No	4649	2100 (45.2)	636 (13.7)	1441 (31.0)	472 (10.2)
Yes	2731	1190 (43.6)	397 (14.5)	831 (30.4)	313 (11.5)
Homeless					
No	6917	3116 (45.0)	969 (14.0)	2135 (30.9)	697 (10.1)
Yes	294	96 (32.7)	44 (15.0)	85 (28.9)	69 (23.5)
Drug use					
No	6822	3110 (45.6)	967 (14.2)	2079 (30.5)	666 (9.8)
Yes	307	75 (24.4)	30 (9.8)	111 (36.2)	91 (29.6)
Alcohol use					
No	6460	2911 (45.1)	908 (14.1)	1979 (30.6)	662 (10.2)
Yes	328	106 (32.3)	42 (12.8)	124 (37.8)	56 (17.1)
Prison					
No	6938	3140 (45.3)	984 (14.2)	2129 (30.7)	685 (9.9)
Yes	225	53 (23.6)	22 (9.8)	78 (34.7)	72 (32.0)
Risk factor count†					
0	6688	3081 (46.1)	950 (14.2)	2020 (30.2)	637 (9.5)
1	508	187 (36.8)	64 (12.6)	180 (35.4)	77 (15.2)
2	159	41 (25.8)	19 (12.0)	58 (36.5)	41 (25.8)
3	84	19 (22.6)	8 (9.5)	26 (31.0)	31 (36.9)
4	19	1 (5.3)	3 (15.8)	6 (31.6)	9 (47.4)
IMD quintile‡					
Mean quintile		2.45	2.40	2.41	2.15

*Drug resistance, resistance to any first-line antibiotic. †Risk factor count, cumulative number of social risk factors reported by each case. ‡IMD quintile, index of multiple deprivation quintile of Lower Super Output Area within London (lowest is most deprived).

Table 4.3: Single variable multinomial logistic regression analysis for risk factors associated with tuberculosis cases in molecular clusters of different sizes in London, 2010-2014

	Number of cases in cluster		
	2	3-20	More than 20
Variable	OR (95% CI)	OR (95% CI)	OR (95% CI)
Sex			
Female	1	1	1
Male	0.94 (0.81-1.08)	1.07 (0.95-1.19)	1.16 (0.99-1.37)*
Age group (years)			
0-14	1.79 (1.08-2.97)*	2.01 (1.35-2.99)*	2.92 (1.82-4.68)*
15-44	1	1	1
45-64	1.14 (0.95-1.36)*	1.07 (0.93-1.23)	1.00 (0.81-1.22)*
65+	0.74 (0.57-0.96)*	0.69 (0.57-0.83)*	0.47 (0.33-0.65)*
Ethnic group			
White	1	1	1
Black-Caribbean	2.01 (1.25-3.22)*	2.56 (1.77-3.72)*	2.92 (1.89-4.50)*
Black-African	1.03 (0.79-1.33)	1.12 (0.92-1.37)	1.04 (0.81-1.34)*
Black-Other	1.39 (0.72-2.69)	1.47 (0.88-2.46)*	1.50 (0.79-2.83)*
Indian	0.70 (0.54-0.89)*	0.61 (0.50-0.73)*	0.31 (0.24-0.40)*
Pakistani	0.79 (0.58-1.08)*	0.64 (0.50-0.82)*	0.65 (0.47-0.90)*
Bangladeshi	0.40 (0.27-0.60)*	0.31 (0.23-0.43)*	0.12 (0.06-0.22)*
Chinese	0.58 (0.29-1.16)*	0.48 (0.28-0.81)*	0.25 (0.10-0.65)*
Mixed /Other	0.81 (0.62-1.06)*	0.69 (0.56-0.85)*	0.41 (0.30-0.55)*
Place of birth			
Non-UK	1	1	1
UK	1.77 (1.44-2.18)*	2.55 (2.18-2.99)*	4.11 (3.38-4.99)*
Time since entry to UK (years)			
0-1	1	1	1
2-4	0.93 (0.73-1.17)	0.85 (0.70-1.03)*	0.76 (0.56-1.02)*
5-9	1.02 (0.80-1.31)	1.14 (0.94-1.39)*	1.20 (0.89-1.61)
10+	1.13 (0.90-1.41)	1.35 (1.13-1.61)*	1.22 (0.93-1.60)*
Occupation			
Other	1	1	1
None	1.09 (0.92-1.29)	1.05 (0.93-1.20)	1.54 (1.27-1.86)*
Education	1.10 (0.88-1.37)	1.14 (0.97-1.35)*	1.54 (1.21-1.96)*
Health care	1.08 (0.75-1.54)	0.89 (0.67-1.19)	0.50 (0.28-0.90)*
Pulmonary disease			
No	1	1	1
Yes	1.47 (1.28-1.70)*	1.58 (1.42-1.77)*	1.92 (1.63-2.26)*
Sputum smear			
Negative	1	1	1
Positive	1.33 (1.11-1.59)*	1.40 (1.22-1.61)*	1.75 (1.44-2.13)*
Previous diagnosis			
No	1	1	1
Yes	1.51 (1.09-2.08)*	1.50 (1.16-1.93)*	1.67 (1.19-2.35)*

Previous treatment			
No	1	1	1
Yes	3.03 (0.35-26.12)	1.20 (0.37-3.86)	-
Drug resistance†			
No	1	1	1
Yes	1.32 (1.05-1.67)*	0.84 (0.68-1.02)*	1.95 (1.55-2.46)*
Inpatient			
No	1	1	1
Yes	1.10 (0.95-1.27)*	1.02 (0.91-1.14)	1.17 (1.00-1.37)*
Homeless			
No	1	1	1
Yes	1.47 (1.02-2.12)*	1.29 (0.96-1.74)*	3.21 (2.33-4.43)*
Drug use			
No	1	1	1
Yes	1.29 (0.84-1.98)	2.21 (1.64-2.98)*	5.67 (4.13-7.78)*
Alcohol use			
No	1	1	1
Yes	1.27 (0.88-1.83)*	1.72 (1.32-2.24)*	2.32 (1.66-3.25)*
Prison			
No	1	1	1
Yes	1.32 (0.80-2.19)	2.17 (1.52-3.09)*	6.22 (4.32-8.95)*
Risk factor count‡			
0	1	1	1
1	1.11 (0.83-1.49)	1.47 (1.19-1.82)*	1.99 (1.51-2.63)*
2	1.50 (0.87-2.60)*	2.16 (1.44-3.23)*	4.84 (3.11-7.52)*
3	1.37 (0.60-3.13)	2.09 (1.15-3.78)*	7.89 (4.43-14.06)*
4	9.73 (1.01-93.64)*	9.15 (1.10-76.07)*	43.53 (5.51-344.20)*
IMD quintile§			
Mean	0.59 (0.54-0.63)*	0.86 (0.81-0.91)*	0.62 (0.57-0.68)*

*p<0.2, included in initial multivariable model.

†Drug resistance, resistance to any first-line antibiotic.

‡Risk factor count, cumulative number of social risk factors (history of homelessness, illicit drug use, alcohol misuse, imprisonment) reported by each case.

§IMD quintile, index of multiple deprivation quintile of Lower Super Output Area within London where lowest is most deprived, included as a continuous variable in multilevel model accounting for random effects of Lower Super Output Area.

At single variable analysis, cases in the largest clusters of more than 20 cases also tended to be younger (OR for age under 15 years compared to 15-44 was 2.92, 95% CI 1.82-4.68). Compared to cases of white ethnicity, black ethnic groups were more likely to be in the largest clusters, whilst other groups (Indian, Pakistani, Bangladeshi, Chinese, and people of mixed/ other ethnicities) were less likely to be in the largest clusters. The strongest association was with black-Caribbean ethnicity (OR 2.92, 95% CI 1.89-4.50). Cases in the largest clusters were also more likely to have disease that was pulmonary, sputum smear positive and resistant to at least one first-line drug. There was a further association with occupation:

Compared with all other occupations, students and those working in education (OR 1.54, 95% CI 1.21-1.96) or with no occupation (OR 1.54, 95% CI 1.27-1.86) were more likely to be in large clusters, whilst those working in healthcare were less likely (OR 0.50, 95% CI 0.28-0.90).

There was a higher preponderance of all social risk factors amongst cases in the largest clusters, with illicit drug use showing the strongest association (OR 5.67, 95% CI 4.13-7.78). Increasing number of social risk factors had progressively stronger associations with being in large clusters; the OR for having three of these risk factors was 7.89 (95% CI 4.43-14.06). The OR for an increased IMD quintile (and therefore decreasing deprivation) was 0.62 (95% CI 0.57-0.68).

Several of these associations were also evident for smaller clusters and some associations became stronger with increased cluster size. For example, for cases of black-Caribbean ethnicity, the OR was 2.01 (95% CI 1.25-3.22) for clusters of two cases, and 2.56 (95% CI 1.77-3.72) for clusters of three to 20 cases.

At multivariable analysis, factors independently associated with cases in the largest clusters after adjustment for sex were age, ethnicity, place of birth, occupation, site of disease, drug resistance, number of social risk factors and IMD (Table 4.4, Figure 4.5). Cases in the oldest age group (65+ years) had an adjusted OR of 0.52 (95% CI 0.35-0.78); aOR for being born in the UK was 2.93 (95% CI 2.28-3.77). The association between black ethnic groups and larger cluster size was maintained (aOR black Caribbean ethnicity 3.64, 95% CI 2.23-5.94), whilst the only ethnic group with significantly lower risk than the white population was Bangladeshi (aOR 0.26, 95% CI 0.13-0.50). Students and those working in education had an increased adjusted odds of being in large clusters (aOR 1.31, 95% CI 1.01-1.70), and in healthcare a decreased adjusted odds (aOR 0.47, 95% CI 0.25-0.87).

Social risk factors were included in the final model as a count, and there was a trend of increased odds with increased number of risk factors, although confidence intervals overlapped (aOR three risk factors 3.75, 95% CI 1.96-7.16; aOR four risk factors 16.64, 95% CI 1.98-139.88). Deprivation was also independently associated with being in a large cluster; the adjusted OR was 0.90 (0.83-0.97) for increased IMD quintile and therefore decreased deprivation level.

Black-Caribbean ethnicity, being born in the UK and pulmonary disease were the only factors that also had significantly elevated odds for clusters of two or 3-20 cases.

Table 4.4: Multivariable multinomial logistic regression analysis for risk factors associated with tuberculosis cases in molecular clusters of different sizes in London, 2010-2014, adjusted for random effects of Lower Super Output Area.

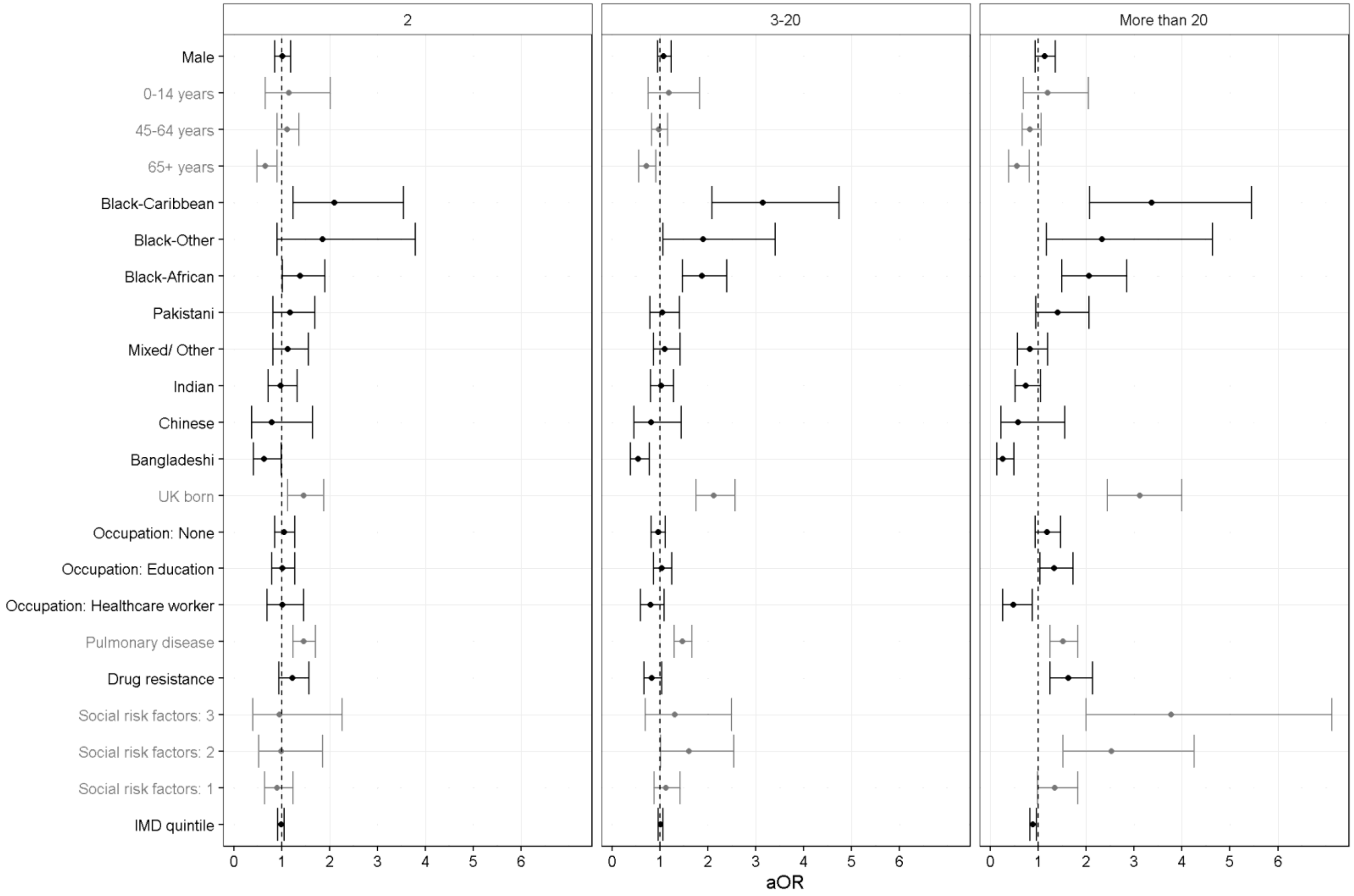
	Number of cases in cluster		
	2	3-20	More than 20
Variable	**aOR (95% CI)**	**aOR (95% CI)**	**aOR (95% CI)**
Sex			
Female	1	1	1
Male	0.98 (0.83-1.16)	1.08 (0.95-1.23)	1.14 (0.94-1.38)
Age group			
0-14	1.15 (0.65-2.02)	1.18 (0.76-1.83)	1.29 (0.75-2.22)
15-44	1	1	1
45-64	1.13 (0.92-1.39)	0.97 (0.82-1.15)	0.82 (0.64-1.04)
65+	0.68 (0.50-0.93)	0.71 (0.56-0.90)	0.52 (0.35-0.78)
Ethnic group			
White	1	1	1
Black-Caribbean	2.1 (1.25-3.55)	3.13 (2.08-4.71)	3.64 (2.23-5.94)
Black-African	1.35 (0.99-1.86)	1.86 (1.46-2.38)	2.09 (1.49-2.91)
Black-Other	1.88 (0.92-3.84)	1.92 (1.07-3.45)	2.29 (1.13-4.66)
Indian	0.95 (0.70-1.30)	1.02 (0.80-1.30)	0.78 (0.55-1.11)
Pakistani	1.08 (0.75-1.55)	1.02 (0.76-1.36)	1.51 (1.02-2.24)
Bangladeshi	0.60 (0.38-0.94)	0.53 (0.37-0.76)	0.26 (0.13-0.50)
Chinese	0.78 (0.37-1.64)	0.78 (0.44-1.40)	0.63 (0.24-1.68)
Mixed / Other	1.12 (0.81-1.54)	1.11 (0.86-1.43)	0.89 (0.61-1.29)
Place of birth			
Non-UK	1	1	1
UK	1.45 (1.12-1.87)	2.13 (1.75-2.58)	2.93 (2.28-3.77)
Occupation			
Other	1	1	1
None	1.04 (0.85-1.27)	0.96 (0.82-1.12)	1.18 (0.94-1.49)
Education	1.04 (0.82-1.31)	1.04 (0.87-1.24)	1.31 (1.01-1.70)
Health care	1.00 (0.69-1.45)	0.82 (0.60-1.11)	0.47 (0.25-0.87)
Pulmonary disease			
No	1	1	1
Yes	1.46 (1.24-1.72)	1.48 (1.30-1.68)	1.47 (1.21-1.79)
Drug resistance*			
No	1	1	1
Yes	1.25 (0.96-1.61)	0.82 (0.65-1.03)	1.75 (1.34-2.28)
Risk factor count†			
0	1	1	1
1	0.83 (0.59-1.15)	1.12 (0.88-1.42)	1.36 (0.99-1.87)
2	1.00 (0.53-1.89)	1.59 (1.00-2.52)	2.46 (1.45-4.18)
3	0.86 (0.35-2.12)	1.28 (0.67-2.45)	3.75 (1.96-7.16)
4	4.35 (0.39-48.5)	4.13 (0.45-37.5)	16.64 (1.98-139.9)
IMD quintile‡	0.98 (0.91-1.04)	1.00 (0.95-1.06)	0.90 (0.83-0.97)

*Drug resistance, resistance to any first-line antibiotic.

†Risk factor count, cumulative number of social risk factors (history of homelessness, illicit drug use, alcohol misuse, imprisonment) reported by each case.

‡IMD quintile, index of multiple deprivation quintile of Lower Super Output Area within London where lowest is most deprived, included in multilevel model accounting for random effects of Lower Super Output Area.

Figure 4.5: Forest plot of adjusted odds ratios and 95% confidence intervals from multivariable multinomial logistic regression analysis for risk factors associated with tuberculosis cases in molecular clusters of different sizes in London, 2010-2014 (Table 4.4).

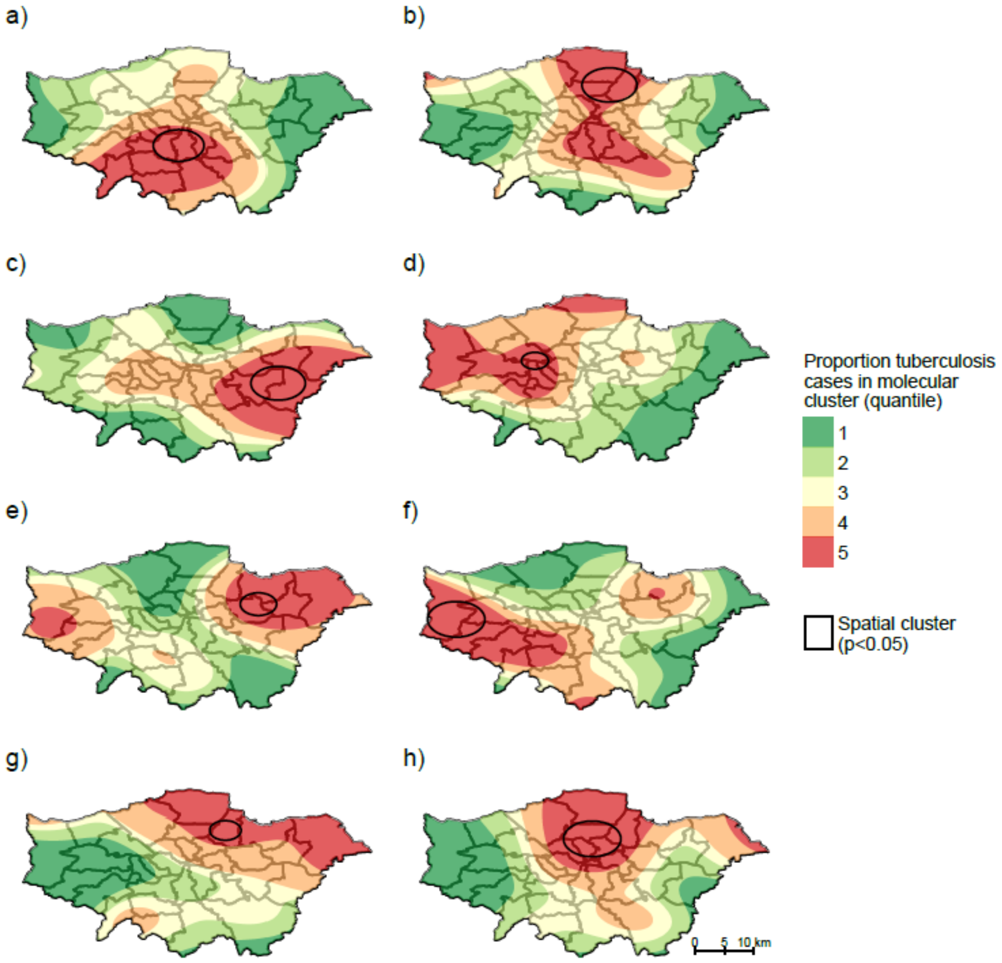


4.5.3 Spatial clusters of cases in large molecular clusters

I used SaTScan to test for spatial clustering in the 20 molecular clusters that had more than 20 cases. A total of 25 significant spatial clusters (p<0.05) were identified, with at least one significant spatial cluster in 18 (90%) of the molecular clusters, and eight of the spatial clusters included more than ten cases. Characteristics of the spatial clusters are presented in Table 4.5. The locations of these eight spatial clusters were in areas of high incidence on the smoothed incidence maps (Figure 4.6). They tended to occur in more deprived areas; the median IMD rank of the 4,970 LSOAs within London for areas within the clusters was 1,110, compared to 2538.5 for areas not in clusters.

Figure 4.6: Locations of significant spatial clusters of cases within eight molecular clusters (denoted a-h) of tuberculosis in London (2010-2014), overlaid on smoothed incidence maps.

Circles represent areas of significant spatial clustering (p<0.05) with more than ten cases of the given molecular cluster, compared to the general distribution of tuberculosis cases. The proportions of cases in the given molecular cluster compared to all other tuberculosis cases represented through kernel density estimation (bandwidth 5km). Contains Ordnance Survey data © Crown copyright and database right 2014.


a)
b)
c)
d)
e)
f)
g)
h)
Proportion tuberculosis cases in molecular cluster (quantile)
1
2
3
4
5
Spatial cluster (p<0.05)
0 5 10 km

4.5.4 Spatial distribution of large clusters

I used the interactive mapping application developed in Chapter 3 to investigate the characteristics of the above eight large molecular clusters with significant spatial clusters including more than ten cases.

Table 4.5 briefly summarises the main features of each of these clusters; screen shots are additionally presented in Appendix 10.4. In most of the clusters, one or two ethnic groups clearly predominated in the area of spatial clustering; and some showed associations with other risk factors such as history of drug abuse or homelessness. Two clusters also showed an apparent change in spatial distribution over time: Cluster c was more spatially aggregated in more recent years, whilst cluster g appeared to become more dispersed over time.

Table 4.5: Description of molecular clusters of tuberculosis with significant spatial clustering in London, 2010-2014.

		Spatial cluster					
Molecular cluster	N total	Radius (km)	N observed	N expected	Relative risk	p	Description
a	28	4.3	20	1.7	38.9	<0.01	Predominantly black-African ethnicity, many of the cases in the area of spatial clustering are either students or work in education.
b	48	4.6	15	3.3	6.14	0.012	Majority black-African or black-Caribbean ethnicity, with an association with history of homelessness in the spatial cluster.
c	24	4.6	11	0.9	21.9	<0.01	Most of these cases were born in southern Asia. The spatial cluster seems to have developed recently, with initial cases more dispersed.
d	102	2.3	13	2.4	6.2	0.016	Overall, the majority of cases are black-African and aged 15-44 years; however, in the spatial cluster in the west, several of the cases are in older age groups.
e	71	3.0	28	7.7	5.4	<0.01	Most cases are of Pakistani or Indian ethnicity, and born outside of the UK.
f	79	4.8	22	7.1	3.9	0.018	Most cases are of Pakistani or Indian ethnicity, and born outside of the UK.
g	28	2.6	11	0.7	25.0	<0.01	Cluster affected a mix of ethnic groups, with no obvious common factors linking them. The majority of cases in the spatial cluster occurred earlier, with latter cases more dispersed.
h	85	4.8	50	7.6	14.6	<0.01	Many of the cases were born in the UK and had a history of drug use.

4.5.5 Prediction of large clusters

There were 260 clusters that met the inclusion criteria for this analysis (i.e. having no cases for the first 12 months of the study period, and at least two cases reported from the start of 2011 to the end of 2012). Of these, 32 (12%) were classified as large, having at least five cases at the end of the 24 month follow-up; 228 were small. The characteristics of the initial two cases in the small and large clusters were similar (Table 4.6). Of the variables considered, only the distance between the cases showed a significant association with progression to large cluster at single variable logistic regression analysis: The initial two cases in large clusters were more likely to be within 2km of each other compared to initial two cases in small clusters (OR for ≤ 2km 2.51; 95% CI 1.06-5.64).

Table 4.6: Characteristics of large (N ≥ 5 cases) and small (N < 5 cases) molecular clusters of tuberculosis and results of single variable logistic regression analysis for cluster-level risk factors associated with cluster size in London, 2010-2014.

Variable		Total clusters	Large clusters N (%)	Small clusters N (%)	OR (95% CI)
Total			**32 (12.3)**	**228 (87.7)**	
Male sex	No cases	41	5 (15.6)	36(15.8)	1
	≥ 1 case	219	27 (84.4)	192 (84.2)	1.01 (0.39-3.14)
Aged 65 years or older	No cases	225	28 (87.5)	197 (86.4)	1
	≥ 1 case	35	4 (12.5)	31 (13.6)	0.91 (0.26-2.51)
Black-Caribbean ethnicity	No cases	242	29 (90.6)	213 (93.4)	1
	≥ 1 case	18	3 (9.4)	15 (6.6)	1.47 (0.33-4.79)
Black-African ethnicity	No cases	174	21 (65.6)	153 (67.1)	1
	≥ 1 case	86	11 (34.4)	75 (32.9)	1.07 (0.47-2.29)
Other black ethnicity	No cases	254	31 (96.9)	223 (97.8)	1
	≥ 1 case	6	1 (3.1)	5 (2.2)	1.44 (0.07-9.31)
Born in UK	No cases	186	23 (71.9)	163 (71.5)	1
	≥ 1 case	74	9 (28.1)	65 (28.5)	0.98 (0.41-2.17)

Occupation: Education	No cases	195	26 (81.2)	169 (74.1)	1
	≥ 1 case	65	6 (18.8)	59 (25.9)	0.66 (0.24-1.59)
Pulmonary disease	No cases	45	5 (15.6)	40 (17.5)	1
	≥ 1 case	215	27 (84.4)	188 (82.5)	1.15 (0.45-3.55)
Drug resistance*	No cases	227	28 (87.5)	199 (87.3)	1
	≥ 1 case	33	4 (12.5)	29 (12.7)	0.98 (0.28-2.73)
Any social risk factor†	No cases	222	26 (81.2)	196 (86)	1
	≥ 1 case	38	6 (18.8)	32 (14.0)	1.41 (0.50-3.51)
Lowest IMD quintile‡	No cases	143	18 (56.2)	125 (54.8)	1
	≥ 1 case	117	14 (43.8)	103 (45.2)	0.94 (0.44-1.98)
Distance between cases	> 2km	215	22 (68.8)	193 (84.6)	1
	≤ 2km	45	10 (31.2)	35 (15.4)	2.51 (1.06-5.64)
Time between cases	> 90 days	178	20 (62.5)	158 (69.3)	1
	≤ 90 days	82	12 (37.5)	70 (30.7)	1.35 (0.61-2.89)

*Drug resistance, resistance to any first-line antibiotic.
†Any social risk factor (history of homelessness, illicit drug use, alcohol misuse, imprisonment) reported by a case in cluster.
‡ Number of cases in lowest IMD quintile, (index of multiple deprivation quintile) of Lower Super Output Area within London, where lowest is most deprived.
CI, confidence interval.

4.6 Discussion

4.6.1 Summary of findings

In this study, I have presented results from the first five years of routine molecular strain typing of tuberculosis by 24 locus MIRU-VNTR in London. There were 20 molecular clusters that had more than 20 cases of tuberculosis notified between 2010 and 2014. These clusters accounted for 795 (11%) of all typed cases notified during this period. Compared with sporadic cases, those in large molecular clusters tended to occur closer together in time and space, and shared epidemiological characteristics. However, characteristics of initial cases were not highly predictive of the number of cases that would occur in a molecular cluster over two years.

4.6.2 Interpretation of results – molecular and spatial clustering

The results of this study indicate that a substantial proportion of new tuberculosis cases in London result from local transmission. Molecular, spatial and epidemiological characteristics were shared by cases in numerous large clusters. One of the molecular clusters described in this study (cluster h) is part of a known outbreak of isoniazid-resistant disease that was first identified in 1999 (and is described in detail in Chapter 5),[180] but this is the first analysis to suggest that multiple similar outbreaks may be ongoing.

Cases in large molecular clusters were more likely to have multiple social risk factors, be of black ethnicities, born in the UK, and live in more deprived areas of London. Small clusters (pairs of cases) and those of intermediate size (3-20 cases) were also associated with black-Caribbean ethnicity, being born in the UK, and pulmonary disease. There was also some association between large clusters and occupation. Large clusters were more likely to include students and those involved in education, which may suggest that outbreaks in schools and universities can spread widely in these settings or in extensive social networks involving students. However, they were less likely to involve healthcare workers. This indicates that there was limited nosocomial transmission, and that when transmission involving a healthcare worker did occur it was usually an isolated incident rather than part of a large outbreak.

The majority of large molecular clusters identified in this study exhibited significant spatial clustering, providing further evidence for transmission. There was a link between large clusters and social deprivation: Spatial clusters tended to be in more deprived areas of London, and the IMD of case residential area was independently associated with being in a large molecular cluster, after accounting for individual risk factors. Studies in other settings including Peru, northern England, Japan, and the USA have also investigated tuberculosis clustering using molecular and spatial data.[190-193] Various methods have been used to assess spatial clustering but all identified areas of likely localised tuberculosis transmission or 'hot spots'.

Previous analyses of MIRU-VNTR strain typing and surveillance data in various settings have sought to determine the proportion of cases that were part of a molecular cluster or identify risk factors for clustering.[186,194-201] A systematic review of 27 articles found that estimates varied from 0 to 63%, and the analysis of strain typing data from London to 2012 identified a clustering rate of 46%.[186,202] This indicates that the rate of molecular clustering identified in this study (56%) was relatively high, and that a higher proportion of cases in London may be linked than formerly understood.

4.6.3 Interpretation of results – prediction of large clusters

The final part of this study attempted to use characteristics of the initial two cases in a molecular cluster to predict which clusters would grow to at least five cases within 24 months. Although there was some evidence that initial cases in large clusters tended to occur closer together in space than those in small clusters, demographic and clinical characteristics were not significantly different. This indicates that prediction of the size of clusters based on the first two linked cases is not reliable in this context.

These results differ from similar attempts to predict cluster growth in the Netherlands, USA and London, in which significant associations with initial cluster characteristics were found.[186,193,203] The previous studies did not assess distance between the initial two cases as a potential predictor of cluster growth, although the USA study used scan statistics to define potential clusters. The studies in the Netherlands and USA used less specific molecular cluster tests (RFLP typing and 12-locus MIRU-VNTR respectively). They also based their analysis on the characteristics of the first three (rather than two) patients, and attempted to predict growth to at least six cases in two years. However, they benefitted from using larger data sets, as the studies covered the whole country and routine molecular strain typing had been ongoing for longer. Similar to the analysis presented here, they also used a 'wash out' period (of 24 months), to exclude cases with links to those before the surveillance period, and a 24 month follow-up.

The previous analysis of data from London was based on only three years of strain typing data and therefore did not include a 'wash-out' or minimum follow-up period.[186] Instead, this study classified the first two cases that were detected after the start of routine molecular strain typing as the first two cases in the molecular cluster. It is therefore likely that several of the cases included were not initial cluster cases, but were in fact linked to cases occurring before the start of the surveillance period. The relationships that they identified may therefore have related to cases occurring later in the cluster rather than initial cases.

4.6.4 Implications for policy and practice

This study can be used to inform practice for tuberculosis cluster investigations. The results imply that detailed investigations of molecular clusters could be beneficial in preventing large chains of transmission through interventions such as contact tracing and screening. The interactive mapping tool developed in Chapter 3 provided descriptions of the clusters and could be used to assist with these investigations. This study also demonstrated the potential benefit of incorporating spatial cluster detection tests, in this case the spatial scan statistic, into routine analysis of molecular cluster data. These analyses could be used to assist with prioritising clusters for further investigation, as use of simple thresholds has previously been ineffective in making these decisions.[164,185]

Further work is needed to determine whether initial case characteristics can reliably be used to predict cluster growth. Cluster investigators should therefore focus efforts on clusters with more than two cases, whilst monitoring these smaller clusters for evidence of rapid increases in numbers of cases.

This study also highlighted populations most at risk of being part of large clusters, and therefore has implications for the control of tuberculosis in London and other high incidence cities. Targeting interventions to deprived areas should be a priority for reducing transmission, whilst efforts should also be made to raise awareness of the disease amongst at-risk groups such as those of black ethnicities born in the UK. An example of such an intervention is the *Find and Treat* mobile radiography unit which actively screens for cases in vulnerable populations in London and provides support to help patients complete treatment.[204] Continued support for this service is therefore a key component of tuberculosis control in London.

4.6.5 Study strengths

This study was based on routine surveillance data and therefore included all cases of tuberculosis in London that were successfully typed by MIRU-VNTR to at least 23 loci from 2010 to 2014. As a result it provides a good representation of the population of tuberculosis cases in the city. There was a low level of missing information in the variables used in the risk factor analysis (Table 4.3). It is therefore unlikely that the results were affected by bias arising from systematic differences in the completion of data for different groups of patients. The use of highly discriminatory molecular typing data allowed cases to be grouped into molecular clusters which may be linked through transmission. These results were reported using relevant parts of the STROME-ID (Strengthening Reporting of Molecular Epidemiology for Infectious diseases) and STROBE (Strengthening Reporting of Observational Studies in Epidemiology) statements.[205,206]

In the analysis of molecular clustering, a multinomial logistic regression approach was used to identify risk factors. The advantage of this method was that it allowed associations to be assessed for different sizes of molecular cluster, as opposed to using a binary 'large or small' outcome. This could potentially enable identification of more subtle changes in risk factors for different sizes of molecular cluster which would not be identified when using a binary outcome. A linear regression approach may have provided even greater distinctions between cluster sizes, but this method could not be used because residuals of these models deviated substantially from normality owing to the skewed distribution of cluster sizes. However, continuous cluster size outcomes were considered through visual comparisons of box plots.

Another strength of this study is that it combined molecular data with spatial clustering analyses. Cases with indistinguishable MIRU-VNTR types are more likely to be linked through transmission than those that do not; but shared molecular strain types may also be indicative of endemic strains. Incorporation of spatial clustering analyses therefore adds further evidence for transmission, amongst at least a proportion of the cases in these large molecular clusters, and may help to prioritise clusters for further investigation. Two methods of spatial analysis were used, scan statistics and kernel density estimation. These outputs showed a high level of correlation, with significant spatial clusters arising in areas with highest density on the smoothed incidence maps. They were also in agreement with the risk factor analysis: Areas with significant spatial clustering tended to have a lower IMD, which was also a risk factor for being part of a large cluster. This indicates that the spatial cluster analyses provided a reliable means of identifying areas with unusually high levels of clustering and therefore potential ongoing transmission.

4.6.6 Study limitations

A limitation of this study is that I had to restrict the analysis to cases of tuberculosis which had been typed by MIRU-VNTR to at least 23 loci. I therefore excluded 7,522 cases that were not culture confirmed or typed, approximately half of the cases notified in London during this time period. For the multinomial logistic regression analysis of factors associated with large molecular clusters, I censored the study period, starting in 2010 when routine molecular strain typing was introduced and was therefore unable to ascertain molecular links with earlier cases. This will have resulted in misclassification of some cases as unclustered which did not have a unique strain, and underestimate of the number of cases in some clusters. Overall, these limitations are therefore likely to have led to an underestimate of the proportion of cases that were in large molecular clusters. Conversely, excluding cases that were not culture confirmed may have led to an overestimate of clustering: Non-culture confirmed cases are less likely to have pulmonary disease, which was found to be a risk factor for being in large clusters.

Another potential limitation of this study is that I did not take into account time between cases in the analysis of clustering, for example through space-time scan statistics. This is because tuberculosis has a highly variable incubation period, meaning that it is difficult to select an informative maximum temporal clustering window.[207] This may have resulted in sporadic cases that occurred close in space being classified as a spatial cluster when they were not linked through transmission. However, I considered the temporal distribution of cases in molecular clusters by examining the median interval between case notification dates. This provided some evidence that cases in larger molecular clusters occurred closer together in time than those in smaller clusters. A true estimate of serial intervals would require ascertainment of epidemiological links between cases to establish chains of transmission.

The risk factor analysis was limited because it was unable to assess the importance of other potential factors affecting tuberculosis transmission, for example HIV status, which are not currently collected in surveillance data. However, estimates show that coinfections of HIV and tuberculosis is relatively uncommon in this setting, contributing less than 5% of tuberculosis cases.[208]

There were also several limitations to the analysis predicting cluster growth. The length of the 'wash-out' period was restricted to 12 months (whereas previous studies have used 24 months[193,203]) because only five years of surveillance data were available. This therefore increases the chance that some of the first cases in molecular clusters that were identified were not the initial cases in the clusters. The characteristics of the initial cases were described at the cluster level (i.e. 'at least one' of the two cases in the cluster having the characteristic compared to neither of the clusters having the characteristic), and the analysis was restricted to characteristics that showed an association with being in a large cluster in the initial analysis. These approaches were made to simplify the analysis and limit the number of comparisons made, but could have resulted in failure to identify associations.

4.6.7 Future directions

This analysis points to potential improvements in systems for tuberculosis cluster investigation. Incorporating mapping and routine spatial cluster analyses such as those described here into surveillance systems could assist with future investigations. A prospective evaluation of this would be required to ascertain the utility of this in practice and help to develop specific recommendations for action.

The analysis predicting cluster size used a cut off of 2 km to define proximity of initial cases. This was an arbitrary value and further work is needed to explore alternative values and ascertain what is most appropriate. It would also be useful to repeat the analysis predicting cluster size using initial case characteristics when more years of strain typing data are available. As well as including more molecular clusters in the analysis, this would allow a longer 'wash out' period to be used, increasing the validity of the assumption that the first case identified by strain typing was the first to have occurred in the cluster.

Other work arising from this study could aim to identify which of the components of the IMD may be contributing to tuberculosis transmission. This would be useful to inform environmental and housing interventions, such as improving ventilation and reducing overcrowding. Finally, whole genome sequencing of tuberculosis isolates could add further resolution to networks suggested by molecular clusters, for example by identifying 'super-spreaders', which are particularly infectious individuals who infect large numbers of secondary cases.[209] This could provide evidence to support or refute transmission in some instances and direct the focus of intensive investigations.[210]

Box 4.1: Summary of Chapter 4.

- Large molecular clusters contribute substantially to the burden of tuberculosis in London, with more than a tenth of cases being part of a cluster of more than 20 cases.
- Cases in large clusters were closer together in space and time than other cases.
- People in large clusters were generally more likely to be of black-African or black-Caribbean ethnicity and born in the UK, have multiple social risk factors such as history of illicit drug use, and live in more deprived areas of London.
- These results can be used to assist with targeting of interventions to at-risk populations.
- Further research is required to ascertain whether characteristics of the first individuals identified in a molecular cluster can be used to make reliable predictions about which clusters will develop into very large clusters.

5 EPIDEMIOLOGY AND SPATIAL ANALYSIS OF A LARGE ISONIAZID-RESISTANT TUBERCULOSIS OUTBREAK

5.1 DESCRIPTION OF CHAPTER CONTENTS

In this chapter, I describe the epidemiology and spatial distribution of a large outbreak of isoniazid-resistant tuberculosis in England and Wales. The outbreak has now been ongoing for 20 years and comprises over 500 cases, placing it as the largest documented outbreak of drug-resistant tuberculosis to date. Here, I combine data from multiple sources to compile a database that allows a complete description of all reported outbreak related cases for the first time. The analysis focuses on characterising the evolution of the outbreak in space and time: I use k-function analysis, kernel density estimation and an animated time series to describe the spatial distribution of the outbreak; and investigate changes in demographic groups affected using proportions and tests for trend. I also summarise treatment outcomes and use data from the pan-London outreach service, Find and Treat, to quantify the impact of its screening and case management system to control of the outbreak. Finally, I conduct a systematic literature review to identify other large drug-resistant tuberculosis outbreaks and compare their characteristics. I discuss implications for control of this and other tuberculosis outbreaks.

5.2 STUDY RATIONALE AND INTRODUCTION

5.2.1 Importance of drug-resistant tuberculosis outbreaks

Development of resistance to anti-tuberculosis drugs has important implications for public health as well as for clinical management of the disease. The most common form of resistance is to isoniazid, and in the UK 6-7% of cases were resistant to this drug over the last ten years.[5,211] Resistance to a single first-line drug such as isoniazid is a precursor for development of multi-drug resistance, which poses a particular threat to disease control because it cannot be managed using standard treatment regimens. Proliferation of drug-resistant strains in outbreaks therefore increases the risk of MDR-tuberculosis spreading widely and may result in patients having extended durations of symptoms, infectiousness, and problems completing treatment.

Drug resistance phenotypes can also be used to identify cases that may be linked through transmission. As discussed in Chapter 4, molecular methods, combined with spatial analyses, can now be used to identify cases that are likely to be part of large outbreaks. Historically, however, highly discriminatory molecular techniques were not routinely available for detailed characterisation of pathogens. Identification of links between cases therefore relied on less specific methods including drug susceptibility testing. In this chapter I describe an outbreak that began before molecular testing was conducted routinely and may therefore not have been identified as promptly if the strain had not been drug-resistant.

5.2.2 History of the outbreak

This outbreak was first identified in a hospital in north London in 2000 where three young men were diagnosed with isoniazid mono-resistant tuberculosis within a week.[180] Subsequent restriction fragment length polymorphism (RFLP) analysis confirmed that the cases shared an indistinguishable strain type of the Euro-American lineage. Analysis of laboratory data showed that the prevalence of isoniazid resistance in strains from this hospital was higher than in other parts of England and Wales, and had been increasing (8% in 1998 and 15% in 1999 at the hospital, compared with 6% nationally).[180] These observations suggested the existence of a local cluster, and an Outbreak Control Committee (OCC) was established. Subsequent retrospective strain typing of isolates available at the time identified a further 15 cases, with the earliest diagnosis in January 1995 in an individual born in Nigeria.[180]

An analysis of the first 70 confirmed cases in London, to end of 2001, found that patients with this strain were more likely than other tuberculosis cases to be male; of white or black Caribbean ethnicity, and born in the UK.[180] There was also a strong geographical link to north London, with one health district in the area accounting for 44% of cases but only 7% of the population of the city. Prison contact and recreational drug use were identified as common features of cases.

Key recommendations made by the OCC at this time included promoting interagency working, expanding use of DOT, and enhancing control in prisons.[212] An analysis of rates of infection amongst patient contacts found high overall rates of transmission of disease in this outbreak (11%) compared with other documented outbreaks (0.7-2%).[213] This led to speculation that the outbreak may have been caused by an unusually virulent strain, and prompted recommendations for enhanced contact tracing to include casual contacts of smear positive cases.

Epidemiological characteristics of this outbreak were last described for cases up to 2006.[214] This case-control study compared characteristics of 293 outbreak-related with 17,743 non-outbreak related cases of tuberculosis, with similar results to the analysis of the first 70 cases. Multivariable logistic regression analysis found that cases were significantly more likely to live in north-central London; be young (aged 15-34 years); UK born; of white or black Caribbean ethnicity; currently in prison; unemployed; a drug dealer or sex worker. Treatment completion gradually improved over time (from 55% for cases diagnosed to the end of 2002 to 65% for cases diagnosed in 2006), but was lower than for controls (83% in 2005).

As cases continue to be reported 20 years after the first identified case, this now represents one of the largest outbreaks of drug-resistant tuberculosis to be documented. A complete analysis of the epidemiology and spatial distribution of all cases reported to date would allow assessment of changes in risk groups over time, and have implications for the control of this and other large outbreaks of tuberculosis.

5.3 AIMS AND OBJECTIVES

Aim: To collate information about the epidemiological characteristics of this large outbreak of isoniazid-resistant tuberculosis in England and Wales, describe its evolution in space and time, and place it in context of other large drug-resistant outbreaks of tuberculosis.

The objectives of this study were to:

1. Combine data from multiple sources to generate a complete outbreak database.
2. Assess changes in the demographics and risk factors of patients affected by outbreak strain.
3. Quantify the contribution of the pan-London outreach service, Find and Treat, to control of the outbreak.
4. Describe the spatial distribution of the outbreak.
5. Quantify the impacts of the outbreak, its clinical features and treatment outcomes.

6. Compare this with other large outbreaks of drug-resistant tuberculosis.
7. Discuss implications for management of this and other large tuberculosis outbreaks.

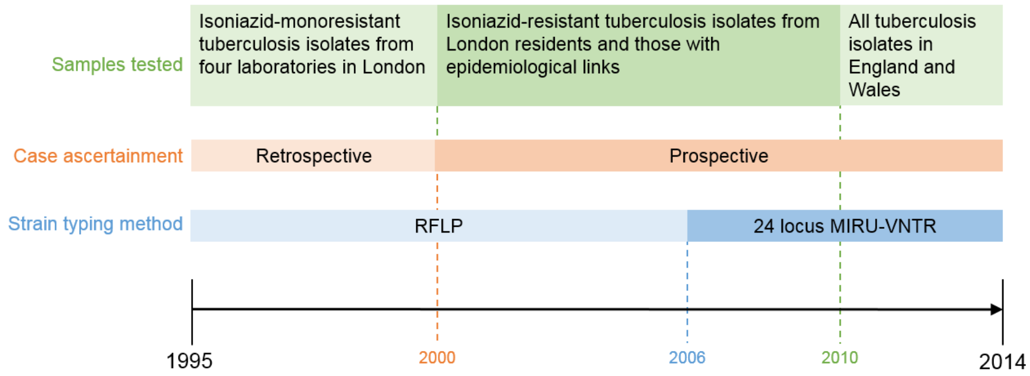
5.4 METHODS

5.4.1 Case definition and data sources

Microbiological testing procedures and surveillance systems for tuberculosis in England and Wales have evolved considerably since this outbreak was initially detected. As a consequence, methods of data collection and typing, and therefore the case definition and patient information collected have changed. This is described below and summarised in Figure 5.1.

Cases were defined as individuals diagnosed from 1995-2014 in England and Wales with an *M. tuberculosis* isolate that was indistinguishable from the outbreak strain. The outbreak strain was initially characterised in 2000 using RFLP analysis. Cases prior to this were ascertained retrospectively: Microbiological databases of four laboratories serving the area where the first cases were reported were reviewed, and isoniazid monoresistant organisms identified were typed by RFLP. After 2000, cases were ascertained prospectively by RFLP typing of all isoniazid-resistant isolates from patients resident in London, and those with known epidemiological links. From 2006, due to a change in routine practice, 24 locus MIRU-VNTR typing was used instead of RFLP typing. From 2010 onwards, strain typing by MIRU-VNTR was conducted on all tuberculosis isolates in England and Wales regardless of links to London. Strain typing was conducted at the PHE (formerly Health Protection Agency) National Mycobacterium Reference Laboratory.

Figure 5.1: Case ascertainment, sampling and strain typing methods for defining cases in the isoniazid-resistant outbreak, 1999-2014.


Samples tested
Isoniazid-monoresistant tuberculosis isolates from four laboratories in London
Isoniazid-resistant tuberculosis isolates from London residents and those with epidemiological links
All tuberculosis isolates in England and Wales
Case ascertainment
Retrospective
Prospective
Strain typing method
RFLP
24 locus MIRU-VNTR
1995
2000
2006
2010
2014

I extracted information on outbreak-related cases from multiple data sources. Demographic, clinical, microbiological and treatment outcome data were derived from a bespoke outbreak database, and from the electronic surveillance systems for London (the London Tuberculosis Register, LTBR) and the rest of England and Wales (ETS). Information on social risk factors (drug use, links to prisons including patients who were in prison at diagnosis, homelessness, alcohol dependence and mental health concerns) were collected initially in the bespoke outbreak database, and from 2009 in surveillance systems.

I also extracted information on outbreak cases from data collected by Find and Treat, a pan-London tuberculosis outreach service.[215] This service aims to identify cases of tuberculosis in 'hard-to-reach' populations, typically those with social risk factors, and support them to complete treatment. I identified outbreak-related cases who had been screened and managed by the service, and used data to supplement information on social risk factors from surveillance systems.

Databases were combined on the basis of unique identifiers, patient names and dates of birth. Patients with multiple episodes of tuberculosis with the outbreak strain were included once only, with the first period retained.

Ethical approval was not required for this study because PHE has Health Research Authority approval to hold and analyse national surveillance data for public health purposes.

5.4.2 Epidemiological analysis

I plotted annual numbers of cases from 1995 to 2014 as an epidemic curve. I described demographic, clinical and microbiological characteristics in counts and proportions, and used χ^2 test for trend to identify changes over time. For social risk factors, I calculated overall proportions and identified changes over time by plotting proportions of cases reporting risk factors by year.

I analysed treatment outcomes at 12 months and at final known outcome. The 12 month treatment outcome analysis included non-MDR cases who were notified between 2002 and 2013. MDR cases were excluded from this analysis because their planned treatment regimen exceeds 12 months; cases before 2002 were excluded as they did not have a date of treatment outcome recorded, and cases reported after 2013 were excluded because outcomes had not been collected at the time of the analysis. Final known treatment outcomes included cases notified before 2002, MDR cases notified before 2013, and incorporated 12- or 24-month treatment follow-up where appropriate. I tested for changes in proportions of cases completing treatment over time and used the χ^2 test to compare proportions of cases with and without social risk factors who completed treatment.

I used Find and Treat data to identify the proportion of cases notified in London who had been screened by the service (cases from 2005 onwards), and referred for case management (cases from 2007 onwards). I calculated proportions of patients referred to Find and Treat who had social risk factors, and used χ^2 tests to compare the rates of treatment completion in patients referred to the service with those who were not.

5.4.3 Spatial analysis

I determined case locations using residential postcodes where available. Prison or clinic postcodes were used where relevant for patients diagnosed whilst in prison or with no fixed abode. I geocoded case postcodes and plotted numbers of cases nationally according to PHE region; and calculated rates within London by borough, using denominator population data from the 2001 census.

To visualise the spatio-temporal progression of the outbreak in London, I produced an animation of case locations by day of notification. This involved creating maps of case locations for each day of the outbreak, and altering the size and opacity of the point on the map according to the time that had passed since the case was notified. I then combined these slides into a video file using ffmpeg.[216] I also compared the spatial intensity of case locations in each five-year period during the outbreak through a series of smoothed incidence maps, produced using kernel density estimation.

I further explored the spatial point pattern of cases in London through k-function analysis. The k-function is a global method for detecting spatial clustering and involves comparison of the observed and expected number of cases within a range of distances from an arbitrary case location.[55] First, I tested the hypothesis that the points were completely spatially random by comparing the k-function of the observed point locations with the function generated by 99 simulated point patterns. I then tested the hypothesis that the locations of cases in the first ten years of the outbreak were part of the same spatial distribution as those in the second decade by calculating their cross k-function. This is the number of points from one distribution within a range of distances of a typical point from the other distribution.[217] The observed function was compared with the functions defined by 99 simulations based on random re-labelling of the joint spatial distribution of points to the two time periods. If the observed function lay within the upper and lower bounds of these limits, this would be consistent with the null hypothesis that the points were part of a common spatial distribution.

Spatial analyses were conducted using the R package *spatstat*.[218]

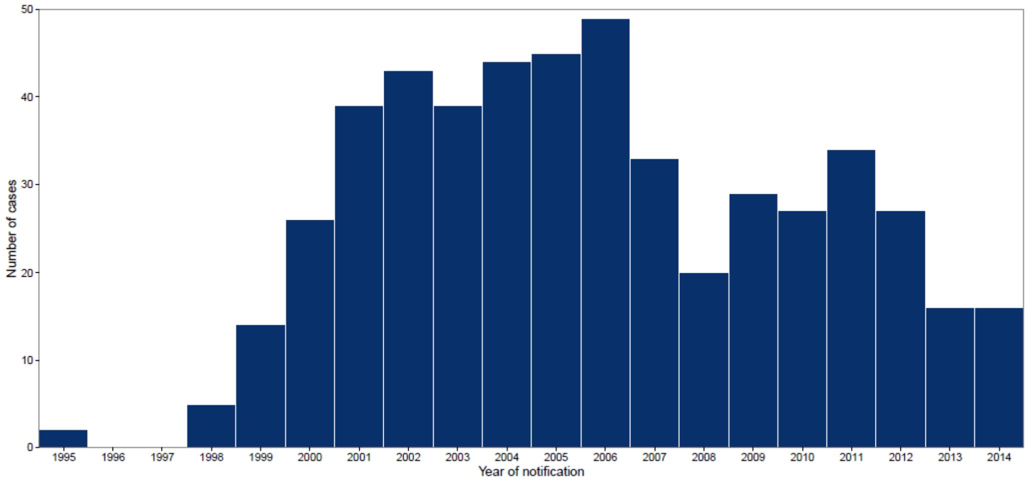
5.4.4 Comparison with other large drug-resistant tuberculosis outbreaks

I searched Embase and Medline for articles that described outbreaks of drug-resistant tuberculosis. The search terms were “drug-resistant tuberculosis”, “multidrug-resistant tuberculosis” or “extensively drug-resistant tuberculosis” combined with “disease outbreaks”, “outbreak”, “epidemic” or “cluster”. The search was run on 16 February 2015 and did not include any restrictions according to language or article type. I sought to identify large outbreaks with known epidemiological links and therefore included only reports describing outbreaks with at least 20 cases, and excluded articles describing clusters linked only through microbiological analysis.

5.5 RESULTS

From 1995 to 2014, 508 cases with the isoniazid-resistant tuberculosis outbreak strain were identified. The epidemic curve (Figure 5.2) shows that, after initial ascertainment of the outbreak in 2000, the number of cases rose steeply, reaching a peak of 49 cases in 2006. After a subsequent decrease in numbers, there appears to have been a second peak in 2011 (34 cases). This was followed by a decline in incidence from 2011 to 2013, and the number of new cases was stable in the last two years (16 cases per year).

Figure 5.2: Number of cases in the isoniazid-resistant tuberculosis outbreak by year, England and Wales, 1995-2014.



5.5.1 Demographics

Table 5.1 presents the demographic characteristics of the cases in this outbreak by five year period. The majority of cases (71%) were male; of white (39%), black Caribbean (26%) or black African (13%) ethnicity; and born in the UK (60%). There were no significant changes in proportions of these characteristics over time (χ^2 trend p=0.97, 0.39, and 0.28 for sex, ethnicity and place of birth respectively).

The median age of the cases increased over time from 25 years (range 6-71, IQR 21-28) in 1995-1999 to 42 in 2010-2014 (range 12-79, IQR 31-49). The breakdown by age groups shows that the proportion of cases aged 25-44 decreased from 81% in the first time period to 31% in the last period. The 45-64 year old age group comprises the largest proportion of the recent cases, a significant increase compared to latter cases (39% in 2010-2014, compared to 4.8% in 1995-1999, χ^2 trend p<0.01).

Table 5.1: Demographic characteristics of cases in the isoniazid-resistant tuberculosis outbreak, England and Wales, 1995-2014.

	All	1995-1999	2000-2004	2005-2009	2010-2014
	N (%)	N (%)	N (%)	N (%)	N (%)
All cases	508	21	191	176	120
Sex					
Male	360 (70.9)	13 (61.9)	139 (72.8)	122 (69.3)	86 (71.7)
Female	148 (29.1)	8 (38.1)	52 (27.2)	54 (30.7)	34 (28.3)
Age (years)					
Median	36	25	35	36	42
<15	9 (1.8)	1 (4.8)	1 (0.5)	5 (2.8)	2 (1.7)
15-24	70 (13.8)	9 (42.9)	31 (16.2)	22 (12.5)	8 (6.7)
25-34	150 (29.5)	8 (38.1)	60 (31.4)	53 (30.1)	29 (24.2)
35-44	145 (28.5)	1 (4.8)	58 (30.4)	55 (31.3)	31 (25.8)
45-64	116 (22.8)	1 (4.8)	34 (17.8)	35 (19.9)	46 (38.3)
>65	18 (3.5)	1 (4.8)	7 (3.7)	6 (3.4)	4 (3.3)
Ethnic group					
White	199 (39.2)	8 (38.1)	59 (30.9)	73 (41.5)	59 (49.2)
Black-Caribbean	134 (26.4)	4 (19.0)	59 (30.9)	49 (27.8)	22 (18.3)
Black African	67 (13.2)	3 (14.3)	34 (17.8)	20 (11.4)	10 (8.3)
Indian	22 (4.3)	2 (9.5)	7 (3.7)	6 (3.4)	7 (5.8)
Black Other	18 (3.5)	1 (4.8)	4 (2.1)	8 (4.5)	5 (4.2)
Other	43 (8.4)	1 (4.8)	18 (9.4)	14 (8.0)	15 (12.5)
Unknown	20 (3.9)	2 (9.5)	10 (5.2)	6 (3.4)	2 (1.7)
UK born					
Yes	306 (60.2)	15 (71.4)	95 (49.7)	114 (64.8)	82 (68.3)
No	172 (33.9)	4 (19.0)	82 (42.9)	51 (29.0)	35 (29.2)
Unknown	30 (5.9)	2 (9.5)	14 (7.3)	11 (6.3)	3 (2.5)
Country/Area of birth if not UK born					
Sub-Saharan Africa*	53 (10.4)	1 (4.8)	24 (12.6)	20 (11.4)	8 (6.7)
Jamaica	32 (6.3)	0 (0)	17 (8.9)	10 (5.7)	5 (4.2)
Ireland	23 (4.5)	1 (4.8)	14 (7.3)	2 (1.1)	6 (5.0)
Indian Subcontinent†	14 (2.8)	1 (4.8)	4 (2.1)	3 (1.7)	6 (5.0)
Other	45 (8.9)	0 (0)	20 (10.5)	15 (8.5)	10 (8.3)
Unknown	35 (6.9)	3 (14.3)	17 (8.9)	12 (6.8)	3 (2.5)

*includes cases born in Angola, Congo, Eritrea, Ethiopia, Gambia, Ghana, Kenya, Liberia, Mauritania, Nigeria, Sierra Leone, Somalia, Tanzania, Uganda, Zambia and Zimbabwe
†includes cases born in India, Bangladesh, Pakistan and Sri Lanka

5.5.2 Clinical features

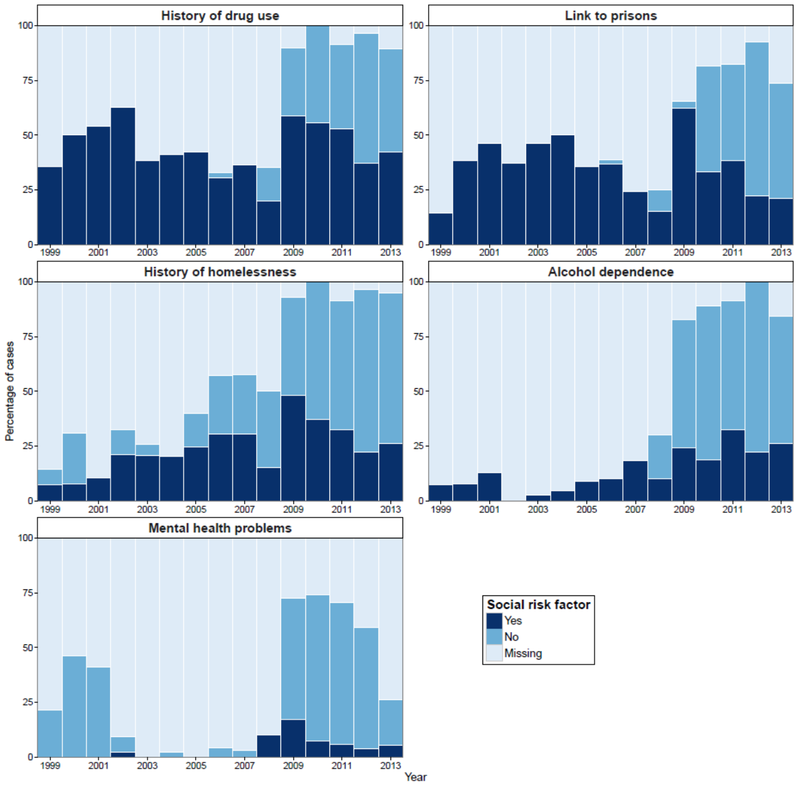
Most cases (85%) had pulmonary tuberculosis, and this proportion did not change over time (χ^2 trend p=0.83). All cases had isoniazid-resistant disease; there were 14 cases of MDR tuberculosis (3%), of which nine were MDR at their initial drug resistance test, and five were initially isoniazid-resistant but subsequently acquired resistance to rifampicin. One MDR case additionally developed pyrazinamide resistance. There were 24 (5%) patients diagnosed with this strain on more than one occasion, half of whom had initially completed treatment. The longest interval between diagnoses was 14 years; median 3.5 years.

5.5.3 Social risk factors

One or more social risk factors were reported for 308 (61%) cases. History of drug use (227, 45%), links to prisons (189, 38%) and homelessness (125, 25%) were most frequently reported. Alcohol dependence (64, 13%) and mental health concerns (13, 3%) were reported less often. Two risk factors were reported for 108 (21%) cases; three for 66 cases (13%), and four for 23 cases (5%).

Figure 5.3 displays proportions of cases by social risk factor and year (excluding years prior to 1999 in which there were fewer than ten cases). This demonstrates the continued importance of drug use, prison links and homelessness over the duration of the outbreak; as well as the change in data collection methods in 2009 resulting in an increased proportion of cases reported with presence or absence of risk factors and decreased proportion with missing data. Prior to this, reports were commonly only made if a risk factor was present and left missing if absent.

Figure 5.3: Percentage of cases in the isoniazid-resistant tuberculosis outbreak by year and social risk factor, 1999-2014.

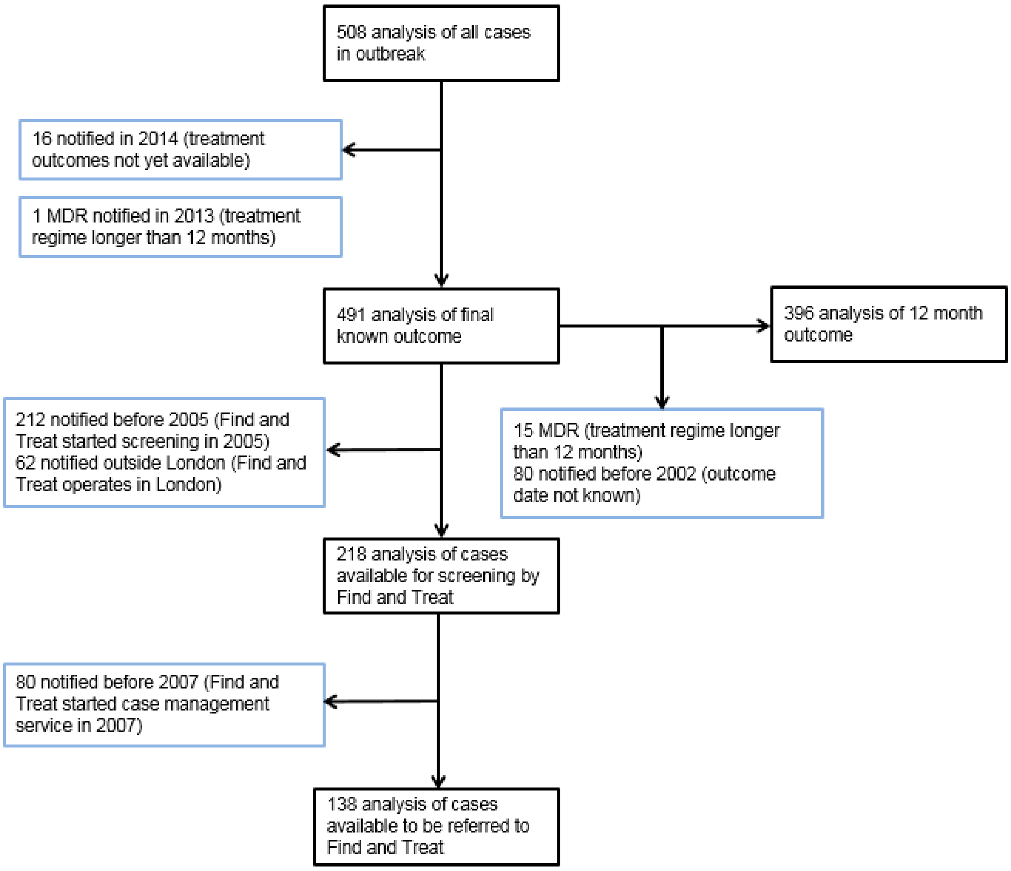


5.5.4 Treatment outcomes and Find and Treat

Eligibility criteria for including cases in these analyses are defined in Figure 5.4.

Figure 5.4: Cases included in analyses of treatment outcomes and contribution of Find and Treat to control of the isoniazid-resistant tuberculosis outbreak.

Black boxes are analyses. Blue boxes are exclusions.


508 analysis of all cases in outbreak
16 notified in 2014 (treatment outcomes not yet available)
1 MDR notified in 2013 (treatment regime longer than 12 months)
491 analysis of final known outcome
396 analysis of 12 month outcome
212 notified before 2005 (Find and Treat started screening in 2005)
62 notified outside London (Find and Treat operates in London)
15 MDR (treatment regime longer than 12 months)
80 notified before 2002 (outcome date not known)
218 analysis of cases available for screening by Find and Treat
80 notified before 2007 (Find and Treat started case management service in 2007)
138 analysis of cases available to be referred to Find and Treat

Treatment was completed by 206 (52%) of 396 eligible patients at 12 months, with a significant increase in this proportion during the outbreak (p<0.01). Cases with at least one social risk factor had a significantly lower percentage of treatment completion at 12 months than those with none or missing information on social risk factors (42% and 67% respectively, p<0.01). Treatment completion was lowest for those with a history of homelessness (42/125, 39%), links to prisons (58/189, 39%), or a history of drug use (72/227, 52%). At final known outcome, 372 (76%) of 491 eligible cases completed treatment, and those with at least one social risk factor also had a lower percentage of treatment completion (71% compared to 82% for those with none or missing information on social risk factors, p<0.01).

Reasons for failing to complete treatment are shown in Table 5.2. Of the 20 (4%) patients who died at final known outcome, tuberculosis contributed to the deaths of three, was incidental for seven, and had an unknown link to the deaths of the remaining ten. There were 34 (9%) cases lost to follow up at 12 month outcome; and 43 (9%) at final known outcome.

Table 5.2: Treatment outcomes of cases in the isoniazid-resistant tuberculosis outbreak, England and Wales.

Treatment outcome	12 month outcome* N (%)	Final known outcome† N (%)
Completed	206 (52.0)	372 (75.8)
Still on treatment	66 (16.7)	11 (2.2)
Lost to follow-up	34 (8.6)	43 (8.8)
Died	12 (3.0)	20 (4.1)
Transferred out	10 (2.5)	17 (3.5)
Unknown/ Not complete – unknown reason	68 (17.2)	28 (5.7)
Total	**396**	**491**

* cases 2002-2013, excluding all multidrug-resistant cases.
† cases 1995-2013, excluding multidrug-resistant case notified in 2013.

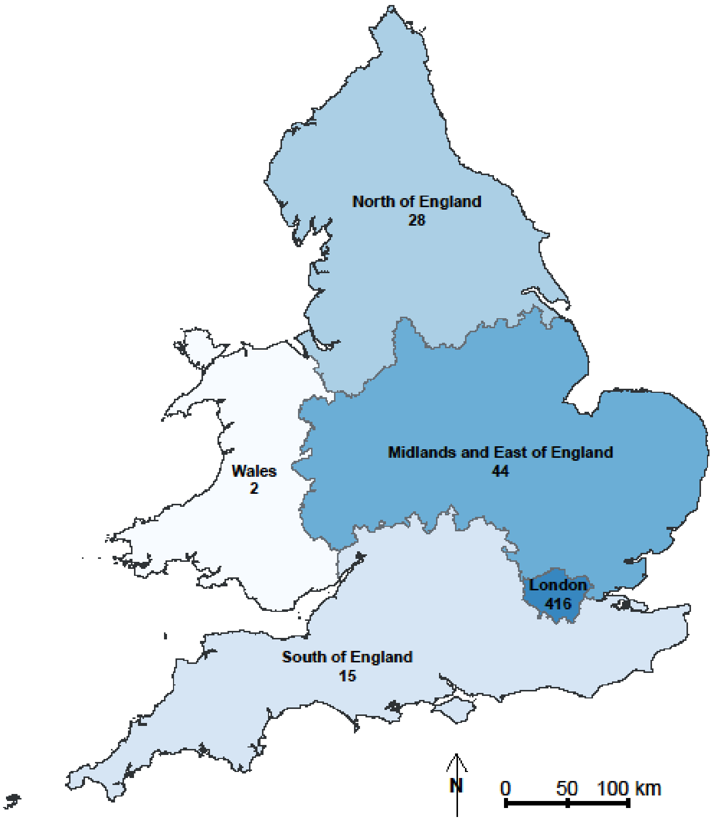
The Find and Treat screening programme aims to identify cases of tuberculosis in 'hard to reach' populations and has been operating in London since 2005. During this period, it screened 11% (25/218) of the individuals who were subsequently found to be part of the outbreak. Since 2007, Find and Treat has also operated a case management service, and one quarter (35/138) of outbreak patients notified in London in this time period have been referred to the service. The majority of these patients had a history of homelessness (30, 86%); drug use (29, 83%) and links to prisons (21, 60%). These patients were significantly less likely to have completed treatment at 12 months (15 completed, 43%) compared to those who were not managed within the service (72 completed, 70%, χ^2 $p<0.01$). However, treatment completion at final known outcome was not significantly different between the two groups (26, 74% for Find and Treat patients compared to 87, 84% for other patients, χ^2 $p=0.27$).

5.5.5 Spatial analysis

All cases were successfully geocoded to locations in England and Wales, with the exception of three which had no location data. The majority of these cases (416, 82%) lived in London; the Midlands and East of England (44, 9%) reported the most cases of other regions (Figure 5.5). Within London, most cases were reported in north east and north central areas, with the highest rates in the boroughs of Hackney and Haringey (45.4 and 33.7 cases per 100,000 population respectively), compared to 5.8 per 100,000 for the whole of London.

Figure 5.5: Numbers of cases in the isoniazid-resistant outbreak in England (by PHE Region) and Wales, 1995-2014.

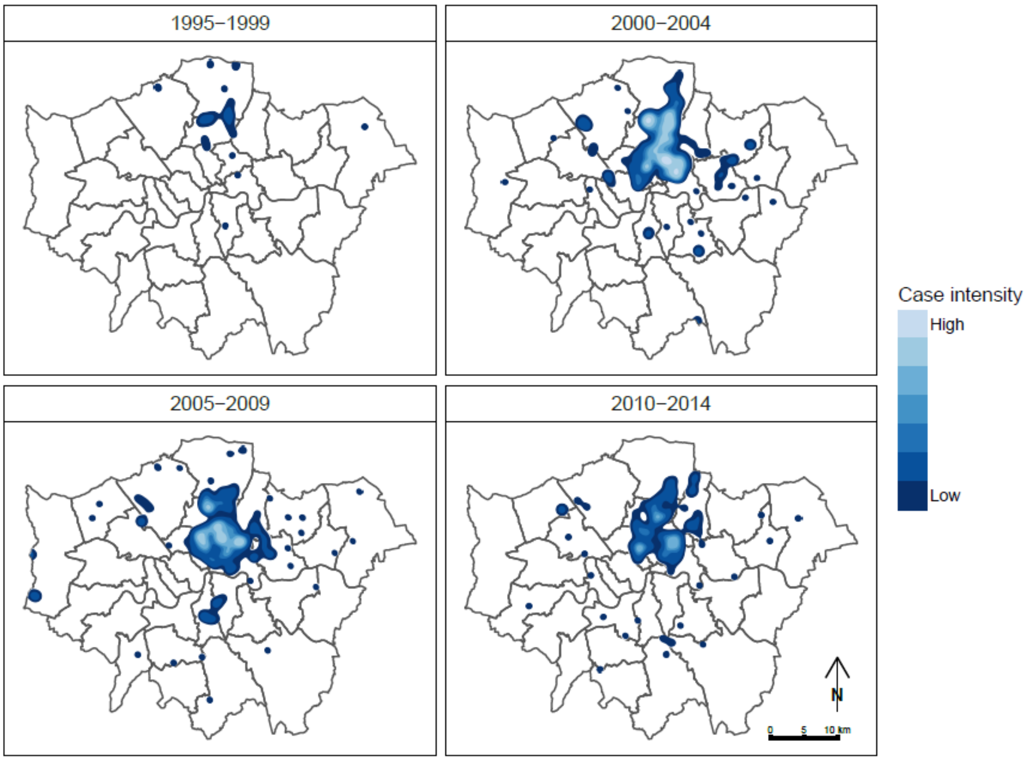
Total cases 505 because three cases had no geographical information. Contains Ordnance Survey data © Crown copyright and database right 2014.



The smoothed incidence maps (Figure 5.6) show that the outbreak has remained largely concentrated in north London, with the highest spatial intensity of cases located in a similar geographic region in all four time periods. Compared to the previous two time periods, cases in 2005-2009 appear to be slightly more dispersed; however in 2010-2014 the outbreak had contracted back to the north-central area. This pattern is also demonstrated by the animation of case locations, which is provided in the CD-ROM attached with this thesis.

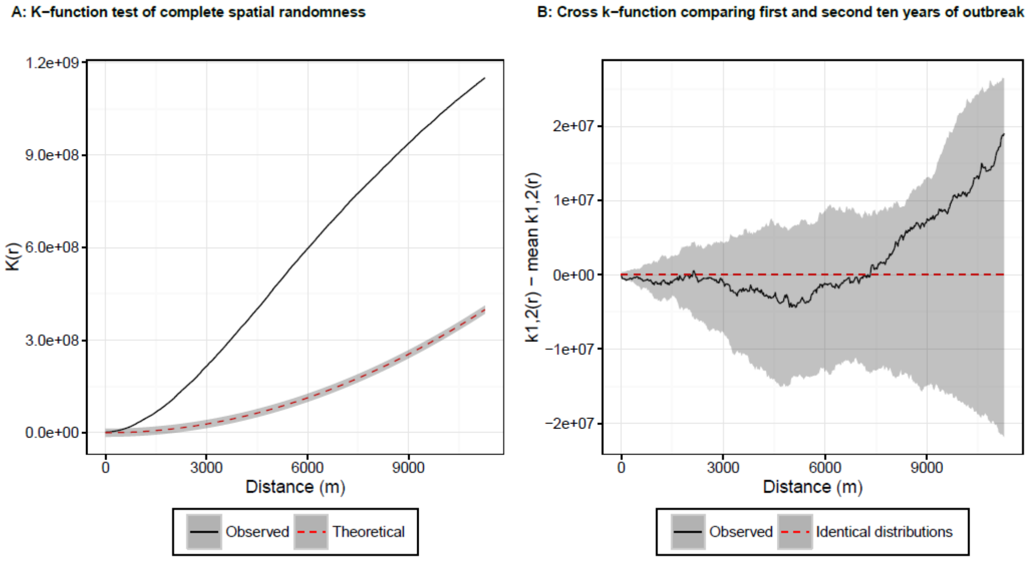
Figure 5.6: Smoothed incidence maps of cases in the isoniazid-resistant tuberculosis outbreak in London, by five-year time period, 1995-2014.

Spatial intensity determined using kernel density estimation, bandwidth 597m. Contains Ordnance Survey data © Crown copyright and database right 2014.



The k-function of the observed point locations (Figure 5.7A) lies clearly outside the simulation envelope representing randomly generated point patterns. This demonstrates that the data show spatial clustering above what would be expected by complete spatial randomness. The cross k-function (Figure 5.7B) compares the spatial distribution of the cases in the first and second ten years of the outbreak. The k-function of the observed data lies within the simulation envelope generated through random labelling of cases to different time periods across all distances. There was therefore no evidence that the spatial distribution of cases in London changed significantly during the outbreak.

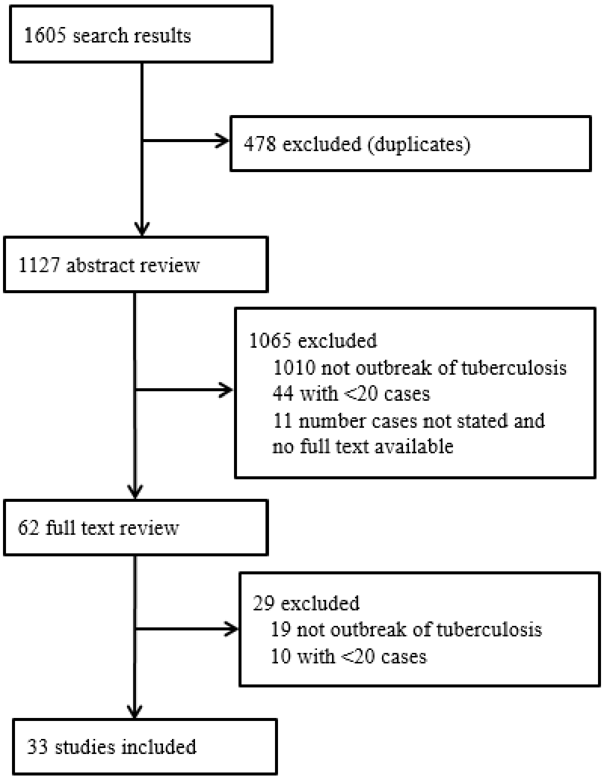
Figure 5.7: K-function analysis of spatial clustering in isoniazid-resistant tuberculosis outbreak in London, 1995-2014.



5.5.6 Comparison with other large drug-resistant tuberculosis outbreaks

The results of the systematic literature search are shown in Figure 5.8. The search yielded 1,127 unique abstracts, and 62 were brought forward for full text review. A total of 35 articles met the inclusion criteria, of which four were previous reports of the isoniazid-resistant tuberculosis outbreak in London.[180,213,214,219] The majority of other reports described outbreaks in the 1990s in America[220-229] and Europe[230-235] that were primarily associated with nosocomial or institutional transmission of MDR strains amongst HIV positive patients. The largest such incident occurred in New York City, and comprised 428 cases from 1990 to 1999.[221,222] A large series of extremely drug-resistant tuberculosis in a rural area of South Africa in 2005-6 was also attributed to multiple generations of nosocomial transmission in HIV positive patients.[223,236,237]

Figure 5.8: Study selection, systematic literature review on large outbreaks of tuberculosis.


1605 search results
478 excluded (duplicates)
1127 abstract review
1065 excluded
1010 not outbreak of tuberculosis
44 with <20 cases
11 number cases not stated and no full text available
62 full text review
29 excluded
19 not outbreak of tuberculosis
10 with <20 cases
33 studies included

Three community outbreaks had strong links to specific risk groups: Outbreaks in Sweden[181,238] and Norway[239] at the end of the twentieth century both centred around populations of migrants from Somalia, with 96 and 20 cases respectively; and in Boston an isoniazid and streptomycin-resistant strain caused 35 cases between 1984 and 1997, mostly amongst homeless populations.[184]

Other outbreaks not attributed to nosocomial transmission included a cluster of a strain resistant to isoniazid, streptomycin and para-aminosalicyclic acid in Mississippi in the 1970s (24 cases);[240] and MDR outbreaks in Tunisia (35 cases, 2001-2006),[241,242] the Federated States of Micronesia (21 cases, 2007-2009),[243-245] and a Hmong refugee camp in Thailand (20 cases, 2005).[246]

5.6 Discussion

5.6.1 Summary of findings

In this study, I have described the epidemiology and spatial distribution of a large outbreak of isoniazid-resistant tuberculosis, which has now been ongoing for 20 years. Over its duration, the outbreak has remained focused in north London, particularly amongst socially marginalised populations. Links to prisons, drug use, and a history of homelessness are important risk factors, and failure of these groups to complete treatment is likely to have perpetuated the outbreak. It is the largest outbreak of drug-resistant tuberculosis that has been documented.

5.6.2 Interpretation of results – characteristics of this outbreak

Major impacts of this outbreak have included tuberculosis disease in over 500 individuals, at least three linked deaths, and reinfection or relapse in 24 cases. Multidrug-resistance has emerged in this strain, and appears to have been transmitted between cases, as nine patients presented with an initial drug resistance test that was MDR. There are also potentially thousands of further individuals who have undetected infections with this strain, given that the lifetime risk of developing active tuberculosis disease following infection is estimated at 10%.[247] This outbreak has contributed considerable economic costs to health and social care services: Management of an uncomplicated case of tuberculosis is estimated to cost £5,000, whilst drug-resistant cases can total more than ten times this amount, before taking into account use of additional resources associated with outbreak investigations such as contact tracing and outbreak control team meetings.[248]

The characteristics of populations in the outbreak remained relatively stable, with the exception of the age of patients at notification, which increased over time. The smoothed incidence maps and k-function analyses demonstrated distinct spatial clustering which persisted in the same region of north London throughout. These observations are consistent with intensive transmission amongst a social cohort approximately 20 years ago, with infections in individuals gradually progressing to active disease. Had a great deal of ongoing transmission occurred in different groups the characteristics of cases would have been likely to change. They would also be likely to have become more widely disseminated in space, with smaller clusters arising in dispersed geographic areas.

The epidemic curve had a two-wave pattern, with an initial peak which may represent cases whose infections rapidly progressed to active disease, and a later peak potentially driven by those presenting with symptoms following a longer period of latency. Alternatively, the second wave of cases may have resulted from a second period of intensive transmission. Whole genome sequencing of isolates could be used to investigate these hypotheses by constructing a phylogenetic tree that identifies likely chains of transmission.[209]

In more recent years, the Find and Treat mobile screening unit has contributed to control of the outbreak. Approximately one in ten outbreak cases were screened by this service since it started operating, and it was also an effective service for managing complex patients: Those managed by Find and Treat had a higher prevalence of social risk factors usually associated with failure to complete treatment, but there was no significant difference in final treatment completion rates in these patients compared to the others in the outbreak.

5.6.3 Interpretation of results – comparison with other outbreaks

The systematic literature search found no documented outbreaks of drug-resistant tuberculosis that were as large or lengthy as this isoniazid-resistant outbreak. Previous incidents of comparable size included outbreaks in New York City and South Africa, both of which were linked to nosocomial transmission amongst HIV positive patients. Unlike these incidents, nosocomial transmission has not been an important factor in the current outbreak, and, although HIV status is not available for all patients, HIV positivity has previously been reported to be comparatively low for this outbreak at approximately 12%.[219] These differences are likely to be due in part to the relatively low prevalence of HIV positivity in injecting drug users in the UK and to advances in HIV care meaning that HIV positive patients are generally now managed as outpatients rather than on dedicated wards.[249]

The current outbreak does, however, share some characteristics with an isoniazid-resistant outbreak in Sweden between 1996 and 2005. As well as being resistant to the same drug, and showing little evolution in drug resistance over time, this Swedish outbreak was also confined to a distinct demographic group (migrants from in the Horn of Africa) in a relatively small geographic area (in Stockholm), and had an epidemic curve with a two-wave pattern.[181,250] Contact tracing revealed that, contrary to the initial assumption that most cases were imported, the outbreak was due to transmission in Sweden. These similarities highlight the importance of community transmission of tuberculosis with in western European cities and the need for focused control measures to affected groups.

5.6.4 Implications for policy and practice

This outbreak proved particularly challenging to control despite efforts overseen by the OCC. It consistently affected mainly socially marginalised groups in north London which are 'hard to reach'. Recommendations implemented by the OCC which attempted to target these groups included extension of contact tracing beyond household contacts to social contacts, dissemination of advice relating to specific drug regimens for treatment, and expanded use of DOT.[212] The OCC met regularly and reviewed progress, and improvements were eventually seen in treatment completion rates, but the decline in cases was slow to occur.

The results of this analysis suggest that control of tuberculosis outbreaks such as this may benefit from screening of at-risk groups to identify individuals with latent infection. Since the outbreak did not significantly disperse geographically or from specific risk populations, it appears that a high proportion of recent cases may have been infected early in the outbreak. This suggests that early screening for latent infection may have had a reasonable yield, and prevented active cases resulting from slow progression of disease. Furthermore, the two-wave shape of the outbreak highlights the need for continued case-finding even after apparent slowing down of case reports. Innovative means to target risk populations, such as the Find and Treat service, should therefore be supported.

This analysis also highlights the importance of enhanced case management to improve treatment outcomes. There were 43 patients lost to follow up and 24 diagnosed with the strain on multiple occasions, but the majority of cases in the outbreak were resistant to isoniazid only. This implies that failure to complete treatment has largely been driven by factors other than drug resistance. Following National Institute for Health and Care Excellence guidance for tackling tuberculosis in hard-to-reach groups should help to improve treatment outcomes.[248] This involves standardised risk assessment for all tuberculosis patients and better recording and monitoring of contact tracing. It also recommends expanded use of DOT, which is used infrequently in London compared to other parts of the world.[251]

Finally, this study demonstrates the importance of regular reviews of epidemiology and spatial distribution of tuberculosis outbreaks. This enables better understanding of the important factors associated with transmission, tracking the extent of spatial dispersion of outbreak strains, and improved targeting of control measures.

5.6.5 Study strengths

One of the strengths of this study was that it combined data from multiple sources, including two surveillance systems, a bespoke outbreak database, and the database from the tuberculosis outreach service, Find and Treat. This enabled description of all reported cases that have been associated with the outbreak, and ensured best possible completeness of variables, exploiting all data that has been collected over two decades by tuberculosis nurses and surveillance staff. It also allowed assessment of changes in the characteristics of cases over time.

The spatial analyses used here, including an animation, a series of smoothed incidence maps, and k-function plots, were another strength of this study. They provided greater understanding of the spatio-temporal progression of the outbreak, demonstrating that the geographic focus of this outbreak has remained stationary, and strengthening the rationale for focused control measures.

A further strength was inclusion of data from Find and Treat, which provides targeted services in the manner that the geographic analyses suggest is necessary. The analysis showed that the mobile screening unit effectively identified individuals at risk of acquiring this strain and therefore provides evidence that this will be an effective component of outbreak control in future.

Finally, the systematic literature search enabled this outbreak to be placed in the context of other large outbreaks of drug-resistant tuberculosis. Many of the previously described large outbreaks primarily involved HIV positive patients and were due largely to nosocomial transmission. A complete description of this unusually large outbreak which, by contrast, involved mostly community-based transmission, therefore contributes to the understanding of the natural history of the disease in a different setting.

5.6.6 Study limitations

A limitation of this study is that, with evolving surveillance systems, equivalence of data fields cannot be guaranteed. This means that comparisons over time, particularly for social risk factors, may be unreliable. Procedures for collection of treatment outcomes also changed during the outbreak: Prior to 2002, outcomes were not collected in routine surveillance data, and these data were therefore derived from the bespoke outbreak database. However, outcomes in this database were not collected specifically at 12 months, which is now the international standard for tuberculosis surveillance. Analysis of these 12 month outcomes could therefore only be performed for cases from 2002 onwards; and final known follow up was the only way to compare outcomes for cases over the entire outbreak.

Owing to the change in routine microbiological testing procedures from RFLP to MIRU-VNTR typing, the case definition for this outbreak also changed. It is therefore possible that these cases are not part of the same outbreak. However, all cases were isoniazid-resistant and epidemiological characteristics of cases did not change significantly following the change in strain typing methodology. This suggests that the updated case definition was appropriate, but this may be confirmed through whole genome sequencing of isolates defined through the two different methods.

It is also inevitable that some cases will have been missed. The analysis used only culture confirmed cases, excluding those without appropriate strain typing information regardless of any epidemiological links to the outbreak. As non-pulmonary cases of tuberculosis are less likely to be culture-confirmed than pulmonary cases, it is therefore likely that the number of non-pulmonary cases identified here was an underestimate. There may also have been undetected cases prior to 2000, when they were only ascertained retrospectively, and outside London prior to 2010, when they were only typed if there were epidemiological links to the city.

The relative importance of social risk factors in this outbreak was not assessed through a formal case control study, because these data have only been collected routinely for non-outbreak cases since 2009. However, the proportion of all London tuberculosis cases reported between 2009 and 2014 who had one or more social risk factor was substantially lower than for cases in this outbreak reported over the same period (approximately 10% and 40% respectively).[252] The shift towards older age groups observed here has also not been observed in London cases more widely; and highest overall rates within London over the last fifteen years have been in north-west rather than north-central areas as in this outbreak.[252] It is therefore likely that the characteristics identified here are specific to this outbreak and do not merely reflect the epidemiology of tuberculosis patients as a whole. Other risk factors such as smoking or being exposed to second-hand smoke and HIV status, which have previously been associated with tuberculosis infection, were not collected in surveillance systems but could have contributed to transmission.[253,254]

There were also limitations to the statistical methods used in this analysis. The chi-square test for trend was used to identify changes in patient characteristics over time. An assumption of this test is that the proportion of cases with a given risk factor did not, for example, increase and then decrease as the outbreak progressed. The spatial analyses were based on point locations provided by residential, or in some instances clinic, postcodes. Point locations provide an incomplete picture of the true spatial distribution of the outbreak, which may be misleading particularly for those with no fixed abode, or who regularly travel large distances from their homes to work or socialise. Some of the spatial analyses also involved division of cases into arbitrary five- or ten-year time periods, and could therefore have missed intra-period changes in distributions. However, the animation of case locations, which did not involve grouping the data into time periods, did not indicate this.

A limitation of the literature review was that the search was unable to capture any outbreaks which have not been documented in the academic literature, or which have had many additional cases subsequent to publication. It is therefore possible that there have been other larger outbreaks which have not been documented.

5.6.7 Future directions

Future work that may be prompted by this study includes further examination of the molecular epidemiology of this outbreak and development of tools to support investigation of similar outbreaks.

For example, whole genome sequencing of isolates in the outbreak may be used to elucidate likely chains of transmission. Since evolution of the genetic sequence occurs during bacterial replication, the strain will change more rapidly when it is passed between individuals than when it is relatively dormant and causing latent infection. A high degree of genetic distance between recent and early cases would therefore indicate passage of the strain through multiple hosts; whereas recent strains similar to earlier cases would indicate earlier acquisition and a longer period of latency. Comparison of results from whole genome sequencing with geographic and social networks (from contact tracing) may validate these results. Closely related strains would be expected to occur in networks of known social contacts, whilst a greater degree of geographic separation may be expected amongst more distantly related strains.

In this study, visualisation of the spatial distribution of cases proved useful to describe the evolution of the outbreak retrospectively. However, maps of cases were not regularly updated in real-time during the outbreak investigation. Improved methods of spatial visualisation which can be easily updated in routine practice, such as the tool developed in Chapter 3, would therefore be beneficial to support ongoing investigations. Contact tracing is another area that may benefit from development of new tools. There are currently challenges with collection of the data, structure, and display of these data, and improvements could help with analysis of social networks.

Box 5.1: Summary of Chapter 5.

- An ongoing outbreak of isoniazid-resistant tuberculosis now comprises over 500 cases and has been circulating for 20 years.
- Populations affected by the outbreak have largely remained stable in terms of demographics, geography, and social risk factors; although the age of the cases affected has increased over time.
- These findings are consistent with a possible intensive historical transmission in a social cohort in the early phase of the outbreak, with subsequent cases arising after varying periods of latency.
- The Find and Treat outreach service has contributed to the control of this outbreak through screening of at risk individuals and managing the treatment of complex patients with numerous social risk factors.
- This is the largest documented outbreak of drug-resistant tuberculosis to date, and is distinct from other very large outbreaks because it is not focussed in HIV positive individuals or strongly linked to nosocomial transmission.

6 Geographic profiling as a tool for targeting tuberculosis control measures

6.1 Description of chapter contents

Geographic profiling is a novel spatial tool that can use locations of linked cases to predict locations of sources, for example of infectious diseases. In this chapter, I use three case studies to investigate the utility of geographic profiling as a tool for targeting measures for tuberculosis control. The first case study uses data from the isoniazid-resistant tuberculosis outbreak described in detail in Chapter 5. For this example I use locations of cases in London to identify potential venues of transmission. The second case study involves a smaller molecular cluster of tuberculosis that occurred over a shorter period of time in a town in England. Here, I compare results of the geographic profile analysis to data from a questionnaire that was conducted as part of the cluster investigation and asked individuals in the cluster to identify venues that they had visited. Finally, in the third case study, I describe an analysis of geographic profiling as a method of spatial targeting for control of bovine tuberculosis in cattle. This analysis is based on historical data from the Randomised Badger Culling Trial, a large scale field trial that tested the effects of different badger culling practices on the incidence of tuberculosis in cattle. I design hypothetical spatially targeted badger culls using (i) a simple circular ring cull approach; and (ii) geographic profiling. I use an adaptation of survival analysis and Cox regression methods to test the efficiency of these designs at identifying badger setts with tuberculosis-infected animals. I summarise the main findings of each case study and discuss implications for the use of geographic profiling in practice.

6.2 Study rationale and introduction

Results presented in Chapters 4 and 5 of this thesis, as well as other studies, have demonstrated that groups of tuberculosis cases can exhibit significant spatial clustering.[190-193] This suggests that spatially-targeted disease control interventions could be more effective and economically efficient than more generalised programmes. However, spatial clusters can cover large areas and current methods do not provide a means of prioritising different areas within clusters to which control measures should be focused. In Chapter 1, I introduced geographic profiling, novel a statistical method that aims to assist with prioritising areas by producing an ordered search strategy.

6.2.1 Previous applications

As with many spatial statistics, geographic profiling was originally developed to approach an analogous problem in a different scientific field, in this instance criminology.[255] In its original application, geographic profiling used locations of instances of serial crime to rank long lists of suspects based on their residential locations. This is a common problem in such investigations: In the investigation into the 'Yorkshire Ripper' murders in the UK in the 1970s, for example, 268,000 names of possible suspects were generated.[256]

Although it has been used routinely in policing for over a decade, application of geographic profiling methods to biological systems is a recent development.[257] To date, examples have included using the locations of foraging sites or to find animal nests and using current locations of invasive species to identify their source populations (Table 6.1).[258-261]

Table 6.1: Application of geographic profiling to criminology, biology and epidemiology.

Adapted from: Faulkner et al. Using geographic profiling to locate elusive nocturnal animals: a case study with spectral tarsiers.[260]

Field of research	Case locations	Source locations	References
Criminology	Linked crime sites (e.g. serial murder)	Areas associated with the offender (e.g. home or workplace)	Rossmo.[255] and many others
Animal foraging	Foraging sites	Possible nests, roosts, dens etc	Le Comber,[258] Martin,[262] Raine [259]
Invasion biology	Current populations of invasive species	Areas associated with source populations	Stevenson,[261] Papini [263]
Epidemiology	Infected individuals residence	Infectious source or transmission venue	Buscema,[264] Le Comber,[102] Verity [265]

Geographic profiling has also shown some promise in application to study of infectious disease outbreaks. As well as successfully identifying the Broad Street pump using data from John Snow's cholera investigation (demonstrated in Chapter

1), it has been applied to a series of cases of malaria in Cairo, Egypt.[266] In this example, the residential addresses of 139 malaria cases were used to rank 59 water bodies that represented potential mosquito breeding sites as possible 'source' locations. The six water bodies that were ranked highest on the geographic profile corresponded with six of eight in which mosquito larvae were identified in an entomological survey.

6.2.2 Mathematical implementation

As well as being applied to a broader range of problems, the mathematical formulation of geographic profiling has also been advanced in recent years.[265] These formulae are described in detail elsewhere.[255,265] Here, I summarise the concepts underpinning the models and their relevance to investigations of infectious disease outbreaks.

The original implementation used the *Criminal Geographic Targeting* (CGT) algorithm, which is based on two concepts, distance decay and the buffer zone.[267] Distance decay is the observation that, since travel incurs a cost (of time and/ or money), crimes are more likely to occur closer to an offender's home. The buffer zone is an area around the home in which crimes are less likely to be committed, partly because of increased risk of being identified and also because the number of criminal opportunities increases with distance from home.[266] In epidemiology, and other biological applications, there is no clear rationale for including the buffer zone concept, and a new implementation of the model was therefore developed which is designed to be more suitable for these applications.[265]

In this new implementation, a Dirichlet Process Mixture (DPM) model is used to assign cases to groups that may have arisen from the same source, with no prior assumptions made about the number of potential sources.[265] Following the grouping step, a migration profile is defined for all possible source locations. The migration profile (a bivariate normal distribution) uses the distance decay concept to describe the probability of a case arising at each location, given the location of the potential source. The probability of a source occurring in each location given the observed set of cases is then calculated by applying Bayes' theorem. Finally, the solution is obtained using Markov Chain Monte Carlo (MCMC) methods, allowing the technique to be applied to large data sets. The DPM model is more likely to be appropriate for epidemiological investigations than the CGT algorithm because it is able to rigorously handle the possibility that cases have arisen from multiple sources.

6.2.3 Application to investigations of tuberculosis clusters

Applying geographic profiling to tuberculosis control raises the possibility of using the locations of cases to identify potential venues in which the disease may have been transmitted. The results may then be used to assist with prioritising specific areas in which to implement location-based control measures such as public health promotion and screening for latent infections.

In this chapter I describe three case studies in which I have applied geographic profiling methods to identify locations that may be associated with tuberculosis transmission. Case studies one and two are both based on clusters of human tuberculosis cases; one using data from the isoniazid-resistant tuberculosis outbreak described in Chapter 5, and one using data from an investigation into a molecular cluster of tuberculosis in a town in England. The third case study involves a more detailed investigation of potential transmission of bovine tuberculosis between badgers and cattle, using data from the Randomised Badger Culling Trial.

6.3 Aims and objectives

The overall aim of this chapter is therefore to assess the utility of geographic profiling as a means of targeting tuberculosis control measures. Specific objectives are listed for each case study.

6.4 Case study 1: Isoniazid-resistant tuberculosis outbreak in London

6.4.1 Introduction

In Chapter 5, I described the epidemiology and spatial distribution of a 20-year outbreak of isoniazid-resistant tuberculosis. One of the striking features of this outbreak is that, despite its long duration, cases have largely remained concentrated in the same region of London. A possible explanation for this is that disease transmission occurred in a localised area throughout the outbreak. In this case study, I demonstrate how geographic profiling could be used to identify potential venues of transmission. The objectives of this case study were to:

1. Use the locations of cases in the isoniazid-resistant tuberculosis outbreak in London to create a geographic profile of the area.
2. Generate hypotheses about potential venues of transmission.

6.4.2 Methods

The data sources and case definition for analysis of this outbreak are described in detail in Chapter 5. In this analysis, I used the locations of the 416 cases who lived in London to create a geographic profile of the outbreak. I used the DPM version of the geographic profile model using the R package *Rgeoprofile*.[265,268] To implement this model, an appropriate value for the prior expectation on the clustering parameter, σ, must be specified. Here, I used a value of 0.01. This corresponds to approximately 1.1 km in the study area, which has previously been found to be appropriate to model human movement in criminology studies.[255]

6.4.3 Results

Figure 6.1A shows the geographic profile derived from the locations of all the cases in the isoniazid-resistant tuberculosis outbreak who resided in London. The geographic profile is an ordered search area, described using hit scores. Areas with the lowest hit scores (in lighter colours in Figure 6.1A) have the highest priority in the search, whilst areas with higher hit scores (in red) have lowest priority.

The geographic profile shows several distinct peaks representing areas with low hit scores. One of these peaks is in the north of the city, amongst the largest concentration of cases, and covers a relatively small geographic area; whilst there are a number of other peaks in areas of fewer cases, which are spread across larger areas. The perspective plot of the probabilities underlying the geographic profile (Figure 6.1B) shows that this peak has a much higher probability than the other peaks.

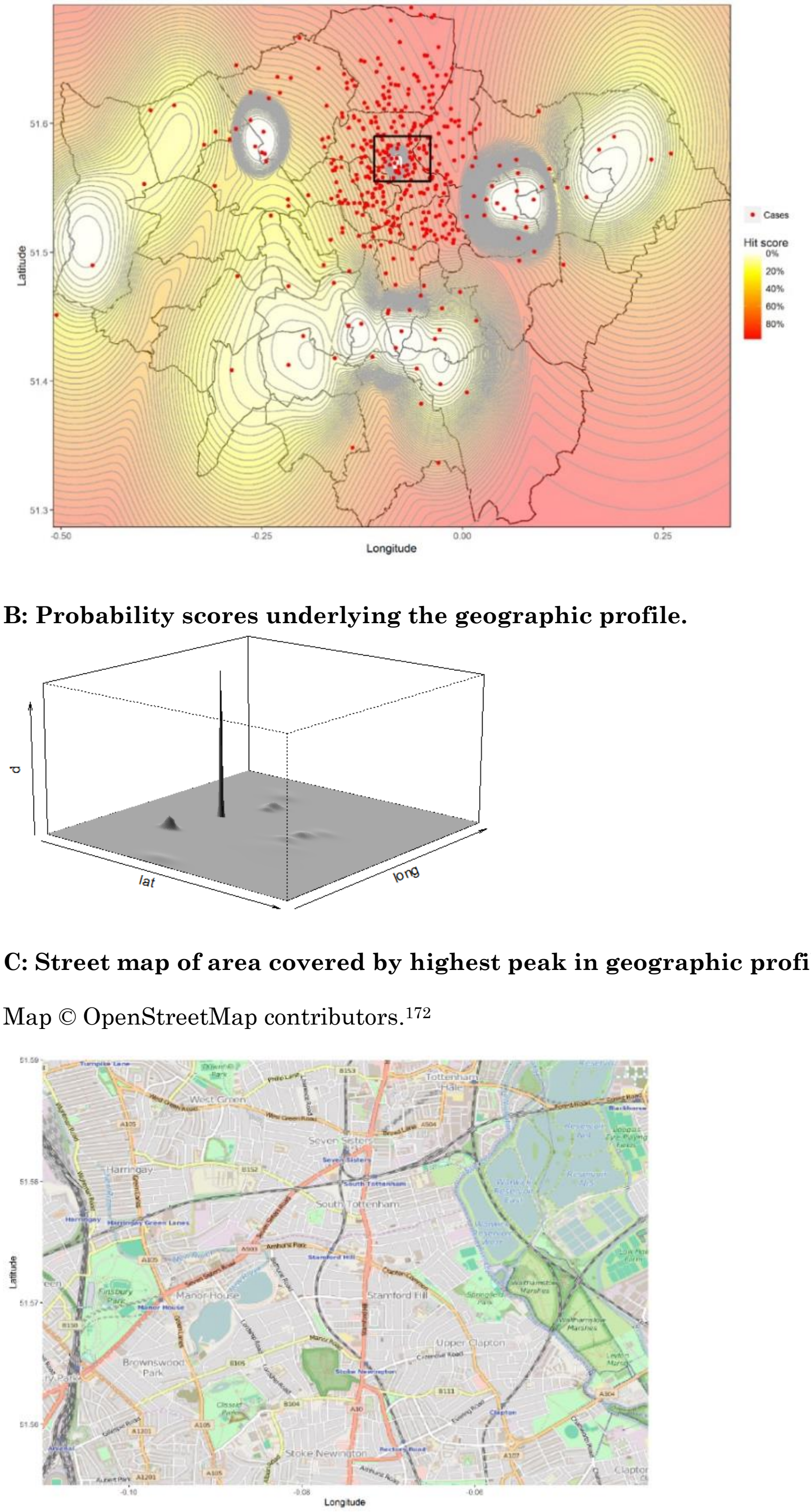
The interpretation of this is that the grouping step of the DPM model has assigned the majority of cases in the north/ central area of London to one common ‘source’, and other smaller aggregations of cases in other areas to separate ‘sources’. The peak in the north-central area covers a relatively small area because it is defined using information from a larger number of cases and there is therefore more certainty about the most likely location of a ‘source’ that would result in this spatial pattern. Figure 6.1C shows a street map of the area in this peak, which would be a sensible starting point for locating and investigating potential venues of transmission.

Figure 6.1: Model results derived from locations of cases in isoniazid-resistant tuberculosis outbreak in London, 1995-2014.

NB: Locations of cases have been randomly altered; dots do not represent actual case residential locations. Contains Ordnance Survey data © Crown copyright and database right 2014.

A: Geographic profile.

Black box shows area of highest probability.



B: Probability scores underlying the geographic profile.

C: Street map of area covered by highest peak in geographic profile.

Map © OpenStreetMap contributors.[172]

6.4.4 Discussion

In this case study, I used geographic profiling to generate hypotheses about the locations of venues of transmission in a large outbreak of isoniazid-resistant tuberculosis in London. The model suggests that the majority of cases are part of one group, which is centred in the Finsbury Park/ Stoke Newington area of north London. It also suggests that several smaller groups of cases may have resulted from transmission in other distinct areas of the city.

The geographic profile output is broadly similar to the smoothed intensity maps produced through kernel density estimation (Chapter 5, Figure 5.6), with peaks and high intensity regions in similar areas. The main advantage of the geographic profile approach in practical terms is that it produces an ordered search area, which can be used to highlight specific regions in which to target interventions. Geographic profiling also uses a more rigorous method for grouping cases and assigning probability than the kernel density method, which is based on distance only.

However, there are limitations of this method. If not interpreted carefully, the geographic profile output may be misleading. In this example, there is a large area of low hit scores in south London, which may at first seem to indicate the most probable ‘source’ location; although the largest aggregation of cases is clearly in the north of the city. This is because the model has assigned the cases in the south of the city to a different ‘source’ to those in the north, and, as there are fewer cases, the precise location of the source is less certain.

Another potential issue is in extrapolating the results of a model of spatial dispersion to disease transmission hypotheses: Whilst the area in the centre of the cases in the north London group may be the most likely ‘source’ location, it may be prudent to investigate other surrounding areas close to this peak before moving on to the ‘peaks’ in the other areas of the city, as would be done if following the results of the model directly.

6.5 Case study 2: Molecular cluster of tuberculosis in a rural town in England

6.5.1 Introduction

In this case study, I apply geographic profiling methods to data from a tuberculosis cluster investigation in a rural town in England. This cluster involved eight cases of tuberculosis disease which shared an indistinguishable 24 locus MIRU-VNTR strain type, and a further 34 people with latent tuberculosis infection who were identified through contact tracing and screening. It occurred in a usually low incidence area, and affected predominantly individuals of white ethnicity who were aged 50-70 years and born in the UK.

PHE cluster investigators hypothesised that the cluster had resulted from extensive recent transmission amongst a social group, which may have occurred in one or more club or commercial venues in the town. A questionnaire was conducted to identify venues that had been visited by multiple members of the cluster, which could therefore warrant targeted interventions. Geographic profiling offers an alternative method of identifying and prioritising potential venues of transmission, and data from this cluster investigation also potentially provides the opportunity to validate the results of the geographic profile model. The objectives of this case study were to:

1. Create a geographic profile based on this cluster and use it to rank potential venues of transmission.
2. Compare the results of the geographic profile to the cluster investigation questionnaire.

6.5.2 Methods

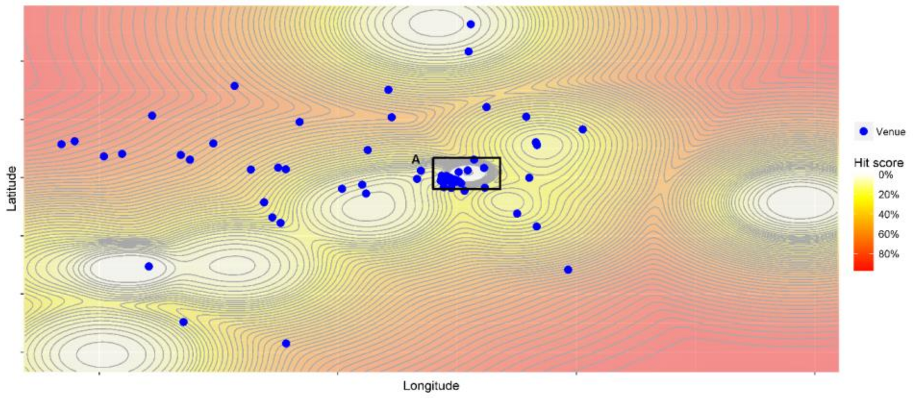
I used the residential locations of the eight cases and 34 latent infections in this cluster to create a geographic profile. I implemented the DPM model using the R package *Rgeoprofile*, and a prior expectation on the clustering parameter, σ, of 0.01, as for case study 1.[265,268] I identified hit scores for 96 venues in the area (including pubs in the town centre, restaurants, sports and social clubs, and out of town pubs) which were identified by the cluster investigation questionnaire. I ranked venues according to hit scores and to the number of people who reported attending them in the questionnaire.

6.5.3 Results

The geographic profile (Figure 6.2) identified the town centre as the most likely focal point of the observed case and latent tuberculosis locations. Of the 42 individuals with tuberculosis disease or latent infection, ten (24%) named at least one venue that they had attended, with a total of 34 venues being visited by at least one case. A social club (venue A) was reported the most times (5), and this was ranked 36/96 venues on the geographic profile. The venue that ranked top on the geographic profile (D) was one of the town pubs, and had been visited by four of the questionnaire respondents. Table 6.2 shows the ranks of hit scores of the venues which were visited by at least two individuals.

Figure 6.2: Geographic profile derived from residential locations of tuberculosis cases and latent infections in molecular cluster of tuberculosis in a rural town in England.

A: Complete geographic profile.



B: Geographic profile zoomed in to show area with lowest hit scores.

Area covered by black box in A.

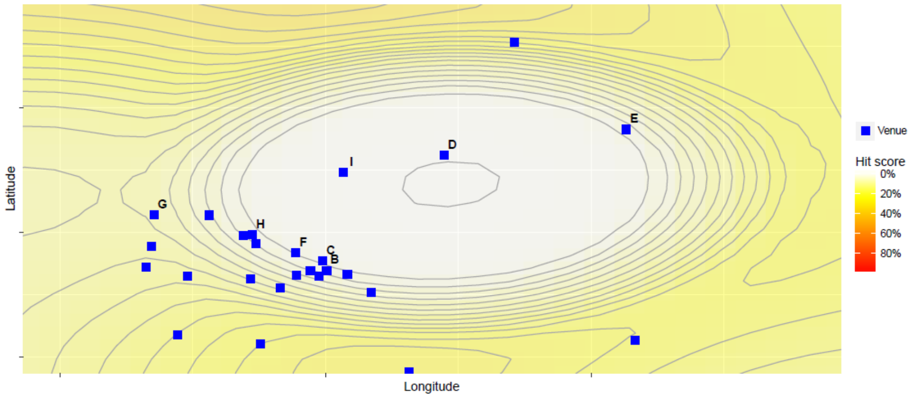


Table 6.2: Tuberculosis cluster investigation questionnaire and geographic profile results (for venues visited by two or more individuals) in molecular cluster of tuberculosis in rural town in England.

Venue ID	Venue type	Number of individuals visited (questionnaire)	Geographic profile rank (total 96 venues)
A	Sports and social clubs	5	36
D	Town pubs	4	1
C	Town pubs	4	8
B	Town pubs	4	13
F	Town pubs	3	12
E	Town pubs	3	21
I	Sports and social clubs	2	4
H	Town pubs	2	10
G	Town pubs	2	35

6.5.4 Discussion

This case study compared geographic profiling with results from a questionnaire to identify potential venues of transmission in a cluster of tuberculosis in a rural town in England. The geographic profile highlighted an area of the town centre in which several pubs were located, and the top-ranked venue on the profile had been visited by four of the ten questionnaire respondents. However, the venue which was visited by most respondents (5) was ranked relatively low on the geographic profile.

These results provide some validation to the geographic profile model, as it identified a region of the town which included plausible locations of transmission. This case study also highlights the potential advantages of this approach: It is much cheaper and less time consuming than administering questionnaires; and, rather than relying on reports from the sub-set of cases who completed questionnaires, it used data for all cases for which a residential location was available. Data obtained from questionnaires may also be unreliable, particularly for infections such as tuberculosis which have a long latency period, when individuals are asked to recall venues that they attended several months previously.

There are also limitations to this analysis. The geographic profile model can be used to rank venues according to their hit scores, but its discriminatory power for venues which are very close together is limited. This means that it is more useful for highlighting the area of the town in which transmission is most likely to have taken place, rather than identifying, for example, a specific pub. The results also indicate that, if control measures were based on model output alone, potentially important venues may be missed, as it did not identify the social club which was cited most often by questionnaire respondents.

6.6 CASE STUDY 3: SPATIAL TARGETING OF BADGER CULLING FOR BOVINE TUBERCULOSIS CONTROL

6.6.1 Introduction

6.6.1.1 Bovine tuberculosis in the UK

Bovine tuberculosis is a chronic infectious disease of cattle caused by *Mycobacterium bovis* that can also infect humans and a wide range of other species. In the past, *M. bovis* is thought to have contributed a substantial proportion of human tuberculosis disease in the UK; it has been estimated to account for approximately 6% of tuberculosis deaths in the 1930s.[269]

During the twentieth century, various control measures aiming to interrupt the chain of transmission from cattle to humans were introduced. These interventions, which included pasteurisation of milk, regular testing of cattle herds, and slaughter of animals showing signs of infection, successfully reduced the incidence of bovine tuberculosis in humans.[270] Today, the risk posed by *M. bovis* to human health in the UK is considered negligible,[270] but it remains a major animal health and economic issue. The incidence of cattle tuberculosis in England has increased over the last 25 years: In 2014, there were 4,713 new herd tuberculosis incidents, compared to 1,075 in 1996, and the average cost of such an incident in a high incidence area is estimated at around £34,000.[271]

The primary focus of bovine tuberculosis control in the UK is surveillance of cattle herds through regular testing. Cattle herds are subject to mandatory surveillance for signs of bovine tuberculosis with testing policies implemented according to the local risk of disease. For example, annual testing is conducted in the high incidence areas in the south west and west of England.[272] Skin test positive animals are removed for slaughter and the herd is placed under movement restrictions with enhanced testing until all remaining cattle have consistently had negative tests.[273]

6.6.1.2 Badger culling as a means of controlling bovine tuberculosis in cattle

Control of bovine tuberculosis in the UK is complicated by the presence of a reservoir species, the Eurasian badger, *Meles meles*. In 1971, a dead badger found in Gloucestershire, an area with high cattle bovine tuberculosis incidence, tested positive for bovine tuberculosis. It has therefore been postulated that transmission between badger and cattle populations is maintaining the disease in the environment.[274,275] On this assumption, various policies to control cattle tuberculosis that include culling of badgers have since been implemented.[270,276-278]

However, there is limited evidence that removal of badgers by culling is effective in controlling cattle tuberculosis. The Randomised Badger Culling Trial (RBCT) was a large scale field trial conducted between 1998 and 2005. It aimed to quantify the contribution of two different culling strategies on the incidence of tuberculosis in cattle: A spatially targeted 'reactive' approach, and a widespread 'proactive' strategy (Table 6.3). Reactive culling involved removal of badgers in small areas around confirmed cattle tuberculosis incidents, whilst proactive culling involved repeated annual culling over all available land. Cattle tuberculosis incidents were monitored in areas in and around trial areas and compared to the incidence in 'survey only' areas with no culling. A key ecological insight that emerged from the RBCT was that removal of badgers by culling disrupted their social groups, causing them to range more widely.[279] As a result of this perturbation behaviour, prevalence of *M. bovis* in badgers increased in the reactive culling areas. An increase in cattle tuberculosis incidence was correspondingly associated with this strategy, which was therefore suspended prior to the conclusion of the trial.[279-282]

Table 6.3: Badger culling strategies used in the current bovine tuberculosis policy in England, the Randomised Badger Culling Trial, and hypothetical spatially targeted culling strategies used in this study.

Badger culling strategy	When used	Description
Pilot culls in Somerset and Gloucestershire	Current policy in England	Culling by industry in licenced areas. The terms of the licences require that culling be widespread and conducted over areas at least 150 km^2; that at least 70% of the land should be accessible for culling; that an effective cull be carried out for a minimum of four years, and that the estimated badger population must be reduced by at least 70% in the first year of the cull.[283]
Proactive culling	Randomised Badger Culling Trial	Annual repeated culling across all accessible land (approximately 100 km^2) in each of ten trial areas.[282]
Reactive culling	Randomised Badger Culling Trial	Local culling on or near farmland where recent cattle tuberculosis incidents had occurred within ten trial areas.[282] Average reactive culling procedure covered 5.3 km^2.
Ring cull	Hypothetical design	Culling across land in circular areas of varying radii around cattle tuberculosis incidents.
Geographic profiling	Hypothetical design	Culling across land in areas identified by novel geographic profiling method as likely sources of bovine tuberculosis incidents.

Badger culling policies are also controversial because, due to limitations in diagnostic tests, they do not discriminate between infected and uninfected badgers. This inevitably results in removal of large numbers of badgers that are not infected with tuberculosis. In spite of these concerns, and the lack of conclusive evidence for the effectiveness of culling from the RBCT, it remains part of contemporary bovine tuberculosis control policy. In 2011, the Bovine Tuberculosis Eradication Programme for England set out plans to allow industry-led culling of badgers by groups of licensed landowners, and the first licences were issued by Natural England in 2012.[284] Subsequent pilot culls carried out in Somerset and Gloucestershire failed to meet the Defra target of culling 70% of badgers, and also fell short of the standards set for humaneness.[283] Nevertheless, the updated strategy for achieving Officially Bovine Tuberculosis Free status for England, published April 2014, includes badger culling in pilot areas (Table 6.3).[278]

Approaches to culling which could lessen the burden on wildlife populations, as well as reduce economic costs, could therefore be attractive. Close spatial correlation between badger and cattle tuberculosis incidence implies that targeting of badger culls to limited geographic areas could provide one such alternative.[285] Spatially-targeted culling approaches based on a ring cull concept have been conducted previously, including the 'clean ring' strategy in the early 1980s.[270] More sophisticated methods of spatial targeting have not been tested.

Geographic profiling presents a new potential approach to the spatial targeting of badger culls for bovine tuberculosis control. Applying geographic profiling to the problem of bovine tuberculosis raises the possibility of using cattle herd tuberculosis incident locations to predict the locations of setts housing tuberculosis-infected badgers (Table 6.3). The results could potentially be used to create a more efficient culling strategy that minimises the spatial extent of culling operations and therefore the number of uninfected badgers culled.

The objectives of this study were to:

1. Use data from the RBCT to design methods of spatial targeting for tuberculosis-infected badgers based on locations of cattle herd tuberculosis incidents, using:
   a. A simple ring cull approach
   b. Geographic profiling
2. Determine the effectiveness of locating setts housing tuberculosis-infected badgers compared to tuberculosis-uninfected badgers using culls based on these designs.
3. Compare the efficiency of the different methods at targeting setts housing tuberculosis-infected badgers.
4. Discuss implications for bovine tuberculosis control.

6.6.2 Methods

6.6.2.1 Study sites and badger and cattle tuberculosis data

This analysis was based on data from the RBCT, in which the incidence of tuberculosis in cattle in areas that had been subjected to different badger culling practices was recorded. The prevalence of the disease in badgers was also estimated through testing of culled animals. The RBCT used a 'triplet' design, in which the three culling policies were tested in regions within ten trial areas (denoted A-J).

Here, I have used results from the proactive trial regions, in which badgers were culled across the entire study region and subsequently tested for tuberculosis. Reactive culling regions were not included in the analysis because culling operations were undertaken only in small areas in close proximity to cattle herd breakdowns, and there is no data on the tuberculosis status of badgers outside these areas. Survey only areas were excluded as no testing of badgers was conducted. I used results of the initial badger cull only because this time period was free from the potentially disruptive effects of previous culls on the spatial distribution of tuberculosis in cattle and badgers.

In 2001 the operations of the RBCT were suspended due to a national foot-and-mouth disease (FMD) epidemic, causing a delay in the enrolment of three trial areas into the study. During the epidemic period, removal of infected cattle from the environment as part of routine bovine tuberculosis control was postponed, and restrictions were placed on cattle movement. This led to increased potential for tuberculosis disease to spread between cattle and from cattle to badgers, and an increased prevalence of *M. bovis* in badgers was observed.[286] In this analysis I therefore used only data from the seven trial areas in which the initial proactive cull took place before the FMD epidemic.

Locations of farms with cattle herds that had bovine tuberculosis breakdowns in the year prior to the initial cull were extracted from the routine herd surveillance database used at the time of the RBCT, VetNet. A breakdown is the term used to describe placement of a herd under movement restrictions following positive tuberculin skin tests from one or more animals in the herd, or the identification of infection in an animal during post-mortem inspection. The tuberculin skin test measures the relative skin reactions to injections of *M. bovis* and another Mycobacterium, *M. avium*: Animals whose reaction to the test meets defined criteria are classified as ‘reactors’ and are compulsorily slaughtered. Breakdowns are confirmed if lesions characteristic of tuberculosis are identified at post mortem or *M. bovis* is cultured from tissue samples, and only confirmed breakdowns were used in this analysis. The one year time window was used because herds in the RBCT trial areas were subject to annual testing and therefore all would be tested during this time period.

Badger sett data were extracted from the RBCT database. During the trial, badger carcasses were examined and cultured for tuberculosis using a standard protocol.[287] For the purposes of this analysis, I classified badger setts as 'infected' or 'uninfected' according to the test results of badgers trapped during the initial cull. If at least one badger captured at the sett was found to be infected with tuberculosis, the sett was classified as infected; if no captured badgers were infected with tuberculosis, the sett was classified as uninfected. Setts at which no badgers were captured were excluded.

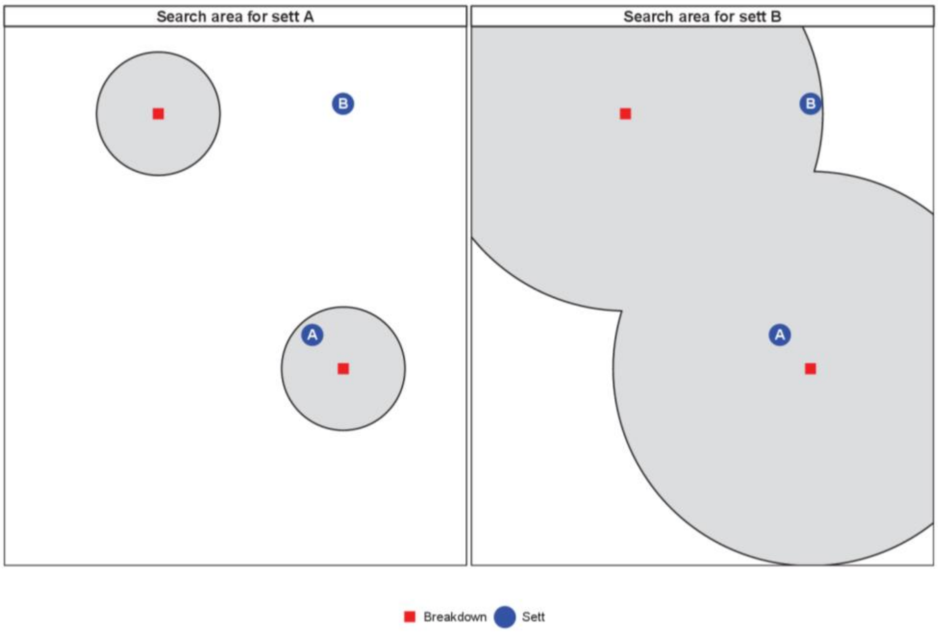
Analysis of each trial area was restricted to a region defined by the rectangular minimum bounding box enclosing the locations of the breakdowns, with a buffer zone extending the lengths of the sides by 5%.

6.6.3 Methods of spatial targeting

I designed search strategies for badger setts in each of the trial areas using two methods of spatial targeting based on breakdown locations, a ring cull and geographic profiling. Search strategies are used to order the space within the trial areas according to the likelihood that the points are locations of infected setts. Success of the strategies at targeting setts was quantified using the hit score, the proportion of the ordered area that would have to be searched before the sett is reached. For example, a sett with a hit score of 10% would be located after searching one tenth of the study area, whilst the other 90% could remain unsearched. The smaller the hit score, the higher the probability that the location is a source of infection.

I calculated hit scores for setts using the ring cull approach using a ring of radius of the minimum distance from any breakdown required to include the sett. For example, a sett 1 km from the closest breakdown would require a ring of radius 1 km to be drawn around all of the breakdowns in the area to be included in the cull. The hit score of this sett is therefore the percentage of the total area covered by all of these rings, after clipping rings to the trial region. To avoid counting the same area twice, rings from different breakdowns that cover overlapping regions are merged. This is depicted in Figure 6.3: the first panel shows the search area that would be required to include a sett at location A; and the second panel the search area required to include sett B.

Figure 6.3: Example search areas using ring cull approach.



I implemented the DPM geographic profiling model. Here, the value of the clustering parameter, σ, relates to the distance over which badgers are likely to interact with, and potentially transmit tuberculosis between, cattle herds. Although there have been no previous geographic profiling analyses of this disease system to guide selection of this parameter, various studies have estimated badger ranging distances and bovine tuberculosis cluster sizes. An analysis of spatial clustering of badger and cattle tuberculosis from the RBCT found that infection in the two species was spatially associated at a scale of 1-2 km.[285] Other published data on badger home range sizes suggest that in areas with high badger density, such as the south west and west of England, badger home range size is generally small, with estimates in the region of 0.2 to 0.8 km^2.[288,289] Bait return studies in the RBCT survey only areas found median return distances in the range of 220-370m, consistent with these relatively small home range size estimates.[279,282] However, badger dispersal has also been demonstrated over longer distances, particularly in tuberculosis-infected badgers.[290]

Given this range in badger dispersal estimates, I used a σ value of 0.02 decimal degrees for the primary analysis. At the latitudes of the trial locations, σ of 0.02 assumes that 68% of breakdowns would occur within approximately 1700m of infected setts if all breakdowns resulted from badger to cattle transmission. This value was selected to be conservatively large, accounting for dispersal at the upper limit of the likely distances.

I performed sensitivity analyses with a range of alternative values of σ (0.004, 0.01 and 0.05, approximately 350m, 900m and 4350m respectively) to assess the effect of altering this parameter on the effectiveness of spatial targeting. I also tested the CGT implementation of the geographic profiling model was also tested, using the distance decay parameter set to 1.2 (the standard value used in criminology). As there was no biological rationale for including a minimum distance from a badger sett to a cattle herd under which transmission would not be expected, the buffer zone parameter was set to zero.

6.6.4 *Statistical analysis*

I converted hit scores into search areas to allow comparison between trial areas of different sizes, using the formula:

$$\text{Search area} = \text{hit score} * \text{total trial area}$$

I used an adaptation of survival analysis to compare the effectiveness of searches over all potential sizes of search area. Survival analysis is a technique typically used to analyse ‘time to event’ data, for example to compare the time to death of patients in a clinical trial of a new cancer therapy versus control. The same logic can be applied to analyse other nonnegative random variables: for example, the number of incidences (infected or uninfected setts in this study) can be considered as a function of increases in search areas (as in this study) as opposed to time (typical survival analysis).[291] For each additional sett included in a search area of increasing size, the spatial survival function was therefore the proportion of setts with a minimum search area of an equal or smaller size. I calculated spatial survival functions for each trial region separately and for the data aggregated across all trials.

I made graphical comparisons of survival functions using Kaplan-Meier curves, which show the number and proportion of setts that would be included in a search of increasing size. I used Cox regression analysis to calculate the relative rate (hazard ratio) of including setts in search areas of increasing sizes, both unadjusted and using a multilevel model that adjusted for the random effects of trail area to account for the clustered design of the study. Where appropriate, I split the search area into sections over which the survivor functions were proportional and calculated stratum-specific Cox regression estimates.

To determine the effectiveness of identifying setts housing badgers that may have been involved in tuberculosis transmission, I compared survival functions for infected and uninfected setts for both spatial targeting methods. I then compared the efficiency of the ring cull and geographic profiling methods using the survival functions for infected setts only.

Analyses were conducted using the R packages *Rgeos*,[292] for spatial analyses, *Rgeoprofile*,[268] for geographic profiles, and *Survival*[293] and *coxme*[294] for survival analysis.

6.6.5 Results

6.6.5.1 Badger and cattle tuberculosis incidence

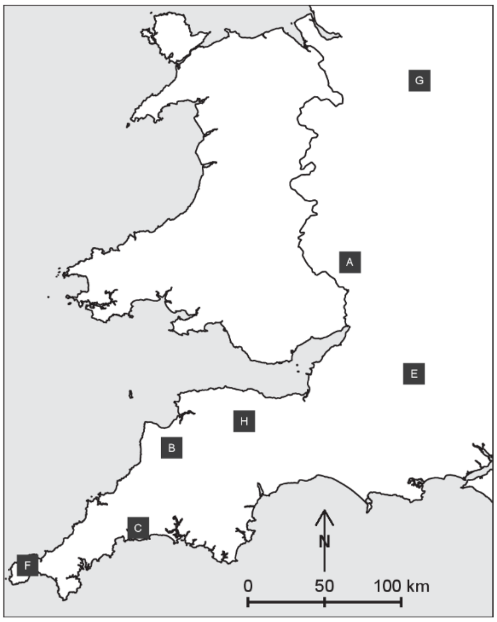
Seven proactive culling trial areas from the RBCT (denoted A3, B2, C3, E3, F1, G2 and H2), in which culling commenced before the FMD epidemic, were included in this analysis. Characteristics of these areas, with results from the initial cull, are shown in Table 6.4.

Table 6.4: Characteristics of included areas from proactive trial regions of first cull of the Randomised Badger Culling Trial.

Proactive trial region	Area of bounding box (km^2) *	Cattle herd breakdowns †	Number of setts	
			Infected ‡	Uninfected
A3	177	15	7	21
B2	169	22	10	97
C3	99.3	12	2	72
E3	152	9	23	90
F1	13.8	4	2	16
G2	92.7	8	15	73
H2	122	16	9	53
Total		*86*	*68*	*422*

* Areas defined by bounding box of breakdown locations plus 5% buffer zone.
† Confirmed cattle herd breakdowns in one year prior to start of initial cull (excluding one outlying breakdown in area E3).
‡ Infected sett, a sett at which at least one tuberculosis-infected badger was captured.

Figure 6.4: Locations of proactive trial areas of Randomised Badger Culling Trial included in this analysis.



In the year prior to the initial cull, a total of 86 confirmed cattle herd breakdowns that occurred across the trial regions were included in the analysis. One outlying breakdown was excluded (in trial area E3), and there was a mean of 12 breakdowns per trial area. Trial areas, defined by the bounding box of the breakdown locations with a 5% buffer region, ranged in size from 14 km^2 (F1) to 177 km^2 (A3). Within these areas, badgers were captured at a total of 490 setts. The proportion of setts that were infected ranged from 3% (C3) to 33% (A3), with 14% infected overall.

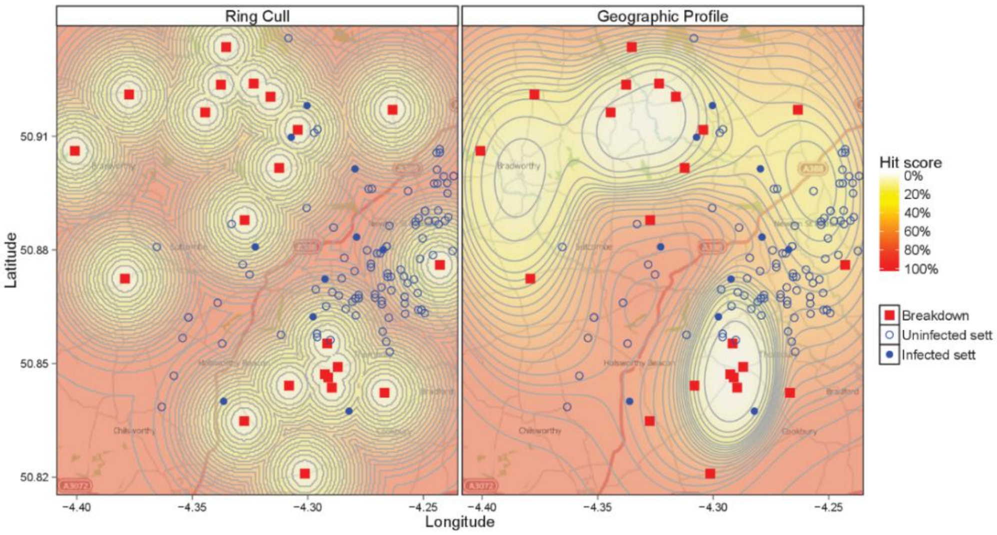
6.6.5.2 *Assessment of search strategies*

Search strategies for setts in each trial area were designed using the ring cull and geographic profiling methods, hit scores calculated, and converted into search areas. Example search strategies for areas B2 and E3 are shown in Figure 6.5.

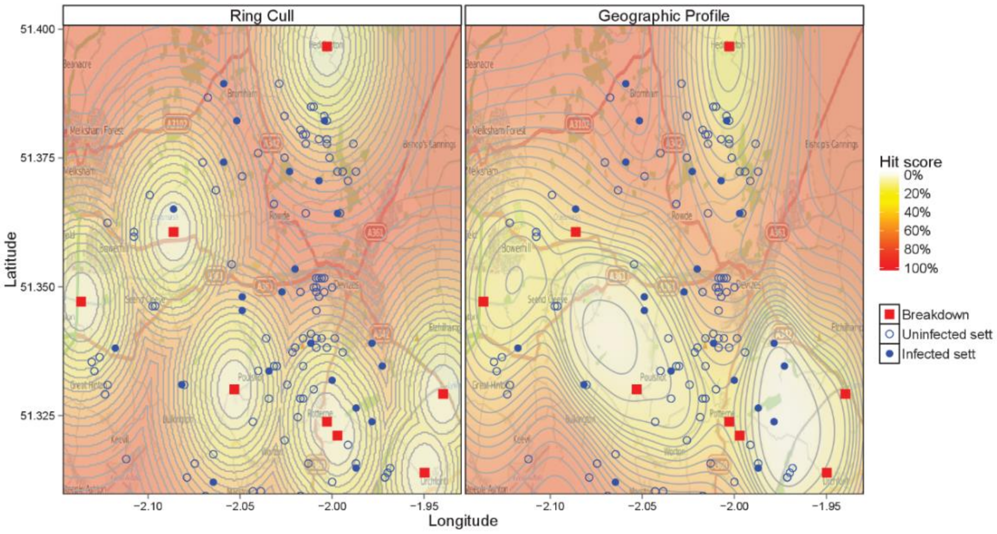
Figure 6.5: Distributions of hit scores around cattle herd breakdowns, designed using ring cull and geographic profiling search strategies.

Map © OpenStreetMap contributors.[172]

Trial region B2.



Trial region E3.



Geographic profiling grouped breakdowns into clusters based on the clustering parameter σ, producing heterogeneous distributions of hit scores with the lowest hit scores, i.e. the locations most likely to be a source, in the centre of apparent clusters of breakdowns. By contrast, the ring cull method produced symmetrically distributed hit scores around all breakdowns, not prioritising areas in the centre of clusters. Infected setts in trial B2 were generally located in areas with lower hit scores for the ring cull design, and in trial E3 were located in areas with lower hit scores for the geographic profiling design.

Spatial survival functions were calculated to compare search strategies for infected and uninfected setts across all trial areas and search sizes. For both the ring cull and geographic profiling methods, the numbers of uninfected setts included in a cull would be much greater than for infected setts at all search sizes (Figure 6.6A). Table 6.5 presents the numbers of setts that would be included at a range of search area thresholds if each threshold was applied across all trial areas. For example, searching the highest probability 10 km^2 of every trial area would include on average three uninfected setts by ring cull, and four by geographic profiling, for every infected sett included. The reactive culling strategy in the RBCT culled across an average of 5.3 km^2.[282] Searching at this threshold would include only 7 (10%) of the infected setts at the expense of 18 (4%) uninfected setts by ring cull and 8 (12%) infected and 23 (5%) uninfected setts by geographic profiling.

Figure 6.6: Kaplan-Meier curves comparing sizes of search areas for setts housing tuberculosis-infected and uninfected badgers, by ring cull and geographic profile methods.

A: Numbers of tuberculosis-infected and uninfected setts; B: Proportions of tuberculosis-infected and uninfected setts.

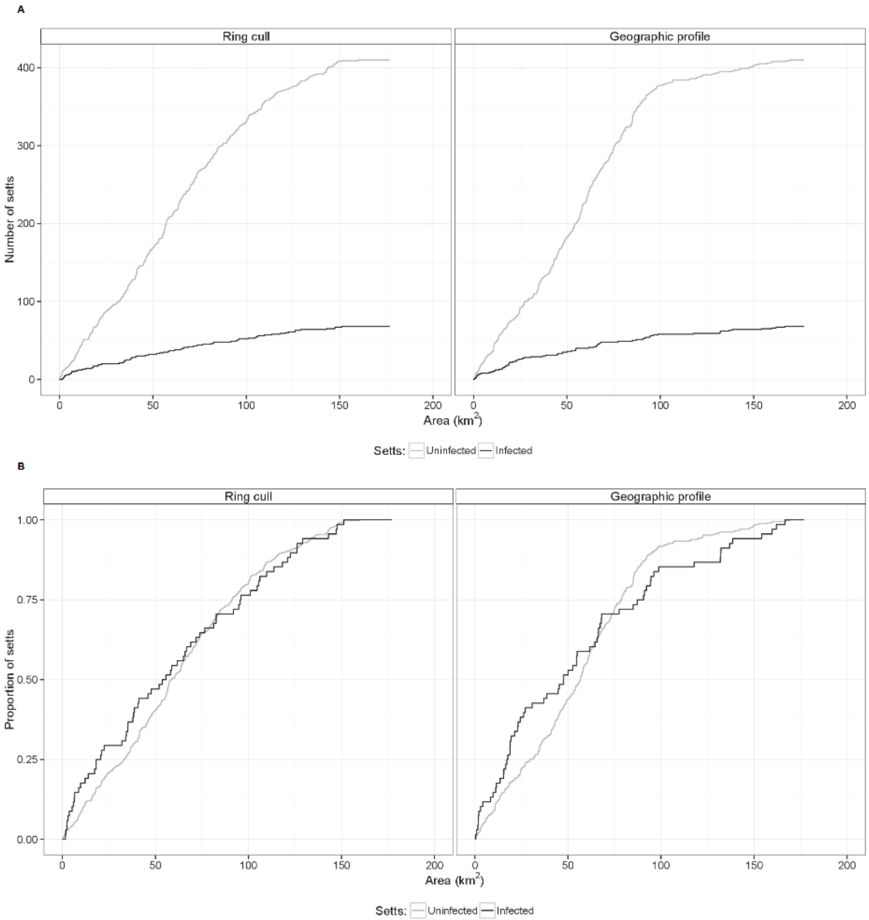


Table 6.5: Infected and uninfected setts included in searches with different thresholds according to search strategy, aggregated across all trial areas.

Search area threshold (km^2)	Ring cull		Geographic profile	
	Infected*	Uninfected	Infected	Uninfected
5.3†	7	18	8	23
10	12	36	10	37
50	32	171	35	186
100	52	339	58	387
150	67	419	64	415
177 ‡	68	422	68	422

*Infected setts, setts at which at least one tuberculosis-infected badger was captured.
† Average size of reactive culling operation in RBCT.
‡ Maximum size of trial area in analysis.

Relative proportions of infected and uninfected setts included by spatially targeted culls were compared using Kaplan-Meier curves (Figure 6.6B). If infected setts were identified more efficiently than uninfected setts, the Kaplan-Meier curve for infected setts would be expected to have a steeper gradient than that for uninfected setts, locating a higher proportion of infected setts within a smaller search area. Using geographic profiling, the Kaplan-Meier curve for the proportion of infected setts lies above the curve for uninfected setts for the first 70 km^2 of the search, and below it for larger search areas. This implies that geographic profiling identified infected setts more efficiently than uninfected setts across smaller search areas only.

Multilevel Cox regression analysis (Table 6.6) gave a hazard ratio (HR, infected/ uninfected setts) of 1.29 (95% CI 0.94-1.77, p=0.11) for areas smaller than 70 km^2. Therefore, for every 1 km^2 increase in search area, the search strategy defined by the model would include a 29% higher proportion of infected than uninfected setts, but the p value indicates that this difference was not significant. For larger search areas, the rate of inclusion of infected setts was significantly lower than for uninfected setts (HR (infected/ uninfected setts) = 0.58, 95% CI 0.36-0.94, p=0.03). Using the ring cull method, Kaplan-Meier curves for infected and uninfected setts were similar (HR (infected/ uninfected setts) =0.94, 95% CI 0.72-1.23, p=0.64). Unadjusted Cox regression analysis produced similar results to the multilevel model for all analyses (Table 6.6).

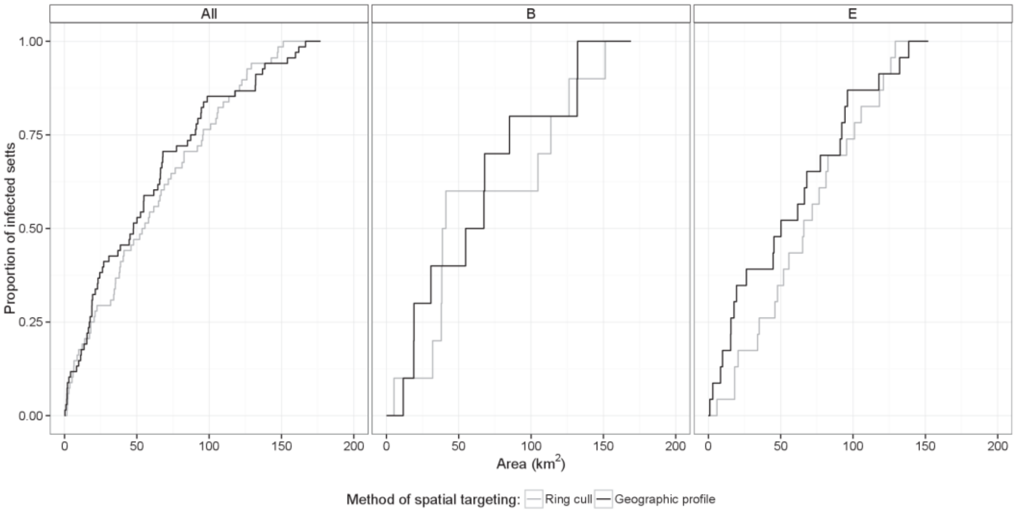
Table 6.6: Cox regression spatial survival analysis comparing search strategies.

Search strategies	Unadjusted			Multilevel*		
	HR	95% CI	p	HR	95% CI	p
Ring cull Infected compared to uninfected setts	0.98	0.76-1.27	0.90	0.94	0.72-1.22	0.64
Geographic profile Infected compared to uninfected setts, <70 km^2	1.18	0.87-1.61	0.29	1.29	0.94-1.77	0.11
Infected compared to uninfected setts, >=70 km^2	0.52	0.33-0.84	<0.01	0.58	0.36-0.94	0.03
Infected setts Geographic profile compared to ring cull	1.00	0.71-1.40	0.99	1.03	0.73-1.45	0.87

*Multilevel Cox regression model adjusted for random effects of trial area; HR, hazard ratio; CI, confidence interval.

Proportions of infected setts included in search areas of increasing sizes using the two different methods of spatial targeting were also compared through Kaplan-Meier plots (Figure 6.7). Although examination of the hit score distributions suggested that the ring cull was more efficient for trial B2, and geographic profiling more efficient for trial E3, the KM curves for these areas did not differ significantly (log-rank trial B2 *p*=0.73, trial E3 *p*=0.66). Aggregating across all trial areas, the curves for the two approaches were not significantly different at unadjusted or multilevel Cox regression analysis (HR (geographic profiling/ ring cull) = 1.03, 95% CI 0.723-1.45, p=0.87). These curves also show that culling over very large areas would be required to capture a large proportion of infected setts. For example, including 70% infected setts would require culling over 88 km^2 (50% total area) and 73 km^2 (42% area) by ring cull and geographic profiling respectively.

Figure 6.7: Kaplan-Meier curves comparing search areas for setts housing TB-infected badgers by ring cull and geographic profile methods for all trial regions; region B2 and region E3.



Results of the sensitivity analyses using different values for the clustering parameter σ and the CGT implementation of the geographic profiling are shown in Appendix 10.5.

6.6.6 Discussion

6.6.6.1 *Summary of findings*

In this study, I used historical data from the RBCT to assess two methods of spatial targeting of tuberculosis-infected badgers, a simple ring cull and a novel geographic profiling approach. Targeting small areas using the geographic profiling method, which accounts for clustering of cattle herd breakdowns, showed a small but non-significant increase in the ratio of setts with tuberculosis infected badgers compared to uninfected badgers included in a cull. However, geographic profiling provided no overall improvement in efficiency at targeting setts with infected badgers compared to the ring cull.

6.6.6.2 *Interpretation of findings*

The results of this study showed that spatial targeting of badger culling could not effectively prioritise culling of tuberculosis-infected badgers. For a search area of increasing size, the rate of inclusion of infected setts did not significantly differ from that for uninfected setts: using the geographic profiling model the hazard ratio comparing infected to uninfected setts was 1.29 (95% CI 0.94-1.77, p=0.11), and using the ring cull it was 0.94 (95% CI 0.72-1.23, p=0.64).

The reactive culling strategy in the RBCT tested a form of spatially targeting similar to a ring cull, and it culled badgers over average of 5.3 km^2.[282] This analysis showed that searching at this threshold using a ring cull could only be expected to include approximately 10% of the tuberculosis-infected badgers in the area. Failure of reactive culls to target sufficiently high proportions of infected badgers during the RBCT led to perturbation of infected badgers to new areas, and ultimately increased cattle tuberculosis incidence.[279,295] The spatially targeted culling designs tested here were no more effective than the reactive cull, and therefore would also risk the detrimental effects of perturbation as well as being harmful to wildlife populations.

The two forms of spatial targeting tested here have been applied in other settings. Ring culling approaches have been used for the control of other diseases of agricultural importance, notably FMD, a viral infection of many cloven-footed mammals including cattle and sheep. In 2001, large scale ring culling was implemented in response to epidemics of FMD in the UK and the Netherlands with the aim of eliminating animals incubating infections that may have spread from the outbreak farms.[296] Modelling during the early phase of the epidemic informed these measures which, in the worst affected regions of the UK, involved culling all sheep within three km of an infected farm.[297] Through combination with other control measures, the epidemics were eventually brought under control. However, the ring cull was very expensive and issues included logistical implications of removal of dead animals, risk of spread of infection during the culling operations, and selection of the appropriate size and shape of the cull.[296] This highlights the difficulty of implementing a ring cull policy, even in a situation such as the FMD epidemic, in which the clear focal nature of the disease and the absence of an intermediate host strengthened the rationale for the approach.

A challenge of applying geographic profiling analysis to a new biological system is the selection of an appropriate value for the dispersal parameter, σ. Here, I used a value of 0.02 decimal degrees, which assumed that 68% of breakdowns arising from badger-cattle transmission would have occurred within approximately 1700m of infected setts. This value was chosen to allow potential dispersal of badgers to the upper limit of their likely home range sizes, which have generally had smaller estimates in the areas where the RBCT was conducted.[282,288,289] Sensitivity analyses using smaller values of σ, potentially reflecting this shorter ranging distance, did not improve the performance of the geographic profiling model (Appendix 10.5).

6.6.6.3 Implications for policy and practice

Given the controversy in the use of badger culling to control cattle tuberculosis, a culling strategy that includes a large proportion of the setts housing tuberculosis-infected badgers whilst minimising the number of setts with uninfected badgers could be favourable. Spatial targeting by a ring cull or geographic profiling would theoretically be able to achieve this by focusing the cull on areas close to cattle herd breakdowns, where a common source of infection is most likely to be located, and excluding less probable areas. However, these results showed that large proportions of uninfected badgers could only be excluded by also excluding many infected badgers.

Licences for recent culling operations in England require that at least 70% of the badger population is culled.[283] Assuming that these targets aim to reduce tuberculosis-infected badger populations by the same proportion, both approaches tested here would require culling over very large areas (88 km^2 by ring cull and 73 km^2 by geographic profiling) to meet these thresholds.

I therefore conclude that cattle tuberculosis incidents are insufficiently clustered around tuberculosis-infected badgers to design a spatially targeted cull that either excludes substantial proportions of uninfected badgers or negates the need for culling over very large areas. Such culls are thus highly inefficient approaches to control of bovine tuberculosis in the UK, and efforts to control the disease should be focused on other measures.

6.6.6.4 Study strengths

This study benefited from using data from the proactive trial areas of the RBCT. Extensive field work during this trial provided the most reliable spatial data available on the relative locations of badgers and cattle tuberculosis incidents. Using the data from the first cull only ensured that the locations of badgers were free from potentially disruptive effects of previous culls. Access to this detailed data set enabled a relatively unbiased assessment of spatial targeting without relying on simulated data or modelling estimates.

I used an adaptation of survival analysis to conduct comparisons of search areas. Survival analyses have been applied to spatial data previously,[291,298] and although these examples used distance to event, rather than area to event, the same logic can be applied. Spatial associations between tuberculosis infection in badgers and cattle have been demonstrated in other studies.[270,285] However, this was the first analysis to assess if these spatial associations could be exploited to design an efficient spatially targeted cull. Comparison of search areas in this way allowed the search strategies to be evaluated across all possible sizes, and therefore avoided the difficulty of arbitrarily selecting areas of different search sizes and counting the number of setts included in each.

6.6.6.5 Study limitations

An important factor that limited the success of spatial targeting in this study is the uncertainty in the degree to which cattle tuberculosis results from badger to cattle transmission. An estimate using data from the RBCT suggested that contribution of badgers to disease in cattle is approximately 50%, but that only around 5-7% of this was a result of direct badger to cattle transmission, with the remainder due to amplification of these events through onward cattle to cattle transmission.[299] Higher contributions of localised sources have also been suggested from other models.[300] My results support low estimates, as geographic profiling would perform more effectively than a ring cull if the majority of tuberculosis was transmitted from badgers to cattle resulting in distinct clustering. Molecular data could be used to test this hypothesis by comparing the strain types of tuberculosis in cattle to those in badgers in high and low probability areas of the geographic profile, although it could not confirm the direction of transmission.

Another limitation inherent in the data used in this analysis was the representation of cattle herds as point locations. These locations are a simplification of the true area that the cattle herds would have occupied, which may have spanned several fields during the study period, and this could have altered the apparent clustering of the breakdowns. Furthermore, a number of farms in the study areas were composed of more than one non-conjoint land parcel, meaning that cattle herds can be located several kilometres from the farm. More detailed information on the locations of herds at different times could potentially refine the analysis, but, given that the time of infection is difficult to determine, a high level of accuracy in exposure location is unlikely to be achieved.

Misclassification of badger and cattle tuberculosis status could also have reduced the apparent success of spatial targeting in this study. Proactive culling trial regions of the RBCT were used in this analysis because badgers were culled, and therefore tested for tuberculosis, across the entire area. However, it is possible that some areas were identified by spatial targeting as likely source locations in which infected badgers were located but were not identified during the RBCT. This could be because badgers had migrated since infecting cattle, evaded capture at setts, or were captured but misclassified as uninfected at post mortem due to deficiencies in diagnostic procedures. The sensitivity of the standard post-mortem to detect tuberculosis in badgers is estimated to be not more than 55%.[287] Similarly, the sensitivity of the tuberculin skin test at the herd level may be as low as 50% if only one animal is infected,[275,301] so some herds may have been misclassified as uninfected, leading to an incomplete representation of the distribution of tuberculosis in cattle herds.

Another limitation of the geographic profiling approach is that it requires a rectangular search area to be defined, which inevitably includes some areas that could never harbour a source. In this example, search areas would ideally match exactly the areas surveyed for badger setts during the RBCT, and therefore exclude, for example, stretches of water and urban areas. I used the same search areas to analyse the ring cull so that valid comparisons between methods could be made, but it would be useful to refine the geographic profiling model to allow exclusion of unsuitable land through integration with GIS.

6.6.7 Future directions

The results of this study imply that limiting badger culling through spatial targeting is not feasible. If badger culling is to continue as part of bovine tuberculosis control, research into alternative means of limiting the number of healthy badgers culled should therefore be expanded. For example, improvement of diagnostic tests for bovine tuberculosis in badgers could allow discrimination between infected and uninfected badgers before they are culled. Cattle-based measures are the alternative approach to bovine tuberculosis control. This is the main focus of current control programmes in the UK, and the results of this study imply that this is likely to be a more appropriate use of resources. Further research in this area could include early identification of infected animals and more effective means of preventing onward transmission.

The form of survival analysis of spatial data used in this study could have other applications in targeting disease control measures. For example, outbreaks of infectious diseases with a suspected environmental point source have been evaluated through 'concentric circle' analysis in which risk of infection is compared in zones of increasing distance from each potential source.[80,112,132] Spatial survival analysis could enhance this method by evaluating risk across a continuous range of search sizes, rather than requiring selection of arbitrary thresholds.

6.7 DISCUSSION

6.7.1 Summary of main findings

In this chapter, I applied geographic profiling methods to three different case studies to investigate their utility in targeting control measures for tuberculosis. The case studies demonstrated some of the scenarios in which geographic profiling could potentially be of use: generating hypotheses about potential transmission venues; ranking venues which had been previously identified to prioritise targeting of interventions, and testing hypotheses that potential sources of infection are closer to cases than expected. However, it has also highlighted some of the limitations of this tool, which requires further validation before it can be used as the basis for public health decision making.

6.7.2 Interpretation of results

The principle advantage of geographic profiling as a means of targeting tuberculosis control measures that emerged in this study is that it can provide a quick, cheap method of generating hypotheses and ranking potential venues of transmission. It uses a different approach to other spatial analyses commonly implemented in cluster or outbreak investigations, as it aims to produce an ordered search strategy rather than to test for significant spatial clustering. This could make the results easier to act upon in an investigation, as the model produces a clear means of prioritising specific targets of intervention through hit scores. This framework also enabled a quantitative approach to examining the effectiveness of spatial targeting in the bovine tuberculosis case study.

This study also identified some of the limitations of geographic profiling. The outputs can be difficult to interpret, partly because the hit score is likely to be an unfamiliar concept to many working in public health. Provided with locations of cases, the model will always produce some areas with very low hit scores, regardless of whether there is a strong degree of certainty in the specific locations, because it aims to identify where best to start a search. This can give the misleading impression that there is a strong spatial pattern in the cases when there is not. Comparing geographic profiles across different disease clusters is also uninformative as the hit score is only relevant in the study area in which it is calculated. Ranks of potential venues of disease transmission are easier to interpret, but can only be produced when there is a set of hypothesised venues.

Geographic profiling also shares limitations with other methods of spatial analysis. For example, it is restricted to the analysis of point locations, which do not reflect the true space that individuals occupy; analyses do not exclude areas of terrain that are unsuitable for sources (such as urban areas in the bovine tuberculosis example), and a clustering parameter must be selected carefully to define the distance over which cases may be grouped into a cluster.

6.7.3 Implications for policy and practice

The results of this study show that whilst geographic profiling alone cannot provide definitive answers about locations of transmission, it can form a useful part of a tuberculosis cluster investigation ‘tool box’. For example, it may be sensible to conduct spatial cluster detection tests prior to geographic profile analysis to ascertain whether there is a significant spatial aggregation in the cases, which may therefore warrant a spatially targeted intervention. Geographic profiles may then be used to pinpoint locations most suitable to intervene. It may be also particularly useful in ‘data-poor’ settings in which it is impractical or unfeasible to conduct a questionnaire of cases to identify likely transmission venues, or where there are few respondents to such a survey.

This study also highlights the need for those undertaking geographic profile analysis to be adequately trained so that they are clear about the limitations and assumptions of the model. If this method is used in tuberculosis cluster investigations in practice it is important that the results are not misinterpreted, as this could lead to false certainty regarding the source of infection, and to interventions being directed inappropriately.

6.7.4 Study strengths and limitations

The strength of this chapter is that I considered three different scenarios in which geographic profiling may be used. Each of these case studies had strengths and limitations, which are discussed individually above.

6.7.5 Future directions

I have demonstrated that geographic profiling has potential to be of use in the context of tuberculosis cluster investigation. To investigate this further, the next step would be to trial integration of this analysis into routine practice to determine whether public health officials involved in cluster investigation find the outputs useful.

Further testing of this model is also needed to validate the method in the context of tuberculosis transmission, and to determine specific scenarios in which it should be used. Ideally, it would be tested on case studies in which the venues of transmission or sources of infection are known, so that the 'true' answer could be compared with the results of the model. However, this is rarely possible in practice. As an alternative, simulated data sets could be produced which mimic the characteristics of a tuberculosis cluster with one or more main venues of transmission. Such simulations could also be used to determine the most appropriate value of the cluster parameter when applied to investigation of tuberculosis clusters, which may vary in different settings.

Box 6.1: Summary of Chapter 6.

- Case study 1: A geographic profile of the isoniazid-resistant tuberculosis outbreak described in Chapter 5 identified a region in the north of London in which venues of transmission may be located.
- Case study 2: In a molecular cluster of 42 cases of tuberculosis disease and latent infection in a town in England, geographic profiling ranked pubs in the town centre as the most likely venues of transmission. A questionnaire of ten individuals in the cluster identified some of the pubs in the town centre but also a social club which was not in the high probability region.
- Case study 3: Using data from the RBCT of the effects of badger culling on bovine tuberculosis incidence in cattle, hypothetical spatially targeted badger culls were designed using a simple ring cull and geographic profiling.
    - A survival analysis based on search area was used to compare the rate at which setts housing tuberculosis-infected and uninfected badgers were targeted using different methods of spatial targeting.
    - Targeting based on a ring cull identified setts of tuberculosis-infected badgers no more efficiently than those of uninfected badgers.
    - Targeting based on geographic profiling showed a small but non significant improvement in efficiency of identifying setts with tuberculosis-infected compared to uninfected badgers for search areas <70 km^2.
    - Targeting using geographic profiling, which accounts for clustering, showed no improvement in efficiency at targeting tuberculosis-infected badger setts compared to the ring cull over all trial areas.
    - Spatially targeted culls are therefore highly inefficient approaches to the control of bovine tuberculosis in the UK and efforts to control the disease should be focused on other measures.
- Geographic profiling shows some utility for generating hypotheses about tuberculosis transmission venues; ranking venues that may be targeted for control measures, and testing whether potential sources are closer to cases than would be expected by chance.
- The tool needs to be validated further to ascertain whether it can be useful in practice.

7 Spatial accessibility of tuberculosis services in London

7.1 Description of chapter contents

In this chapter I discuss the spatial organisation of tuberculosis services in London. I use the residential locations of tuberculosis cases in London to investigate spatial accessibility of 29 tuberculosis clinics in the city. I derive estimates of the travel time from each patient residential location to each clinic using data from the Transport for London Journey Planner service. I use these data to describe the distribution of travel times to different clinics, and calculate the difference in mean travel times if patients used their closest clinic. I then investigate the impact of rationalisation of tuberculosis clinics on spatial accessibility: First, by calculating the change in mean travel times if each of the individual clinics in the city was removed and patients moved to their next closest clinics. Then, by using a combinatorial optimisation algorithm to determine which configuration of different numbers of clinics could provide the minimum overall patient travel time. Findings are discussed in the context of proposals for moving to a pan-London model for commissioning tuberculosis services.

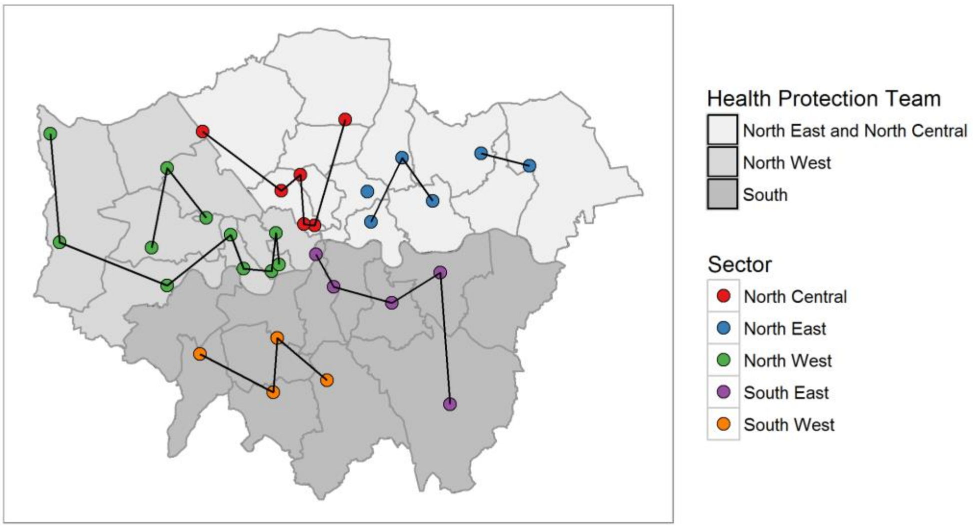
7.2 Study rationale and introduction

7.2.1 Tuberculosis service provision in London

Provision of high quality services is an important aspect of tuberculosis control. In England, Clinical Commissioning Groups (CCGs) are responsible for commissioning and delivering tuberculosis services.[178] There are 32 CCGs in London, and patients diagnosed within a CCG are referred to a local tuberculosis clinic with which the CCG has an arrangement. Until 2014 there were 33 tuberculosis clinics operating in London (one has subsequently closed). Clinics interact with CCGs through various networks including PHE Health Protection Teams (of which there are three in London); tuberculosis 'sectors' (of which there are four); and cohort review groups (of which there are eight). Cohort reviews are systematic quarterly reviews of the management of every case of tuberculosis for monitoring of treatment and contact investigations.[302] Figure 7.1 shows the locations of the clinics and working relationships between them.

Figure 7.1: Locations of tuberculosis clinics in London by Health Protection Team, Clinical Commissioning Group, Sector, and cohort review Group. Small areas are CCG boundaries.

Lines link clinics in the same cohort review group. Contains Ordnance Survey data © Crown copyright and database right 2014.



The theoretical advantage of this highly local approach to commissioning is the ability to provide services that are accessible to local populations and tailored to meet their specific needs. However, it also means that some clinics will see relatively small numbers of patients and therefore operate with small clinical teams. As a result, the level of expertise in tuberculosis care may suffer in some areas, leading to variation in the quality of services across the city, as well as inefficiency and extra costs resulting from duplication of effort. This is particularly relevant for tuberculosis because it has a relatively low incidence in the population but needs to be managed by experienced specialists.

A report by the London Assembly Health Committee in 2015, *Tackling TB in London*, concluded that commissioning of tuberculosis services at the city level could improve quality, make them more consistent, and save money.[178] It therefore recommended exploring a pan-London approach to commissioning. The British Thoracic Society has also recommended that tuberculosis services be funded on a collaborative basis between CCGs.[303] As well as improving quality, an additional advantage of this funding model would be increased accessibility of clinics to patients. Rather than being restricted to clinics that have links with their CCG, patients could choose to attend the clinic that was most convenient for them, for example the one to which they would have the shortest travel time.

The pan-London commissioning model has successful precedents including Find and Treat, the mobile screening unit, and the London Tuberculosis Extended Contact Tracing service, which carries out mass screening in response to local outbreaks of tuberculosis.[215,304] The recent Collaborative Tuberculosis Strategy for England recommended establishment of a tuberculosis Control Board for London.[305] This board will oversee all aspects of tuberculosis control in London, and represents a potential mechanism through which tuberculosis services could be commissioned at the city level.

The *Tacking TB in London* report also suggested that this commissioning model may provide an opportunity for rationalisation of services.[178] The aim of this would be to provide high quality, specialist services at a smaller number of centres. This approach has been supported for other rare diseases including some cancers in the NHS *Five Year Forward View*.[306] As an example of the potential benefits of concentrating care, this report cites the consolidation of 32 stroke units to eight specialist centres in London, which achieved a 17% reduction in 30-day mortality and a 7% reduction in patient length of stay.

Other cities in Western Europe operate their tuberculosis services using a more centralised model than London. For example, Paris has five clinics whilst Amsterdam, Rotterdam and Barcelona each have one clinic, compared to more than thirty in London.[178] The epidemic of tuberculosis in New York City in the 1990s, which had a peak of over 3,700 cases in 1992, was also controlled with fewer clinics.[307,308] Small scale rationalisation of tuberculosis services has been implemented in the North Central London tuberculosis sector. Services from one site in the sector (University College London Hospital, UCLH) were combined with those at Whittington Hospital, where a new specialist centre was built, the North Central London (NCL) South Hub.[5] Consolidation of services for tuberculosis therefore seems feasible, and may provide opportunity for improved services, provided that they are planned in a strategic way.

7.2.2 Accessibility of services

An important consideration when planning health care services is their accessibility to patients. Health care accessibility is a multi-dimensional concept that is influenced by spatial and aspatial factors.[309] It has been defined as comprising five dimensions: availability, accessibility, affordability, acceptability and accommodation.[310] Availability and accessibility are inherently spatial factors describing, respectively, the number of services in comparison to the number of potential users of services, and the burden of travel between locations.[311] The latter three dimensions, conversely, reflect financial arrangements and cultural attitudes and are therefore largely aspatial.[312]

Spatial accessibility can be considered as a measure of the 'friction' or cost of travelling between locations.[311] It can be quantified in various ways including Euclidian (straight line) or network (along a path) distances, and travel time. Distance-based measures can be estimated easily in a GIS, but, particularly in cities such as London with extensive public transport networks, travel time offers a more accurate representation of the cost of travel.[311]

Travel time is influenced by a number of factors such as the time of day, weather conditions and mode of transport. In London, estimates of the travel time between locations can be made using the online *Journey Planner* service maintained by Transport for London (TfL), the local government body responsible for the transport system in the city.[313] This service calculates travel time using various different modes of transport as well as walking, and determines the quickest means of travel between the two points. It can therefore be used to estimate the length of time that a patient would have to travel to access the tuberculosis clinics in the city, and thus represents a measure of the spatial accessibility of the clinics for patients.

7.2.3 Aims and objectives

The aim of this study is to investigate spatial accessibility to tuberculosis services in London using travel time data. The objectives were to:

1. Estimate travel time from patient residential locations to each clinic in London.
2. Describe the distribution of travel times from patients' residential locations to the clinics that they used, and to the clinics providing shortest travel times.
3. Assess the impact of closure of the UCLH tuberculosis clinic on patient travel times.
4. Calculate the potential impact of removing individual clinics on patient travel times.

5. Determine the optimum configurations of clinics to minimise patient travel times.

7.3 METHODS

7.3.1 Tuberculosis case location and travel time data

This was an analysis of cases of tuberculosis who resided in London and were notified between 1 January 2010 and 31 December 2013. I extracted the residential locations and the clinic that was attended by the patient from the ETS system. In this analysis, I aimed to generate a realistic assessment of accessibility of clinics that patients could attend. I therefore excluded small or specialist clinics and the patients who attended them. This included children aged under 18 years (who are eligible to attend a specialist children's hospital); and three clinics that had served fewer than 30 patients over the four years of the study. A total of 29 clinics and the patients attending them were therefore included. I excluded patients whose postcode was the same as that of the clinic, because this postcode is used when the residential address of the patient is unknown.

I estimated travel times from each patient residential location to each clinic using data from the TfL *Journey Planner* service which allows users to calculate approximate travel times from any location served by the London public transport system to any other location (https://tfl.gov.uk/plan-a-journey/). This service can also be accessed through an Application Programming Interface (API), which allows the Journey Planner to be queried programmatically through HTTP requests. I accessed the TfL Journey Planner API using the R package *XML*[314] and used it to estimate the minimum travel time from each patient residential location to each tuberculosis clinic in London. I chose an arbitrary weekday date (Wednesday 24th June 2015) and time outside of usual rush hour services (10:30 am), under the assumption that the majority of clinic appointments would be available at these times.

Travel times were calculated as part of a service quality assessment, and were made through a series of sequential anonymous requests to the automated system. These requests comprised postcodes only, with no accompanying data to indicate the context in which they were being made. They did not involve any batch data uploads and included postcode- rather than street-level data. Requests were made over the course of a week, during which millions of similar requests would be made to the TfL API. Internal risk assessment therefore concluded that there was no risk of compromising patient confidentiality through this process.

7.3.2 Comparison of used and catchment clinics

I defined the clinic that the patient was assigned to in ETS as their 'used' clinic; and the clinic which would provide the shortest travel time as their 'catchment' clinic (i.e. the one for which they were in the catchment area based on travel times). If a patient had the same minimum travel time to multiple clinics, I assigned the catchment clinic as the same as the used clinic, or selected a clinic at random if they used none of these clinics.

I plotted residential locations of patients according to their used and catchment clinics on a map, and investigated the differences between used and catchment clinic travel times at the patient- and clinic-levels. For patients, I calculated the overall mean travel time to the used and catchment clinics, and compared them using a t-test. I also plotted the distribution of travel times for patients' used and catchment clinics and summarised the difference in number of patients within different travel time thresholds. For clinics, I reported the numbers of patients using each clinic and the numbers of patients for which the clinic was their catchment clinic. I determined the impact on caseload if clinics served only the patients in their catchments. I also summarised the distribution of used and catchment patient travel times by clinic.

7.3.3 Effects of removing individual clinics on travel time

To assess the impact of the closure of the UCLH clinic and relocation of patients to the NCL South Hub clinic, I calculated the change in mean travel times for the patients affected. I also estimated the effects of removing each individual clinic on patient travel time under the assumption that all patients would then attend their next closest clinic in time.

7.3.4 Optimum clinic configurations for travel time

I investigated the optimal theoretical combinations of subsets of clinic locations using a combinatorial optimisation algorithm. The aim of this analysis was to determine, for each potential set of n clinics in London, which group of n clinics would provide the minimum overall patient travel time. For example, if there were to be seven clinics in London, which group of seven of the 29 clinics in the city would minimise overall patient travel time. In theory, this could be determined by calculating the total travel time for all possible combinations of seven clinics. However, this is not computationally feasible in practice because the number of combinations of clinics becomes very high with increasing numbers of clinics being selected. This can be calculated as follows:

The number of combinations of *n* distinct clinics, taken *r* at a time is:

$$\mathrm{nCr} \ = \ \mathrm{n}!/\ \mathrm{r}!\ \ (\mathrm{n}-\mathrm{r})!$$

For example, choosing from 29 clinics in groups of seven:

$$\mathrm{n} = 29;\ \mathrm{r} = 7$$

$$29\mathrm{C}7 \ = \ 29!/\ 7!\,(29-7)! \ = \ 1{,}560{,}780$$

Therefore, to test all possible combinations of seven clinics, minimum travel times would have to be determined for each patient for more than 1.5 million different sets.

Optimisation algorithms are designed to solve problems such as these in which an exhaustive search is not feasible. A combinatorial optimisation algorithm was required in this case because it uses discrete variables that can represent quantities that can only be integers, such as clinics. This is distinct from a continuous optimisation problem in which the solution represents, for example, the mass of an object, and may take any value, often within a bounded range.[315] There are several classes of combinatorial optimisation algorithm, and here I used a genetic algorithm implemented in the R package *genoud*.[316]

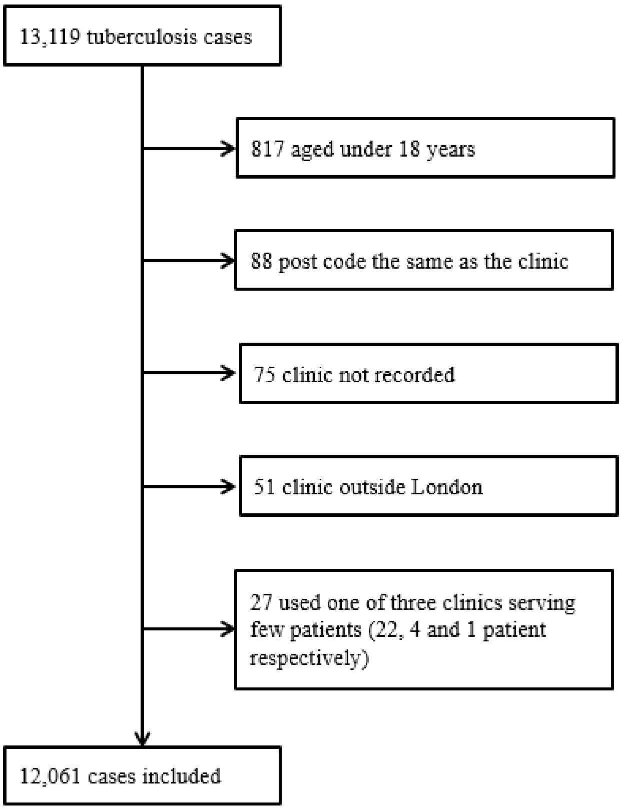
Genetic algorithms begin the search for the optimum result with a random sample of candidate solutions, termed a ‘population’. The population is then ‘evolved’ though multiple generations towards better solutions, using logical operations based on the evolutionary processes of mutation, crossover and selection.[317] With each generation the average ‘fitness’ (the closeness to the solution) of the population generally increases, and the process is repeated through multiple generations until a termination condition is reached. Therefore in this application of the algorithm, the optimal solution is the combination of clinics that has the minimum total patient travel time. The first population is a random selection of different combinations of clinics, and the combinations with the higher fitness (shortest travel time) are selected to create the next generation. Mutation and crossover introduce random changes to the combinations of clinics in the next generation. The process is repeated until several generations pass without any combinations with shorter overall travel time being produced.

To identify an optimum set of clinics based on minimum patient travel time, I used travel times for a random sample of 1,000 patients (8.3% total). The random sample was used to limit the computation time. I derived a function in R which calculated the minimum patient travel time for a given set of clinics, and then used the *genoud* package to test multiple combinations of clinics and find the optimum set. I repeated this process for each possible total number of clinics (sets of one to 28 subsets of the 29 total clinics). I determined the impact on travel time for all patients, i.e. including those not in the random sample on which the optimisation algorithm was run. I mapped the locations of each optimum set of clinics and calculated the numbers of patients that would attend each based on their catchments.

7.4 Results

There were 13,119 cases of tuberculosis in patients residing in London from 2010 to 2013. Of these, 817 were excluded from the analysis because they were aged under 18 years; 88 because their post code was recorded as the same as the clinic; 75 because their clinic was not recorded; 51 because they used a clinic outside London, and 27 because they used one of three tuberculosis clinics that served fewer than 30 patients (Figure 7.2). A total of 12,061 tuberculosis patients with viable post codes attending 29 clinics in London were therefore included.

Figure 7.2: Cases included in analysis of tuberculosis service accessibility in London, 2010-2013.


13,119 tuberculosis cases
817 aged under 18 years
88 post code the same as the clinic
75 clinic not recorded
51 clinic outside London
27 used one of three clinics serving few patients (22, 4 and 1 patient respectively)
12,061 cases included

7.4.1 Comparison of used and catchment clinics

The locations of patients according to their used and catchment clinic are shown in Figure 7.3. Although broadly similar, there are some areas in which the used clinic for a group of patients is different to the catchment clinic. One such area is circled in Figure 7.3: In the *Used* panel, the patients in this area represented by green dots are using the clinic to the west of the circle; whereas in the *Catchment* panel, many of these patients are now represented by pink dots because they are in the catchment of the clinic to the south of the circle.

Figure 7.3: Tuberculosis patients and clinics in London, 2010-2013. Cases are colour coded according to the tuberculosis clinic that they used (*Used* panel) and to their nearest clinic based on travel time (*Catchment* panel).

Circle highlights an area in which many of the used and catchment clinics for patients differ. NB: Exact locations of cases have been altered; they do not show actual case residential locations. Contains Ordnance Survey data © Crown copyright and database right 2014.

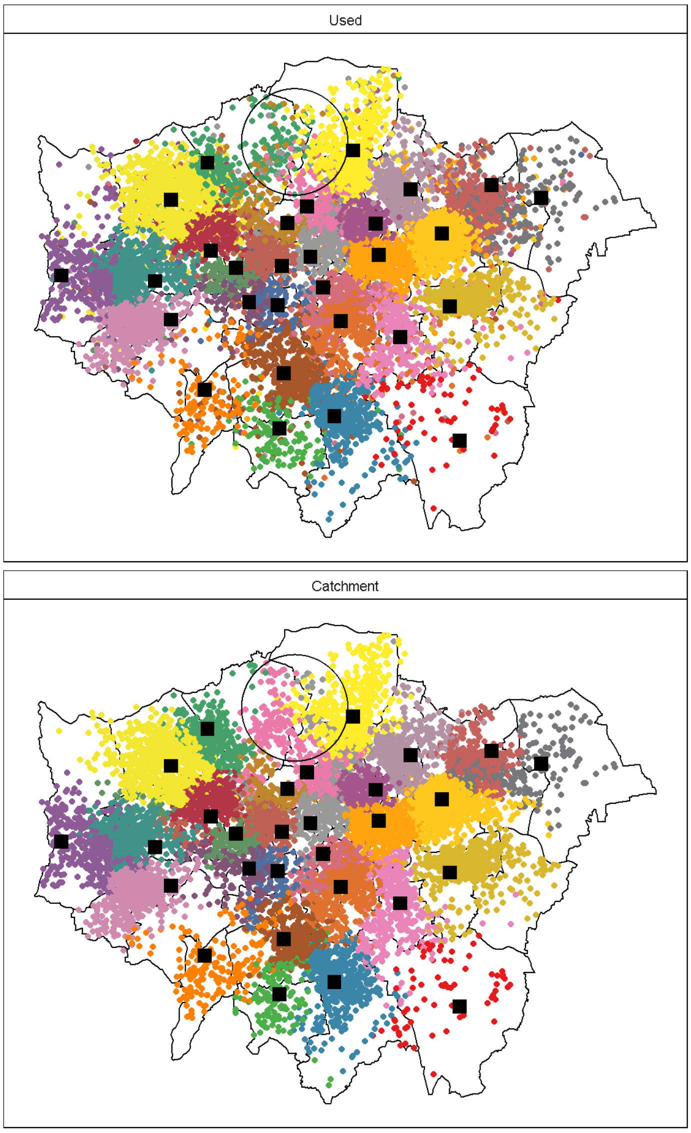
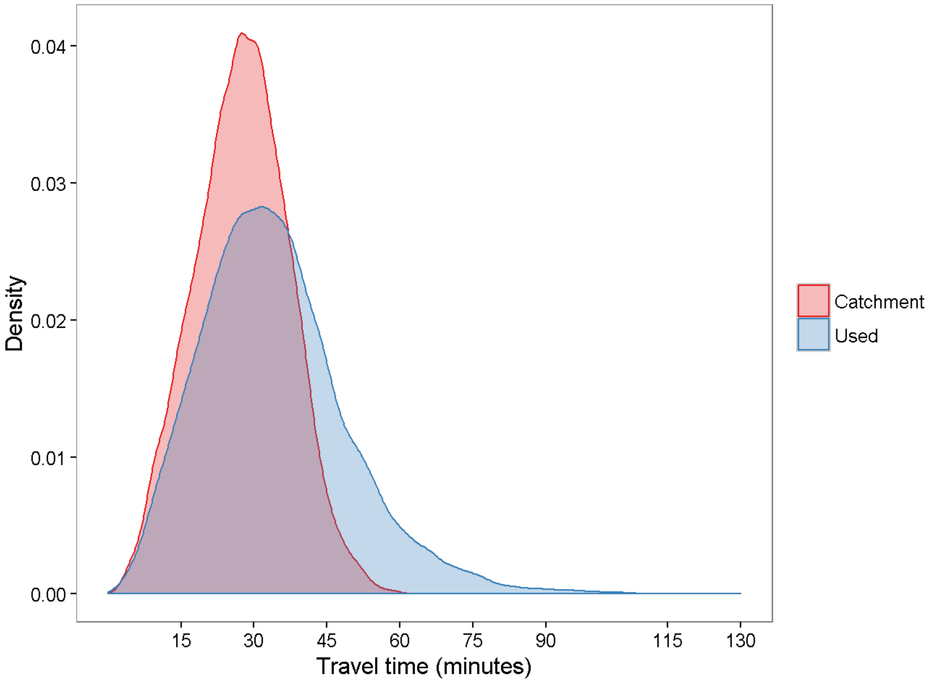


Figure 7.4 shows the distribution of travel times for used and catchment clinics. There was a small but significant decrease in average patient travel times to catchment clinics compared to used clinics (27.5 minutes for catchment clinics, sd 9.6 minutes, compared to 33 minutes, sd 15.1 minutes, t-test p<0.01). A total of 7,337 (61%) patients used their catchment clinic; 2,130 (18%) used a clinic more than 15 minutes further than their catchment clinic; 767 (6%) more than 30 minutes, and 59 (0.5%) more than 60 minutes.

Figure 7.4: Distribution of travel times for tuberculosis patients to their used and catchment clinics, London, 2010-2013.

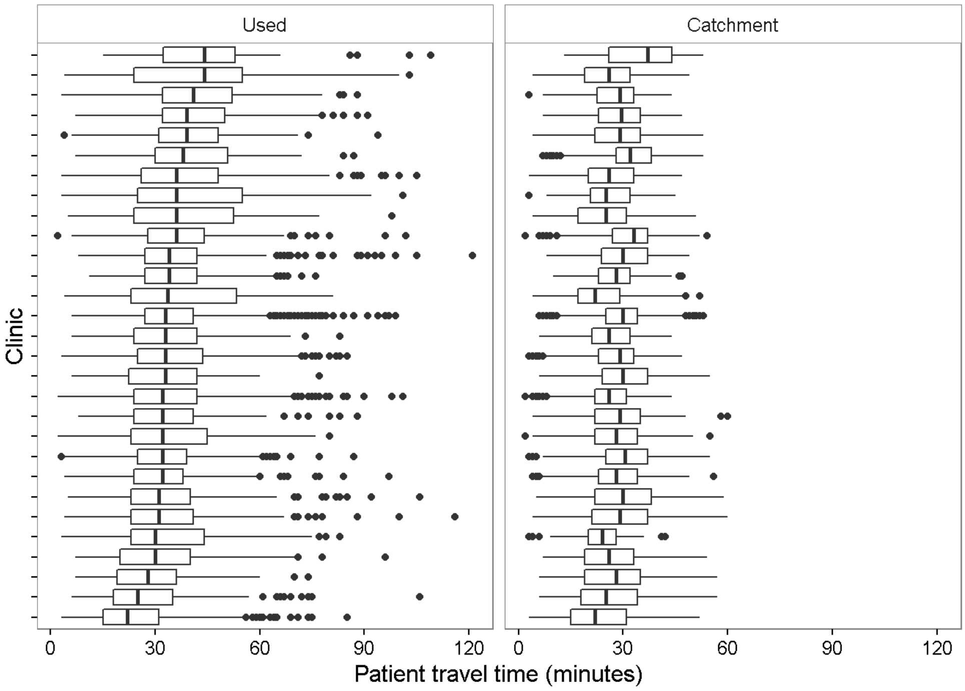


The median total number of patients using each clinic over the four years of the study was 369 (IQR 252-416), and in clinic catchments was 400 (IQR 205-510). There were 12 clinics that served more patients than are in their catchment, 16 served fewer than in their catchment, and one served the same number. The clinic with the largest change in caseload would have served 368 more patients if it included all those in its catchment, a 35% increase.

Box plots of the numbers of patients by clinic show that assigning patients by catchment produces a smaller range in travel times by clinic, and patient travel times would be more consistently less than 30 minutes (Figure 7.5).

Figure 7.5: Distribution of tuberculosis patient travel times by clinic, London, 2010-2013.

NB: Locations of cases have been randomly altered; dots do not represent actual case residential locations

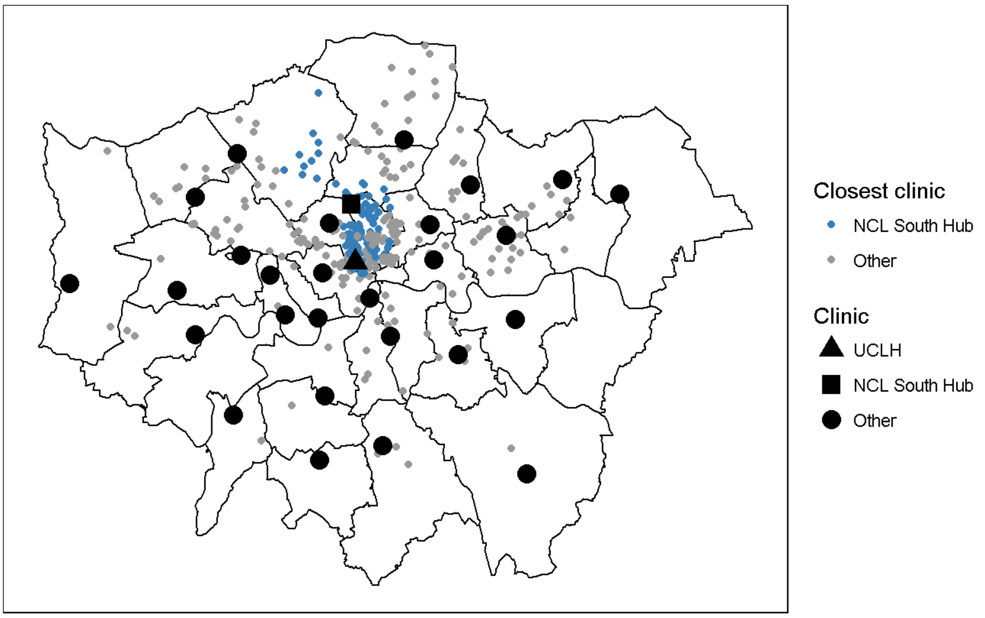


7.4.2 Effects of removing individual clinics on travel time

There were 386 patients treated at UCLH during the study period, with an average travel time of 34 minutes (sd 15 minutes). If all of these patients instead used the NCL South Hub clinic, their average travel time would increase to 41 minutes (sd 16 minutes). However, 261 (68%) patients could travel for less time to reach an alternative clinic. If patients attending UCLH were assigned to their nearest clinic (regardless of whether it was the NCL South Hub), average patient travel time would have decreased to 26 minutes (sd 9 minutes). Figure 7.6 shows the residential locations of patients attending UCLH. Those in blue are the patients for whom the NCL South Hub would be the nearest alternative clinic; those in grey are closer to a different clinic.

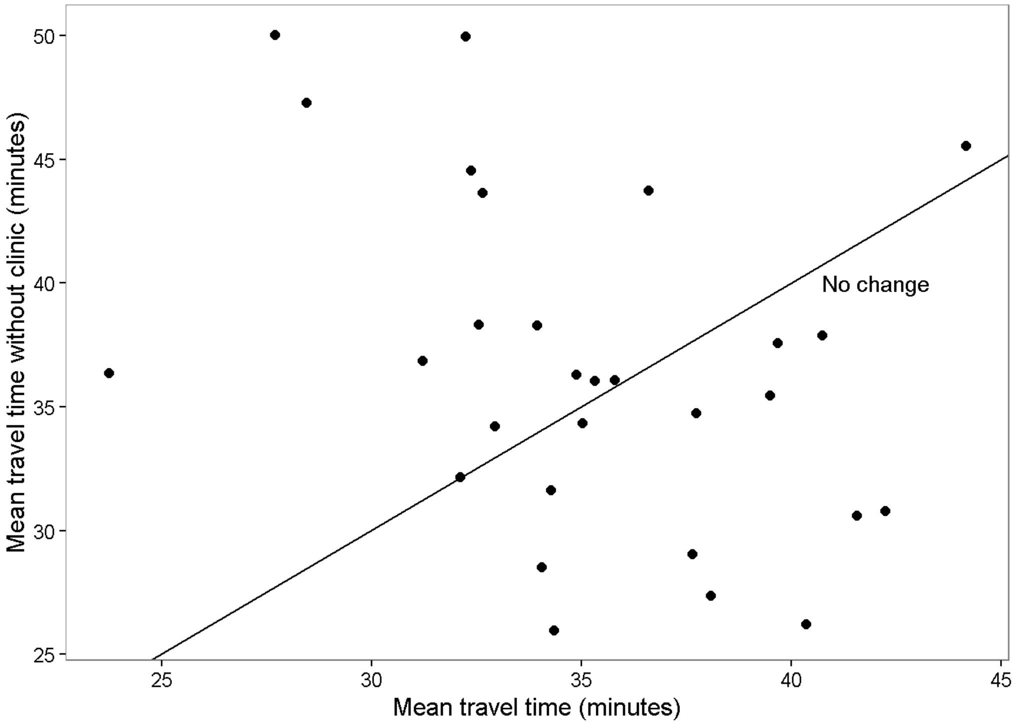
Figure 7.6: Tuberculosis patients who attended University College London Hospital clinic, 2010-2013.

NB: Locations of cases have been randomly altered; dots do not represent actual case residential locations. Contains Ordnance Survey data © Crown copyright and database right 2014.



The effects on average travel time of removing an individual clinic and assigning patients to their next closest clinic in time were highly varied (Figure 7.7). Since a substantial proportion of patients did not use their catchment clinic, removing clinics led to an overall decrease in average travel time in 13 instances. The average decrease in patient travel time for these clinics was 5.7 minutes. Conversely, patient travel time would increase if one of 16 clinics was removed, by an average of 7.5 minutes. Overall, this suggests that removing a single given clinic would not have very substantial effects on spatial accessibility, assuming that patients were able to attend their next most convenient clinic.

Figure 7.7: Impact of removing each tuberculosis clinic on average patient travel times, London, 2010-2013.



7.4.3 Optimum clinic configuration for travel time

The locations of the optimum configurations of clinics to minimise total patient travel time are shown in Figure 7.8, and the average patient travel time and distributions of the travel times for each set of clinics in Figure 7.9. As would be expected, if only one clinic was used, its optimum location is in the centre of the city, but the average travel times are long (53 minutes). Adding additional clinics to this one has a substantial impact on average travel time up to approximately ten clinics (mean travel time 34 minutes), after which the benefits in accessibility become more marginal with additional clinics. Assuming that patients use their catchment clinic, a mean travel time of less than 45 minutes could be achieved with three clinics; and of less than 30 minutes with 18 clinics.

Figure 7.8: Optimum configurations of tuberculosis clinics to minimise patient travel time, by number of clinics included, London, 2010-2013.

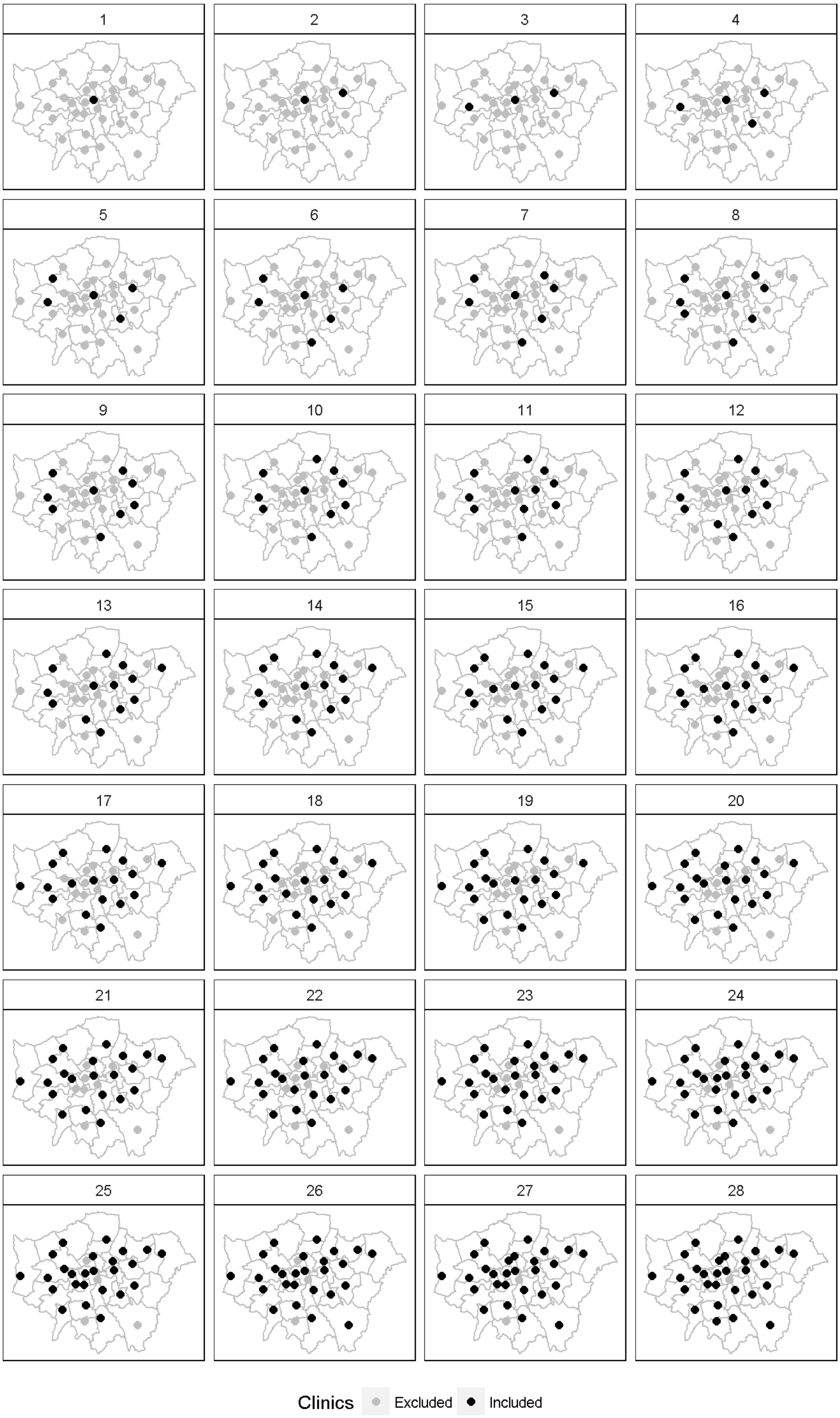
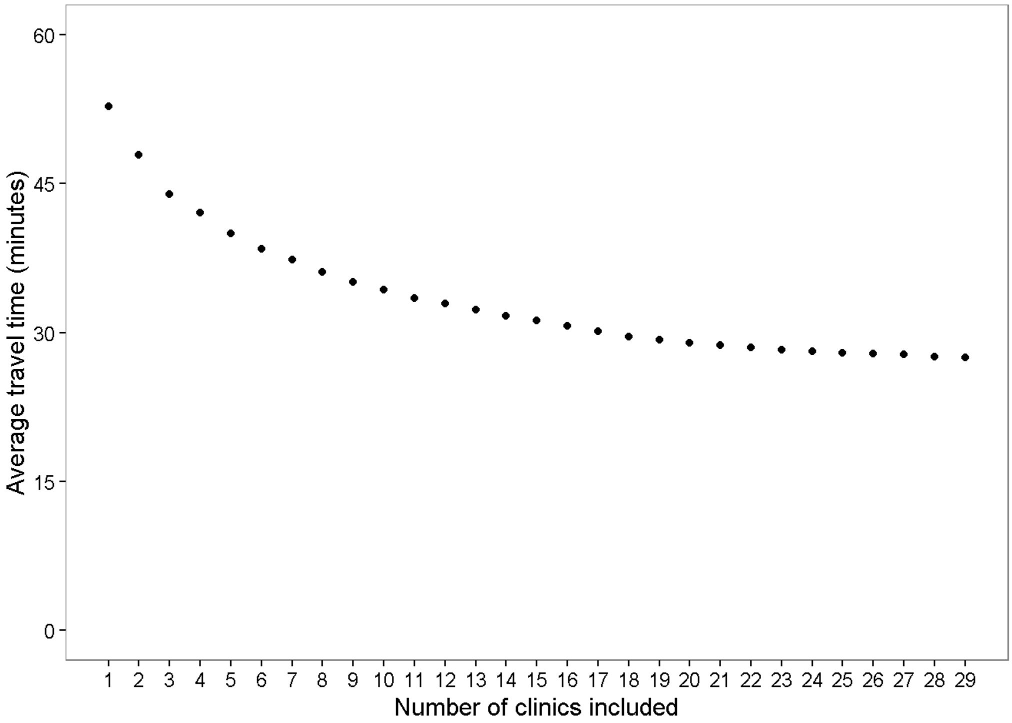


Figure 7.9: Patient travel times to tuberculosis clinics for different numbers of optimum clinic configurations, London, 2010-2013. Patients assigned to clinics based on minimum travel times.

A: Average travel times.



B: Distributions of travel times.

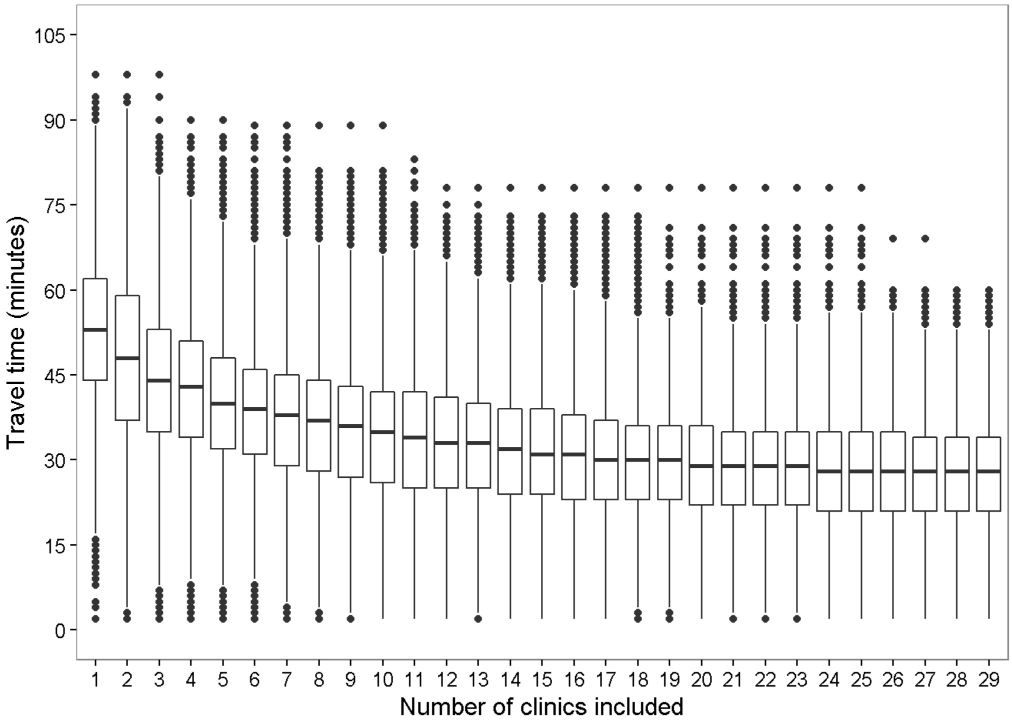
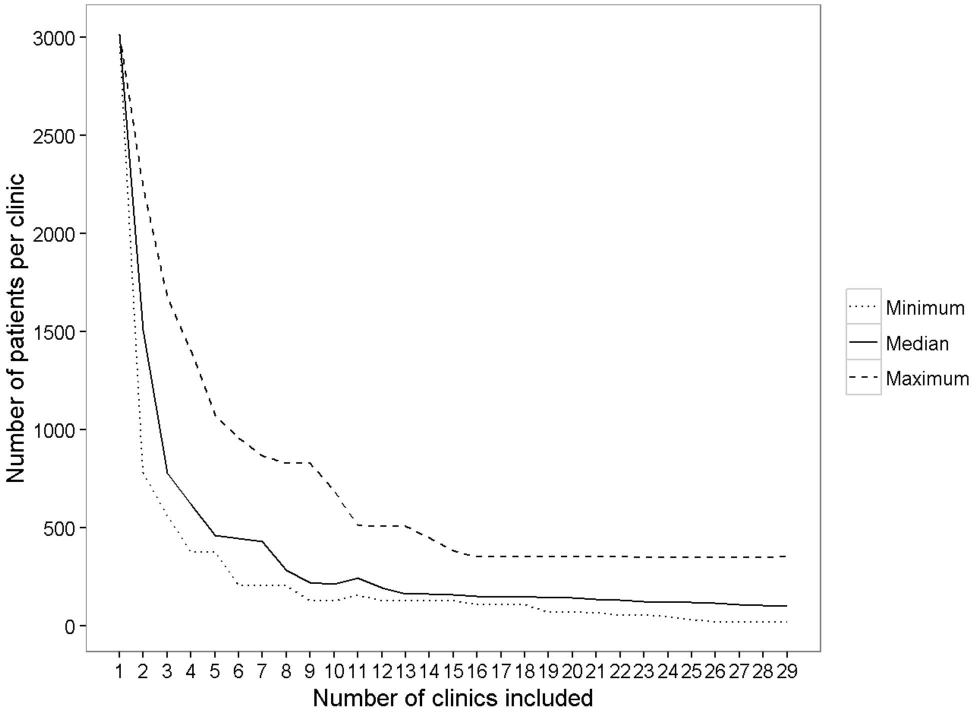


Figure 7.10 summarises the average annual number of patients that would have attended each clinic for each number of clinics included. The graph shows a similar

pattern to the change in travel times with increased numbers of clinics included. A sharp decline in the number of patients per clinic is evident with addition of clinics, up to approximately ten clinics (median 855 patients, IQR 715-1571).

Figure 7.10: Average annual number of tuberculosis patients per clinic for different numbers of optimum clinic configurations, London, 2010-2013. Patients assigned to clinics based on minimum travel times.



7.5 Discussion

7.5.1 Summary of main findings

In this study I have estimated the spatial accessibility of clinics for tuberculosis patients in London using travel time data. More than one third of patients could have used a clinic with a shorter travel time if patients were assigned on this basis rather than through CCG networks; 16% patients travelled at least 30 minutes longer than necessary. If all patients used their closest clinic, the mean patient travel time would have a small but significant decrease, to less than 30 minutes. Services could be rationalised by reducing the number of clinic sites without substantial impacts on mean travel times, provided that patients were able to attend their most convenient clinic regardless of CCG configurations.

7.5.2 Interpretation and implications for policy and practice

Overall, this study supports a move towards commissioning of tuberculosis services using a pan-London approach. This would enable patients to attend their closest clinic rather than being assigned a clinic according to established commissioning arrangements, and therefore improve spatial accessibility of clinics for patients in the short term. It would also raise the possibility of strategic rationalisation of services for system-wide improvements and cost savings.

Providing increased flexibility in the clinics to which patients are referred could allow them to access services that are more convenient. This analysis suggests that one third of patients could travel for less time to reach an existing tuberculosis clinic. Spatial accessibility is particularly important for tuberculosis patients because of their long treatment duration (six months standard regimen) requiring multiple journeys to the clinic. Patients on DOT have to travel to clinics three to five times per week, which is burdensome and potentially stigmatising. Reducing travel time could therefore have implications for treatment completion rates as well as lowering economic costs for patients.[318]

Operating tuberculosis services from a smaller number of well-resourced clinics could also reduce costs to the National Health Service. The results of this study show that, provided locations were selected strategically, this could be achieved without substantial negative impacts on spatial accessibility. Concentration of services would also simplify tuberculosis care networks, and bring together specialists including ancillary services such as translators and support workers for hard to reach groups. This could both improve patient care and lead to greater efficiency, for example by reducing the number of groups involved in the cohort review process. Under this model, it would remain important to ensure that diagnostic services for tuberculosis were widely available, because patients may still present with symptoms at any site.

7.5.3 Study strengths and limitations

A strength of this study was the novel method of estimating travel time from patient residential locations to clinics. By accessing the TfL Journey Planner, I obtained estimates of travel time that made use of different modes of transport including walking, buses and the London Underground system. Using travel time data, as opposed to distances, provided a more realistic approximation of the cost of the journey to the patient.

Another strength was that this analysis was based on comprehensive surveillance data that included residential locations of all patients notified with tuberculosis in London over a four-year period. This means that the spatial accessibility of the clinics was assessed in relation to the locations of people with tuberculosis, as opposed to the general population, which has a different distribution.

This study also had several limitations, primarily that the analysis was based on estimated rather than actual travel times. By selecting the minimum travel time from residential locations to clinics, I assumed that patients would always have a preference for the shortest route in time. In reality, there are other factors that contribute to the choice of journey such as the number of changes between modes of transport and price of the journey. Furthermore, patients may opt to travel to the clinic from a location other than their home, for example from their workplace. These issues could alter the realised accessibility of the clinics.

Another assumption of this study was that the TfL Journey Planner provides an accurate estimate of the travel time between two locations. Walking speeds, for example, will vary between individuals and may be longer for people who are suffering with tuberculosis than the general population. I may therefore have underestimated travel times for which a large proportion of the journey would be made on foot. Travel times may also be affected by the time of day, for example due to the number of services available and how busy they are. I set the journey time at 10.30 am on a week day, which may have underestimated travel times when compared to more typically busy periods.

The combinatorial optimisation approach that I used provided an objective means of assessing optimal combinations of clinics, which could be used to inform service rationalisation. A limitation of this analysis was that, since the algorithm does not test every combination, it cannot be guaranteed to have found the optimum. For example, the algorithm may converge on local optima rather than the global optimum solution.[317] However, in this study, I re-ran the optimisation algorithm for each possible number of total clinics (from 1 to 28). For each increasing number of clinics, the optimum combination identified included the subset of clinics in the previous combination, as opposed to a new subset. It is therefore unlikely that the algorithm had become 'stuck' in local optima as it would have had to occur in the same way on multiple occasions. Another limitation was that this analysis was based on a random sample of 1,000 patients (a restriction used to limit computation time, which in this study took more than two days). Using a different sample may therefore have resulted in different combinations of clinics.

Finally, the analysis presented here does not take into account other dimensions of health care accessibility which are important in planning services. For example, clinic facilities, space and staffing may influence whether it could accommodate additional patients. Previous analyses of accessibility in other contexts have developed methods which incorporate some of these other factors. Gravity models, for instance, account for the diminishing attractiveness of services with increased distance and demand from the population for limited services.[319] Provider population relationship measures include a simple ratio of provider supply to patient demand.[320] The two-step floating catchment area method is a more sophisticated technique which consists of overlapping catchments of both service provision and resident utilisation.[309] It brings together elements of provider population relationship and gravity models, and includes availability relative to demand and distance between services and residents. However, these models are based on distance and were developed for analysis of primary care usage, and may not be suitable for specialised services such as tuberculosis, or for incorporation of public transport travel time data. They were designed to identify areas with poor or good accessibility, rather than to inform service rationalisation.

7.5.4 Future directions

This study may prompt further research in tuberculosis service planning and other areas. For example, to determine whether the assumptions of this study regarding travel time and spatial accessibility are appropriate, a qualitative study of patient preferences may be useful. This could investigate attitudes towards different modes of transport; large specialist clinics versus smaller local centres, and access from home or other locations.

The methods used in this study could also be applied to other healthcare service planning problems. Although here I used the optimisation algorithm to consider the effects of rationalising services, it could also be used to consider different alternative combinations of new locations of services, such as selection of a new site for a tuberculosis clinic given a set of potential options. Incorporation of travel time data into routine health systems could also be used to improve patient care by referring to centres that are most convenient.

Funding restrictions in the NHS mean that opportunities for improving quality whilst reducing costs are attractive. This has been demonstrated by the NHS *Five Year Forward View* which promotes concentration of services for rare diseases, following the successful outcomes when this was implemented for stroke services.[306] Service planning using the approaches outlined here presents a potential means of informing concentration of services, and could be considered for any rare disease.

Box 7.1: Summary of Chapter 7.

- Tuberculosis services in London are currently commissioned individually by the 32 CCGs in the city, and patients diagnosed with the disease are referred to clinics commissioned by the CCG. Moving to a pan-London commissioning model would enable patients to travel to their nearest clinic, rather than using these established arrangements.
- More than one third of patients could have travelled for a shorter time to reach a tuberculosis clinic if they were referred on the basis of travel time rather than CCG networks, and 16% of patients travelled at least 30 minutes longer than necessary.
- Removing individual clinics would have small overall impacts on travel time, provided that patients were then able to attend the clinic with the shortest travel time.
- Using an optimum combination of 18 of the 29 clinics in London could provide mean travel times of less than 30 minutes; ten clinics could provide a mean travel time of 34 minutes, and only three clinics would be required for mean travel times of less than 45 minutes.
- A smaller number of specialised clinics could therefore theoretically provide services for tuberculosis patients in London without impacting spatial accessibility.

8 SUMMARY OF RESEARCH AND MAIN FINDINGS, RECOMMENDATIONS AND FINAL CONCLUSIONS

8.1 DESCRIPTION OF CHAPTER CONTENTS

In this chapter I review the background and rationale for this thesis, and summarise the research undertaken. I outline the key findings of the work, combining evidence from this thesis with other research, and summarise strengths and limitations. I make recommendations for the control of tuberculosis. I also develop a framework for application of spatial methods during prospective outbreak investigations, and outline examples of how they could be applied in investigations of sexually transmitted and foodborne infections. Finally, I indicate areas for future research resulting from this work.

8.2 SUMMARY OF RESEARCH UNDERTAKEN

8.2.1 Background and rationale

Tuberculosis remains a leading infectious cause of death worldwide.[1] In low incidence countries such as England, the problem is concentrated in large cities; and in London the highest rates are found amongst deprived populations.[5,7] Although there have been great advances in control of the disease worldwide in the last 20 years, rates in England have not declined substantially since 2000.[2,5] With the publication of the WHO's *End TB Strategy* in 2015, there has been a change in emphasis from control of the disease to elimination.[36] To achieve this, locally tailored responses must be developed which are informed by appropriate data including spatial information.[37] There are numerous methods available which use spatial data for visualisation, description, cluster detection and modelling of diseases. Effective use of these methods is important to support control of tuberculosis as well as other diseases, for example by identifying areas of intensive transmission and directing interventions.

The work presented in this thesis aimed to explore the use of spatial methods to support local tuberculosis investigations, with a particular focus on control of the disease in London.

8.2.2 Summary of methods and results

A systematic literature review (Chapter 2) described spatial methods that have been used in previous infectious disease outbreak investigations, and demonstrated that they can generate important insights. It also showed that there was scope for much wider implementation of spatial methods, and development of new tools which enable spatial data to be explored more easily. One such tool, an interactive, open-source application for producing disease point maps was developed with the aim of supporting tuberculosis cluster investigations (Chapter 3). Spatial data were combined with molecular strain typing information to investigate transmission of tuberculosis in London (Chapter 4). Results of this analysis suggested that long chains of transmission contribute a substantial proportion of the cases of tuberculosis in London, and highlighted the importance of social deprivation as a risk factor for transmission. A detailed analysis of one of these clusters (Chapter 5) showed that, despite cases being reported for 20 years, the strain did not spread significantly from specific risk populations or geographic areas. Geographic profiling, a novel tool for identifying sources of infectious diseases, showed some utility for prioritising areas for control during tuberculosis cluster investigations (Chapter 6), but further research is required to validate this method. An analysis of patient travel times to tuberculosis clinics in London (Chapter 7) demonstrated that a city-wide commissioning model, which allowed patients to use their nearest clinic, could increase spatial accessibility. It could also lead to improvements in services, by concentrating expertise, without impacting travel times.

This work has generated new knowledge about tuberculosis transmission and the utility of spatial methods for investigation of infectious diseases (see list of publications, page 7). Key findings, strengths and limitations are summarised below, and placed in the context of previous studies.

8.3 Key findings

8.3.1 Tuberculosis transmission and control

1. Large numbers of tuberculosis cases in London have resulted from local transmission.

Supporting evidence

i. Analysis of cases of tuberculosis linked through molecular strain typing showed that more than one in ten cases in London were part of a large cluster (of more than 20 cases).

ii. Large molecular clusters showed pronounced spatial clustering identified through scan statistics and visualised using smoothed incidence maps. This indicates that common strains occur amongst spatially aggregated groups, supporting the hypothesis of transmission.

iii. There was some evidence that the initial cases in large molecular clusters occurred closer together in space than cases in smaller molecular clusters; cases located no more than 2km apart had 2.5 times the unadjusted odds of growing to a large cluster than a small cluster.

iv. An outbreak of isoniazid-resistant disease showed significant spatial clustering which persisted over 20 years, and affected specific risk groups. This was demonstrated through k-function analysis and visualised using smoothed incidence maps. It was also supported by geographic profiling analysis, which revealed a spatial peak which had very high probability of being an area in which transmission is likely to have been ongoing.

v. A similar large outbreak in another European city (Stockholm) involved community transmission of tuberculosis amongst distinct risk groups in a small region.[181] Previous studies of molecular clustering in other settings have also showed spatial clustering of linked cases, indicative of ongoing transmission.[190-193]

vi. Previous research using whole genome sequencing of community-based clusters of tuberculosis in the Midlands of England indicated substantial community transmission.[209] The study identified likely 'super-spreaders', which are particularly infectious individuals who infect large numbers of secondary cases.[209]

Strengths and limitations

i. Use of routine surveillance data enabled identification of cases in molecular and spatial clusters who were likely to have been linked through transmission. However, full contact investigations would be required to confirm epidemiological links, as molecular clusters could also reflect common endemic strains.

ii. Molecular strain typing was based on 24-locus MIRU-VNTR, the most specific method available at the time. Some variations within strains may not have been detected by this method.

iii. A strength of the analysis of the isoniazid-resistant outbreak was that it combined data from multiple sources and enabled linking of cases prior to the start of routine MIRU-VNTR typing. However, this also meant that the case definition for the outbreak evolved over time, and it is therefore possible that all cases were not part of the same outbreak.

2. Social complexity and deprivation are associated with tuberculosis transmission.

Supporting evidence

i. An increased count of individual-level social risk factors (history of homelessness, imprisonment, misuse of drugs or alcohol) was associated with being in a large molecular cluster. Having one of these risk factors increased the odds of being in a large cluster by 1.36 times, having four risk factors resulted in more than 16 times the odds.

ii. Area-level deprivation was also independently associated with being in large clusters. The odds of being in a large cluster increased by 10% with increasing quintile of deprivation.

iii. Spatial clusters of cases occurred in more deprived areas of London. The median IMD rank of LSOAs in spatial clusters was 1,110 compared to 2538.5 for LSOAs not in spatial clusters.

iv. Almost two thirds of cases in the isoniazid-resistant outbreak had at least one social risk factor. These risk factors remained important as the outbreak persisted for two decades.

v. Cases with social risk factors in the isoniazid-resistant outbreak were less likely to complete treatment (42% completed at 12 months) than those with no social risk factors (67% completed at 12 months), increasing the duration of infectiousness for these patients and therefore the likelihood that they would transmit the infection.

vi. Routine surveillance has shown that the rate of tuberculosis in the most deprived decile of England was more than eight times the rate in the least deprived decile.[5]

vii. Previous research has indicated that the link between social deprivation and increased risk of tuberculosis is related to poor ventilation and overcrowding; malnutrition; social and economic barriers and stigma.[321]

Strengths and limitations

i. Use of routine surveillance data for the analysis of molecular clusters ensured that there was relatively little missing data about social risk factors (less than 5% for homelessness, prison history and drug use; less than 10% for alcohol misuse).

ii. A limitation of the IMD is that it is an area-level measure of deprivation and may therefore not accurately reflect individual characteristics.

iii. Analysis of the isoniazid-resistant outbreak benefitted from combining data from multiple sources. Supplementing routine data with information from Find and Treat enabled improved estimates of the prevalence of social risk factors in this population. However, the relative importance of social risk factors in the outbreak was not assessed in a formal case control study because data from the same sources were not available for non-outbreak cases to make a reliable comparison.

iv. Further work is required to determine how actions on reducing social determinants will impact on rates of tuberculosis.[321]

3. Pan-London commissioning could improve tuberculosis services by enhancing spatial accessibility.

Supporting evidence

i. Analysis of travel time data showed that if patients were able to travel to their closest tuberculosis clinic, rather than being assigned a clinic based on pre-existing commissioning arrangements, the mean travel time would be estimated to reduce by more than five minutes (from 33 minutes to 27.5).

ii. An optimum combination of 18 of the 29 clinics in London could ensure mean travel times of less than 30 minutes; and mean travel times of 34 minutes could be achieved with an optimum combination of ten clinics.

iii. Find and Treat and the London Extended Contact Tracing team are examples of successful services that have been funded by pan-London commissioning.[215,304]

Strengths and limitations

i. The analysis was based on travel time, which provides a more realistic assessment of spatial accessibility of clinics in London than distance-based measures. However it used estimated rather than actual travel times, and did not consider other factors that may affect travel time such as price of journey, preference for different modes of transport, or travel from locations other than homes.

ii. Use of an optimisation algorithm to define sets of clinics that would minimise travel times provided an objective means of selecting service locations in London, but was based on a random sample of 1,000 cases.

iii. The analysis did not take into account other measures of healthcare accessibility such as the current resources of the clinics in terms of facilities, space, and staffing, which could influence whether the clinic would be able to accommodate larger numbers of patients if it was developed into a specialist centre.

8.3.2 Use of spatial methods

1. Spatial methods provide an important complementary tool to epidemiological analyses.

Supporting evidence

i. The systematic literature review demonstrated that spatial methods are useful for investigations of a wide range of infectious diseases; outbreaks in various contexts, and in different types or stages of investigations. They provided important insights that informed public health actions. Maps were particularly valued as a means of communicating findings to health officials, policy makers and the public.

ii. Spatial clustering analysis of molecular clusters of tuberculosis in London pinpointed groups of cases that were likely to be linked through transmission. This could assist with prioritising molecular clusters for further investigation.

iii. Analysis of the spatial distribution of the isoniazid-resistant outbreak contributed to understanding of how the outbreak is likely to have progressed. A hypothesis consistent with the data is that the majority of cases were infected relatively early in the outbreak and have gradually progressed from latent to active disease.

iv. Interactive dot maps, produced using the dot mapping application developed in Chapter 3, were used to generate hypotheses about epidemiological links between cases in molecular clusters of tuberculosis. Rapid interrogation of data using this tool enabled identification of shared characteristics of cases that were close in space.

v. Analysis of two methods of spatial targeting, geographic profiling and a ring cull, showed that incidents of bovine tuberculosis in cattle were insufficiently clustered around setts with infected badgers to design an efficient spatially targeted cull. This added further weight to the conclusion of the RBCT which stated that badger culling would be unlikely to contribute usefully to the control of cattle tuberculosis in Britain.[282]

Strengths and limitations

i. The systematic review benefitted from a broad set of inclusion criteria, encompassing any infectious disease outbreak, and therefore identified a wide range of applications to investigations. It used a robust methodology following relevant sections of the PRISMA guidelines. However, it may have suffered from publication bias if spatial methods were only reported when they were found to be useful.

ii. Spatial clustering of molecular clusters of tuberculosis in London was demonstrated using two complementary methods (scan statistics and smoothed incidence maps) which indicated similar regions of likely transmission. A limitation of these analyses is that they did not take into account time, so it is possible that sporadic cases that occurred close in space were not a true cluster resulting from transmission.

iii. The dot mapping application was designed to assist with tuberculosis cluster investigations and similar epidemiological analyses, and incorporated feedback from potential users during development. However, an online survey about the application received few respondents and evaluation of the final version was therefore limited.

2. Novel spatial tools are required to support epidemiological investigations.

Supporting evidence

i. In the systematic review, it was estimated that less than half a percent of published outbreak investigations used spatial methods, despite the improved understanding that they can provide.

ii. Spatial methods were used particularly infrequently to investigate outbreaks in developing counties. There were ten reports included in the systematic review of outbreaks in Africa, the same number as in the UK alone. These methods were also implemented less frequently for investigations without a suspected environmental point source, such as foodborne and sexually transmitted infections.

iii. Barriers preventing further use of existing tools included the expense of specialist GIS software; requirement for specialist training to operate it, and lack of flexibility in tools available.

iv. A previous systematic review found that users have preference for dynamic, interactive graphics for data exploration.[157] Technology for generating such graphics is now widely available, but most tools for displaying spatial data produce static, non-interactive graphics.

v. Simple dot maps are the most frequently used form of spatial visualisation. They are relatively easy to produce, but are limited because they do not take into account the distribution of the underlying population. Spatial scan statistics are the analytic method used most often. They are easily implemented using the free programme, SaTScan. This suggests that there is a demand for more sophisticated analyses if they can be implemented easily.

vi. The dot mapping application developed in Chapter 3 had mostly positive feedback. It provides a free tool with interactive interface that does not require training to operate and can be integrated into routine practice.

Strengths and limitations

i. The systematic literature review included studies published concerning any infectious disease and in any language. It may have been limited if investigations of outbreaks occurring in different countries had a different likelihood of being published.

ii. The degree to which spatial methods are implemented may also be limited by their perceived utility. Limitations inherent to these methods include lack of specificity in results and requirement to select clustering parameters. They can also be subject to the ecological fallacy (the assumption that an individual's characteristics are the same as those for the group to which the individual belongs), and the modifiable aerial unit problem (that patterns are sensitive to changes in the boundaries into which they are grouped), which make visualisations of aggregated spatial data vulnerable to misinterpretation.

iii. Confidentiality issues may also prevent investigators accessing, creating or publishing presentations of spatial data. Provided that spatial data are available, precise visualisations can be shared within secure networks. They can also be shared more widely if maps are presented at a larger scale or by introducing random error to alter exact positions of points and preserve anonymity.

3. Geographic profiling may assist with epidemiological investigations of infectious diseases in some circumstances.

Supporting evidence

i. Analysis of the isoniazid-resistant tuberculosis outbreak using geographic profiling identified an area in which there was high probability that the infection had been transmitted. This could be used to design a strategy for spatial targeting of interventions.

ii. Geographic profiling of a second molecular cluster of tuberculosis was used to rank potential venues of transmission in a smaller town in England. These results showed some correlation with those from a cluster investigation questionnaire in which cases were asked to list venues that they had attended.

iii. The case study of bovine tuberculosis in cattle and badgers demonstrated how this method could be used when data are available about the potential sources of infection. In this example, different hypotheses about disease transmission were tested, and showed that cattle tuberculosis incidents did not appear to cluster around badgers that were infected with tuberculosis.

iv. A previous study has demonstrated the utility of geographic profiling in analysis of a series of cases of malaria, from which mosquito breeding sites were successfully identified.[266]

v. It is a quick, cheap method and may therefore be useful in the hypothesis generation stage of an investigation, when information from cases about potential transmission venues is not available.

vi. Interpreting the results of geographic profiling analysis can be challenging, because the hit score measures are not comparable between studies, i.e. a low hit score in one study area does not indicate the same level of certainty as a low hit score in another study area.

vii. Case studies presented here also demonstrated that geographic profiling alone cannot provide definitive conclusions about locations of transmission.

Strengths and limitations

i. The geographic profile of the isoniazid-resistant outbreak highlighted an area of likely transmission but there was no data on venues visited by cases with which to validate this hypothesis.

ii. In the investigation of a molecular cluster of tuberculosis the results of a geographic profile analysis were compared to a cluster investigation questionnaire. Theoretically, this could be used to validate the results of the geographic profile analysis, however, the questionnaire had responses from less than a quarter of those in the tuberculosis cluster so these results were not highly reliable.

iii. The previous study of cases of malaria benefitted from availability of data on potential sources from an entomological survey of water bodies.[266] However, this survey was not contemporaneous with reporting of cases. The analysis was also based on a convenience sample of cases reported in the area and therefore will not have included all cases of malaria over the study period. Results from the geographic profile were compared with simple measures of spatial central tendency. However, the utility of geographic profiling could have been tested through a more rigorous comparison with an alternative method of spatial targeting, such as the ‘ring cull’ approach demonstrated in the analysis of bovine tuberculosis (Chapter 6)

8.3.3 Summary of evidence generated

Overall this work has described several important features of tuberculosis transmission in London and highlighted how spatial methods can assist with investigations of outbreaks. It has shown, for the first time, that there are multiple large clusters of tuberculosis in London which appear to be linked through transmission. It has also added further evidence of the importance of social deprivation in promoting transmission of the disease, and suggested a model for commissioning of services to improve quality. Spatial analyses have provided evidence to support these conclusions; and this research has also identified ways in which further use of spatial methods could improve control of infectious diseases more generally. Implications and recommendations for policy and practice derived from this work are described below.

8.4 IMPLICATIONS AND RECOMMENDATIONS FOR POLICY AND PRACTICE

8.4.1 Control of tuberculosis

i. Molecular clusters of tuberculosis that show significant spatial clustering should be prioritised for detailed cluster investigation. Routine integration of spatial scan statistics to analysis of molecular strain typing data would enable these clusters to be identified promptly.

ii. Spatial and epidemiological characteristics of cases in molecular clusters should be interrogated simultaneously to identify cases that may be linked through transmission. The interactive dot mapping application developed in Chapter 3 provides a means of doing this that does not require specialist expertise in GIS.

iii. Cluster investigations should focus on groups of more than two cases, because characteristics of the initial two cases in molecular clusters are not highly predictive of cluster growth.

iv. Interventions aiming to raise awareness of the risks of tuberculosis should be targeted to populations at risk of transmitting the disease. These include individuals in black ethnic groups, who were born in the UK, and live in more deprived areas of London. Health professionals, social support services, and relevant community groups in these areas should also be made aware of the increased risk of tuberculosis disease in these populations to improve the chances of an early diagnosis.

v. Further efforts to control the disease in individuals with social risk factors are required. Reducing incidence in these populations may have a substantial effect on the overall incidence of tuberculosis because individuals in these groups are more likely to be part of long chains of transmission. The main example of such a targeted service is Find and Treat, which uses a mobile radiography unit to screen for cases actively in vulnerable populations in London and supports patients to complete treatment. Examples of other initiatives could include screening in prisons, which is an opportunity to contact some of the individuals in these 'hard-to-reach' groups.

vi. Screening for latent infection as well as active disease should be considered. To achieve elimination of tuberculosis, cases must be identified before they become infectious. The results from this work suggest that large outbreaks can be confined to relatively predictable populations, indicating that screening in appropriate groups could effectively identify instances of latent infection.

vii. A pan-London commissioning model for tuberculosis services should be considered. It would provide the opportunity to focus treatment services in high quality centres without impacting spatial accessibility for patients.

viii. Measures for control of bovine tuberculosis in cattle should not include spatially targeted badger culls. Evidence from this study shows that these culls are highly inefficient means of locating badgers that are infected with tuberculosis, and therefore unlikely to remove substantial numbers of infected badgers without culling over extremely large areas. Previous studies also show that these culls may lead to an increase in cattle tuberculosis incidence, as well as being expensive and damaging to wildlife populations.[282]

8.4.2 Use of spatial methods

This thesis has shown that there are numerous spatial methods that can be used to support outbreak investigations of infectious diseases such as tuberculosis. Table 8.1 synthesises findings into a framework for application of spatial methods during prospective outbreak investigations. Steps in outbreak investigations are adapted from the ECDC Field Epidemiology manual (Chapter 2, Box 2.1).[66]

Table 8.1: Application of spatial methods to steps in outbreak investigation.

1. Establish the existence of an outbreak	Visualise case distribution (e.g. dot map). Identify and confirm clustering (e.g. spatial scan statistics; point process modelling).
2. Confirm diagnosis	Spatial methods alone cannot confirm diagnoses. However, spatial epidemiology of infection should be considered to develop preliminary diagnostic hypotheses.
3. Define and identify outbreak cases	Set geographic limits in which cases are considered part of the outbreak (e.g. post code area; hospital ward). Select controls in case-control study based on same geographic limits. Use maps to assist with active case finding.
4. Describe cases and develop hypotheses	Visualise distribution of cases in relation to known risk factors or potential sources (e.g. rate map, thematic maps). Describe progression of outbreak (e.g. using interactive dot mapping application, smoothed incidence maps). Identify centre of outbreak (e.g. spatial mean). Identify high-risk areas (e.g. attack rates in zones at different distances from potential sources). Assess likelihood of potential sources (e.g. geographic profiling).
5. Evaluate hypotheses and draw conclusions	Test for overall clustering (e.g. k-function analysis). Locate significant clusters (e.g. spatial scan statistic). Identify significant trends in attack rates with distance from potential sources (e.g. linear regression of attack rates).
6. Compare with established facts	Calculate maximum dispersal distance from probable source to cases. Model concentrations of infected particles to understand transmission dynamics (e.g. computational fluid dynamics; atmospheric modelling).
7. Execute prevention measures	Spatial targeting of interventions to control outbreak (e.g. order to boil water in area served by contaminated reservoir).

	Spatial targeting of health promotion campaigns (e.g. using post codes on social networks). Identify geographic areas at risk of future outbreaks (e.g. risk mapping).
8. Communicate findings	Use maps to communicate results to health officials/ policy makers, to the public, and in scientific journals.

Although the work that has been presented in this thesis has focused on the control of tuberculosis in London, many of the methods could be applied to other infectious diseases in different contexts. Below I provide examples of how spatial methods could be used to support investigations of STIs and foodborne infections.

8.4.2.1 Example: Application to investigations of STIs

Outbreaks of STIs are a continued public health problem in the UK. Recent incidents of concern have included outbreaks of drug-resistant gonorrhoea, and emerging or re-emerging infections such as shigella and syphilis.[322-324] The spread of some of these infections is thought to have been exacerbated by increased use of social networking applications to find sexual partners.[325,326] These applications, which are often based on geographical proximity, can increase the size and connectivity of sexual networks.[327] Consideration of geographical information is therefore important when investigating STI outbreaks.

Provided that accurate location data were available, STI outbreak investigations could be assisted by implementing many of the methods described in Table 8.1. For example, in a recent gonorrhoea outbreak in England that affected more than 300 people, it was hypothesised that the infection was circulating in young heterosexuals in a localised area.[328] A test of spatial clustering could have been used to investigate this hypothesis by comparing the geographical concentration of cases in this group with those in a control group, such as men who have sex with men. In this investigation, an innovative means of communicating health promotion messages was used which involved targeting users on social media sites based on their age and post code of residence. Inclusion of simple maps within these messages that highlight the areas most at risk may be an effective approach to further increase engagement in similar campaigns in the future.

Spatial targeting tools may also help to identify areas in which STIs have been transmitted. For example, in an investigation of hepatitis B infection in men who have sex with men, a truck stop was implicated as a possible transmission venue.[329] Geographic profiling may have corroborated results from interviews with cases, strengthening the rationale for health promotion interventions.

Travel time data could be used to inform planning of STI services. For example, it could be used to investigate spatial accessibility of genitourinary medicine clinics and determine optimum locations for new services. It could also be used to direct individuals identified through contact tracing to their nearest services for screening.

8.4.2.2 Example: Application to investigations of foodborne infections

Investigations of foodborne illness have some challenges that are distinct from those for tuberculosis or STIs. For example, the incidence is much higher, with an estimated 17 million annual cases of infectious intestinal disease in the UK,[330] and a smaller proportion of cases are likely to present to healthcare services. Outbreaks tend to be shorter owing to reduced incubation periods, and must therefore be detected more rapidly. There is also a wider range of pathogens that can cause similar symptoms, and the burden of disease that is foodborne, rather than resulting from, for example, institutional transmission, is not well understood.[330] In spite of these challenges, similar spatial methods can be used to assist with investigations.

In detection of outbreaks, statistical methods that combine measures of spatial proximity with temporal aggregation can be effective. For example, a point process model that incorporates seasonality has been used to identify spatio-temporal anomalies in reports of non-specific gastrointestinal infections in the UK.[331] This method, which is based on consultations from NHS Direct, a telephone clinical advice service, has been incorporated into an online surveillance system which could allow earlier detection than relying on confirmed laboratory reports.

Following identification of an outbreak, tools such as the interactive mapping application developed in Chapter 3 could be used to track its progression and characteristics. For example, recent multinational outbreaks of *Salmonella* have involved complex investigations of epidemiological, molecular, food-chain, and environmental information.[332,333] Use of the interactive mapping tool could enable consideration of all of these data in their geographic context, and at different geographic scales, allowing links to be identified.

In smaller-scale outbreaks in which a point source is suspected, it may be of value to trial geographic profiling methods. These investigations often aim to identify suppliers of contaminated products, or premises from which they have been purchased.[333,334] However, as with tuberculosis cluster investigations, it can be challenging to acquire high rates of responses to questionnaires about food exposures. In such situations, geographic profiling may assist by narrowing a search to a set of likely vendors, potentially enabling more rapid identification of implicated premises, and therefore implementation of control measures.

8.5 FUTURE DIRECTIONS

8.5.1 Control of tuberculosis

i. Recommendations for improvements to tuberculosis molecular cluster investigations have emerged from this work (Section 8.4.1). They have included approaches to prioritise clusters for detailed investigation and the use of mapping to aid identification of epidemiological links. A prospective evaluation of these methods could enable development of specific guidelines, such as thresholds for investigation of spatial clusters, and assessment of their utility in practice.

ii. Predicting the growth of molecular clusters based on the characteristics of initial cases in the cluster is an attractive possibility. Repeating this analysis when more data are available would increase the power of the study and may therefore reveal undetected associations.

iii. Social deprivation has been highlighted as an important factor for tuberculosis transmission. Further work in this area could aim to identify which of the elements of the IMD are important in promoting transmission. This could be used to inform environmental or housing interventions, such as improving ventilation and reducing overcrowding.

iv. Whole genome sequencing has been used in previous work to identify potential super-spreaders, who represent the molecular 'centres' of tuberculosis clusters.[209] It would be useful to combine these results with maps and spatial clustering analyses to determine the extent to which these molecular 'centres' correlate with geographic centres of clusters.

8.5.2 Use of spatial methods

i. Provision of training for individuals working in public health, for example through short courses, may increase the regularity with which spatial tools are used in practice.

ii. Protection of patient confidentiality is an important consideration when using spatial data. Maintaining secure systems through which public health professionals can share and interrogate sensitive data is therefore vital. For example, hosting the interactive mapping application developed in Chapter 3 on a secure server would enable sharing of data and could therefore allow it to be explored by multiple users.

iii. Consistency of collection and reporting of spatial data could be improved by extension of the STROBE statement. This could include, for example, items relating to the precision of the location data collected; how it was collected; sources for population denominators, and guidelines for presentation of maps.

iv. Novel tools should be developed to improve the use of spatial data in cluster investigations. For example, integrating map interfaces into data collection systems could support tuberculosis cluster investigations by enabling patients to point out exact locations that they had visited and may have transmitted the disease. Real-time visualisations of data linked with surveillance systems would also support investigations by improving timeliness with which data are assessed.

v. Engagement of potential users and gathering feedback is a challenge for integration of novel tools into practice. This was demonstrated by the low numbers of responses to the evaluation survey for the mapping application in Chapter 3. Improving the uptake of such tools requires an environment which fosters innovation. In organisations such as PHE, a possible way of achieving this would be establishment of a special interest group of 'beta-testers' to test new tools and suggest improvements.

vi. Integration of novel spatial data into epidemiological analysis is likely to be an increasing challenge in the coming years. For example, GPS-enabled devices including watches and fitness trackers as well as smartphones, are becoming increasingly prevalent, and the connectivity of other everyday items is also increasing ('the internet of things'). Data on individuals' movements rather than simply point locations are therefore likely to become available. Such data have many potential applications in epidemiology, particularly in measurement of exposures, but methods for analysis are not well developed.

8.6 CONCLUSIONS

This thesis has improved understanding of the epidemiology and transmission of tuberculosis in London. The studies provide evidence that tuberculosis is maintained in large chains of transmission in distinct populations in the city. They point to specific recommendations for improvement of cluster investigations, identification of new cases, and organisation of services.

The thesis also explored the use of spatial methods for investigation of outbreaks of infectious diseases including tuberculosis. It has demonstrated that spatial analyses can make many valuable contributions to these investigations, particularly when synthesised with other information on time and person. Simple maps alone provide fundamental insights about the distribution of cases. However, advancements in GIS technology and increasing availability of good quality spatial data provide an opportunity for development and implementation of more sophisticated techniques. Adoption of these new techniques, and wider use of existing methods, have the potential to support more effective investigations and therefore limit the public health impacts of infectious disease outbreaks in future.

9 REFERENCES

1. World Health Organization. Tuberculosis: Fact sheet No. 104. http://www.who.int/mediacentre/factsheets/fs104/en/ (accessed 30 August 2016).

2. World Health Organization. Global Tuberculosis Report 2015. 2015. http://apps.who.int/iris/bitstream/10665/191102/1/9789241565059_eng.pdf (accessed 30 August 2016).

3. Selwyn PA, Hartel D, Lewis VA, et al. A prospective study of the risk of tuberculosis among intravenous drug users with human immunodeficiency virus infection. *The New England journal of medicine* 1989; **320**(9): 545-50.

4. Van den Broek J, Borgdorff MW, Pakker NG, et al. HIV-1 infection as a risk factor for the development of tuberculosis: a case-control study in Tanzania. *International journal of epidemiology* 1993; **22**(6): 1159-65.

5. Public health England. Tuberculosis in England: 2015 report version 1.1: Public Health England: London, 2015.

6. Public Health England. Tuberculosis in the UK: 2014 report. London, 2014.

7. de Vries G, Aldridge RW, Cayla JA, et al. Epidemiology of tuberculosis in big cities of the European Union and European Economic Area countries. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* 2014; **19**(9): pii: 20726.

8. Public Health England. Tuberculosis in London: Annual review (2014 data): Public Health England: London, 2015.

9. Story A, Murad S, Roberts W, Verheyen M, Hayward AC. Tuberculosis in London: The importance of homelessness. problem drug use and prison. *Thorax* 2007: 667-71.

10. Daniel TM. The history of tuberculosis. *Respiratory Medicine* 2006; **100**(11): 1862-70.

11. Lawn SD, Zumla AI. Tuberculosis. *The Lancet* 2011; **378**(9785): 57-72.

12. World Health Organization. Treatment of tuberculosis: guidelines. Geneva: World Health Organization, 2010.

13. Nienhaus A, Schablon A, Costa JT, Diel R. Systematic review of cost and cost-effectiveness of different TB-screening strategies. *BMC Health Services Research* 2011; **11**(1): 1-10.

14. Smith C, Abubakar I, Thomas HL, Anderson L, Lipman M, Reacher M. Incidence and risk factors for drug intolerance and association with incomplete treatment for tuberculosis: analysis of national case registers for England, Wales and Northern Ireland, 2001-2010. *Thorax* 2014; **69**(10): 956-8.

15. Dye C, Williams BG. Eliminating human tuberculosis in the twenty-first century. (1742-5689).

16. Styblo K. The relationship between the risk of tuberculosis infection and the risk of developing infectious tuberculosis. *Bull Int Union Tuberc Lung Dis* 1985; **60**(3-4): 117-9.

17. Centers for Disease Control and Prevention. The Difference Between Latent TB Infection and TB Disease. http://www.cdc.gov/tb/publications/factsheets/general/ltbiandactivetb.htm (accessed 30 August 2016).

18. Addiss DG, Davis JP, LaVenture M, Wand PJ, Hutchinson MA, McKinney RM. Community-acquired Legionnaires' disease associated with a cooling tower: evidence for longer-distance transport of Legionella pneumophila. *American journal of epidemiology* 1989; **130**(3): 557-68.

19. Boehme CC, Nabeta P, Hillemann D, et al. Rapid Molecular Detection of Tuberculosis and Rifampin Resistance. *New England Journal of Medicine* 2010; **363**(11): 1005-15.

20. Boehme CC, Nicol MP, Nabeta P, et al. Feasibility, diagnostic accuracy, and effectiveness of decentralised use of the Xpert MTB/RIF test for diagnosis of tuberculosis and multidrug resistance: a multicentre implementation study. *The Lancet* 2011; **377**(9776): 1495-505.

21. Penz E, Boffa J, Roberts DJ, et al. Diagnostic accuracy of the Xpert(R) MTB/RIF assay for extra-pulmonary tuberculosis: a meta-analysis. *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease* 2015; **19**(3): 278-84, i-iii.

22. Iseman MD. Evolution of drug-resistant tuberculosis: a tale of two species. *Proceedings of the National Academy of Sciences of the United States of America* 1994; **91**(7): 2428-9.

23. Shah NS, Wright A, Bai G-H, et al. Worldwide Emergence of Extensively Drug-resistant Tuberculosis. *Emerging Infectious Diseases* 2007; **13**(3): 380-7.

24. Jagielski T, van Ingen J, Rastogi N, et al. Current Methods in the Molecular Typing of Mycobacterium tuberculosis and Other Mycobacteria. *BioMed Research International* 2014; **2014**: 21.

25. Mathema B, Kurepina NE, Bifani PJ, Kreiswirth BN. Molecular Epidemiology of Tuberculosis: Current Insights. *Clinical Microbiology Reviews* 2006; **19**(4): 658-85.

26. Hayward AC. Restriction fragment length polymorphism typing of Mycobacterium tuberculosis. *Thorax* 1995; **50**(11): 1211-8.

27. Centers for Disease Control and Prevention. Core Curriculum on Tuberculosis: What the Clinician Should Know. 2013. https://www.cdc.gov/tb/education/corecurr/ (accessed 30 August 2016).

28. Xie X, Li Y, Chwang ATY, Ho PL, Seto WH. How far droplets can move in indoor environments – revisiting the Wells evaporation–falling curve. *Indoor Air* 2007; **17**(3): 211-25.

29. World Health Organization. Definitions and reporting framework for tuberculosis - 2013 revision. Geneva: World Health Organization,, 2013.

30. Karumbi J, Garner P. Directly observed therapy for treating tuberculosis. *Cochrane Database of Systematic Reviews* 2015; (5).

31. Story A, Aldridge RW, Abubakar I, et al. Active case finding for pulmonary tuberculosis using mobile digital chest radiography: an observational study. *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease* 2012; **16**(11): 1461-7.

32. Fox GJ, Barry SE, Britton WJ, Marks GB. Contact investigation for tuberculosis: a systematic review and meta-analysis. *The European Respiratory Journal* 2013; **41**(1): 140-56.

33. World Health Organization. Tuberculosis vaccine development. http://www.who.int/immunization/research/development/tuberculosis/en/ (accessed 30 August 2016).

34. Yates TA, Tanser F, Abubakar I. Plan Beta for tuberculosis: it's time to think seriously about poorly ventilated congregate settings. *International Journal of Tuberculosis and Lung Disease* 2016; **20**(1): 5-10.

35. Li Y, Leung GM, Tang JW, et al. Role of ventilation in airborne transmission of infectious agents in the built environment - a multidisciplinary systematic review. *Indoor Air* 2007; **17**(1): 2-18.

36. World Health Organization. The End TB Strategy. 2015. http://www.who.int/tb/strategy/en/ (accessed 30 August 2016).

37. Theron G, Jenkins HE, Cobelens F, et al. Data for action: collection and use of local data to end tuberculosis. *The Lancet* 2015; **386**(10010): 2324-33.

38. Fenner F, Henderson D, Arita I, Jezek Z, Ladnyi I. Smallpox and its eradication. Geneva: World Health Organization, 1988.

39. World Health Organization. Global Polio Eradication Initiative Strategic Plan: 2004-2008: World Health Organization, 2003.

40. Koch T. Disease Maps: Epidemics on the Ground: University of Chicago Press; 2011.

41. Koch T. 1831: the map that launched the idea of global health. *International journal of epidemiology* 2014; **43**(4): 1014-20.

42. Bingham P, Verlander NQ, Cheal MJ. John Snow, William Farr and the 1849 outbreak of cholera that affected London: a reworking of the data highlights the importance of the water supply. *Public Health* 2004; **118**: 387-94.

43. Registrar-General. Report on the mortality of cholera in England 1848-49: Her Majesty's Stationery Office, 1852.

44. Snow J. On the Mode of Communication of Cholera. London: John Churchill; 1855.

45. World Health Organization. Foodborne Disease Outbreaks: Guidelines for Investigation and Control. Geneva, Switzerland, 2008.

46. Public Health England. Communicable Disease Outbreak Management: Operational guidance. London, UK: Public Health England, 2014.

47. European Centre for Disease Prevention and Control. Toolbox for investigation and response to Food and Waterborne Disease Outbreaks with an EU dimension. http://www.ecdc.europa.eu/en/healthtopics/food_and_waterborne_disease/toolkit/Pages/index.aspx (accessed 30 September 2014).

48. Centers for Disease Control and Prevention. Multistate and Nationwide Foodborne Outbreak Investigations: A Step-by-Step Guide. 2013. http://www.cdc.gov/foodsafety/outbreaks/investigating-outbreaks/investigations/index.html (accessed 30 September 2014).

49. Pfeiffer DU, Robinson T, Stevenson M, Stevens KB, Rogers D, Clements AC. Spatial analysis in epidemiology. Oxford, UK: Oxford University Press; 2008.

50. Yamada I. Thiessen Polygons. International Encyclopedia of Geography: People, the Earth, Environment and Technology: John Wiley & Sons, Ltd; 2016.

51. Bull M, Hall IM, Leach S, Robesyn E. The application of geographic information systems and spatial data during Legionnaires disease outbreak

responses. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* 2012; **17**(49).

52. Tanser F, Gijsbertsen B, Herbst K. Modelling and understanding primary health care accessibility and utilization in rural South Africa: An exploration using a geographical information system. *Social Science & Medicine* 2006; **63**(3): 691-705.

53. Cuzick J, Edwards R. Spatial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society Series* 1990; **52**: 73-104.

54. Reader S. Using survival analysis to study spatial point patterns in geographical epidemiology. *Social Science & Medicine* 2000; **50**(7): 985-1000.

55. Ripley BD. Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society* 1977; **Series B**(39): 172-212.

56. Moran PAP. Notes on Continuous Stochastic Phenomena. *Biometrika* 1950; **37**(1): 17-23.

57. Kulldorff M. A spatial scan statistic. *Communications in statistics: theory and methods* 1997; **26**: 1481-96.

58. Kulldorff M, Heffernan R, Hartman J, Assunção RM, Mostashari F. A space-time permutation scan statistic for the early detection of disease outbreaks. *PLoS Medicine* 2005; **2**: 216-24.

59. Jacquez GM. Disease cluster statistics for space-time interaction *Statistics in Medicine* 1994; **15**: 873-85.

60. Knox EG. The detection of space-time interactions. *Journal of Applied Statistics* 1964; **13**: 24-30.

61. Kulldorff M, Hjalmars U. The Knox method and other tests for space-time interaction. *Biometrics* 1999; **55**(2): 544-52.

62. Grimson RC, Aldrich TE, Drane JW. Clustering in sparse data and an analysis of rhabdomyosarcoma incidence. *Statistics in Medicine* 1992; **11**(6): 761-8.

63. Pigott DM, Golding N, Mylne A, et al. Mapping the zoonotic niche of Ebola virus disease in Africa. *eLife* 2014; **3**: e04395.

64. Messina JP, Kraemer MUG, Brady OJ, et al. Mapping global environmental suitability for Zika virus. *eLife* 2016; **5**: e15272.

65. Carter R, Mendis KN, Roberts D. Spatial targeting of interventions against malaria. *Bulletin of the World Health Organization* 2000; **78**(12): 1401-11.

66. European Centre for Disease Prevention and Control. Outbreak Investigations. 06/08/2014 2014.

https://wiki.ecdc.europa.eu/fem/w/wiki/outbreak-investigations.aspx (accessed 15 October 2014).

67. Abellan JJ, Martinez-Beneito MA, Zurriaga O, Jorques G, Ferrandiz J, Lopez-Quilez A. [Point processes as a tool for analyzing possible sources of contamination]. *Gac Sanit* 2002; **16**(5): 445-9.

68. Acheson P, McGivern M, Frank P, et al. An ongoing outbreak of heterosexually-acquired syphilis across Teesside, UK. *Int J STD AIDS* 2011; **22**(9): 514-6.

69. Affolabi D, Faihun F, Sanoussi N, et al. Possible outbreak of streptomycin-resistant Mycobacterium tuberculosis Beijing in Benin. *Emerging Infectious Diseases* 2009; **15**(7): 1123-5.

70. Ali M, Wagatsuma Y, Emch M, Breiman RF. Use of a geographic information system for defining spatial risk for dengue transmission in Bangladesh: Role for Aedes albopictus in an urban outbreak. *American Journal of Tropical Medicine and Hygiene* 2003; **69**(6): 634-40.

71. Angulo JJ, Pederneiras CA, Sakuma ME, Takiguti CK, Megale P. Contour mapping of the temporal-spatial progression of a contagious disease. *Bull Soc Pathol Exot Filiales* 1979; **72**(4): 374-85.

72. Bali S, Kar SS, Kumar S, Ratho RK, Dhiman RK, Kumar R. Hepatitis E epidemic with bimodal peak in a town of north India. *Indian J Public Health* 2008; **52**(4): 189-93, 99.

73. Barcellos C, Sabroza PC. Socio-environmental determinants of the leptospirosis outbreak of 1996 in western Rio de Janeiro: a geographical approach. *Int J Environ Health Res* 2000; **10**(4): 301-13.

74. Barreto ML. The dot map as an epidemiological tool: a case study of Schistosoma mansoni infection in an urban setting. *International journal of epidemiology* 1993; **22**(4): 731-41.

75. Bartels SA, Greenough PG, Tamar M, VanRooyen MJ. Investigation of a cholera outbreak in Ethiopia's Oromiya Region. *Disaster med* 2010; **4**(4): 312-7.

76. Bessong PO, Odiyo JO, Musekene JN, Tessema A. Spatial distribution of diarrhoea and microbial quality of domestic water during an outbreak of diarrhoea in the Tshikuwi community in Venda, South Africa. *J Health Popul Nutr* 2009; **27**(5): 652-9.

77. Blondin N, Baumgardner DJ, Moore GE, Glickman LT. Blastomycosis in indoor cats: suburban Chicago, Illinois, USA. *Mycopathologia* 2007; **163**(2): 59-66.

78. Boccia D, Oliver CI, Charlett A, et al. Outbreak of a new Salmonella phage type in South West England: alternative epidemiological investigations are needed. *Commun Dis Public Health* 2004; **7**(4): 339-43.

79. Bowie WR, King AS, Werker DH, et al. Outbreak of toxoplasmosis associated with municipal drinking water. *The Lancet* 1997; **350**(9072): 173-7.

80. Brown CM, Nuorti PJ, Breiman RF, et al. A community outbreak of Legionnaires' disease linked to hospital cooling towers: an epidemiological method to calculate dose of exposure. *International journal of epidemiology* 1999; **28**(2): 353-9.

81. Carr R, Warren R, Towers L, et al. Investigating a cluster of Legionnaires' cases: public health implications. *Public Health* 2010; **124**(6): 326-31.

82. Chadee DD, Lee R, Ferdinand A, Prabhakar P, Clarke D, Jacob B. Meningococcal meningitis outbreak in Trinidad, 1998. *European Journal of General Medicine* 2006; **3**(2): 49-53.

83. Chadee DD, Williams FLR, Kitron UD. Impact of vector control on a dengue fever outbreak in Trinidad, West Indies, in 1998. *Trop Med Int Health* 2005; **10**(8): 748-54.

84. Chung WM, Buseman CM, Joyner SN, et al. The 2012 West Nile encephalitis epidemic in Dallas, Texas. *Jama* 2013; **310**(3): 297-307.

85. de Moura L, Bahia-Oliveira LMG, Wada MY, et al. Waterborne toxoplasmosis, Brazil, from field to gene. *Emerging Infectious Diseases* 2006; **12**(2): 326-9.

86. Epp T, Argue C, Waldner C, Berke O. Spatial analysis of an anthrax outbreak in Saskatchewan, 2006. *Canadian Veterinary Journal-Revue Veterinaire Canadienne* 2010; **51**(7): 743-8.

87. Fang L-Q, Li X-L, Liu K, et al. Mapping spread and risk of avian influenza A (H7N9) in China. *Sci* 2013; **3**: 2722.

88. Fernandez MAL, Mason PR, Gray H, Bauernfeind A, Fesselet JF, Maes P. Descriptive spatial analysis of the cholera epidemic 2008-2009 in Harare, Zimbabwe: a secondary data analysis. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 2011; **105**(1): 38-45.

89. Fevre EM, Coleman PG, Odiit M, Magona JW, Welburn SC, Woolhouse MEJ. The origins of a new Trypanosoma brucei rhodesiense sleeping sickness outbreak in eastern Uganda. *The Lancet* 2001; **358**(9282): 625-8.

90. Firestone SM, Ward MP, Christley RM, Dhand NK. The importance of location in contact networks: Describing early epidemic spread using spatial social network analysis. *Prev Vet Med* 2011; **102**(3): 185-95.

91. Fitzpatrick G, Ward M, Ennis O, et al. Use of a geographic information system to map cases of measles in real-time during an outbreak in Dublin, Ireland, 2011. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* 2012; **17**(49): 19-29.

92. Garcia-Fulgueiras A, Navarro C, Fenoll D, et al. Legionnaires' disease outbreak in Murcia, Spain. *Emerging Infectious Diseases* 2003; **9**(8): 915-21.

93. Gubbels S-M, Kuhn KG, Larsson JT, et al. A waterborne outbreak with a single clone of Campylobacter jejuni in the Danish town of Koge in May 2010. *Scand J Infect Dis* 2012; **44**(8): 586-94.

94. Hackert VH, van der Hoek W, Dukers-Muijrers N, et al. Q fever: single-point source outbreak with high attack rates and massive numbers of undetected infections across an entire region. *Clin Infect Dis* 2012; **55**(12): 1591-9.

95. Hyland JM, Hamlet N, Saunders C, Coppola J, Watt J. Outbreak of Legionnaires' disease in West Fife: review of environmental guidelines needed. *Public Health* 2008; **122**(1): 79-83.

96. Jansa JM, Cayla JA, Ferrer D, et al. An outbreak of Legionnaires' disease in an inner city district: importance of the first 24 hours in the investigation. *International Journal of Tuberculosis and Lung Disease* 2002; **6**(9): 831-8.

97. Keramarou M, Evans MR. A community outbreak of Legionnaires' disease in South Wales, August-September 2010. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* 2010; **15**(42).

98. Kirrage D, Reynolds G, Smith GE, Olowokure B, Hereford Legionnaires Outbreak Control T. Investigation of an outbreak of Legionnaires' disease: Hereford, UK 2003. *Respiratory Medicine* 2007; **101**(8): 1639-44.

99. Kistemann T, Dangendorf F, Krizek L, Sahl HG, Engelhart S, Exner M. GIS-supported investigation of a nosocomial Salmonella outbreak. *International Journal of Hygiene and Environmental Health* 2000; **203**(2): 117-26.

100. Lai PC, Wong CM, Hedley AJ, et al. Understanding the spatial clustering of severe acute respiratory syndrome (SARS) in Hong kong. *Environmental Health Perspectives* 2004; **112**(15): 1550-6.

101. Lai P-c, Kwong K-h. Spatial Analysis of the 2008 Influenza Outbreak of Hong Kong. Computational Science and Its Applications - Iccsa 2010, Pt 1, Proceedings; 2010: 374-88.

102. Le Comber SC, Rossmo DK, Hassan AN, Fuller DO, Beier JC. Geographic profiling as a novel spatial tool for targeting infectious disease control. *International Journal of Health Geographics* 2011; **10**.

103. Le H, Poljak Z, Deardon R, Dewey CE. Clustering of and Risk Factors for the Porcine High Fever Disease in a Region of Vietnam. *Transboundary and Emerging Diseases* 2012; **59**(1): 49-61.

104. Liang W, McLaws ML, Liu M, Mi J, Chan DKY. Hindsight: a re-analysis of the severe acute respiratory syndrome outbreak in Beijing. *Public Health* 2007; **121**(10): 725-33.

105. Luquero FJ, Banga CN, Remartinez D, Palma PP, Baron E, Grais RF. Cholera Epidemic in Guinea-Bissau (2008): The Importance of "Place". *PLoS ONE* 2011; **6**(5).

106. Manfredi Selvaggi T, Rezza G, Scagnelli M, et al. Investigation of a Q-fever outbreak in northern Italy. *Eur J Epidemiol* 1996; **12**(4): 403-8.

107. McKee KT, Shields TM, Jenkins PR, Zenilman JM, Glass GE. Application of a geographic information system to the tracking anti control of an outbreak of shigellosis. *Clin Infect Dis* 2000; **31**(3): 728-33.

108. Miranda ME, Yoshikawa Y, Manalo DL, et al. Chronological and spatial analysis of the 1996 Ebola Reston virus outbreak in a monkey breeding facility in the Philippines. *Experimental animals / Japanese Association for Laboratory Animal Science* 2002; **51**(2): 173-9.

109. Mongoh MN, Dyer NW, Stoltenow CL, Khaitsa ML. Risk factors associated with anthrax outbreak in animals in North Dakota, 2005: A retrospective case-control study. *Public Health Reports* 2008; **123**(3): 352-9.

110. Morrison AC, Getis A, Santiago M, Rigau-Perez JG, Reiter P. Exploratory space-time analysis of reported dengue cases during an outbreak in Florida, Puerto Rico, 1991-1992. *American Journal of Tropical Medicine and Hygiene* 1998; **58**(3): 287-98.

111. Neira-Munoz E, Okoro C, McCarthy ND. Outbreak of waterborne cryptosporidiosis associated with low oocyst concentrations. *Epidemiology and Infection* 2007; **135**(7): 1159-64.

112. Nguyen TM, Ilef D, Jarraud S, et al. A community-wide outbreak of legionnaires disease linked to industrial cooling towers--how far can contaminated aerosols spread? *The Journal of infectious diseases* 2006; **193**(1): 102-11.

113. Nisha V, Gad SS, Selvapandian D, et al. Geographical information system (GIS) in investigation of an outbreak. *J Commun Dis* 2005; **37**(1): 39-43.

114. Nishiguchi A, Kobayashi S, Ouchi Y, Yamamoto T, Hayama Y, Tsutsui T. Spatial analysis of low pathogenic H5N2 avian influenza outbreaks in Japan in 2005. *J Vet Med Sci* 2009; **71**(7): 979-82.

115. Norstrom M, Pfeiffer DU, Jarp J. A space-time cluster investigation of an outbreak of acute respiratory disease in Norwegian cattle herds. *Prev Vet Med* 1999; **47**(1-2): 107-19.

116. Nygard K, Schimmer B, Sobstad O, et al. A large community outbreak of waterborne giardiasis-delayed detection in a non-endemic urban area. *BMC Public Health* 2006; **6**: 141.

117. Nygard K, Werner-Johansen O, Ronsen S, et al. An outbreak of legionnaires disease caused by long-distance spread from an industrial air scrubber in Sarpsborg, Norway. *Clin Infect Dis* 2008; **46**(1): 61-9.

118. Orsi A, Alicino C, Patria AG, et al. Epidemiological and molecular approaches for management of a measles outbreak in Liguria, Italy. *Journal of preventive medicine and hygiene* 2010; **51**(2): 67-72.

119. Parkinson R, Rajic A, Jenson C. Investigation of an anthrax outbreak in Alberta in 1999 using a geographic information system. *Can Vet J* 2003; **44**(4): 315-8.

120. Pasma T. Spatial epidemiology of an H3N2 swine influenza outbreak. *Can Vet J* 2008; **49**(2): 167-76.

121. Passos AD, Castro e Silva AA, Ferreira AH, Maria e Silva J, Monteiro ME, Santiago RC. [Rabies epizootic in the urban area of Ribeirao Preto, Sao Paulo, Brazil]. *Cadernos de Saude Publica* 1998; **14**(4): 735-40.

122. Pfister JR, Archer JR, Hersil S, et al. Non-rural point source blastomycosis outbreak near a yard waste collection site. *Clinical medicine & research* 2011; **9**(2): 57-65.

123. Rivas AL, Chowell G, Schwager SJ, et al. Lessons from Nigeria: The role of roads in the geo-temporal progression of avian influenza (H5N1) virus. *Epidemiology and Infection* 2010; **138**(2): 192-8.

124. Rivas AL, Smith SD, Sullivan PJ, et al. Identification of geographic factors associated with early spread of foot-and-mouth disease. *American Journal of Veterinary Research* 2003; **64**(12): 1519-27.

125. Roquet D, Diallo A, Kodio B, Daff BM, Fenech C, Etard JF. [Cholera epidemic in Senegal in 1995-1996: an example of geographic approach to health]. *Sante* 1998; **8**(6): 421-8.

126. Rotela C, Fouque F, Lamfri M, et al. Space-time analysis of the dengue spreading dynamics in the 2004 Tartagal outbreak, Northern Argentina. *Acta Trop* 2007; **103**(1): 1-13.

127. Roy M, Benedict K, Deak E, et al. A large community outbreak of blastomycosis in wisconsin with geographic and ethnic clustering. *Clin Infect Dis* 2013; **57**(5): 655-62.

128. Saha T, Murhekar M, Hutin YJ, Ramamurthy T. An urban, water-borne outbreak of diarrhoea and shigellosis in a district town in eastern India. *Natl Med J India* 2009; **22**(5): 237-9.

129. Sanson RL, Gloster J, Burgin L. Reanalysis of the start of the UK 1967 to 1968 foot-and-mouth disease epidemic to calculate airborne transmission probabilities. *Veterinary Record* 2011; **169**(13): 336-U44.

130. Sarkar R, Prabhakar AT, Manickam S, et al. Epidemiological investigation of an outbreak of acute diarrhoeal disease using geographic information systems. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 2007; **101**(6): 587-93.

131. Sasaki S, Suzuki H, Igarashi K, Tambatamba B, Mulenga P. Spatial analysis of risk factor of cholera outbreak for 2003-2004 in a peri-urban area of Lusaka, Zambia. *American Journal of Tropical Medicine and Hygiene* 2008; **79**(3): 414-21.

132. Schimmer B, Veenstra T, Ter Schegget R, et al. The use of a geographic information system to identify a goat dairy farm as the most likely source of an Urban Q fever outbreak. *Clinical Microbiology and Infection* 2010; **16**: S391-S2.

133. Siddiqui FJ, Bhutto NS, von Seidlein L, et al. Consecutive outbreaks of Vibrio cholerae O139 and V-cholerae O1 cholera in a fishing village near Karachi, Pakistan. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 2006; **100**(5): 476-82.

134. Sowmyanarayanan TV, Mukhopadhya A, Gladstone BP, Sarkar R, Kang G. Investigation of a hepatitis A outbreak in children in an urban slum in Vellore, Tamil Nadu, using geographic information systems. *Indian Journal of Medical Research* 2008; **128**(1): 32-7.

135. Sze-To GN, Chao CYH. Use of Risk Assessment and Likelihood Estimation to Analyze Spatial Distribution Pattern of Respiratory Infection Cases. *Risk Analysis* 2011; **31**(3): 351-69.

136. Tenzin, Sharma B, Dhand NK, Timsina N, Ward MP. Reemergence of rabies in Chhukha district, Bhutan, 2008. *Emerging Infectious Diseases* 2010; **16**(12): 1925-30.

137. Turcios-Ruiz RM, Axelrod P, St. John K, et al. Outbreak of Necrotizing Enterocolitis Caused by Norovirus in a Neonatal Intensive Care Unit. *Journal of Pediatrics* 2008; **153**(3): 339-44.

138. Ulugtekin N, Alkoy S, Seker DZ. Use of a geographic information system in an epidemiological study of measles in Istanbul. *Journal of International Medical Research* 2007; **35**(1): 150-4.

139. van der Hoek W, van de Kassteele J, Bom B, et al. Smooth incidence maps give valuable insight into Q fever outbreaks in The Netherlands. *Geospat Health* 2012; **7**(1): 127-34.

140. Varani S, Cagarelli R, Melchionda F, et al. Ongoing outbreak of visceral leishmaniasis in Bologna Province, Italy, November 2012 to May 2013. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* 2013; **18**(29): 20530.

141. Waldron LS, Ferrari BC, Cheung-Kwok-Sang C, Beggs PJ, Stephens N, Power ML. Molecular Epidemiology and Spatial Distribution of a Waterborne Cryptosporidiosis Outbreak in Australia. *Applied and Environmental Microbiology* 2011; **77**(21): 7766-71.

142. Wallensten A, Moore P, Webster H, et al. Q fever outbreak in Cheltenham, United Kingdom, in 2007 and the use of dispersion modelling to investigate the possibility of airborne spread. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* 2010; **15**(12).

143. White PS, Graham FF, Harte DJG, Baker MG, Ambrose CD, Humphrey ARG. Epidemiological investigation of a Legionnaires' disease outbreak in Christchurch, New Zealand: the value of spatial methods for practical public health. *Epidemiology and Infection* 2013; **141**(4): 789-99.

144. Wong BCK, Lee N, Li Y, et al. Possible role of aerosol transmission in a hospital outbreak of influenza. *Clin Infect Dis* 2010; **51**(10): 1176-83.

145. Yu ITS, Wong TW, Chiu YL, Lee N, Li YG. Temporal-spatial analysis of severe acute respiratory syndrome among hospital inpatients. *Clin Infect Dis* 2005; **40**(9): 1237-43.

146. Davis GS, Sevdalis N, Drumright LN. Spatial and temporal analyses to investigate infectious disease transmission within healthcare settings. *Journal of Hospital Infection* 2014; **86**: 227.

147. European Food Safety Authority, European Centre for Disease Prevention and Control. The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2013. *European Food Safety Authority Journal* 2015; **13**(1).

148. Centers for Disease Control and Prevention. Surveillance for Foodborne Disease Outbreaks, United States, 2012, Annual Report. Atlanta, Georgia: US Department of Health and Human Services, CDC, 2014.

149. Tanser FC, Le Sueur D. The application of geographical information systems to imporant public health problems in Africa. *International Journal of Health Geographics* 2002; **1**(4).

150. Bhaduri B. Encyclopedia of GIS. New York: Springer; 2008.

151. Martin D. Directions in Population GIS. *Geography Compass* 2011; **5**(9): 655-65.

152. Tango T. A test for spatial disease clustering adjusted for multiple testing. *Statistics in Medicine* 2000; **19**(2): 191-204.

153. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for reporting observational studies. *Preventive Medicine* 2007; **45**(4): 247-51.

154. Moher D, Liberati A, Tetzlaff J, Altman DG, The PG. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med* 2009; **6**(7): e1000097.

155. Kulldorff M, Information Management Services Inc. SaTScan™ v9.4.2: Software for the spatial and space-time scan statistics. 2015.

156. Sickweather. 2014. http://www.sickweather.com/ (accessed 15 October 2014).

157. Carroll LN, Au AP, Detwiler LT, Fu TC, Painter IS, Abernethy NF. Visualization and analytics tools for infectious disease epidemiology: a systematic review. *Journal of biomedical informatics* 2014; **51**: 287-98.

158. Chester KG. BioSense 2.0. *Online Journal of Public Health Informatics* 2013; **5**(1): e200.

159. Chang W, Cheng J, J.J. A, J. X, J. M. shiny: Web Application Framework for R. 0.12.2 ed; 2015.

160. ESRI. ArcGIS Desktop. 10.2 ed. Redlands, CA: Environmental Systems Research Institute; 2013.

161. QGIS Development Team. QGIS Geographic Information System. 2.10.1 ed: Open Source Geospatial Foundation Project; 2015.

162. European Centre for Disease Prevention and Control. ECEC Map Maker (EMMa). 2015.

163. Public Health England. Tuberculosis in the UK 2014 report. London, 2014.

164. Public Health England. TB Strain Typing and Cluster Investigation Handbook. London, 2014.

165. Mears J, Abubakar I, Crisp D, et al. 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* 2014; **14**(1): 1023.

166. Mears J, Vynnycky E, Lord J, et al. Evaluation of the Tuberculosis Strain Typing Service (TB-STS) in England. *The Lancet* 2013; **382(Supplement 3)**(S73).

167. World Health Organization. Electronic recording and reporting for tuberculosis care and control. Geneva, 2012.

168. Centers for Disease Control and Prevention DoTE. Reported Tuberculosis in the United States. Atlanta, GA: U.S. Department of Health and Human Serices, CDC, 2014.

169. van Walle I. ECDC starts pilot phase for collection of molecular typing data. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* 2013; **18**(3).

170. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; 2013.

171. Cheng J, Xie Y. leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library. 1.0.0.9999 ed; 2015.

172. © OpenStreetMap contributors. OpenStreetMap®. 2015.

173. Neuwirth E. RColorBrewer: ColorBrewer palettes. 1.0-5 ed; 2011.

174. Kahle D, Wickham H. ggmap: A package for spatial visualization with Google Maps and OpenStreetMap. 2.3 ed; 2013.

175. Wickham H. ggplot2: elegant graphics for data analysis. Springer New York; 2009.

176. Aragón T. epitools: Epidemiology Tools. 0.5-7 ed; 2012.

177. RStudio. SuperZip example. 2015.

178. London Assembly Health Committee. Tackling TB in London. London, UK, 2015.

179. Maguire H, Brailsford S, Carless J, et al. Large outbreak of isoniazid-monoresistant tuberculosis in London, 1995 to 2006: case-control study and recommendations. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* 2011; **16**(13): pii: 19830.

180. Ruddy M, Davies A, Yates M, et al. Outbreak of isoniazid resistant tuberculosis in north London. *Thorax* 2004; **59**: 279 - 85.

181. Ghebremichael S, Petersson R, Koivula T, et al. Molecular epidemiology of drug-resistant tuberculosis in Sweden. *Microbes and Infection* 2008; **10**(6): 699-705.

182. Edlin BR, Tokars JI, Grieco MH, et al. An outbreak of multidrug-resistant tuberculosis among hospitalized patients with the acquired immunodeficiency syndrome. *New England Journal of Medicine* 1992; **326**(23): 1514-21.

183. Frieden TR, Sherman LF, Maw KL, et al. A multi-institutional outbreak of highly drug-resistant tuberculosis: Epidemiology and clinical outcomes. *Journal of the American Medical Association* 1996; **276**(15): 1229-35.

184. Miller AC, Butler WR, McInnis B, et al. Clonal relationships in a shelter-associated outbreak of drug-resistant tuberculosis: 1983-1997. *International Journal of Tuberculosis and Lung Disease* 2002; **6**(10): 872-8.

185. Mears J, Vynnycky E, Lord J, et al. The prospective evaluation of the TB strain typing service in England: a mixed methods study. *Thorax* 2015.

186. Hamblion EL, Le Menach A, Anderson LF, et al. Recent TB transmission, clustering and predictors of large clusters in London, 2010–2012: results from first 3 years of universal MIRU-VNTR strain typing. *Thorax* 2016; **71**(8): 749-56.

187. Department of Communities and Local Government. The English Indices of Deprivation 2015. London, 2015.

188. Kwak C, Clayton-Matthews A. Multinomial Logistic Regression. *Nursing Research* 2002; **51**(6): 404-10.

189. Kleinman K. rsatscan: Tools, Classes, and Methods for Interfacing with SaTScan Stand-Alone Software. 0.3.9200 ed; 2015.

190. Zelner JL, Murray MB, Becerra MC, et al. Identifying Hotspots of Multidrug-Resistant Tuberculosis Transmission Using Spatial and Molecular Genetic Data. *The Journal of infectious diseases* 2016; **213**(2): 287-94.

191. Saavedra-Campos M, Welfare W, Cleary P, et al. Identifying areas and risk groups with localised Mycobacterium tuberculosis transmission in northern England from 2010 to 2012: spatiotemporal analysis incorporating highly discriminatory genotyping data. *Thorax* 2015.

192. Izumi K, Ohkado A, Uchimura K, et al. Detection of Tuberculosis Infection Hotspots Using Activity Spaces Based Spatial Approach in an Urban Tokyo, from 2003 to 2011. *PLoS ONE* 2015; **10**(9): e0138831.

193. Althomsons SP, Kammerer JS, Shang N, Navin TR. Using Routinely Reported Tuberculosis Genotyping and Surveillance Data to Predict Tuberculosis Outbreaks. *PLoS ONE* 2012; **7**(11): e48754.

194. Biadglegne F, Merker M, Sack U, Rodloff AC, Niemann S. Tuberculous lymphadenitis in Ethiopia predominantly caused by strains belonging to the Delhi/CAS lineage and newly identified Ethiopian clades of the mycobacterium tuberculosis complex. 2015; **10**((Biadglegne) College of Medicine and Health Sciences, Bahir Dar University, Bahir Dar, Ethiopia).

195. Ojo OO, Sheehan S, Corcoran DG, et al. Molecular epidemiology of Mycobacterium tuberculosis clinical isolates in Southwest Ireland. *Infection, Genetics and Evolution* 2010; **10**(7): 1110-6.

196. Lim LK-Y, Sng LH, Win W, et al. Molecular Epidemiology of Mycobacterium tuberculosis Complex in Singapore, 2006-2012. *PLoS ONE* 2013; **8**(12): e84487.

197. Tuite AR, Guthrie JL, Alexander DC, et al. Epidemiological evaluation of spatiotemporal and genotypic clustering of Mycobacterium tuberculosis in Ontario, Canada. *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease* 2013; **17**(10): 1322-7.

198. Chen KS, Liu T, Lin RR, Peng YP, Xiong GC. Tuberculosis transmission and risk factors in a Chinese antimony mining community. 2016; **20**((Chen) Key Laboratory of Medical Molecular Virology, Fudan University, Shanghai, China): 57-62.

199. Goldblatt D, Rorman E, Chemtob D, et al. Molecular epidemiology and mapping of tuberculosis in Israel: Do migrants transmit the disease to locals? 2014; **18**((Goldblatt, Freidlin, Cedar, Kaidar-Shwartz) National Mycobacterium Reference Laboratory, National Public Health Laboratory, Ministry of Health, 69 Ben Tzvi Blvd, Tel-Aviv, 61082, Israel): 1085-91.

200. Toit K, Altraja A, Acosta CD, et al. A four-year nationwide molecular epidemiological study in Estonia: Risk factors for tuberculosis transmission. 2014; **4**((Toit, Kummik) United Laboratories Tartu, University Hospital, Tartu, Estonia): S34-S40.

201. Yang C, Shen X, Peng Y, et al. Transmission of Mycobacterium tuberculosis in China: A Population-Based Molecular Epidemiologic Study. 2015; **61**((Yang, Luo, Sun, Li, Qiao, Gao) Key Laboratory of Medical Molecular Virology, Fudan University, Ministries of Education and Health, 138 Yi Xue Yuan Road, Shanghai 200032, China): 219-27.

202. Mears J, Abubakar I, Cohen T, McHugh TD, Sonnenberg P. Effect of study design and setting on tuberculosis clustering estimates using Mycobacterial

Interspersed Repetitive Units-Variable Number Tandem Repeats (MIRU-VNTR): a systematic review. *BMJ Open* 2015; **5**(1): e005636.

203. Kik SV, Verver S, van Soolingen D, et al. Tuberculosis outbreaks predicted by characteristics of first patients in a DNA fingerprint cluster. *American journal of respiratory and critical care medicine* 2008; **178**(1): 96-104.

204. Jit M, Stagg HR, Aldridge RW, White PJ, Abubakar I. Dedicated outreach service for hard to reach patients with tuberculosis in London: observational study and economic evaluation. *BMJ* 2011; **343**: d5376.

205. Field N, Cohen T, Struelens MJ, et al. Strengthening the Reporting of Molecular Epidemiology for Infectious Diseases (STROME-ID): an extension of the STROBE statement. *The Lancet Infectious Diseases* 2014; **14**(4): 341-52.

206. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *The Lancet* 2007; **370**(9596): 1453-7.

207. Vynnycky E, Fine PEM. Lifetime Risks, Incubation Period, and Serial Interval of Tuberculosis. *American journal of epidemiology* 2000; **152**(3): 247-63.

208. Public Health England. HIV in the United Kingdom 2014 report: data to end 2013. London: Public Health England 2014.

209. Walker TM, Ip CLC, Harrell RH, et al. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *The Lancet Infectious Diseases* 2013; **13**(2): 137-46.

210. Walker TM, Lalor MK, Broda A, et al. Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: an observational study. *The Lancet Respiratory Medicine* 2014; **2**(4): 285-92.

211. Van Rie A, Warren R, Richardson M, et al. Classification of drug-resistant tuberculosis in an epidemic area. *The Lancet* 2000; **356**(9223): 22-5.

212. Davies A, Ruddy MC, Neely F, Ruggles R, Maguire H. Outbreak of isoniazid resistant Mycobacterium tuberculosis in north London 1999-2004. Executive summary report key points and recommendations. London: Health Protection Agency.

213. Neely F, Maguire H, Le Brun F, Davies A, Gelb D, Yates S. High rate of transmission among contacts in large London outbreak of isoniazid mono-resistant tuberculosis. *Journal of Public Health* 2010; **32**(1): 44-51.

214. Maguire H, Ruddy M, Bothamley G, et al. Multidrug resistance emerging in North London outbreak. *Thorax* 2006; **61**(6): 547-8.

215. Jit M, Stagg HR, Aldridge RW, White PJ, Abubakar I. Dedicated outreach service for hard to reach patients with tuberculosis in London: observational study and economic evaluation. *BMJ* 2011; **343**.

216. ffmpeg.

217. Diggle PJ. Statistical analysis of spatial point patterns. 2nd ed. London: Arnold; 2003.

218. Baddeley A, Turner R. spatstat: An R Package for Analyzing Spatial Point Patterns. *Journal of Statistical Software* 2005; **12**(6): 1-42.

219. Maguire H, Brailsford S, Carless J, et al. Large outbreak of isoniazid-monoresistant tuberculosis in London, 1995 to 2006:Case-control study and recommendations. *Eurosurveillance* 2011; **16**(13).

220. Moss AR, Alland D, Telzak E, et al. A city-wide outbreak of a multiple-drug-resistant strain of Mycobacterium tuberculosis in New York. *International Journal of Tuberculosis and Lung Disease* 1997; **1**(2): 115-21.

221. Edlin BR, Tokars JI, Grieco MH, et al. An outbreak of multidrug-resistant tuberculosis among hospitalized patients with the acquired immunodeficiency syndrome. *New England Journal of Medicine* 1992; **326**(23): 1514-21.

222. Centers for Disease Control. Outbreak of hospital acquired multidrug resistant tuberculosis. *Commun Dis Rep CDR Wkly* 1995; **5**(34): 161.

223. Frieden TR, Sherman LF, Maw KL, et al. A multi-institutional outbreak of highly drug-resistant tuberculosis: epidemiology and clinical outcomes. *Jama* 1996; **276**(15): 1229-35.

224. Munsiff SS, Nivin B, Sacajiu G, Mathema B, Bifani P, Kreiswirth BN. Persistence of a highly resistant strain of tuberculosis in New York City during 1990-1999. *The Journal of infectious diseases* 2003; **188**(3): 356-63.

225. Fischl MA, Uttamchandani RB, Daikos GL, et al. An outbreak of tuberculosis caused by multiple-drug-resistant tubercle bacilli among patients with HIV infection. *Annals of Internal Medicine* 1992; **117**(3): 177-83.

226. Valway SE, Greifinger RB, Papania M, et al. Multidrug-resistant tuberculosis in the New York State prison system, 1990-1991. *The Journal of infectious diseases* 1994; **170**(1): 151-6.

227. Ritacco V, Di Lonardo M, Reniero A, et al. Nosocomial spread of human immunodeficiency virus-related multidrug-resistant tuberculosis in Buenos Aires. *The Journal of infectious diseases* 1997; **176**(3): 637-42.

228. Ritacco V, Lopez B, Ambroggi M, et al. HIV infection and geographically bound transmission of drug-resistant Tuberculosis, Argentina. *Emerging Infectious Diseases* 2012; **18**(11): 1802-10.

229. Gonzalez Montaner LJ, Alberti F, Palmero D. [Multidrug-resistant tuberculosis associated with AIDS (kinetics of nosocomial epidemics of multidrug-resistant tuberculosis associated with AIDS. Possible transformation into endemic disease]. *Bull Acad Natl Med* 1999; **183**(6): 1085-94; discussion 94-6.

230. Portugal I, Covas MJ, Brum L, et al. Outbreak of multiple drug-resistant tuberculosis in Lisbon: Detection by restriction fragment length polymorphism analysis. *International Journal of Tuberculosis and Lung Disease* 1999; **3**(3): 207-13.

231. Hannan MM, Peres H, Maltez F, et al. Investigation and control of a large outbreak of multi-drug resistant tuberculosis at a central Lisbon hospital. *Journal of Hospital Infection* 2001; **47**(2): 91-7.

232. Centers for Disease Control. Multidrug-resistant tuberculosis outbreak on an HIV ward--Madrid, Spain, 1991-1995. *MMWR Morb Mortal Wkly Rep* 1996; **45**(16): 330-3.

233. Moro ML, Gori A, Errante I, et al. An outbreak of multidrug-resistant tuberculosis involving HIV-infected patients of two hospitals in Milan, Italy. Italian Multidrug-Resistant Tuberculosis Outbreak Study Group. *Aids* 1998; **12**(9): 1095-102.

234. Moro ML, Errante I, Infuso A, et al. Effectiveness of infection control measures in controlling a nosocomial outbreak of multidrug-resistant tuberculosis among HIV patients in Italy. *International Journal of Tuberculosis and Lung Disease* 2000; **4**(1): 61-8.

235. Rivero A, Marquez M, Santos J, et al. High rate of tuberculosis reinfection during a nosocomial outbreak of multidrug-resistant tuberculosis caused by Mycobacterium bovis strain B. *Clin Infect Dis* 2001; **32**(1): 159-61.

236. Moodley P, Shah NS, Tayob N, et al. Spread of extensively drug-resistant tuberculosis in Kwazulu-Natal Province, South Africa. *PLoS ONE* 2011; **6**(5).

237. Gandhi NR, Moll A, Sturm AW, et al. Extensively drug-resistant tuberculosis as a cause of death in patients co-infected with tuberculosis and HIV in a rural area of South Africa. *The Lancet* 2006; **368**(9547): 1575-80.

238. Ghebremichael S, Koivula T, Hoffner S, et al. Resistant tuberculosis is spreading in Sweden. Molecular epidemiological strain identification by "fingerprinting" can make the infection tracing easier. [Swedish]. *Lakartidningen* 2002; **99**(23): 2618-9, 22-23.

239. Dahle UR, Sandven P, Heldal E, Mannsaaker T, Caugant DA. Deciphering an outbreak of drug-resistant Mycobacterium tuberculosis. *Journal of Clinical Microbiology* 2003; **41**(1): 67-72.

240. Reves R, Blakey D, Snider DE, Jr., Farer LS. Transmission of multiple drug-resistant tuberculosis: report of a school and community outbreak. *American journal of epidemiology* 1981; **113**(4): 423-35.

241. Namouchi A, Haltiti R, Hawari D, Mardassi H. Re-emergence of the progenitors of a multidrugresistant outbreak strain of mycobacterium tuberculosis among the post-outbreak case patients. *The Journal of infectious diseases* 2010; **201**(3): 390-8.

242. Mardassi H, Namouchi A, Haltiti R, et al. Tuberculosis due to resistant Haarlem strain, Tunisia. *Emerging Infectious Diseases* 2005; **11**(6): 957-61.

243. Brostrom R, Fred D, Heetderks A, et al. Islands of hope: building local capacity to manage an outbreak of multidrug-resistant tuberculosis in the Pacific. *American journal of public health* 2011; **101**(1): 14-8.

244. Centers for Disease Control. Two simultaneous outbreaks of multidrug-resistant tuberculosis--Federated States of Micronesia, 2007-2009. *MMWR Morb Mortal Wkly Rep* 2009; **58**(10): 253-6.

245. Fred D, Desai M, Song R, et al. Multi-drug resistant tuberculosis in Chuuk State Federated States of Micronesia, 2008-2009. *Pacific health dialog* 2010; **16**(1): 123-7.

246. Oeltmann JE, Varma JK, Ortega L, et al. Multidrug-resistant tuberculosis outbreak among US-bound Hmong refugees, Thailand, 2005. *Emerging Infectious Diseases* 2008; **14**(11): 1715-21.

247. World Health Organization. Tuberculosis Fact Sheet. 2014. http://www.who.int/mediacentre/factsheets/fs104/en/ (accessed 10 February 2015).

248. National Institute for Health and Clinical Excellence. Identifying and managing tuberculosis among hard-to-reach groups. Manchester, 2012.

249. Public Health England, Health Protection Scotland, Public Health Wales, Ireland PHAN. Shooting Up: Infections among people who inject drugs in the United Kingdom 2013. London: Public Health England, 2014.

250. Kan B, Berggren I, Ghebremichael S, et al. Extensive transmission of an isoniazid-resistant strain of Mycobacterium tuberculosis in Sweden.

International Journal of Tuberculosis and Lung Disease 2008; **12**(2): 199-204.

251. Goodburn A, Drennan V. The use of directly observed therapy in TB: a brief pan-London survey. *Nursing standard (Royal College of Nursing (Great Britain) : 1987)* 2000; **14**(46): 36-8.

252. Public Health England. Tuberculosis in London: Annual review (2013 data). London, 2014.

253. Patra J, Bhatia M, Suraweera W, et al. Exposure to Second-Hand Smoke and the Risk of Tuberculosis in Children and Adults: A Systematic Review and Meta-Analysis of 18 Observational Studies. *PLoS Med* 2015; **12**(6): e1001835.

254. Narasimhan P, Wood J, MacIntyre CR, Mathai D. Risk Factors for Tuberculosis. *Pulmonary Medicine* 2013; **2013**: 11.

255. Rossmo DK. Geographic Profiling. Boca Raton, Florida, USA: CRC Press; 2000.

256. Doney R. The aftermath of the Yorkshire Ripper: the response of the United Kingdom Police Service. In *Serial Murder: An Elusive Phenomenon.* New York: Praeger; 1990.

257. Rossmo DK. Recent Developments in Geographic Profiling. *Policing* 2012; **6**(2): 144-50.

258. Le Comber SC, Nicholls B, Rossmo DK, Racey PA. Geographic profiling and animal foraging. *Journal of theoretical biology* 2006; **240**(2): 233-40.

259. Raine NE, Rossmo DK, Le Comber SC. Geographic profiling applied to testing models of bumble-bee foraging. *Journal of the Royal Society, Interface / the Royal Society* 2009; **6**(32): 307-19.

260. Faulkner SC, Stevenson MD, Verity R, et al. Using geographic profiling to locate elusive nocturnal animals: a case study with spectral tarsiers. *Journal of Zoology* 2015; **295**(4): 261-8.

261. Stevenson MD, Rossmo DK, Knell RJ, Le Comber SC. Geographic profiling as a novel spatial tool for targeting the control of invasive species. *Ecography* 2012; **35**(8): 704-15.

262. Martin RA, Rossmo DK, Hammerschlag N. Hunting patterns and geographic profiling of white shark predation. *Journal of Zoology* 2009; **279**(2): 111-8.

263. Papini A, Mosti S, Santosuosso U. Tracking the origin of the invading Caulerpa (Caulerpales, Chlorophyta) with Geographic Profiling, a

criminological technique for a killer alga. *Biological Invasions* 2013; **15**(7): 1613-21.

264. Buscema M, Grossi E, Breda M, Jefferson T. Outbreaks source: A new mathematical approach to identify their possible location. *Physica A: Statistical Mechanics and its Applications* 2009; **388**(22): 4736-62.

265. Verity R, Stevenson MD, Rossmo DK, Nichols RA, Le Comber SC. Spatial targeting of infectious disease control: identifying multiple, unknown sources. *Methods in Ecology and Evolution* 2014; **5**(7): 647-55.

266. Le Comber SC, Rossmo DK, Hassan AN, Fuller DO, Beier JC. Geographic profiling as a novel spatial tool for targeting infectious disease control. *International journal of health geographics* 2011; **10**: 35.

267. Le Comber SC, Stevenson MD. From Jack the Ripper to epidemiology and ecology. *Trends in ecology & evolution* 2012; **27**(6): 307-8.

268. Stevenson MD, Verity R. Rgeoprofile: Generates geograpic profiles from point pattern data. 1.1 ed; 2013.

269. Hardie RM, Watson JM. Mycobacterium bovis in England and Wales: past, present and future. *Epidemiology and Infection* 1992; **109**(1): 23-33.

270. Krebs JR, Anderson R, Clutton-Brock T, Morrison I, Young D, Donnelly CA. Bovine tuberculosis in cattle and badgers. London: Ministry of Agriculture, Fisheries and Food (MAFF) Publications, 1997.

271. Department for Environment Food and Rural Affairs. Monthly publication of National Statistics on the Incidence of Tuberculosis (TB) in Cattle to end December 2014 for Great Britain, 2015.

272. Animal Health and Veterinary Laboratories Agency. TB Testing Interval review England. 2014.

273. Department for Environment Food and Rural Affairs. Changes to TB Cattle Movement Controls. London, 2012.

274. Murhead RH, Burns KJ. Tuberculosis in wild badgers in Gloucestershire: epidemiology. *Veterinary Record* 1974; **95**(552-555).

275. Godfray HC, Donnelly CA, Kao RR, et al. A restatement of the natural science evidence base relevant to the control of bovine tuberculosis in Great Britain. *Proceedings Biological sciences / The Royal Society* 2013; **280**(1768): 20131634.

276. Dunnet GM, Jones DM, McInerney JP. Badgers and Bovine Tuberculosis: Review of Policy. London: HMSO, 1986.

277. Zuckerman S. Badgers, Cattle and Tuberculosis. London: HMSO, 1980.

278. Department for Environment Food and Rural Affairs. The Strategy of achieving Officially Bovine Tuberculosis Free status for England, 2014.

279. Woodroffe R, Donnelly CA, Cox DR, et al. Effects of culling on badger Meles meles spatial organization: implications for the control of bovine tuberculosis. *Journal of Applied Ecology* 2006; **43**(1): 1-10.

280. Donnelly CA, Woodroffe R, Cox DR, et al. Impact of localized badger culling on tuberculosis incidence in British cattle. *Nature* 2003; **426**(6968): 834-7.

281. Vial F, Donnelly CA. Localized reactive badger culling increases risk of bovine tuberculosis in nearby cattle herds. *Biology letters* 2012; **8**(1): 50-3.

282. Bourne FJ, Donnelly CA, Cox DR, et al. Final Report of the Independent Scientific Group on Cattle TB, 2007.

283. Independent Expert Panel. Pilot Badger Culls in Somerset and Gloucestershire. London, 2014.

284. Department for Environment Food and Rural Affairs. Bovine TB Eradication Programme for England. London: Department for Environment Food and Rural Affairs, 2011.

285. Woodroffe R, Donnelly CA, Johnston WT, et al. Spatial association of Mycobacterium bovis infection in cattle and badgers Meles meles. *Journal of Applied Ecology* 2005; **42**(5): 852-62.

286. Woodroffe R, Donnelly CA, Jenkins HE, et al. Culling and cattle controls influence tuberculosis risk for badgers. *Proc Natl Acad Sci U S A* 2006; **103**(40): 14713-7.

287. Crawshaw TR, Griffiths IB, Clifton-Hadley RS. Comparison of a standard and a detailed postmortem protocol for detecting Mycobacterium bovis in badgers. *Veterinary Record* 2008; **163**(16): 473-7.

288. Rogers LM, Cheeseman CL, Mallinson PJ, Clifton-Hadley R. The demography of a high-density badger (Meles meles) population in the west of England. *Journal of Zoology* 1997; **242**(4): 705-28.

289. Kauhala K, Holmala K. Landscape features, home-range size and density of northern badgers (Meles meles). *Annales Zoologici Fennici* 2011; **48**: 221-32.

290. Pope LC, Butlin RK, Wilson GJ, et al. Genetic evidence that culling increases badger movement: implications for the spread of bovine tuberculosis. *Molecular Ecology* 2007; **16**(23): 4919-29.

291. Carruthers JI, Lewis S, Knaap G-J, Renner RN. Coming undone: A spatial hazard analysis of urban form in American metropolitan areas*. *Papers in Regional Science* 2010; **89**(1): 65-88.

292. Bivand R, Rundel C. rgeos: Interface to Geometry Engine - Open Source (GEOS). 0.3-3 ed; 2014.

293. Therneau T. A Package for Survival Analysis in S. A Package for Survival Analysis in S ed; 2014.

294. Therneau T. coxme: Mixed Effects Cox Models.; 2012.

295. Bielby J, Donnelly CA, Pope LC, Burke T, Woodroffe R. Badger responses to small-scale culling may compromise targeted control of bovine tuberculosis. *Proceedings of the National Academy of Sciences* 2014; **111**(25): 9193-8.

296. Sutmoller P, Barteling SS, Olascoaga RC, Sumption KJ. Control and eradication of foot-and-mouth disease. *Virus research* 2003; **91**(1): 101-44.

297. Ferguson NM, Donnelly CA, Anderson RM. The foot-and-mouth epidemic in Great Britain: pattern of spread and impact of interventions. *Science (New York, NY)* 2001; **292**(5519): 1155-60.

298. Reader S. Using survival analysis to study spatial point patterns in geographical epidemiology. *Social science & medicine (1982)* 2000; **50**(7-8): 985-1000.

299. Donnelly CA, Nouvellet P. The contribution of badgers to confirmed tuberculosis in cattle in high-incidence areas in England. *PLoS Currents* 2013.

300. Green DM, Kiss IZ, Mitchell AP, Kao RR. Estimates for local and movement-based transmission of bovine tuberculosis in British cattle. *Proceedings Biological sciences / The Royal Society* 2008; **275**(1638): 1001-5.

301. Conlan AJ, McKinley TJ, Karolemeas K, et al. Estimating the hidden burden of bovine tuberculosis in Great Britain. *PLoS computational biology* 2012; **8**(10): e1002730.

302. Royal College of Nursing. Tuberculosis case management and cohort review. 2012. https://www2.rcn.org.uk/__data/assets/pdf_file/0010/439129/004204.pdf (accessed 16 September 2016).

303. British Thoracic Society. Defining a model for a gold standard for a TB MDT group and associated networks, 2014.

304. Public Health England. LTBEx - a new model for management of TB incidents in metropolitan areas. 2016. https://www.nice.org.uk/sharedlearning/ltbex-%E2%80%93-a-new-model-for-management-of-tb-incidents-in-metropolitan-areas (accessed 16 September 2016).

305. Public Health England, NHS England. Collaborartive Tuberculosis Strategy for England 2015 to 2020. London, 2015.

306. National Health Service. Five Year Forward View, 2014.

307. Hayward AC, Coker RJ. Could a tuberculosis epidemic occur in London as it did in New York? *Emerging Infectious Diseases* 2000; **6**(1): 12-6.

308. Frieden TR, Fujiwara PI, Washko RM, Hamburg MA. Tuberculosis in New York City--turning the tide. *The New England journal of medicine* 1995; **333**(4): 229-33.

309. Wang F, Luo W. Assessing spatial and nonspatial factors for healthcare access: towards an integrated approach to defining health professional shortage areas. *Health & place* 2005; **11**(2): 131-46.

310. Penchansky R Fau - Thomas JW, Thomas JW. The concept of access: definition and relationship to consumer satisfaction. *Medical care* 1981; **19**(2): 127-40.

311. Delamater PL, Messina JP, Shortridge AM, Grady SC. Measuring geographic access to health care: raster and network-based methods. *International Journal of Health Geographics* 2012; **11**(1): 1-18.

312. Guagliardo MF. Spatial accessibility of primary care: concepts, methods and challenges. *International Journal of Health Geographics* 2004; **3**(1): 1-13.

313. Transport for London. Plan a Journey. https://tfl.gov.uk/plan-a-journey/ (accessed 22 September 2016).

314. Lang DT. XML: Tools for parsing and generating XML within R and S-Plus. R package version 3.98-1.1 ed; 2013.

315. Mullen KM. Continuous Global Optimization in R. *Journal of Statistical Software* 2015; **60**(6): 1-45.

316. Mebane WR, Jr, Sekhon JS. Genetic Optimization Using Derivatives: The rgenoud Package for R. *Journal of Statistical Software* 2011; **42**(11): 1-26.

317. Sekhon JS, Mebane WR, Jr. Genetic Optimization Using Derivatives: Theory and Application to Nonlinear Models. *Political Analysis* 1998; **7**: 189-213.

318. Alistair Story, Richard S. Garfein, Andrew Hayward, et al. Monitoring Therapy Adherence of Tuberculosis Patients by using Video-Enabled Electronic Devices. *Emerging Infectious Disease journal* 2016; **22**(3): 538.

319. McGrail MR, Humphreys JS. Spatial access disparities to primary health care in rural and remote Australia. *Geospat Health* 2015; **10**(2).

320. Nesbitt RC, Gabrysch S, Laub A, et al. Methods to measure potential spatial access to delivery care in low- and middle-income countries: a case study in rural Ghana. *International Journal of Health Geographics* 2014; **13**(1): 1-13.

321. Hargreaves JR, Boccia D, Evans CA, Adato M, Petticrew M, Porter JDH. The Social Determinants of Tuberculosis: From Evidence to Action. *American journal of public health* 2011; **101**(4): 654-62.

322. BBC News. 'Super-gonorrhoea' outbreak in Leeds. 2015. http://www.bbc.co.uk/news/health-34269315 (accessed 12 October 2015).

323. Simms I, Field N, Jenkins C, et al. Intensified shigellosis epidemic associated with sexual transmission in men who have sex with men--Shigella flexneri and S. sonnei in England, 2004 to end of February 2015. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* 2015; **20**(15).

324. Simms I, Wallace L, Thomas DR, et al. Recent outbreaks of infectious syphilis, United Kingdom, January 2012 to April 2014. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* 2014; **19**(24).

325. Lehmiller JJ, Ioerger M. Social Networking Smartphone Applications and Sexual Health Outcomes among Men Who Have Sex with Men. *PLoS ONE* 2014; **9**(1): e86603.

326. Beymer MR, Weiss RE, Bolan RK, et al. Sex on demand: geosocial networking phone apps and risk of sexually transmitted infections among a cross-sectional sample of men who have sex with men in Los Angeles County. *Sexually Transmitted Infections* 2014; **90**(7): 567-72.

327. Hull P, Mao L, Prestage G, Zablotska I, de Wit J, Holt M. The use of mobile phone apps by Australian gay and bisexual men to meet sex partners: an analysis of sex-seeking repertoires and risks for HIV and STIs using behavioural surveillance data. *Sexually Transmitted Infections* 2016; **92**(7): 502-7.

328. Foster K, Cole M, Hotonu O, et al. How to do it: lessons identified from investigating and trying to control an outbreak of gonorrhoea in young heterosexual adults. *Sexually Transmitted Infections* 2016.

329. Shankar AG, Mandal S, Ijaz S. An outbreak of hepatitis B in men who have sex with men but identify as heterosexual. *Sexually Transmitted Infections* 2016.

330. Tam CC, Rodrigues LC, Viviani L, et al. Longitudinal study of infectious intestinal disease in the UK (IID2 study): incidence in the community and presenting to general practice. *Gut* 2012; **61**(1): 69-77.

331. Diggle P, Rowlingson B, Su T-l. Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics* 2005; **16**(5): 423-34.

332. Bayer C, Bernard H, Prager R, et al. An outbreak of Salmonella Newport associated with mung bean sprouts in Germany and the Netherlands, October to November 2011. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* 2014; **19**(1).

333. Inns T, Lane C, Peters T, et al. A multi-country Salmonella Enteritidis phage type 14b outbreak associated with eggs from a German producer: 'near real-time' application of whole genome sequencing and food chain investigations, United Kingdom, May to September 2014. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* 2015; **20**(16).

334. Launders N, Byrne L, Adams N, et al. Outbreak of Shiga toxin-producing E. coli O157 associated with consumption of watercress, United Kingdom, August to September 2013. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* 2013; **18**(44).

10 APPENDICES

APPENDIX 10.1: DETAILS OF STUDIES INCLUDED IN SYSTEMATIC LITERATURE REVIEW ON SPATIAL METHODS FOR INFECTIOUS DISEASE OUTBREAK INVESTIGATIONS (CHAPTER 2).

Listed by continent, country, author.

First author	Infectious disease	Location	Publication year	Context	Prospective/ retrospective	Spatial methods used	Stage of investigation*	Outcome summary
Europe								
Acheson[68]	Syphilis	UK	2011	STI	P	Dot map; Rate map	4,7	Some clusters found in high deprivation areas. Adverts placed on social networks linked to users' postcodes.
Boccia[78]	Salmonellosis	UK	2004	Food	P	Dot map; Spatial case definition; Source proximity	3,4,5	No significant difference between closest case and control to suspect outlets.
Carr[81]	Legionnaires' disease	UK	2010	Environmental	P	Dot map; Case movement map; Spatial case definition	3,4	Identified no hot spots; concluded pseudo-cluster.
Hyland[95]	Legionnaires' disease	UK	2008	Environmental	P	Dot map; Case movement map;	3,4,5,6,7,8	Sullage tanks identified as source. Review of national guidelines.

						Spatial case definition		
Keramarou[97]	Legionnaires' disease	UK	2010	Environmental	P	Dot map; Case movement map; Spatial case definition	3,4	Two distinct spatio-temporal clusters identified but no definitive source.
Kirrage[98]	Legionnaires' disease	UK	2007	Environmental	P	Dot map; Case movement map; Spatial case definition; Source proximity	3,4,5,7	Identified cluster of cooling towers as likely source. Closed and cleaned one of the towers.
Neira-Munoz[111]	Cryptosporidiosis	UK	2007	Water	P	Dot map; Thematic map; Spatial case definition	3,4,6	Hypothesis that low level contamination of drinking water caused outbreak. Potential change in water monitoring suggested.
Sanson[129]	Foot and mouth disease	UK	2011	Farm	R	Dot map; Spatial case definition; Source proximity; Case-case distance; Air dispersion modelling	5,6	Distance and direction from index farm significant predictors of infection status. Minimum infective dose might be less than previously established.

Wallensten[142]	Q-fever	UK	2010	Farm	P	Dot map; Spatial case definition; Air dispersion modelling	3,4,5	Air from each of suspected farms may have exposed town, couldn't rule any out as potential sources.
Le Comber[102]	Cholera & malaria	UK & Egypt	2011	Vector/ water	R	Dot map; Spatial average; Geographic profiling	4,5	Identified most likely locations of sources of infection.
Manfredi Selvaggi[106]	Q-fever	Italy	1996	Farm	P	Rate map; Thematic map; Spatial case finding	3,4	Infected individuals tended to live closer to sheep migration route.
Orsi[118]	Measles	Italy	2010	Community	P	Dot map	4	Identified worst affected areas.
Varani[140]	Leishmaniasis	Italy	2013	Vector	P	Dot map	3,4	Most patients in hilly, rural areas
Norstrom[115]	Acute respiratory disease	Norway	1999	Farm	R	Dot map; Smoothed incidence map; Spatial case definition; Space-time scan statistic; K-nearest neighbour test; Knox test	3,4,5	Described progression of outbreak. Identified cluster. Supports hypothesis of single common source of infection.

Nygard[117]	Legionnaires' disease	Norway	2008	Environmental	P	Dot map; Case movement map; Spatial case definition; Source proximity; Air dispersion modelling	3,4,5,6,7	Identified industrial air scrubber as source of outbreak. Scrubber closed, new routines for cleaning and national regulations implemented.
Nygard[116]	Giardiasis	Norway	2006	Water	P	Dot map; Thematic map; Spatial case definition	3,4,5,7	Higher attack rate in zone supplied by water supply A. Boil water notice issued.. Flushed distribution system.
Abellan[67]	Legionnaires' disease	Spain	2002	Environmental	R	Dot map; Smoothed incidence map; K-function	4,5	Cases more aggregated than controls; confirmed environmental origin of outbreak.
Garcia-Fulgueiras[92]	Legionnaires' disease	Spain	2003	Environmental	P	Rate map; Spatial case definition; Source proximity	3,4,5,6	Zone of exposure around hospital associated with illness; replaced cooling tower. Legionella may be able to spread over larger distances from source than previously thought.
Jansa[96]	Legionnaires' disease	Spain	2002	Environmental	P	Dot map; Spatial case definition	3,4	Cooling towers identified as source.

Hackert[94]	Q-fever	Netherlands	2012	Farm	R	Dot map; Smoothed incidence map; Spatial case definition; Source proximity	3,4,5,6	Incidence increased with proximity to index farm. Cases scattered in wedge shape area downwind of farm.
Schimmer[132]	Q-fever	Netherlands	2010	Farm	R	Dot map; Thematic map; Spatial case definition; Source proximity; Spatial average	3,4,5	Gradual diminishing risk from certain farms, identified as probable sources
van der Hoek[139]	Q-fever	Netherlands	2012	Farm	R	Dot map; Rate map; Thematic map; Smoothed incidence map; Source proximity	4,8	Identified 5 hot spots, all around infected dairy goat farms.
Gubbels[93]	Campylobacteriosis	Denmark	2012	Water	P	Dot map; Thematic map; Spatial case definition	3,4,7	Cases lived across entire water supply area; concluded contamination of central water supply. Implemented boiling order.

Nguyen[112]	Legionnaires' disease	France	2006	Environmental	P	Dot map; Rate map; Case movement map; Spatial case definition; Spatial case finding; Air dispersion modelling	3,4	Dispersion of plumes from cooling tower correlated with geographical distribution of cases. Spread over longer distance than previously thought possible.
Kistemann[99]	Salmonellosis	Germany	2000	Hospital	P	Thematic maps; Schematic map; Spatial case definition	1,3,4,7,8	Identified functional relationship between cases. Measures introduced to prevent future outbreaks.
Fitzpatrick[91]	Measles	Ireland	2012	Community	P	Dot map; Thematic map	4,7	Identified emergence of cluster during outbreak in real time. Intervention in high rate area - expediated MMR vaccine schedule/ catch up campaign.
Ulugtekin[138]	Measles	Turkey	2007	Community	P	Dot map; Thematic maps	4	Identified high incidence areas.
Asia								
Lai[101]	Influenza	Hong Kong	2010	Community	R	Dot map; Smoothed incidence map; Standard deviation ellipse; Moran's	4,5	Identified hot spots and directional trend.

						I; Getis-Ord Gi* statistic		
Lai[100]	SARS	Hong Kong	2004	Community	R	Dot map; Rate map; Smoothed incidence map; Standard deviation ellipse; Origin-destination plots; Moran's I; Nearest neighbour analysis	4,5	Clear clustering identified. Directional bias and radius of spread of superspreading events demonstrated.
Sze-To[135]	Varicella	Hong Kong	2011	Hospital	R	Schematic map; Air dispersion modelling	4,5	Model matches epidemiological distribution of cases.
Wong[144]	Influenza	Hong Kong	2010	Hospital	R	Schematic map; Spatial case definition; Source proximity; Air dispersion modelling	3,4,5,6	Proximity to air purifier associated with infection. Suggests possible role for aerosol transmission.

Yu[145]	SARS	Hong Kong	2005	Hospital	R	Schematic map; Spatial case definition; Air dispersion modelling	3,4,5	Attack rates higher in bays closer to index patient. Suggests airborne transmission played important role.
Bali[72]	Hepatitis E	India	2008	Water	P	Dot map; Spatial case finding; Spatial case definition	3,4	Cases mapped to water supply distribution area.
Nisha[113]	Dengue	India	2005	Vector	P	Dot map; Spatial case finding; Scan statistic	3,4,7	Identified cluster. Fogging and larval reduction. Drawing up standard protocol for GIS in outbreaks.
Saha[128]	Shigellosis	India	2009	Water	P	Rate map; Spatial case definition	3,4	Incidence higher downstream of damaged pipeline.
Sarkar[130]	Diarrhoea	India	2007	Water	P	Dot map; Thematic map; Spatial case definition; Source proximity; Spatial case finding; Spatial scan statistic	3,4,5,7,8	Showed dispersed nature of outbreak; no significant clustering. Funds released to improve drainage network.
Sowmyanarayanan[134]	Hepatitis A	India	2008	Water	P	Dot map; Spatial scan statistic	4,5	Cluster not significant Outbreak generalised across area.

Fang[87]	Influenza	China	2013	Community	R	Dot map; Thematic map; Environmental risk prediction model	4	Identified high incidence areas. Predicted areas with high risk to inform future control efforts.
Liang[104]	SARS	China	2007	Community	R	Rate map	4	Rate increased with distance from city centre, supported spatial quarantining of city for future outbreaks.
Ali[70]	Dengue	Bangladesh	2003	Vector	R	Dot map; Thematic map; Smoothed incidence map; Source proximity; Spatial case finding; Kriging	3,4,5	Clusters identified, generally closer to major hospitals. Spatial association between clusters and vector populations.
Tenzin[136]	Rabies	Bhutan	2010	Community	R	Dot map; Spatial average; Standard deviation ellipse	4	Visualised spread of outbreak. Seemed to follow road network that had many free-roaming dogs.
Nishiguchi[114]	Influenza	Japan	2009	Farm	R	Dot map; Scan statistic	3,4,5	Identified cluster and factors associated with farms inside cluster.

Siddiqui[133]	Cholera	Pakistan	2006	Water	R	Dot map; Spatial case definition; K-nearest neighbour test	3,4,5	Clustering in one of the outbreaks investigated; water reservoir identified as likely source.
Miranda[108]	Ebola	Philippines	2002	Breeding facility	R	Schematic map	4	Documented progression of outbreak.
Le[103]	Porcine high fever disease	Vietnam	2012	Farm	R	Dot map; Smoothed incidence map; Spatial & space-time scan statistic; K-nearest neighbour test; Knox test; Space-time k function	4,5	Little evidence for clustering; thought not to be important in this outbreak.
North America								
Addiss[18]	Legionnaires' disease	USA	1989	Environmental	P	Dot maps; Spatial case definition; Source proximity	3,4,5	Rate decreased with distance from one cooling tower. Implicated as probable source of outbreak.
Blondin[77]	Blastomycosis	USA	2007	Environmental	R	Dot map; Thematic map; Source proximity	4,5	No common source identified; infection likely to have been acquired close to homes.

Brown[80]	Legionnaires' disease	USA	1999	Environmental	P	Dot map; Thematic map; Spatial case definition; Source proximity	3,4,5	Transmission mostly in close proximity to cooling towers.
Chung[84]	West Nile Virus	USA	2013	Vector	R	Dot map; Rate map; Getis-Ord Gi* statistic	4,5	As outbreak progressed it became clustered and hot spot identified.
McKee[107]	Shigellosis	USA	2000	Water	P	Dot map; K-nearest neighbour test	4,7	Space-time clustering found; identified communal wading pools as probable source. Targeted information campaigns and education.
Mongoh[109]	Anthrax	USA	2008	Farm	R	Dot map; Thematic map	4	Displayed spatial distribution of premises with cases in study.
Pfister[122]	Blastomycosis	USA	2011	Environmental	P	Dot map; Spatial case definition; Spatial average	3,4	Centre of outbreak identified, north of river. Yard waste disposal identified as likely source.
Roy[127]	Blastomycosis	USA	2013	Environmental	P	Dot map; Spatial case definition; Scan statistic	1,3,4	Confirmed the presence of the outbreak.

Bowie[79]	Toxoplasmosis	Canada	1997	Water	P	Dot map; Thematic map	4,5	Outbreak related cases in area served by water distribution system.
Epp[86]	Anthrax	Canada	2010	Farm	R	Thematic map; Smoothed incidence map; Velocity vector map; Spatial case definition; Space time scan statistic; K-nearest neighbour test; K-function; Oden's ipop	4,5	Three separate movements of spread identified; clusters located.
Parkinson[119]	Anthrax	Canada	2003	Farm	R	Dot map; Thematic maps	4	Described physical characteristics of outbreak and documented spatial descriptive patterns.
Pasma[120]	Influenza	Canada	2008	Farm	R	Thematic map; Spatial average; Standard deviation ellipse; k-nearest neighbour test; Spatial scan	4,5,6	Identified and located clusters. Outbreak established in densely populated areas, moved to less densely populated areas. Suggests focus for surveillance.

						statistic; Knox test; Nearest neighbour analysis; Mantel's test		
Morrison[110]	Dengue	Puerto Rico	1998	Vector	R	Dot map; Knox test; K-function analysis; Barton and David test	3,4	Significant case clustering within households over short periods of time; but in general cases had pattern similar to population as a whole. Control measures need to be applied to entire municipality.
Chadee[82]	Meningococcal meningitis	Trinidad	2006	Community	P	Dot map; Case-case distance	1,4	Revealed two clusters.
Chadee[83]	Dengue	West Indies	2005	Vector	R	Dot map; K-nearest neighbour test	4	Cases occurred in clusters when mosquito densities were high enough.
Africa								
Affolabi[69]	Tuberculosis	Benin	2009	Community	R	Dot map; Case movement map	1,4	Identified potential cluster.
Bartels[75]	Cholera	Ethiopia	2010	Water	P	Dot map	4	Cases mapped along river; thought to be most likely source.

Luquero[105]	Cholera	Guinea-Bissau	2011	Water	P	Dot map; Rate map; Smooth incidence maps; Spatial case finding; Scan statistic; K-function	3,4,5,7,8	Two clusters identified. Improved sanitation systems and hygiene collection in affected area.
Rivas[123]	Influenza	Nigeria	2010	Farm	R	Dot map; Thematic maps; Spatial case definition; Case-case distance; Risk factor proximity	3,4,5	Supports hypothesis that major highway network promoted epidemic spread.
Roquet[125]	Cholera	Senegal	1998	Water	R	Dot map; Rate map	4	Identified high incidence areas
Bessong[76]	Diarrhoea	South Africa	2009	Water	P	Dot map; Spatial case finding	3,4	Identified hot spot of outbreak. Two water extraction points implicated.
Fevre[89]	Trypanosomiasis	Uganda	2001	Vector	R	Dot map; Spatial case definition; Source proximity; Spatial scan statistic	3,4,5	Significant cluster detected. Distance from market significant risk factor.

Sasaki[131]	Cholera	Zambia	2008	Water	R	Dot map; Rate map; Voronoi diagram; Nearest neighbour analysis; Moran's I	4,5	Significant clustering found in areas with lower coverage of latrines and effective drainage systems.
Fernandez[88]	Cholera	Zimbabwe	2011	Water	R	Dot map; Thematic map; Rate map; Empirical Bayes smoothing	4,5,7	Spatial pattern linked to historical social construction of city characterised by distinct regions of socioeconomic status.
South America								
Angulo[71]	Variola minor	Brazil	1979	Community	R	Dot map; Smoothed incidence map	4	Demonstrated importance of schools in epidemic spread.
Barcellos[73]	Leptospirosis	Brazil	2000	Water	R	Dot map; Thematic maps; Risk factor proximity	4	Concentration of cases observed; identified characteristics of areas.
Barreto[74]	Schistosomiasis	Brazil	1993	Vector	P	Dot map; Thematic maps	4	Children with frequent water contact around open bodies of water, no sewage disposal, absence of water supply associated with infection.

de Moura[85]	Toxoplasmosis	Brazil	2006	Water	P	Dot map; Rate map	4,6,7	Cases more likely to be served by water reservoir A than B. Closed reservoir.
Passos[121]	Rabies	Brazil	1998	Community	R	Dot map	4	Cases corresponded to parts of city with most slums and lower income populations.
Rotela[126]	Dengue	Argentina	2007	Vector	R	Dot map; Smoothed incidence map; Knox test; Environmental risk prediction model	4,5	Identified clusters and developed predictive risk model.
Rivas[124]	Foot and mouth disease	Uruguay	2003	Farm	R	Dot map; Thematic map; Source proximity	4,5	Generated hypothesis that early epidemic virus disseminated taking advantage of road network, then spread outwards.
Australia								
Firestone[90]	Influenza	Australia	2011	Farm	R	Dot map; Smoothed incidence map; Spatial social network analysis; Space-time scan	4,5,6	Local spread through contact network to distance of 15km. Identified 5 significant clusters.

						statistic; Kriging		
Waldron[141]	Cryptosporidiosis	Australia	2011	Water	R	Dot map	4	Identified hot spots and movement of cluster over time.

White[143]	Legionnaires' disease	New Zealand	2013	Environmental	P	Dot map;
						Thematic map;
						Scan statistic;
						Moran's I
							4,5,6,8	Identified clusters; case distribution consistent with plume effect from probable source.

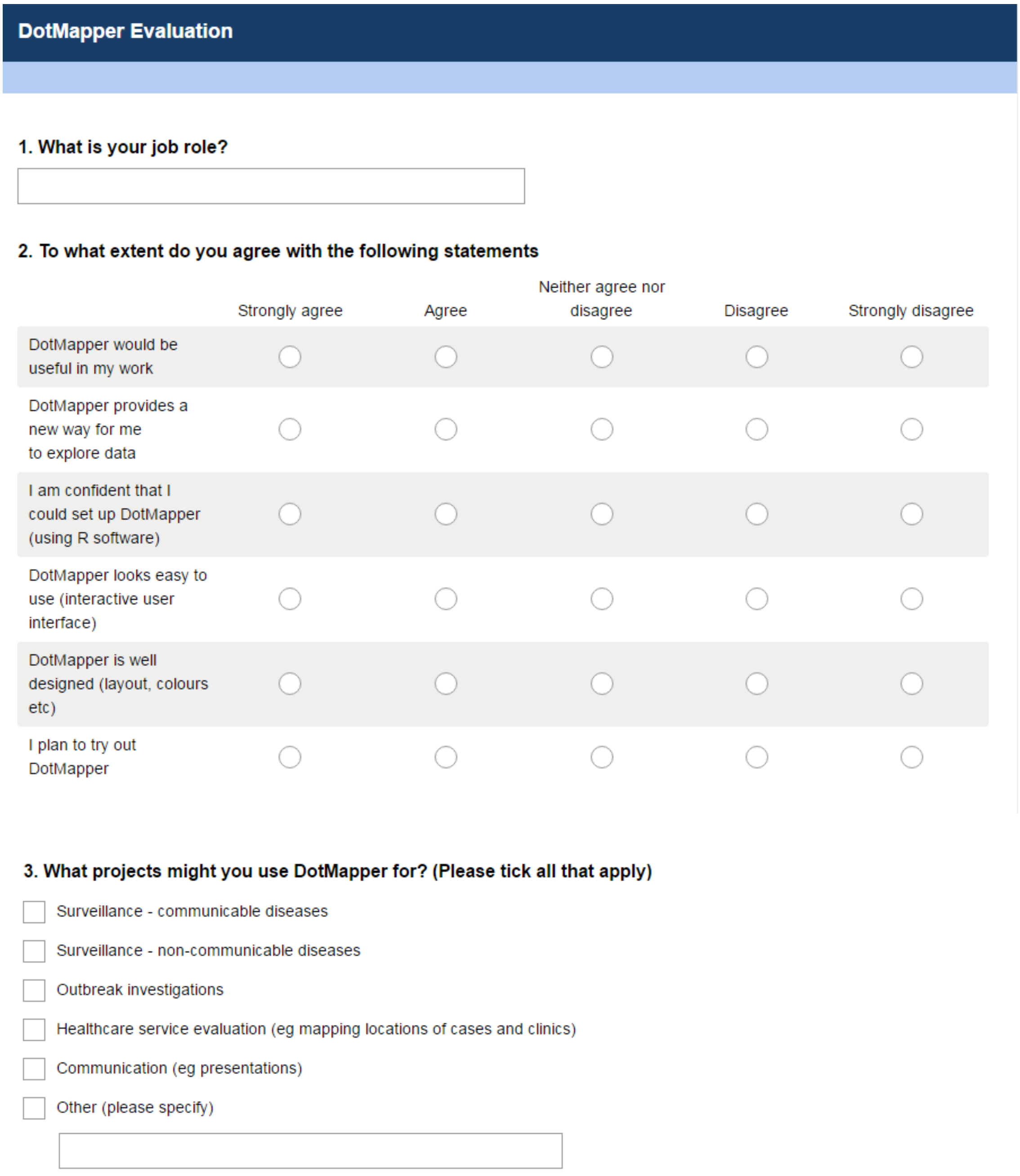
Unknown location								

Turcios-Ruiz[137]	Necrotizing enterocolitis	Location not stated	2008	Hospital	P	Schematic map;
						Spatial case definition;
						Grimson test
							3,4,5	Clustering identified. Suggested possible association with caregivers working in affected area.

*Stages in outbreak investigations defined as: 1. Establishing existence of an outbreak; 2. Confirming diagnosis; 3. Defining and identifying outbreak cases; 4. Describing cases and developing hypotheses; 5. Evaluating hypotheses and drawing conclusions; 6. Comparing with established facts; 7. Executing prevention measures; 8. Communicating findings.

MMR: measles-mumps-rubella vaccine; P: prospective; R: retrospective; SARS: severe acute respiratory syndrome.

APPENDIX 10.2: QUESTIONNAIRE USED IN EVALUATION OF INTERACTIVE MAPPING APPLICATION (CHAPTER 3).



DotMapper Evaluation

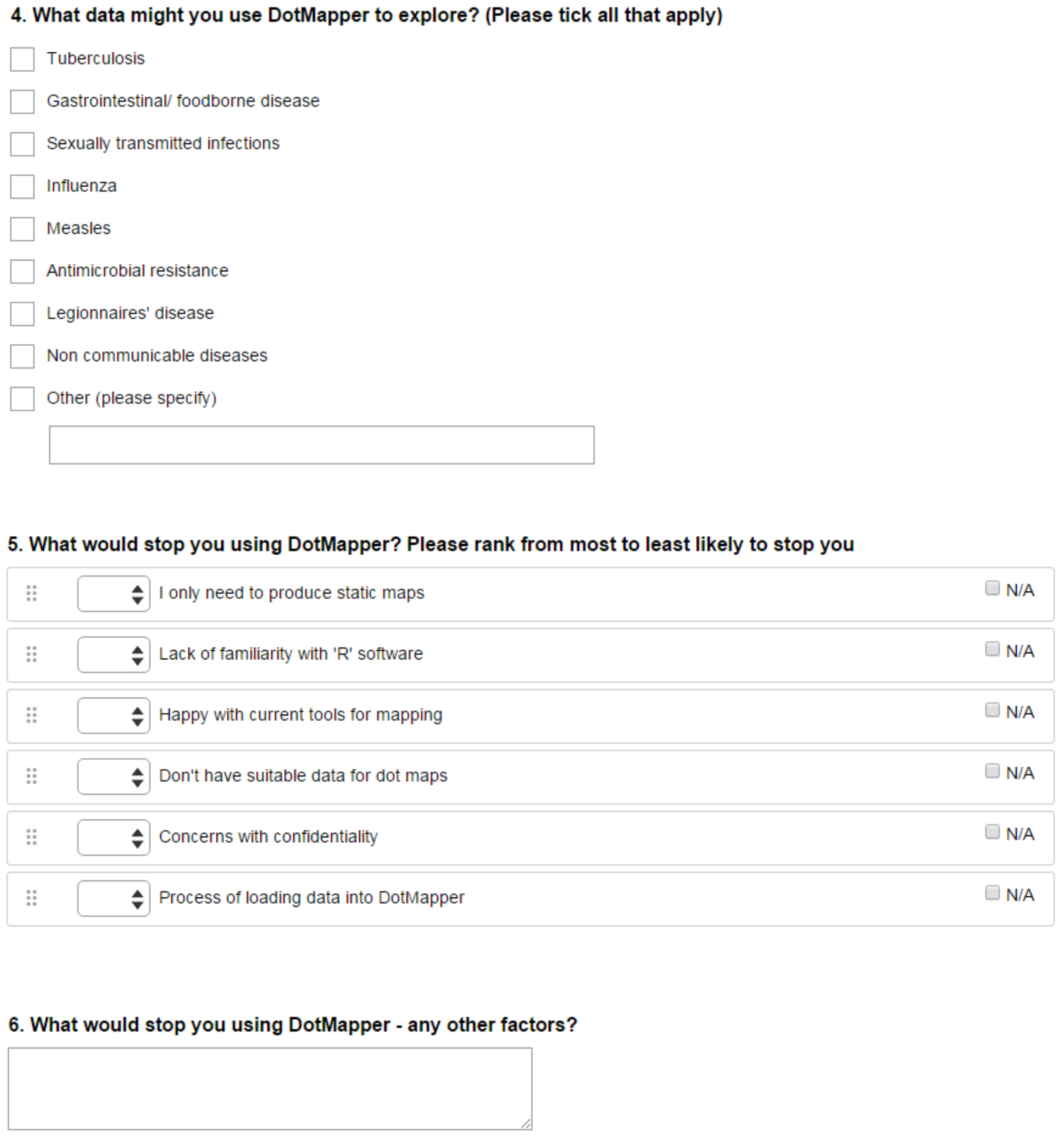
1. What is your job role?

2. To what extent do you agree with the following statements

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
DotMapper would be useful in my work	○	○	○	○	○
DotMapper provides a new way for me to explore data	○	○	○	○	○
I am confident that I could set up DotMapper (using R software)	○	○	○	○	○
DotMapper looks easy to use (interactive user interface)	○	○	○	○	○
DotMapper is well designed (layout, colours etc)	○	○	○	○	○
I plan to try out DotMapper	○	○	○	○	○

3. What projects might you use DotMapper for? (Please tick all that apply)

- □ Surveillance - communicable diseases
- □ Surveillance - non-communicable diseases
- □ Outbreak investigations
- □ Healthcare service evaluation (eg mapping locations of cases and clinics)
- □ Communication (eg presentations)
- □ Other (please specify)



4. What data might you use DotMapper to explore? (Please tick all that apply)

□ Tuberculosis

□ Gastrointestinal/ foodborne disease

□ Sexually transmitted infections

□ Influenza

□ Measles

□ Antimicrobial resistance

□ Legionnaires' disease

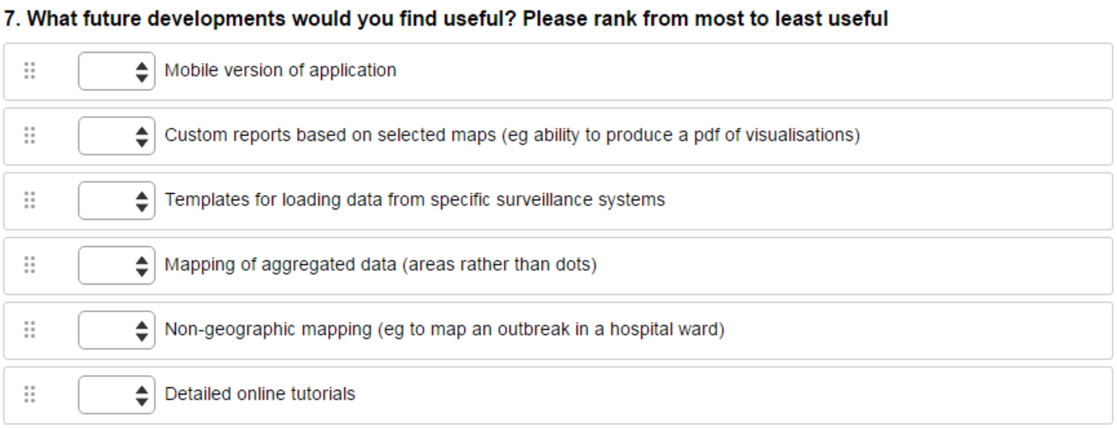
□ Non communicable diseases

□ Other (please specify)

5. What would stop you using DotMapper? Please rank from most to least likely to stop you

⇕	I only need to produce static maps	□ N/A
⇕	Lack of familiarity with 'R' software	□ N/A
⇕	Happy with current tools for mapping	□ N/A
⇕	Don't have suitable data for dot maps	□ N/A
⇕	Concerns with confidentiality	□ N/A
⇕	Process of loading data into DotMapper	□ N/A

6. What would stop you using DotMapper - any other factors?



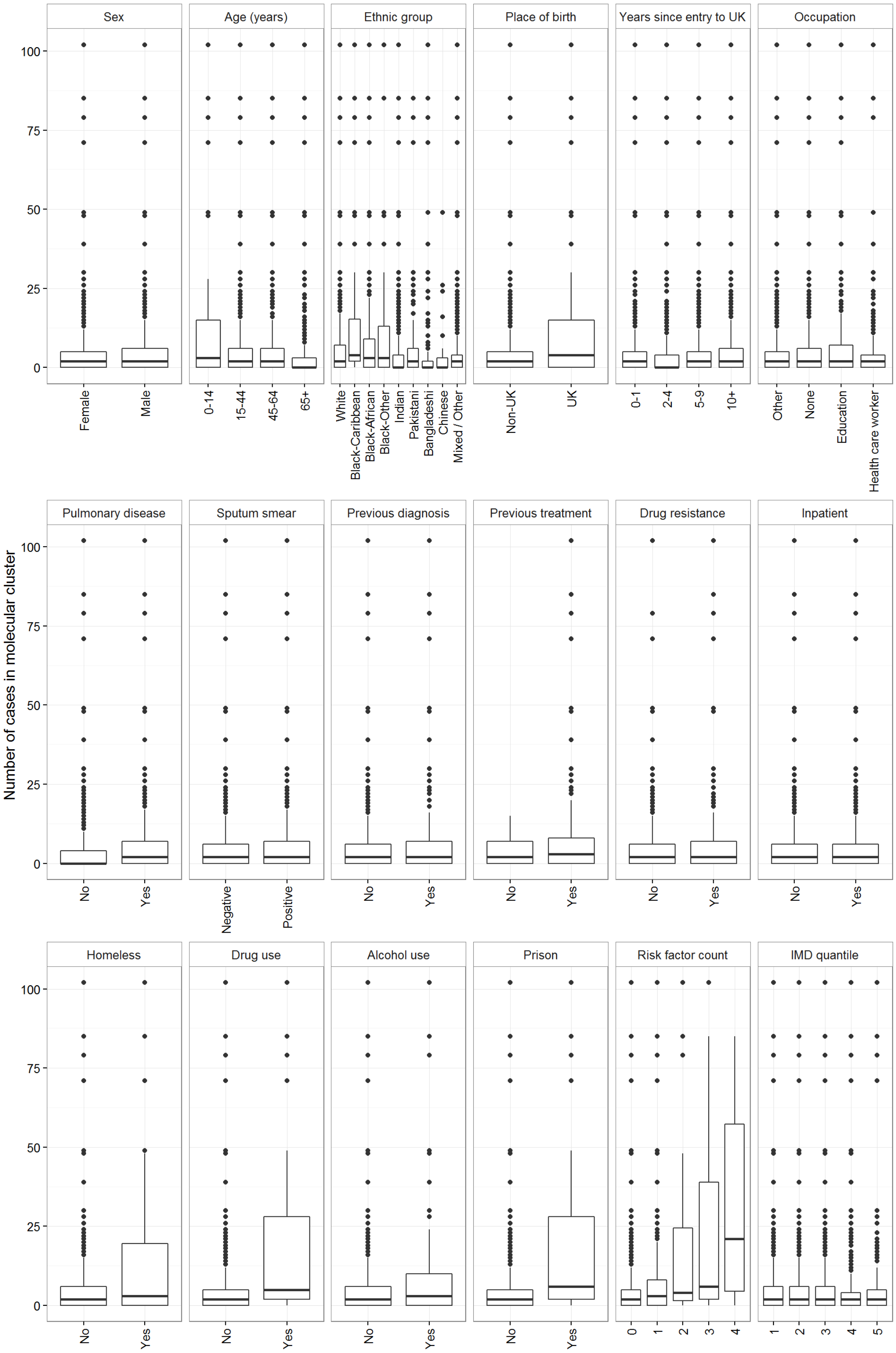
7. What future developments would you find useful? Please rank from most to least useful

⇕	Mobile version of application
⇕	Custom reports based on selected maps (eg ability to produce a pdf of visualisations)
⇕	Templates for loading data from specific surveillance systems
⇕	Mapping of aggregated data (areas rather than dots)
⇕	Non-geographic mapping (eg to map an outbreak in a hospital ward)
⇕	Detailed online tutorials

8. Other ideas for future developments

9. Please use this box for any other comments about DotMapper

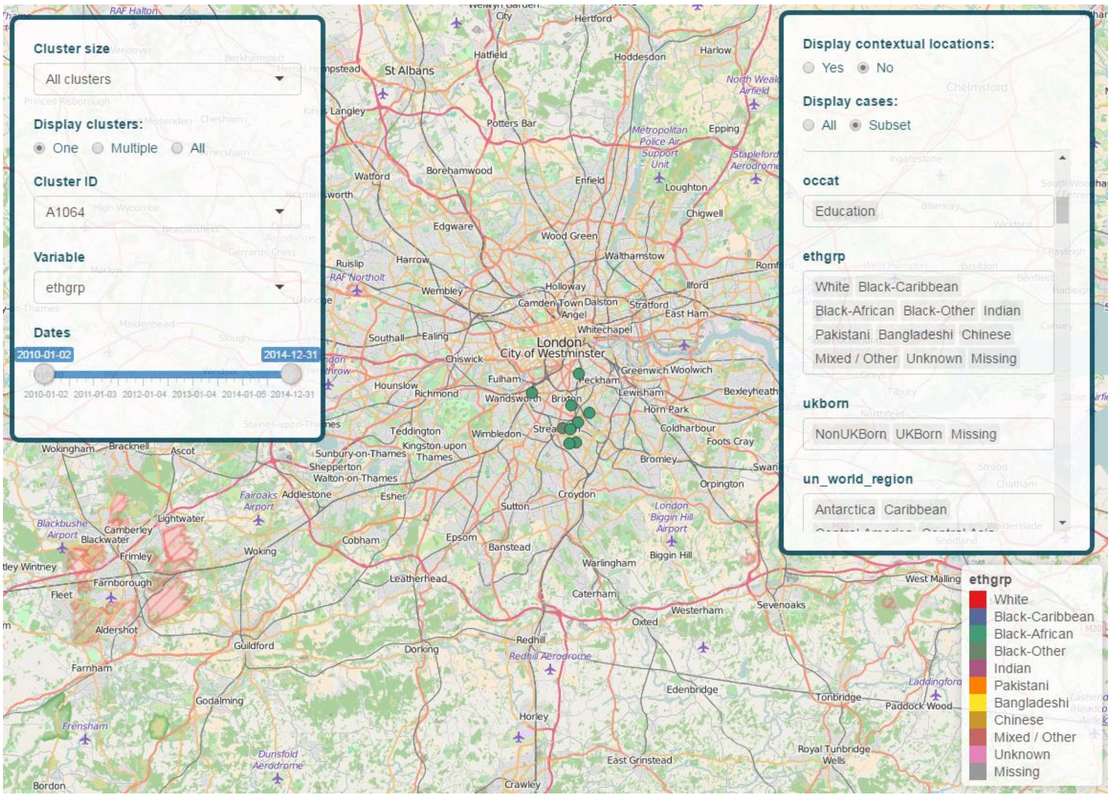
Appendix 10.3: Box plots of distribution of number of cases in molecular tuberculosis cluster by risk factor, London, 2010-2014 (Chapter 4).



Appendix 10.4: Screen shots from interactive mapping application showing key characteristics of molecular clusters of tuberculosis with significant spatial clustering in London, 2010-2014 (Chapter 4).

NB: Locations of cases in screen shots have been altered; they do not show actual case residential locations. Maps © OpenStreetMap contributors.[172]

A:



Cluster size: All clusters

Display clusters: ● One ○ Multiple ○ All

Cluster ID: A1064

Variable: ethgrp

Dates: 2010-01-02 – 2014-12-31

Display contextual locations: ○ Yes ● No

Display cases: ○ All ● Subset

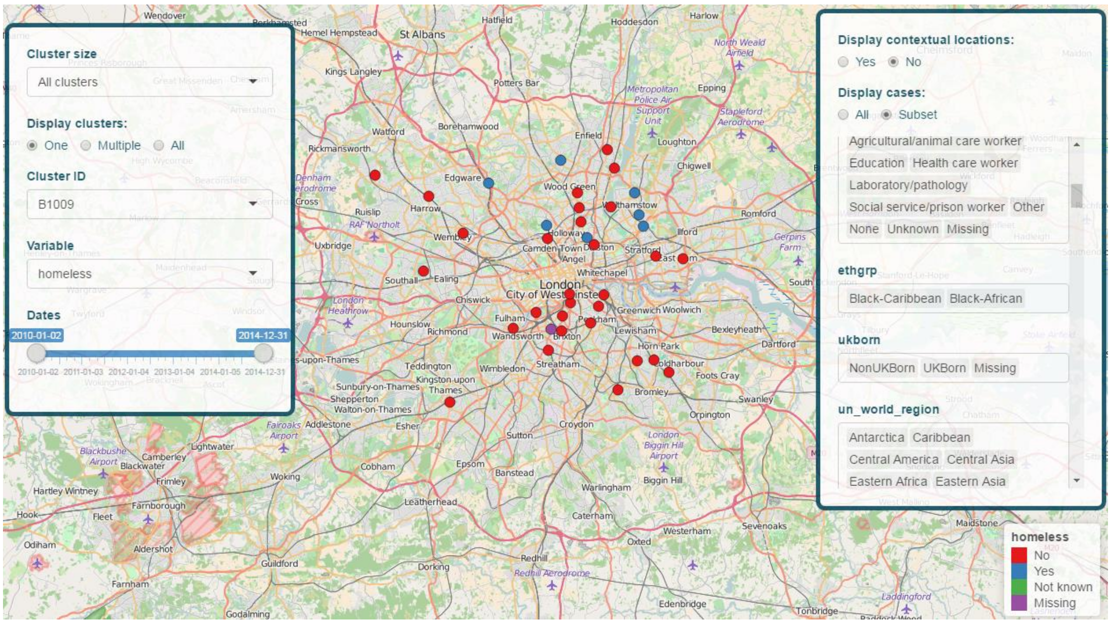
occat: Education

ethgrp: White, Black-Caribbean, Black-African, Black-Other, Indian, Pakistani, Bangladeshi, Chinese, Mixed / Other, Unknown, Missing

ukborn: NonUKBorn, UKBorn, Missing

un_world_region: Antarctica, Caribbean

B:



Cluster size: All clusters

Display clusters: ● One ○ Multiple ○ All

Cluster ID: B1009

Variable: homeless

Dates: 2010-01-02 – 2014-12-31

Display contextual locations: ○ Yes ● No

Display cases: ○ All ● Subset

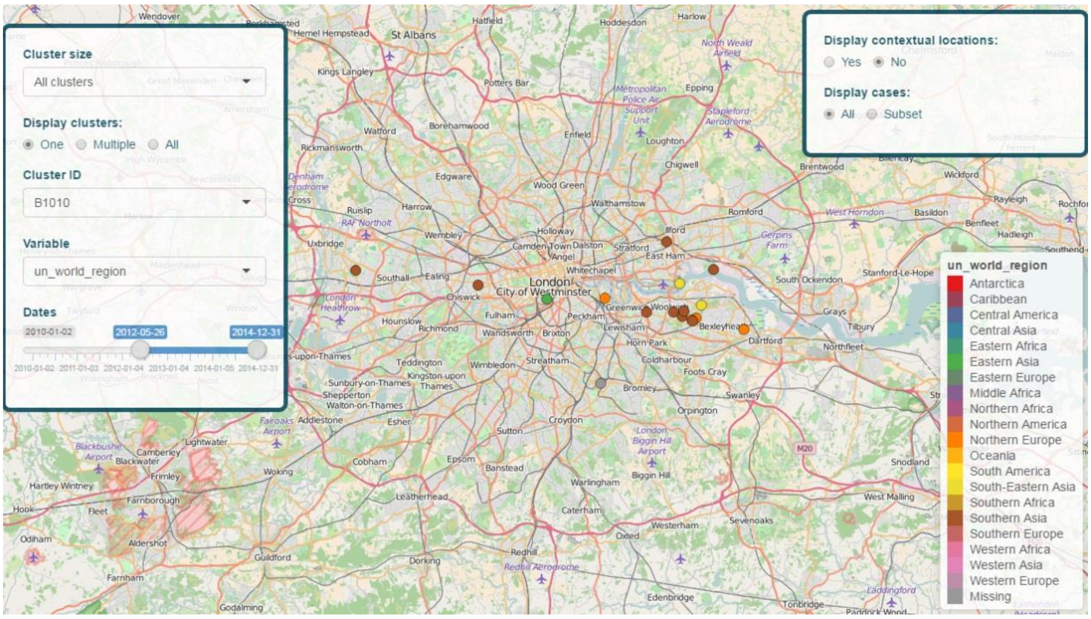
occat: Agricultural/animal care worker, Education, Health care worker, Laboratory/pathology, Social service/prison worker, Other, None, Unknown, Missing

ethgrp: Black-Caribbean, Black-African

ukborn: NonUKBorn, UKBorn, Missing

un_world_region: Antarctica, Caribbean, Central America, Central Asia, Eastern Africa, Eastern Asia

C:



Cluster size

All clusters

Display clusters:

● One ○ Multiple ○ All

Cluster ID

B1010

Variable

un_world_region

Dates

2010-01-02 2012-05-26 2014-12-31

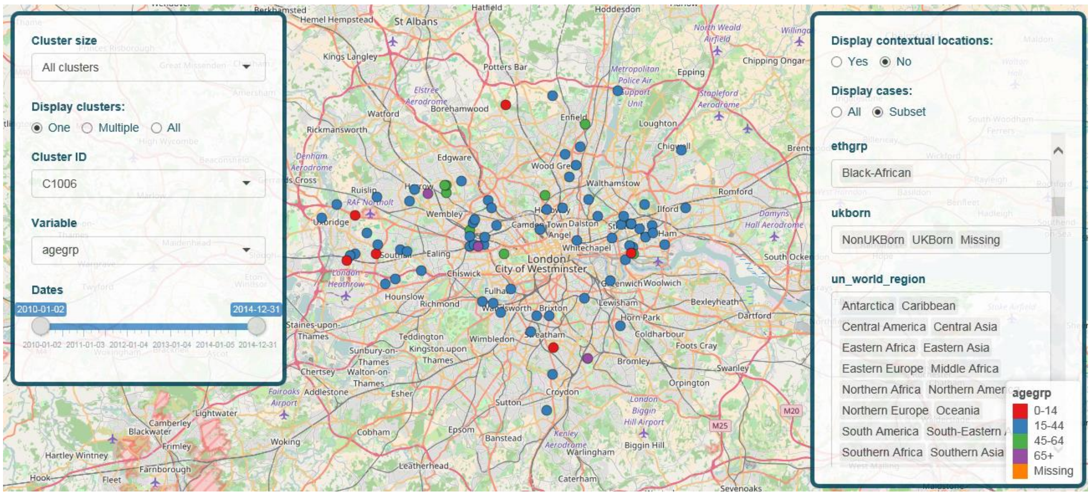
Display contextual locations:

○ Yes ● No

Display cases:

● All ○ Subset

D:



Cluster size

All clusters

Display clusters:

● One ○ Multiple ○ All

Cluster ID

C1006

Variable

agegrp

Dates

2010-01-02 2014-12-31

Display contextual locations:

○ Yes ● No

Display cases:

○ All ● Subset

ethgrp

Black-African

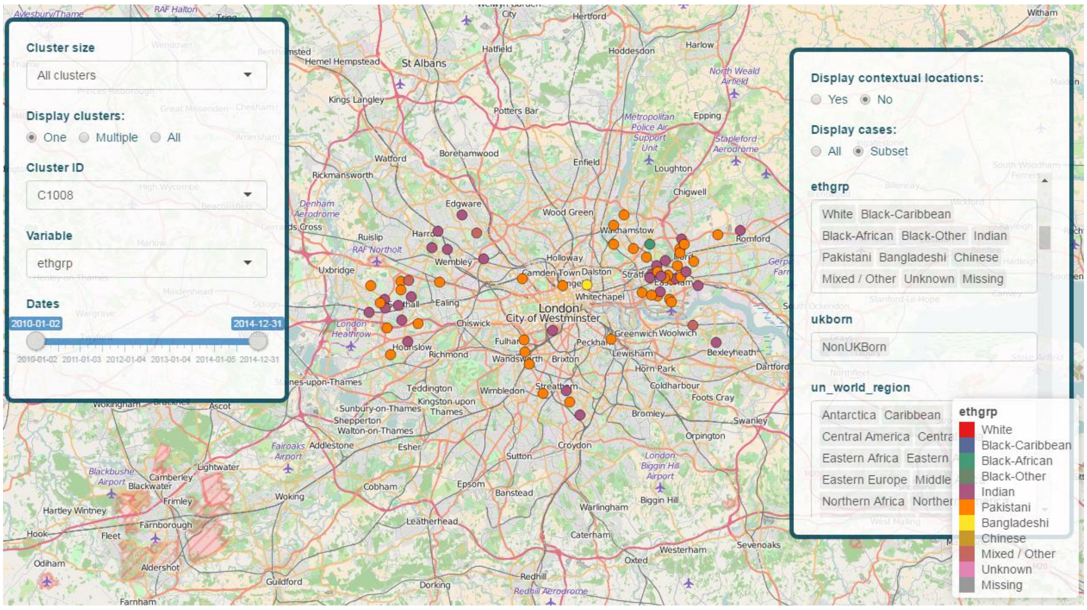
ukborn

NonUKBorn UKBorn Missing

un_world_region

Antarctica Caribbean Central America Central Asia Eastern Africa Eastern Asia Eastern Europe Middle Africa Northern Africa Northern Amer Northern Europe Oceania South America South-Eastern / Southern Africa Southern Asia

E:



Cluster size

All clusters ▾

Display clusters:

● One ○ Multiple ○ All

Cluster ID

C1008 ▾

Variable

ethgrp ▾

Dates

2010-01-02 – 2014-12-31

Display contextual locations:

○ Yes ● No

Display cases:

○ All ● Subset

ethgrp

White Black-Caribbean Black-African Black-Other Indian Pakistani Bangladeshi Chinese Mixed / Other Unknown Missing

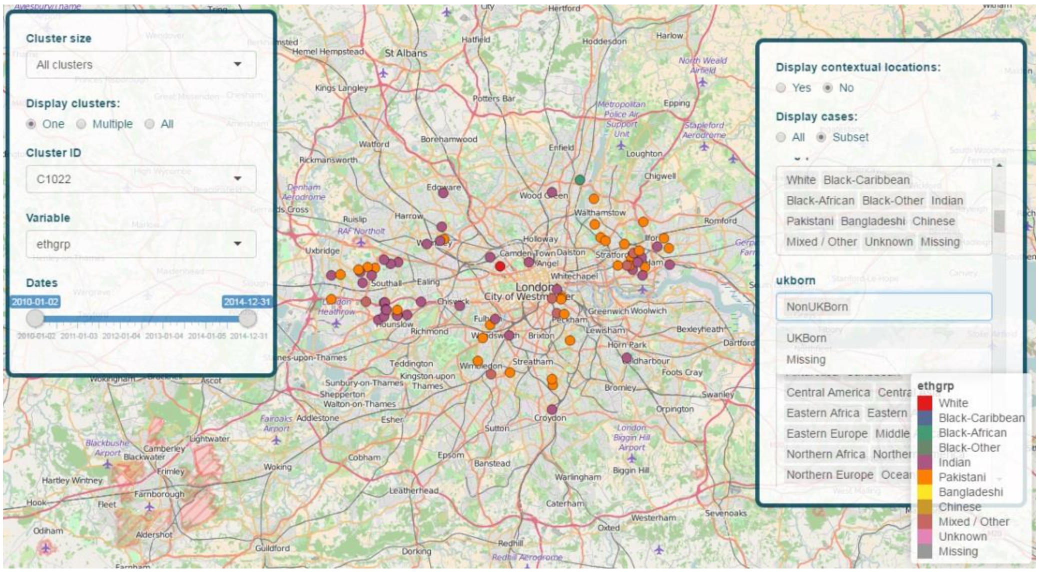
ukborn

NonUKBorn

un_world_region

Antarctica Caribbean Central America Centra Eastern Africa Eastern Eastern Europe Middle Northern Africa Norther

F:



Cluster size

All clusters ▾

Display clusters:

● One ○ Multiple ○ All

Cluster ID

C1022 ▾

Variable

ethgrp ▾

Dates

2010-01-02 – 2014-12-31

Display contextual locations:

○ Yes ● No

Display cases:

○ All ● Subset

White Black-Caribbean Black-African Black-Other Indian Pakistani Bangladeshi Chinese Mixed / Other Unknown Missing

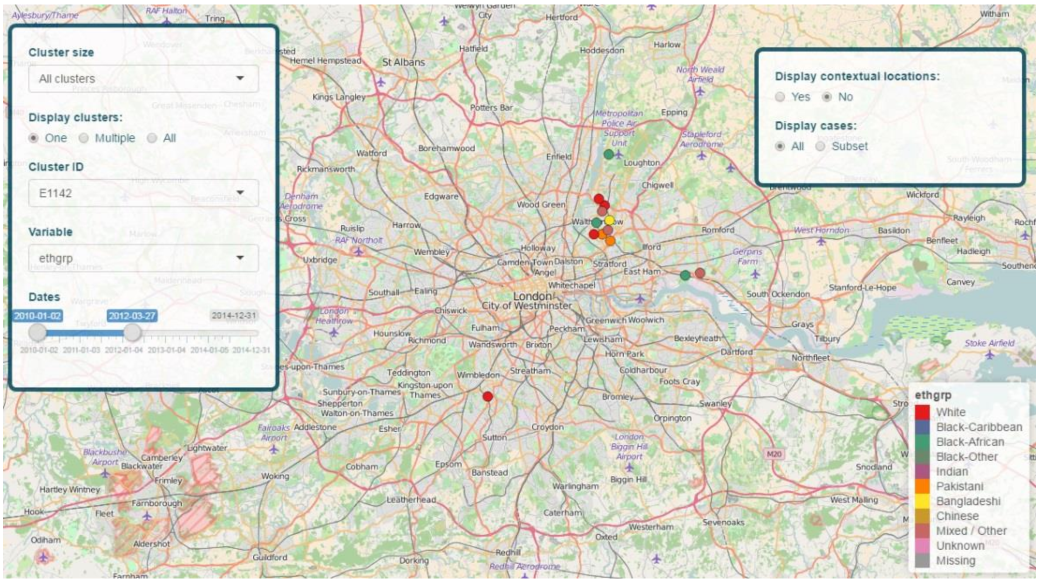
ukborn

NonUKBorn

UKBorn
Missing

Central America Centra Eastern Africa Eastern Eastern Europe Middle Northern Africa Norther Northern Europe Ocear

G:



Cluster size

All clusters ▾

Display clusters:

● One ○ Multiple ○ All

Cluster ID

E1142 ▾

Variable

ethgrp ▾

Dates

2010-01-02 2012-03-27 2014-12-31

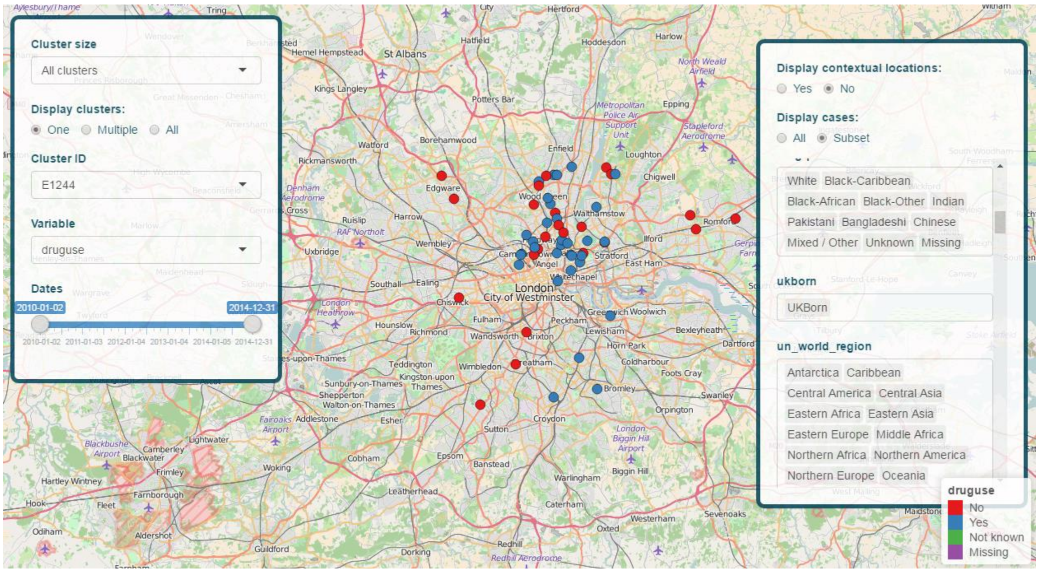
Display contextual locations:

○ Yes ● No

Display cases:

● All ○ Subset

H:



Cluster size

All clusters ▾

Display clusters:

● One ○ Multiple ○ All

Cluster ID

E1244 ▾

Variable

druguse ▾

Dates

2010-01-02 2014-12-31

Display contextual locations:

○ Yes ● No

Display cases:

○ All ● Subset

White Black-Caribbean Black-African Black-Other Indian Pakistani Bangladeshi Chinese Mixed / Other Unknown Missing

ukborn

UKBorn

un_world_region

Antarctica Caribbean Central America Central Asia Eastern Africa Eastern Asia Eastern Europe Middle Africa Northern Africa Northern America Northern Europe Oceania

APPENDIX 10.5: SENSITIVITY ANALYSIS USING DIFFERENT IMPLEMENTATIONS OF GEOGRAPHIC PROFILE MODEL FOR ANALYSIS OF RANDOMISED BADGER CULLING TRIAL DATA (CHAPTER 6).

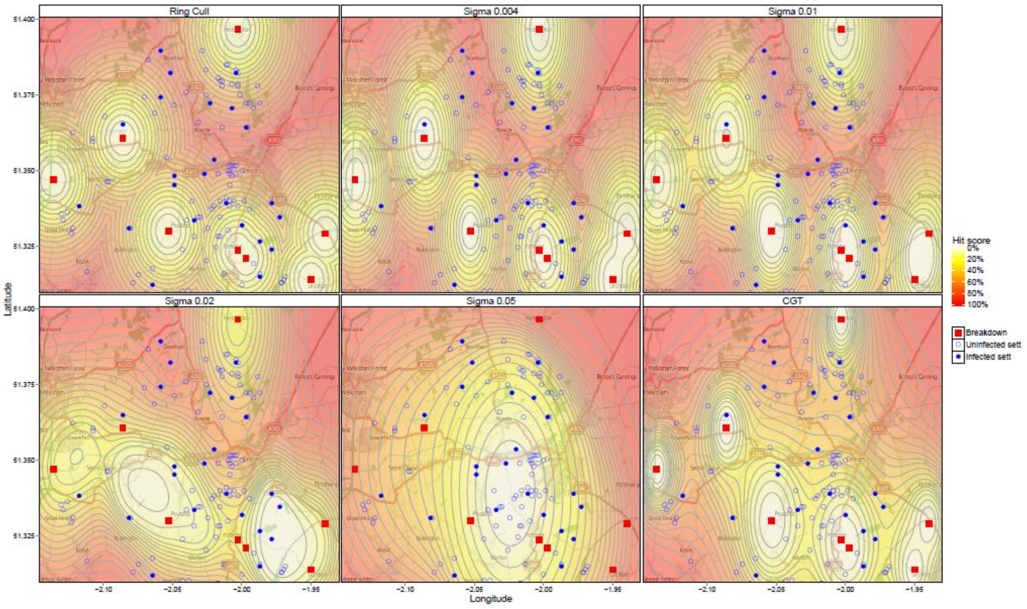
Alternative implementations of the geographic profiling model, using different values of the DPM geographic profiling clustering parameter, σ, and the CGT algorithm, were used to assess the sensitivity of the results to the model used. Search strategies for these methods for trial regions B2 and E3 are shown in Figure 10.5.1. Comparing the surfaces for the different values of σ clearly demonstrates the effect of altering this parameter. The smallest value of σ, 0.004, does not result in formation of many groups of breakdowns that may have arisen from the same source, and therefore closely approximates the ring cull design. Increasing this value (σ=0.01 and 0.02) produces some small clusters of breakdowns, with lower hit scores in the centre. The largest value of σ forms one large cluster of all breakdowns in the study area and therefore places the lowest hit scores in the centre of surface.

The CGT model uses a different algorithm to generate clusters. It produced a similar number of clusters to the DPM model with a σ value of 0.01, but the distribution of hit scores between clusters differs: For the CGT model, the majority of the highest probability is focused in the centre of the whole surface, whereas for the DPM model hit scores are symmetrically spread out from the clusters to the peripheral areas of the surface as well as the centre.

Figure 10.5.1: Distributions of hit scores around cattle herd breakdowns, designed using ring cull and different implementations of the geographic profiling model.

Map © OpenStreetMap contributors.[172]

Trial region B2.



Trial region E3.

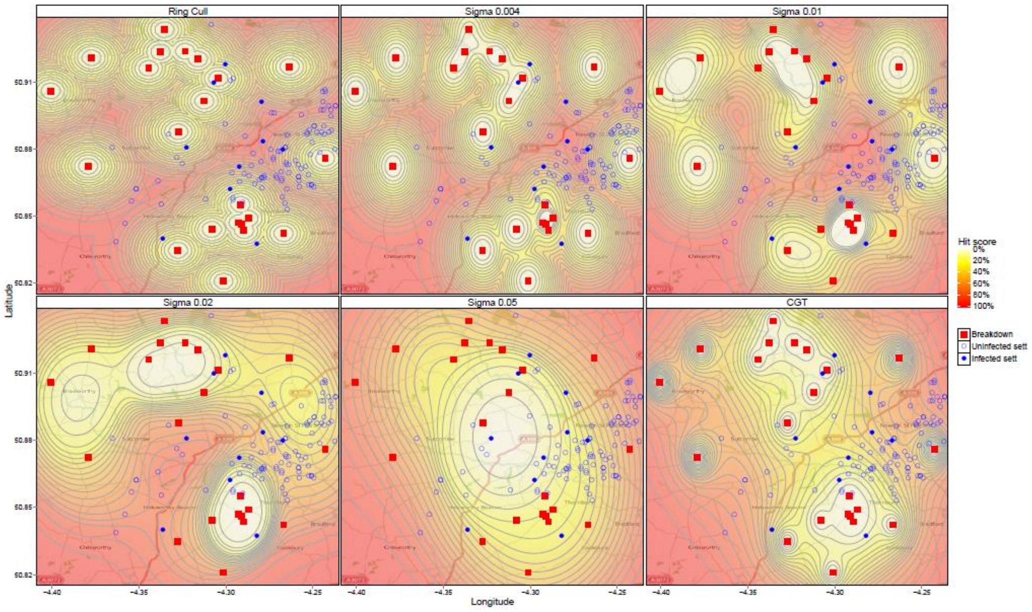


Figure 10.5.2 compares the survival functions for infected and uninfected setts for each geographic profiling model, aggregated over all trial areas. As anticipated from inspection of the hit score distributions, the functions for σ 0.04 closely resemble those for the ring cull search. The functions for σ 0.01 are similar to those

for σ 0.02 (used in the primary analysis), with infected setts identified slightly more efficiently for smaller search areas. The CGT and σ 0.05 models apparently identified infected setts more efficiently than uninfected setts at all search area sizes, with the Kaplan-Meier curve for infected setts remaining above that for uninfected setts. Cox regression analysis (Table 10.5.1) confirms these results: Infected setts were identified at a higher rate than uninfected setts for the σ 0.05 (HR (infected/ uninfected setts) = 1.35, 95% CI 1.04-1.77, p=0.025) and CGT models (HR (infected/ uninfected setts) = 1.30, 95% CI 0.99-1.69, p=0.055).

Figure 10.5.2: Kaplan-Meier curves comparing sizes of search areas for setts housing tuberculosis-infected and uninfected badgers, by ring cull and different implementations of the geographic profile model.

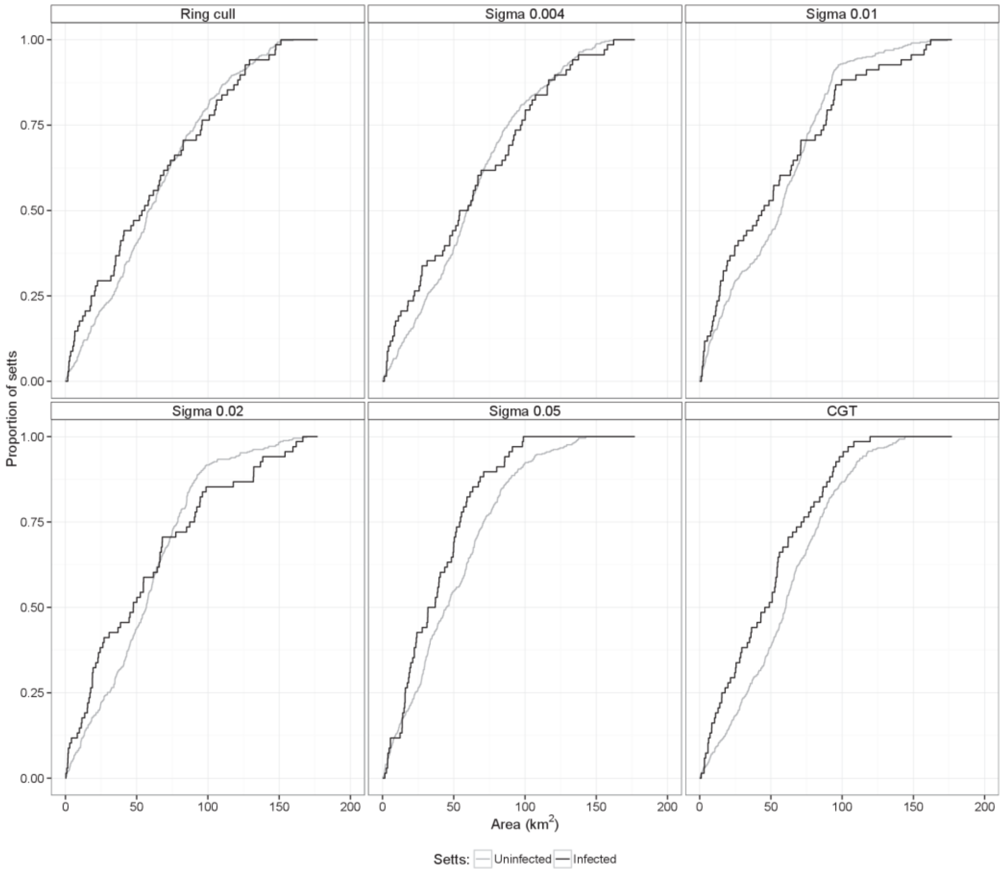


Table 10.5.1: Cox regression spatial survival analysis comparing search strategies with different implementations of the geographic profile model.

Model	Comparison	Multilevel HR*	95% CI	p
σ=0.04	Infected compared to uninfected setts, <70 km^2	1.29	0.92 – 1.80	0.14
	Infected compared to uninfected setts, >=70 km^2	0.63	0.41 – 0.98	0.039
σ=0.1	Infected compared to uninfected setts, >=70 km^2	1.20	0.86 – 1.67	0.29
	Infected compared to uninfected setts, >=70 km^2	0.72	0.46 – 1.13	0.15
σ=0.2	Infected compared to uninfected setts, <70 km^2	1.29	0.94 – 1.77	0.11
	Infected compared to uninfected setts, >=70 km^2	0.58	0.36 – 0.94	0.026
σ=0.5	Infected compared to uninfected setts, all areas	1.35	1.04 – 1.77	0.025
CGT	Infected compared to uninfected setts, all areas	1.30	0.99 – 1.69	0.055

*Multilevel Cox regression model adjusted for random effects of trial area; HR, hazard ratio; CI, confidence interval.

Using the largest σ value of 0.05, there was an apparent slight improvement in the rate of identification of infected setts. However, inspection of the surfaces showed that using this parameter produced only one large cluster in the centre of the area. Its apparent ability to exclude uninfected setts was therefore simply due to excluding peripheral areas, rather than any true spatial discrimination based on clustering. The value of 0.02 used in the primary analysis (Chapter 6) therefore appears to be a reasonable selection for this disease system.

The classic approach to geographic profiling, using the CGT algorithm, was also tested in the sensitivity analysis. Over small search areas, it performed similarly to the DPM model with σ set at 0.02. However, at larger search areas, it behaved similarly to the DPM with σ 0.05 by focusing the highest probability areas in the centre of the surface. Again, this is not likely to have been a true spatial discrimination and there was no evidence that this model could provide an effective means of targeting tuberculosis-infected setts. The advantages of the more mathematically rigorous DPM model, which uses a Bayesian framework, over the classic CGT model, have been demonstrated in previous simulations and case studies.[265] However it is not clear from the results of this study if it is more effective in practice.

APPENDIX 10.6: ETHICAL APPROVAL APPLICATION EXEMPTION LETTER.

From: GradSch.Ethics
Sent: 13 January 2014 15:02
To: Smith, Catherine
Cc: Hayward, Andrew
Subject: EXEMPT - Ethics application form project id 5277/001

Dear Catherine

The Chair of the UCL REC has reviewed your application and is of the view that your study is exempt from the requirement to obtain ethical approval given that no participants are being recruited. Databases are being used and there is no identification of the data subjects.

With best wishes, Helen

Helen Dougal
Research Ethics Co-ordinator
UCL Graduate School
North Cloisters
Wilkins Building
Gower Street
London WC1E 6BT
Tel: 020 7679 7844 (ext 37844)
Email: ethics@ucl.ac.uk

Working hours: Mon-Fri 8am-3.30pm