

University of Southampton Research Repository ePrints Soton

Copyright © and Moral Rights for this thesis are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holders.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given e.g.

AUTHOR (year of submission) "Full thesis title", University of Southampton, name of the University School or Department, PhD Thesis, pagination

UNIVERSITY OF SOUTHAMPTON
FACULTY OF ENGINEERING, SCIENCE & MATHEMATICS
School of Geography

**Regression modelling of cervical cancer and Chlamydia incidence
in the context of national screening programmes**

by

Man Ying Edith Cheng

Thesis for the degree of Doctor of Philosophy

February 2009

DECLARATION OF AUTHORSHIP

I, Man Ying Edith Cheng, declare that the thesis entitled Regression modelling of cervical cancer and Chlamydia incidence in the context of national screening programmes and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- ☐ this work was done wholly or mainly while in candidature for a research degree at this University;
- ☐ where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- ☐ where I have consulted the published work of others, this is always clearly attributed;
- ☐ where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- ☐ I have acknowledged all main sources of help;
- ☐ where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- ☐ none of this work has been published before submission, **or** [delete as appropriate] parts of this work have been published as: [please list references]

Signed:

Date:

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

**FACTLTY OF ENGINEERING, SCIENCE & MATHEMATICS
SCHOOL OF GEOGRAPHY**

Doctor of Philosophy

**REGRESSION MODELLING OF CERVICAL CANCER AND
CHLAMYDIA INCIDENCE IN THE CONTEXT OF
NATIONAL SCREENING PROGRAMMES**

by

Man Ying Edith Cheng

Prevention of cervical cancer development or reduction in undetected Chlamydia incidence and further onward Chlamydia transmission can be achieved through regular screening. Early detection through a regular screening programme is essential to achieve this goal. A well established screening policy is needed to improve screening efficiency.

This PhD study demonstrated the use of mathematical and spatial modelling to explore the risk factors through various regression models, to explore the relation between socio-economic conditions and disease incidence, and also other techniques including classification analysis, decision models, and simulation to evaluate screening options. Based on the risk factors and risk grouping, different groups may have different screening policies. Alternatively, geographical differences can be taken into account by dividing areas into a few parts; the population living in each part may be considered to have different risks of developing cervical cancer or Chlamydia in their life time. Therefore, different screening programmes and services could be provided to those populations according their location or the risk groups which they belong to.

Contents

Abstract	i
Contents	ii
List of Figures, List of Tables	viii
Acknowledgments.....	xiii
Principal Acronyms	xiv
Chapter 1 Introduction	1
1.1 Introduction	1
1.2 Research plan	2
1.3 Research objectives	4
1.4 Cervical cancer	5
1.4.1 NHS cervical cancer screening programme in the UK	7
1.4.2 Diagnostic tests within the screening programme in the UK	7
1.4.3 Treatments	10
1.5 Chlamydia	11
1.5.1 National Chlamydia screening programme (NCSP)	12
1.5.2 Chlamydia screening tests	12
1.5.3 Chlamydia treatments	13
1.6 Summary	13
Chapter 2 Methods	16
2.1 Introduction	16
2.2 Definition of variables	16
2.2.1 Incidence rate	17
2.2.2 Direct standardised expected cases	18
2.2.3 Indirectly standardised incidence ratio (SIR)	19
2.2.4 Deprivation	19
2.2.4.1 Townsend index Z	21

2.2.5 Disease mapping	22
2.3 Regression modelling	23
2.3.1 Regression models and their use	23
2.3.2 Generalised linear regression model (GLM)	23
2.3.3 Logistic and Binomial regression model	24
2.3.4 Poisson regression model	25
2.3.5 Bayesian hierarchical models	25
2.3.5.1 Non-spatial model	26
2.3.5.2 Spatial model	27
2.3.5.3 Bayesian model measurement	28
2.3.6 Geographically weighted regression (GWR) modelling	29
2.3.6.1 Model structure	29
2.3.6.2 Weighted Function: Kernel	31
2.3.6.3 Model measurement	32
2.3.7 Multilevel regression models	34
2.3.7.1 Model structure	35
2.3.7.2 Parameters estimation process	37
2.3.7.3 Model comparison	37
2.3.8 Available software for regression modelling	37
2.4 Classification and Regression Trees (CART) analysis	38
2.4.1 CART algorithm	39
2.5 Decision theory	40
2.5.1 Decision tree model structure	40
2.6 Simulation model	41
2.6.1 Markov and Semi-Markov models	42
 Chapter 3: Data	 44
3.1 Introduction	44
3.2 Cervical cancer data	44
3.2.1 Cervical cancer national data 2004	44
3.2.2 Socio-economic national data at district or unitary authority level in 2001	45
3.2.3 Socio-economic data at Census area statistics (CAS) ward data ..	47

	Contents
3.3 Chlamydia data	47
3.3.1 Chlamydia individual clinical data 1999-2001	47
3.3.2 Socio-economic data at OA level from the Census 2001	49
3.3.3 Socio-economic data at Census area statistics CAS ward	49
data ...	
3.4 Data problem	50
3.5 Summary	50
 Chapter 4 Model for early detection of cervical cancer through cervical cancer screening programme	 51
4.1 Introduction	51
4.2 Objectives	52
4.3 Current screening programme	53
4.3.1 Potential problems	53
4.4 Disease states	54
4.5 Data analysis	55
4.5.1 CART analysis and Colposcopy risk groupings	55
4.5.2 CART analysis with Port	56
4.5.3 CART results	57
4.6 Decision tree model	61
4.6.1 Model for evaluating cervical cancer screening options	63
4.6.2 Decision tree results	65
4.7 Simulation model	68
4.7.1 Simulation results	70
4.8 Summary	71
 Chapter 5 Cervical cancer regression study	 72
5.1 Introduction	72
5.2 Data	73
5.3 Analysis	77
5.3.1 Exploratory analysis	77
5.3.2 Townsend index	80
5.4 Generalised linear regression model	81

5.4.1 Truncated missing data	81
5.4.2 Methods	81
5.4.3 Generalised linear modelling	82
5.5 Bayesian regression model	92
5.5.1 Truncated missing data	92
5.5.2 Methods	93
5.5.3 Model definition	93
5.5.4 Model assumptions	94
5.5.5 Non-spatial model	94
5.5.6 Spatial model	95
5.5.7 Analysis and results	98
5.6 Geographically weighted Poisson regression model (GWPR)	108
5.6.1 Truncated missing data	110
5.6.2 Prediction mean	110
5.6.3 Prediction variance	110
5.6.4 Analysis and results	111
5.6.5 Variables definition	111
5.6.6 Global Poisson regression model	112
5.6.7 GWPR results	112
5.7 What does this mean for the design of screening programmes?	118
5.7.1 CART analysis based on regression results	119
5.7.2 Risk grouping	124
5.7.2.1 CART based on the Townsend index score	124
5.8 Summary	126
 Chapter 6 Chlamydia study	 128
6.1 Introduction.....	128
6.2 Data	129
6.3 Age	129
6.3.1 Age category	129
6.3.2 Age distribution	130
6.4 Decision tree model	134
6.4.1 Decision tree structure	137
6.5 Simulation	139

6.5.1 Simulation model structure	139
6.5.2 Simulation results	143
6.6 Summary	144
 Chapter 7 Chlamydia regression models	 145
7.1 Introduction	145
7.2 Data	147
7.3 Exploratory analysis	152
7.3.1 Age	152
7.3.2 Ethnic origin	152
7.3.3 Location	153
7.3.4 Townsend index	156
7.4 Generalised regression model	156
7.4.1 Generalised linear regression model at Output Area level	156
7.4.2 Generalised linear regression model at CAS Ward level	158
7.5 Multilevel regression model	161
7.5.1 Multilevel regression model structure	161
7.5.2 Model	163
7.5.3 Analysis	164
7.6 Geographically weighted regression model	167
7.6.1 Output Area level	168
7.6.2 CAS Ward level	171
7.7 What does this mean for the design of screening programmes?	174
7.7.1 CART analysis	175
7.7.2 Risk grouping	176
7.7.2.1 CART based on the observed Chlamydia data	177
7.7.2.2 CART based on the Townsend index score	180
7.8 Summary	182
 Chapter 8 Discussion	 184
8.1 Introduction	184
8.2 Cervical cancer	185
8.3 Chlamydia	190

8.4 Summary	192
Chapter 9 Conclusion	194
9.1 Introduction	194
9.2 Cervical cancer	194
9.3 Chlamydia	196
9.4 Summary	197
 Appendices	
Appendix A – Medical Terms Explanation	199
Appendix B – Cervical Cancer National Screening Guidelines	201
Appendix C – Model for Evaluating Cervical Cancer Screening Options ...	202
Appendix D – Cervical Cancer Simulation Model	208
Appendix E – Cervical cancer GWPR results	210
Appendix F - Chlamydia Decision Tree Model for Evaluating Screening Options	213
Appendix G – Chlamydia Simulation Model	215
 Presentations	217
 References	218

List of Figures

1.1 The female cervix	5
1.2 Smear procedure	8
1.3 Colposcope and video colposcope	9
1.4 The accessory tools for colposcopy	9
1.5 Video colposcopy image	10
1.6 Chlamydia bacteria at 0.15 microns	11
2.1 Regression point i and data point j in GWR model	31
2.2 CART analysis of cardiac patients	38
2.3 Decision tree for folic acid supplementation decision	41
3.1 Linking the patient's details, results and explanatory data together	48
4.1 Problems associated with unnecessary colposcopies	54
4.2 CART tree and number of risk groups for coploscopy results	57
4.3 The decision tree for the national cervical cancer screening programme	62
4.4 Schematic diagrams showing the basic calculation within the decision tree model	65
4.5 Cervical cancer disease model	69
5.1 Socio-economic variables maps	76
5.2 Incidence plotted against the expected cases for England 2004	77
5.3 Matrix plot showed the relation between variables	78
5.4 Box-plot of cervical cancer incidence and mortality per PHO England 2004	79
5.5 Townsend index, England 2001	80
5.6 Observed cervical cancer cases and SIR per region	83
5.7 Map of final model with Townsend index score	86
5.8 Map of final model with proportion of female married population	87
5.9 Map of final model with proportion of female single population	88
5.10 Map of final model with proportion of household with lone parent	89
5.11 Map of final model with proportion of household with female lone parent	90

5.12 Map of final model with proportion of low social grade population	91
5.13 The link between variables and parameters in the non-spatial model ...	95
5.14 The link between variables and parameters in the BYM model	97
5.15 The link between variables and parameters in the MIX model	98
5.16 Map of final non-spatial Bayesian Poisson models	104
5.17 Map of final BYM Bayesian Poisson models	106
5.18 Map of final MIX Bayesian Poisson models	107
5.19 Map of final GWPR models	114
5.20 Map of estimated parameters from GWPR models	115
5.21 Kernel bandwidth and the corresponding AICc values	117
5.22 Pre-cervical cancer and cervical cancer disease process	120
5.23 Decision three model	121
5.24 Decision three model based on the observed national cervical cancer data	122
5.25 Decision tree with the best split from CART	123
5.26 Decision tree model with two risk groups	125
5.27 Risk grouping by the Townsend index score	126
6.1 Q-Q plot, to compare the goodness of fit between the observed risk and expected risk with two distributions	132
6.2 The cumulative distribution function (CDF) curve of observed and expected distribution	132
6.3 The observed and expected probability per age classes	133
6.4 <i>Logit</i> of the expected risk per age group	134
6.5 Chlamydia disease states	135
6.6 Simple version of Chlamydia disease states	136
6.7 Chlamydia decision tree model	137
6.8 Transition time and probability from state i to state j	139
6.9 Chlamydia disease system	141
6.10 Chlamydia simulation model	142
7.1 Study area at CAS ward level 2001	146
7.2 STD clinics in the Portsmouth area	147
7.3 Socio-economic variables map at OA level	150
7.4 Socio-economic variables map at CAS ward level	151
7.5 Probabilities of positive test results plotted against age	152
7.6 Positive rate per postcode sector	155

7.7 Positive rate per CAS ward	155
7.8 Positive rate per local authorities	155
7.9 Map of final GLM model at Output Areas levels	158
7.10 Map of final GLM models at CAS Ward level	160
7.11 Unit diagrams	162
7.12 <i>Logit</i> of risk per groups	164
7.13 Map of final GWPR models at Output Area level	169
7.14 Map of estimated parameters at Output Area level	170
7.15 Map of final GWPR models at CAS Ward level	173
7.16 Chlamydia disease process	177
7.17 Decision tree model	178
7.18 Decision tree model based on the Chlamydia data	178
7.19 Potential decision tree	179
7.20 Decision tree model with two risk groups	181
7.21 Risk grouping by the Townsend index score	182

List of Tables

1.1 The research plan	3
1.2 Risk types of HPV	6
1.3 NHS cervical cancer screening frequency	7
2.1 Summary of variables in cervical cancer and Chlamydia studies	17
2.2 Deprivation indexes and indicators	21
3.1 Public Health Observatories (PHO) in England	45
3.2 Summary of explanatory variables used as indicators in the regression analysis	46
3.3 Chlamydia study summary at unitary authority level	49
4.1 CART tree nodes summary for colposcopy results	59
4.2 Probabilities of vary risk groups	66
4.3 Decision tree results for different screening options	67
4.4 Summary of simulation results	70
5.1 Correlations between incidence and socio-economic variables	78
5.2 Summary of GLM	84
5.3 Summary of non-spatial Bayesian Poisson regression results	99
5.4 Summary of BYM CAR Bayesian Poisson regression results	100
5.5 Summary of MIX CAR Bayesian Poisson regression result	101
5.6 Summary statistics of GWPR model comparisons	116
5.7 Test for indicating non-stationary variables	117
5.8 Decision tree outcome	122
5.9 Potential decision tree results	126
6.1 Kolmogrov-Smirnov test results from both Normal and Log-Normal distributions	131
6.2 Examples of screening options for Chlamydia	138
6.3 Simulation results from a simple Chlamydia simulation model	143
7.1 Positive cases by ethnic origin group	153
7.2 Positive rates by CAS Ward levels	154
7.3 Summary of intercept and coefficient value of GLM models at Output	

Area level	157
7.4 Summary of final GLM models at Output Area level	157
7.5 Summaries of the GLM models at CAS Ward level	159
7.6 The intercept and coefficient value of GLM models at CAS Ward level .	159
7.7 Summaries of multilevel regression model structure	162
7.8 Data used in the multilevel regression model	162
7.9 Global regression models with age categories	165
7.10 Global regression models measurement with age categories	165
7.11 Global regression models with <i>Logit</i> of risk per age	165
7.12 Global regression models measurement with <i>Logit</i> of risk per age	166
7.13 CAR regression models	166
7.14 CAR regression models measurement	166
7.15 GWPR results at Output Areas levels	168
7.16 GWPR results at CAS Ward levels	172
7.17 Test for indicating stationary variables	174
7.18 Decision tree model outcome	179
7.19 Potential decision tree	181

ACKNOWLEDGEMENTS

Many thanks and appreciation is due to my supervisors Dr Arjan Shahani and Prof. Peter Atkinson for offering me the chance to study this PhD, passing on their knowledge, and giving me their invaluable guidance, ideas, inspiration to the project, help, motivation and support in every aspect throughout the whole research period. This research would not have been possible without the contribution from both of my respected academic supervisors. With the oversight of my supervisors, editorial advice has been sought. No changes of intellectual content were made as a result of this advice.

I am grateful to the sincere support from Dr Harindra, St Mary hospital, Portsmouth, who provided the individual Chlamydia and colposcopy data and Mrs Susan Walrond from the North East Public Health Observatory, who provided the national cervical cancer data: because of their kindness and contribution, it allows this PhD to become real and possible to carry out.

I would like to give thanks to the GeoData institute for their GIS support, it made the GIS work go very smoothly. The practical and great advice from Prof David Martin and Prof Graham Moon throughout the study period is greatly appreciated. The maps within this work are based on data provided through EDINA UKBORDERS with the support of the ESRC and JISC and uses boundary material which is copyright of the Crown.

Finally, my unlimited and continued love goes to my brothers, sister in laws and my three lovely nephews and niece and friends for their support and motivation. Particularly, I would like to dedicate my PhD to my wonderful parents, for their continued love and brought me into this world, giving me the opportunity to study abroad. Thanks to my Lord Jesus Christ who guides me throughout my PhD study and my life.

I hope that this PhD research can benefit the general female population, and in particular cervical cancer and Chlamydia patients.

Principal acronyms

<i>AIC</i>	Akaike information criterion
<i>APHO</i>	Association of Public Health Observatories
<i>BIC</i>	Bayesian information criterion
<i>BYM</i>	Besag, York and Mollie's
<i>CAR</i>	Conditionally autoregressive
<i>CART</i>	Classification and regression trees analysis
<i>CAS ward</i>	Census area statistics ward
<i>CDF</i>	Cumulative distribution function
<i>CIN</i>	Cervical intraepithelial neoplasia
<i>DIC</i>	Deviance information criterion
<i>DOH</i>	Department of Health
<i>GLM</i>	Generalised linear regression model
<i>GUM</i>	Genito-Urinary Medicine
<i>GWR</i>	Geographically weighted regression model
<i>HPA</i>	Health protection agency
<i>HPV</i>	Human papilloma virus
<i>LBC</i>	Liquid based cytology
<i>MIX</i>	Bayesian mixed model
<i>NCSP</i>	National Chlamydia screening programme
<i>NHS</i>	National Health Service
<i>OA</i>	Output Area
<i>OS</i>	Original Squamous
<i>OSE</i>	Original Squamous Epithelium
<i>pD</i>	penalty value
<i>PHO</i>	Public Health Observatories
<i>PID</i>	pelvic inflammation disease
<i>SIR</i>	Standardised incidence ratio
<i>SMR</i>	Standardised mortality ratio
<i>STD/STI</i>	Sexually transmitted infection disease/ Sexually transmitted infection
<i>WHO</i>	World Health Organization

Chapter 1 Introduction

1.1 Introduction

This PhD study concentrates on two diseases of interest, cervical cancer and Chlamydia. This research investigates geographical variation in the incidence of cervical cancer and Chlamydia in England and the links between these and deprivation, social grade status and family structure (i.e. factors which may reflect sexual behaviour). This study is mainly concerned with the use of regression models to identify the risk factors relating to development of the diseases and any spatial variation in disease patterns that may be related to these risk factors. A number of mathematical methods were applied to understand the disease processes and to identify the associated risk factors concerning their development. Various regression models were used as the main method to explore the relationships between observed incidence cases and deprivation social grade status and family structure variables and other methods including: (i) classification analysis was used to partition the population into different risk groups according to the probabilities of developing disease, and from this, risk groups were created, (ii) a decision tree model was used to evaluate the screening options, and (iii) a simulation model was used to study the natural history of disease processes. Disease mapping techniques were used to display disease incidence patterns, which allowed for visualization of these patterns. The regression models can increase current disease knowledge and potentially can be used to increase screening efficiency by suggesting changes from the current fixed national policies to more adaptive regional policies. Thus, different risk groups or different regions may be allocated different screening policies, such as screening tests and test intervals, based on need. The data that are required for these analyses and modelling are described in detail in Chapters 2 and 3.

The reasons for having such a flexible screening programme are to increase efficiency in managing resources; to increase the probability of detecting abnormal cells at pre-cancer states (and to reduce the number of undiagnosed Chlamydia cases) by targeting high risk groups, encouraging them to take screening tests at potentially more frequent regular periods (e.g. every six

months). Simulation models are used to study the natural history of disease processes. The models provide more knowledge of the natural pre-cervical cancer processes (and Chlamydia and related infertility, caused by untreated Chlamydia cases). It is possible to estimate the number of patients in each disease state and the length of period of stay in each disease state through simulation models. Ultimately, based on analysis of the data collected, the population will be divided into different risk groups, which can each be allocated different screening frequencies and specific screening tests according to their needs from the results obtained from the analysis. Finally, disease mapping techniques were used to display disease incidence patterns and also patterns in the estimated model parameters, allowing researchers to examine any spatial variation over space.

1.2 Research plan

Table 1.1 shows the summary of the research plan, the required data for each unit of analysis, and model. This thesis is organised into nine chapters with the following structure:

Chapter 1: General introduction,

Chapter 2: The methods that were used in this research study,

Chapter 3: Data section: A description of all the necessary data in detail,

Chapter 4: Explanatory analysis of cervical cancer,

Chapter 5: Regression models for cervical cancer,

Chapter 6: Explanatory analysis of Chlamydia,

Chapter 7: Regression models for Chlamydia,

Chapter 8: Discussion

Chapter 9: Conclusion

Table 1.1 The research plan.

	CERVICAL CANCER	CHLAMYDIA
NATIONAL DATA (REGRESSION)	<p>1. Cervical cancer national regression model (Chapter 5)</p> <p>Data 1: Data acquired on 354 Districts or unitary authorities from Association of Public Health Observatories (APHOs).</p> <p>Data 2: Deprivation, social grade status and family structure variables from census data acquired at District or unitary authorities levels 2001.</p> <p>EDA: Townsend index etc.</p> <p>Model 1: Generalised linear regression (GLM) model for incidence counts and standardised incidence ratio (SIR).</p> <p>Model 2: Bayesian hierarchical model for incidence rates.</p> <p>Model 3: Geographically weighted regression (GWR) model for incidence counts and SIR.</p>	
LOCAL DATA (REGRESSION)		<p>3.Chlamydia local regression models (Chapter 7)</p> <p>Data 4: Individual Chlamydia data acquired from Portsmouth St Mary Hospital (Postcode information is available).</p> <p>Data 5: Output Area and Census area statistics (CAS) Ward deprivation, social grade status and family structure variables from the Census data from Portsmouth, plus postcode headcounts to allow redistribution to postcodes.</p> <p>EDA: Townsend Index.</p> <p>Model 1: GLM Regression.</p> <p>Model 2: Multilevel regression model..</p> <p>Model 3: GWR.</p>
LOCAL (ANALYSIS OF SCREENING PROGRAMME)	<p>2. Model for early detection of cervical cancer through cervical cancer screening programme, (Chapter 4)</p> <p>Data 3: Individual Colposcopy data 1998 – 2006.</p> <p>Model 4 Classification and Regression Trees (CART) analysis (to split patients from Data 3 into several risk groups).</p> <p>Model 5: Decision tree and simulation model which can be used to evaluate different options for cervical cancer screening. These models describe the natural history of the disease process, and can evaluate various intervention options, including screening, for cervical cancer.</p>	<p>4. Models for early detection and reduction of the number of undetected Chlamydial infections, (Chapter 6)</p> <p>Data 4: Individual Chlamydial infection data (e.g. patients' age)</p> <p>Model 5.Explanatory analysis</p> <p>Model 6: Decision tree and simulation models for Chlamydial infection. This model describes the natural history of the infection. This model will evaluate various intervention options, including screening, for Chlamydia.</p>

1.3 Research objectives

There are five specific research objectives: (i) to study cervical cancer at the national level and Chlamydia at local levels, (ii) to undertake geographical mapping of cervical cancer across England and Chlamydia incidence patterns in Portsmouth at various spatial levels (e.g. Output Areas level and Ward level), to examine the spatial patterns, (iii) to determine the relationships between disease incidence and a range of deprivation indicators, social grade and family structure factors through various regression models. Such factors that have relationships with disease incidence may be considered as high risk factors or associated risk factors as identified through regression models. It is important to understand that associated factors do not directly cause the development of cervical cancer or Chlamydia, but are associated with the development of these diseases, (iv) to examine any spatial variation that may exist, and finally, identify those risk factors that would help to inform planning, and target national and local screening.

The aims of this research were to understand the key aspects of pre-cervical cancer processes and Chlamydia infection through modelling, with the desired outcome of promoting the understanding of, and justification for, targeted screening policies and intervention events (for example, sex education for the young sexually active population). An additional aim was to increase the chances of preventing cancer development and detecting early pre-cancer and/or cancer cases through national cervical cancer screening programmes. In the case of Chlamydia, it was aimed to reduce the number of undetected asymptomatic cases, which in turn would help to reduce further, or onward, transmission by the infected patients and further complications for the patients. When the cervical cancer and Chlamydia cases can be diagnosed early enough, the appropriate treatments can be provided to patients at an early disease stage, which can increase the chance of recovering from cervical cancer and or Chlamydia. Thus, an effective and efficient screening programme is needed.

1.4 Cervical cancer

Cancer can happen to anyone at any time in their life; it is a common cause of death worldwide. The World Health Organisation (WHO, 2008a) reported that approximately 13% of deaths worldwide in 2007 were caused by cancer. Cancer can be a long term disease caused by abnormal and uncontrollable cells; a tumour arises from the abnormal and uncontrolled cells (Cooper, 1993). The process of a cancerous tumour developing takes a long time; therefore, it is often possible to detect any abnormal cells at the early disease stages and even treat and/or remove them before a tumour arises. Once the cancerous tumour forms the cells may invade and destroy the surrounding tissues, and in time, also begin to spread to other parts of the body through the bloodstream or the lymphatic channels (Martin, 2000). Details of medical terms can be found in Appendix A.

For women, the most common cancers worldwide are (i) breast, (ii) lung, (iii) stomach, (iv) colorectal and (v) cervical cancer (WHO, 2008a). Cervical cancer occurs in the cervix, or neck of the womb (Figure 1.1). Some research has shown evidence of an association between Human Papilloma Virus (HPV) and cervical cancer development (Singer and Monaghan, 2000). In particular, HPV 16 and 18 are highly related to cervical cancer development (Jenkins *et al.*, 1996; Arias-Pulido *et al.*, 2006). Nearly 100% of cervical cancer cases presented with various HPV; and 70% of the cases were associated with HPV 16 and 18. HPV 16 and 18 were considered as the high risk HPV (WHO, 2007), some examples of high risk HPV are listed in Table 1.2 (Moore-Higgs *et al.*, 2000; Patnick, 2008).

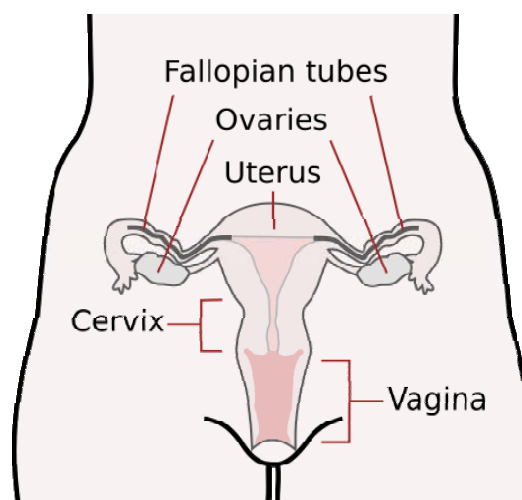


Figure 1.1 The female cervix (Wikimedia, 2007)

Table 1.2 Risk types of HPV (Moore-Higgs *et al.*, 2000)

Risk level	HPV
Low risk	6, 11
Medium risk	31, 33, 35
High risk	16, 18, 31, 45, 56

The chance of developing cervical cancer is also closely related to the patient's sexual behaviour, reproductive history and family history. In addition, 70% of all cancer deaths in 2007 occurred in low and middle-income countries (WHO, 2008b). Deaths from cancer worldwide are projected to continue rising, with an estimated 11.4 million deaths predicted for 2030 (WHO, 2008b). On the other hand cancer is commonly preventable and curable (but not always) at early disease states, and the WHO reported that over 40% of all cancers can be prevented (WHO, 2008b).

Three main elements of early detection of cervical cancer are (i) regular screening, (ii) the types of diagnostic tests used in screening programmes, and (iii) cervical cancer vaccination. The National Health Service (NHS) reported that there is evidence of a decreasing number of cervical cancer cases since the national screening programme became available in the 1960s (Patnick, 2004; 2008). Some studies showed that the use of the HPV test can increase the chance of detecting pre-cancer patients (Jenkins *et al.*, 1996), and other studies have shown that the combination of Liquid Based Cytology (LBC) and HPV tests may increase the detection rates, but may cause a higher chance of the inappropriate use of colposcopies (Sherlaw-Johnson and Philips, 2004). In addition, the length of screening interval does affect the detection rate (Sherlaw-Johnson *et al.*, 1997). Thus, consideration of which diagnostic test should be used in screening programmes may affect the cost-effectiveness of the screening programme. Vaccination is 100% effective in protecting a woman from getting infected by HPV 16 and 18. These HPVs are assumed to be highly related to the development of cervical cancer. However, it is estimated that 70% of cervical cancer incidence was associated with HPV 16 and 18. The other 30% incidence was associated with other types of HPV, which are not prevented by the vaccination, so screening is still needed to prevent the other 30% of cases (WHO, 2007). In practice, it is best to vaccinate the female population between the ages of nine to twenty-six

and/or those members of the population who have not yet been affected by HPV 16 and HPV 18. The cost of the full vaccination is approximately £500 per person (Cancer Research UK, 2008). A UK based study summarised that the HPV 16/18 vaccine led to an estimated reduction of 66% of the prevalence of high grade pre-cancerous lesions and a 76% reduction in cervical cancer deaths (Kohli *et al.*, 2007). In addition, the vaccines are the most cost effective ways of preventing cervical cancer, and increasing the chances of surviving from cervical cancer, and constitute the best use of limited resources (Ferko *et al.*, 2007; Hammerschmidt *et al.*, 2007).

1.4.1 NHS cervical cancer screening programme in the UK

In the UK context, cervical screening began in Britain in the mid-1960s. The NHS Cervical Screening Programme in the UK was set up in 1988 when the Department of Health (DOH) instructed all health authorities to introduce computerised call-recall systems. The screening programme was, and still is, available free-of-charge to women in the UK between the ages of 25-64 to attend once every three to five years (Table 1.3). The National Coordination Office was set up in 1994, based in Sheffield. This office is responsible mainly for the improvement of the overall performance of the programme (Patnick, 2004).

Table 1.3 NHS cervical cancer screening frequency

Age group (years)	Frequency of screening
25	First invitation
25-49	Three yearly
50-64	Five yearly

1.4.2 Diagnostic tests within the screening programme in the UK

The principal tool for detection in the Cervical Screening Programme is the smear test. Smear tests and LBC examine the cells of the cervix to identify abnormalities and any changes within the cells. The first stage of the screening is either the smear test or LBC.

(i) The smear test is a process of taking a sample of cells from the cervix for analysis. A speculum (an instrument, see Figure 1.2 and 1.4) is used to open the woman's vagina and a spatula is used to sweep around the cervix; then the sample is smeared on a slide. Therefore the test is called "smear" or "sweep" test. Finally, the sample is sent to the laboratory for examination.

(ii) The LBC is more accurate than the smear test and it is available in the screening programme. Other similar tests are available and are also applied in different societies and/or organizations. The LBC is very similar to the smear test; the only difference is that the head of the spatula is broken off into a small glass vial containing preservative fluid or rinsed directly into the preservative fluid rather than smearing the sample on a slide.

(iii) The test for HPV is used to examine whether HPV DNA is present or not, and NHS staff have been trained to use the new HPV technology and techniques in NHS cervical cancer screening (Patnick, 2008) and also other countries.

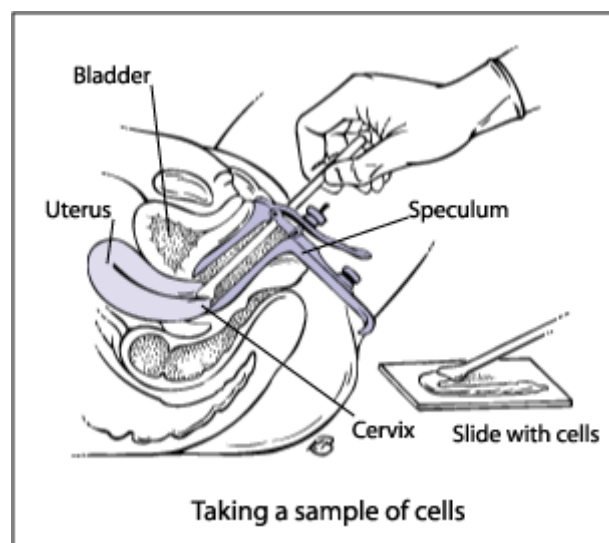
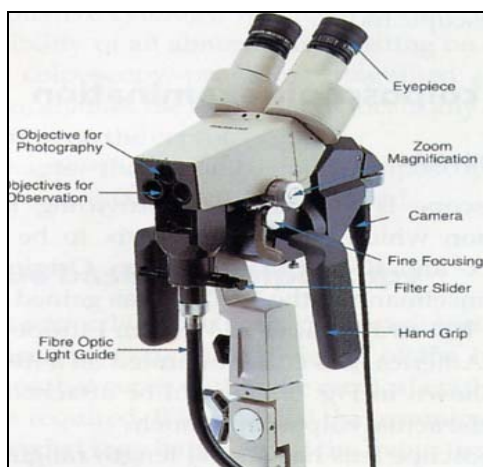


Figure 1.2 Smear procedure (American Academy of Family Physicians, 2008)

The second stage of the NHS cervical cancer screening programme uses the Colposcopy, which is a more advanced tool used for the diagnosis of cervical pre-cancer, to determine the cause of abnormalities found in smear test results. A colposcope is used in this process; this tool (Figure 1.3a) is a microscope providing illuminated magnification, which allows the viewing of the cervix at 6-fold to 40-fold magnification. It was invented by Hans Hinselmann in 1925

(Singer and Monaghan, 2000). There is also a video colposcope available (Figure 1.3b), which utilises a camera and a colposcope with an electronic green filter, motorised zoom magnification and fine focus controls, combined within a single unit. The Colposcopy accessories are shown in Figures 1.4 and the process can be found from Singer and Monaghan (2000). The samples are sent to laboratories for full examination including cervical biopsy. If patients have been diagnosed with cervical precancerous lesions (CIN 1, 2 and 3), treatments will be given to remove these (Figure 1.5).

(a)



(b)



Figure 1.3 (a) colposcope and (b) video colposcope (Singer and Monaghan, 2000).



Figure 1.4 The accessory tools for colposcopy: 1. pots with solution (acetic acid, saline, and Lugol's iodine), 2. vaginal speculum, 3. sponge holding forceps, 4. Desjardin's endocervical forceps, 5. three-pronged probe for retraction, 6. cotton tipped fine swab sticks, 7. Aylesbury cytology spatula, 8. larger cotton tipped

swab sticks, 9. local anaesthetic syringe, 10. silver nitrate sticks for hemostasis, 11. endocervical brushes, 12. antibiotic cream, 13. cotton swabs, 14. diathermy electrodes for diagnosis or treatment, 15. Monsell's solution, 16. local anaesthetic ampules, 17. Eppendorfer cervical biopsy forceps (Singer and Monagham, 2000).

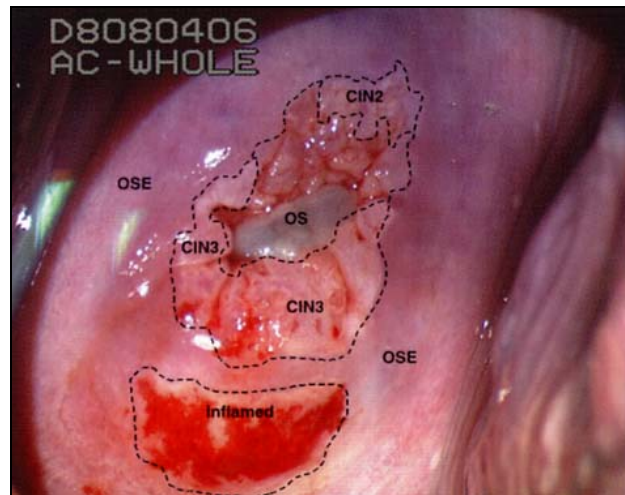


Figure 1.5 This is a video colposcopy image, which can be stored with a patients' history. It allows for the capture and measuring of the CINs from the image. It is clear this patient had moderate Cervical Intraepithelial Neoplasia (CIN2), severe Cervical Intraepithelial Neoplasia (CIN3), Original Squamous (OS) and Original Squamous Epithelium (OSE); therefore treatment can be given based on such an image (Singer and Monagham, 2000).

1.4.3 Treatments

Once cervical cancer is diagnosed, treatment is given by the NHS. Cervical cancer treatments include (i) surgery, (ii) radiation therapy, (iii) chemotherapy, (iv) hormone therapy, and/or (v) combination of therapies. The most suitable treatment might vary between patients; it really depends on the disease state and the patient's response to the treatment. If cancer cells have not yet spread to other parts of the body, it is possible to apply surgery to remove the tumour; otherwise another therapy or combination of therapies can be adopted.

1.5 Chlamydia

Chlamydia trachomatis is a bacteria (Figure 1.6), which is the most common sexually transmitted infection (STI) disease in England (Primarolo, 2006). The national positive rate is estimated at about 10% (Primarolo, 2006). It can be treated easily by antibiotics. The rate of new diagnoses of cases of Chlamydia in those who attended Genito-Urinary Medicine (GUM) clinics greatly increased from 116 to 175 per 100000 patients between 2000 and 2004 (National Statistics, 2006), and rose by 5% (in terms of cases) between 2004 and 2005 (NHS, 2008a). However, over 70% (Health Protection Agency, 2007) of female infected patients and 50% (Health Protection agency, 2007) of male infected patients are asymptomatic at the early period. If Chlamydia remains undetected long enough, it can lead to complications, such as increasing the risk of developing pelvic inflammation disease (PID) for women, causing ectopic pregnancy, and even infertility. While the condition remains undetected, the patient is at risk, and approximately 10-40% of infected, untreated women develop Pelvic Inflammation Disease (PID) (Health Protection Agency, 2007). In England, 75% of Chlamydia cases are found in the young population between the ages of 16 to 24, and only 25% of cases are present in those over 25 (Health Protection Agency, 2006). Details of medical terms can be found in Appendix A.

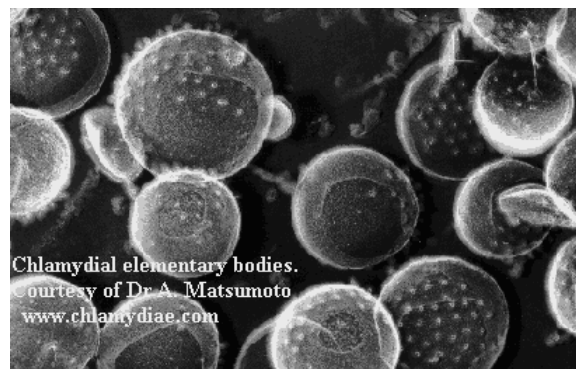


Figure 1.6 Chlamydia bacteria at 0.15 microns (www.chlamydiae.com).

For female patients who develop Chlamydia during pregnancy, the patient should seek treatment; otherwise, the mother can transmit Chlamydia to the baby through the birth process. Once the baby has been infected by the mother, the baby remains asymptomatic for a certain period (about 40 days), and this can cause great damage to the baby's health (NHS, 2007).

Therefore, the earlier the disease is detected, the easier it is to treat it by using antibiotics. During the treatment period, all sexual partners should be treated at the same time to avoid any re-infection or spread of Chlamydia.

1.5.1 National Chlamydia screening programme (NCSP)

The national Chlamydia screening programme was introduced in England in 2003; its objective is to prevent asymptomatic infection cases remaining undetected. Optimally, it can reduce onward transmission to sexual partners and avoid further complications or damage. Thus, the national positive incidence rate can be reduced. Since most cases are found in the young age group under 25 (i.e. 16-24), the programme targets those young populations as the highest risk group (NHS, 2007).

The aim of the NCSP is to control and prevent undetected cases across England through screening, and it provides access to services, screening, and treatment (NHS, 2007; Low *et al.*, 2006).

1.5.2 Chlamydia screening tests

The Chlamydia test is free of charge in England; patients only need to provide a urine sample. For female patients, a swab test is also available, but the test result only shows whether Chlamydia is positive or negative; it does not indicate any other positive STIs. However, the test site or centres can provide information about other STI tests and necessary help for patients. The Chlamydia test is available from most of the NHS local GUM and other sexual health services. For example, in Southampton, the GUM department, located in the Royal South Hants hospital, provides a free Chlamydia test. Further information about local Chlamydia test centres and treatments can be found at the following website: <http://www.chlamydiaSCREENING.nhs.uk>.

1.5.3 Chlamydia treatments

Treatments are needed for patients testing positive for Chlamydia. If a patient has positive Chlamydia only, it can be treated by antibiotics in the form of a single dose of Azithromycin, and doxycycline twice a day for seven days in total (NHS, 2008b). If any patients have PID, it can be treated by antibiotics; for example Amoxicillin, Streptomycin and Erythromycin, for 14 days (NHS, 2008b; c). Treatments are very straight forward, simple and effective.

1.6 Summary

Large proportions (70%) of cancer deaths occur in the low and middle-income countries; thus it is possible to infer a link between cervical cancer incidence and deprivation conditions, social grade status and family structure. This research demonstrates the use of various statistical techniques to explore the relation between the incidence of cervical cancer and deprivation condition, social grade status and family structure variables in the UK.

The majority of Chlamydia cases occur in the young population, and it remains asymptomatic at the early stages, until the patients reach further stages. Therefore, it is necessary to ensure that the young population is aware of this. It is interesting to explore the relationships, if any, between the incidence of the disease and prevailing deprivation indicator, social grade status and family structure variables, which would increase the current knowledge about Chlamydia and provide strategies for targeting the high risk populations.

An effective national screening programme might increase the early detection rate for cervical cancer and increase awareness of Chlamydia infection. The quality of human life can increase through screening programmes. However, designing an effective programme requires us to have a better understanding of the disease process, to identify risk factors, associated factors and to examine spatial variation. Combining all the above factors will lead to a greater understanding of the disease and the link with deprivation indicator, social grade status and family structure factors, and provide information on how to improve and increase the efficiency of screening with a given amount of (limited) resources. To summarise:

The overall aims were

- To understand pre-cervical cancer processes and Chlamydia infection through modelling, with the desired outcome of promoting the understanding of and justification for, targeted screening policies and intervention events,
- To increase the chances of detecting the early pre-cervical cancer or cervical cancer cases (and to reduce the number of undetected Chlamydia cases), so that the chances of detecting the diseases at early stages increase.
- Allowing the necessary and appropriate treatments to be provided to the patients at the earliest possible stage, which can increase the chances of recovery from the disease.
- Therefore, a more effective and efficient screening programme is needed in order to improve the quality of screening.
- The overall aims will be investigated using two available datasets.
 - Cervical cancer in England
 - Chlamydia in Portsmouth.

The key objectives

- Study cervical cancer and Chlamydia at national and local levels, respectively.
- Map cervical cancer and Chlamydia incidence patterns in England and Portsmouth, respectively.
- Determine the relationships between incidence and deprivation indicator, social grade and family structure factors.
- Determine the associated factors through regression modelling, which can help to plan and target screening programmes and policies.

There is a CD attached at the back of the thesis, which contain four models listed below:

- (i) Cervical cancer decision tree model, details can be found from Chapter 4 and Appendix C
- (ii) Cervical cancer simulation model, details can be found from Chapter 4 and Appendix D

- (iii) Chlamydia decision tree model, details can be found from Chapter 6 and Appendix F
- (iv) Chlamydia simulation model, details can be found from Chapter 6 and Appendix G

The simulation insert requires generator, which is called “ITVCMath.dll”. Please follow the instructions in Appendix D and G in order to link the generator and the models together.

Chapter 2 Methods

2.1 Introduction

Disease modelling is a large topic with many mathematical methods available. However, the choice of which method to use depends on the main purposes of the study. To achieve the goal of this PhD study, various methods were applied such as (i) regression models, (ii) classification analysis, (iii) decision theory, and (iv) simulation. The details of each method are discussed and explained in this chapter.

Regression models were used to explore the relationship between variables, whereas classification techniques were used to divide a population into groups according to potential risk (e.g. low risk, high risk groups etc.). Both decision theory and simulation were used to understand the natural disease process; how patients flow through (i.e. transfer from one disease state to another) the disease system, and to evaluate the screening options. A summary of the results from the above mentioned methods will help to identify the associated risk factors for cervical cancer development and Chlamydia infection and transmission. Incorporating those associated risk factors into the national screening system may increase the level of prevention of cervical cancer development and reduce the number of undetected asymptomatic Chlamydia cases, and avoid further or onward Chlamydia infection. In this study, regression models are the main method of analysis. Other methods were included as a demonstration of what is possible, but were not applied in detail due to a lack of suitable data.

2.2 Definition of variables

General variables were used for both cervical cancer and Chlamydia as listed in Table 2.1. The total number of study regions is denoted as N , for each $i = 1, 2, \dots, N$. Let e_i be number of expected cases, Y_i the incidence (i.e., number of cervical cancer cases or number of positive Chlamydia cases), D_i the mortality (i.e., number of deaths due to cervical cancer) and S_i the survival rate (i.e., number of

survivors from cervical cancer), where $i = 1, 2, \dots, N$. All variables are listed in Table 2.1 below.

Table 2.1 Summary of variables in cervical cancer and Chlamydia studies

Variables	Cervical cancer	Chlamydia	Region general	Population general
No. of study regions	N_{cc}	N_{ch}	N	
Expected cases	e_{cci}	e_{chi}	e_i	e
Incidence	Y_{cci}	Y_{chi}	Y_i	Y
Mortality	D_{cci}	---	D_i	D
Survival	S_{cci}	---	S_i	S
Region	i	i	i	i

2.2.1 Incidence rate

The full definition of incidence is the number of events (e.g. positive cases) happening within a given time period, for example, between 2003 and 2004 (Waller and Gotway, 2004). Therefore, the incidence rate θ_i measures the proportion of positive cases within a given population. Alternatively, it represents the probability of a person contracting the disease within a specified time period or a given time period, which might be determined by some personal characteristics and/or attributes; e.g. gender, age, family history, occupation and so on. An assumption might be needed to calculate the rate; for example, the at-risk population is assumed stationary and fixed. The calculation is defined below:

$$\theta_i = \frac{Y_i}{p_i} \quad \text{where } i=1, 2, \dots, N \quad (2.1)$$

The overall incidence rate is defined as:

$$\theta = \frac{\sum_{i=1}^N Y_i}{\sum_{i=1}^N p_i} \quad \text{where } i=1, 2, \dots, N \quad (2.2)$$

Mortality and survival rates which measure the probabilities of death and survival from disease follow the same principles. Mortality rate d and survival rate s are defined below:

$$d = \frac{\sum_{i=1}^N D_i}{\sum_{i=1}^N p_i} \quad \text{where } i = 1, 2, \dots, N \quad (2.3)$$

$$s = \frac{\sum_{i=1}^N S_i}{\sum_{i=1}^N Y_i} = \frac{\sum_{i=1}^N (Y_i - D_i)}{\sum_{i=1}^N Y_i} \quad \text{where } i = 1, 2, \dots, N \quad (2.4)$$

when the incidence is large, the number of survivors can also be expected to be large. Therefore, the number of survivors is standardised by incidence to produce a survival rate s . Note that to aid clear interpretation, the word ‘rate’ is used whenever variables are standardised.

2.2.2 Direct standardised expected cases

Some variables can be estimated directly, when the expected cases provide information which can give an idea as to the number of expected cases for a particular location over a particular time period. The standardised expected case e_i is defined below:

$$e_i = \frac{\sum_j Y_j}{\sum_j p_j} p_i \quad \text{where } i = 1, 2, \dots, N \quad (2.5)$$

e_i is the number of expected incidence cases, and p_i represents the total population in region i and j represents the age group. If age information is available from the observed incidence cases Y_i then the age-specific expected cases can be calculated based on the same equation (2.5).

2.2.3 Indirectly standardised incidence ratio (SIR)

In practice, the standardised incidence ratio (SIR) and standardised mortality ratio (SMR) are commonly used to measure and compare regional incidence and death rates. In this study, the property of interest is the incidence rate rather than mortality ratio, and so the standardised incidence rate is used. The regional SIR and SMR are defined below (Waller and Gotway, 2004):

$$SIR_i = \frac{Y_i}{e_i} \quad \text{where } i = 1, 2, \dots, N \quad (2.6)$$

$$SMR_i = \frac{D_i}{e_i} \quad \text{where } i = 1, 2, \dots, N \quad (2.7)$$

where the expected number of cases e_i was defined in equation (2.5). Since SIR and SMR are standardised indicators of incidence and mortality rates, which vary around one, if the rate is above one, the observed incidence is greater than expected; if the rate is less than one, the observed incidence is less than expected (Waller and Gotway, 2004).

2.2.4 Deprivation

It is important to distinguish between deprivation and poverty. Poverty describes the population with low income and lack of material resources; deprivation is defined as a population or society having disadvantage in local society, for example, poor access to social services (e.g. health services) affecting life chances (Rees *et al.*, 2002). Deprivation could vary over space and over time. The Black report (Townsend *et al.*, 1982) discussed that higher income populations commonly make better use of health services and there are significant social inequalities in using local health services. Therefore, deprivation is highly related to ill-health condition (Townsend *et al.*, 1982).

Deprivation can be used to measure local welfare and behaviour which can be useful in health care studies, because some studies have shown that there is a linkage between health condition and welfare condition. Health condition may also be related to human behaviour which is a function of background as well as

educational level. There are many deprivation indices or measurements (Table 2.2) (Rees *et al.*, 2002). Different indices may use different deprivation and different numbers of indicators. The index is commonly measured in different areal units which depend on the purpose of the study and the areal units available for study. Since the deprivation indicators are measured in areal units and the units are likely to vary over space, some of the regions might be relatively larger than others. For this reason the index may be more sensitive to the size of the denominators.

Most of the deprivation indicators are available from the UK census. However, there are still certain limitations in terms of data coverage. For example, all the data are aggregated into certain areal levels (i.e. not available at individual level) and some information is not available (e.g. personal income, environmental conditions). In most healthcare studies, a deprivation index is used to measure deprivation at the regional level, which is used to analyse and examine the relation between health and social grade status, family structure variables and deprivation. A deprivation index may also be useful in assessing public health services, health policy and to target resources at the regional level.

In this study, the Townsend index was chosen to measure deprivation. It is one of the more common choices and it has been well used in health studies. The four indicators are the common choice within most indices. There is a disadvantage of using the Townsend index. It only takes account of socio-economic information, but no family structure information is included. Therefore, in chapters 5 and 7 some other variables were added into the regression to represent family structure and social status.

Table 2.2 Deprivation indexes and indicators.

Deprivation indexes	Deprivation indicators
Townsend	Unemployment, Households with no car, Over crowded housing, Households not owned.
Carstairs	Male unemployment, Households with no car, Over crowded housing, Social classes IV or V.
Jarman UPA	Unemployment, Over crowded housing, Single pensioners, Lone parents, Unskilled manual, Ethnicity, Children under 5, Residential mobility.
DoE81	Unemployment, Over crowded housing, Lacks basic amenities, Single pensioners, Lone parents, Ethnicity.

2.2.4.1 Townsend index Z

The Townsend index is used to measure relative deprivation (Townsend *et al.*, 1988). Deprivation is thought to be strongly related to ill-health (i.e., personal health condition is related to personal behaviour; for example, lifestyle, diet, smoking etc.). Social deprivation is very important to the investigation of small area health studies (Townsend *et al.*, 1982; Townsend *et al.*, 1988; McCullagh and Nelder, 1952). The Townsend index is a combination of four socio-economic indicators, which are:

- i. Percentage of unemployed population,
- ii. Percentage of households with no car or van,
- iii. Percentage of households not owned,
- iv. Percentage of over-crowded housing (over one person sharing one bedroom).

The calculation of the Townsend score for each variable is defined below. Let V_{ih} be the value of socio-economic variables, for variables $h = 1$ to 4 and $i = 1$ to N area units in the data. The Townsend score z_{ih} is a standardised measure for each of the four deprivation variables obtained by subtracting from V_{ih} the mean m_{ih} and dividing by the standard deviation σ_{ih} as below.

$$z_{ih} = \frac{V_{ih} - m_{ih}}{\sigma_{ih}} \quad \text{where } i = 1, 2, \dots, N \text{ and } h = 1, 2, 3, 4 \quad (2.8)$$

Both variables (i) unemployed population and (iv) over-crowded housing were transformed by a natural $\log y = \ln(x+1)$, where y is the value after the transformation and x is the observed value of the socio-economic variables, to make the variables approximately normally distributed.

The Townsend index is calculated from the sum of z_{ih} as follows:

$$Z_i = \sum_{h=1}^4 z_{ih} \quad \text{where } i = 1, 2, \dots, N \text{ and } h = 1, 2, 3, 4 \quad (2.9)$$

The greater the Z value the greater the deprivation.

2.2.5 Disease mapping

Disease mapping can be used in many applications; for example, describing spatial variation and patterns in observed incidence, mortality, survival and risk and helping in the understanding of disease aetiology. It allows for the display of information graphically and can be visual; therefore, the general population can understand the information more easily. Disease mapping has been used widely in disease applications, such as in identifying the relationship between explanatory factors and health condition, and analysing communicable diseases (Elliott *et al.*, 2000). Disease mapping has also been used in the cases of cervical cancer and Chlamydia, where the incidence maps summarise complex geographical variation. Such information can be used for simple descriptive purposes, to assess whether health targets are being met or whether a new set of policies is needed. Furthermore, such maps reveal spatial patterns, which may help to highlight structure in spatial variation.

2.3 Regression modelling

Regression is a well-known statistical tool for exploring relations between target and explanatory variables. Various types of regression techniques are available and are commonly used to model the relations between variables in ecological and disease studies. Examples of global regression models are: generalised linear regression models, Bayesian models for small area studies (Green and Richardson, 2002), multi-level models and local regression models (McCullagh and Nelder, 1952; Fotheringham *et al.*, 2002).

2.3.1 Regression models and their use

In this study, regression modelling was used to explore the relationships between the target variables (cervical cancer incidence and Chlamydia incidence) and the explanatory variables (measures of deprivation). The outcomes from the models provide a summary of the complex relations between variables, and, if spatial information is available and included, then it can highlight the geographical relations that exist.

2.3.2 Generalised linear regression model (GLM)

Generalised linear regression modelling (GLM) is a popular statistical modelling tool for exploring relationships between target and explanatory variables (Gatrell and Bailey, 1996; Elliott *et al.*, 2000). A generalised linear regression model is described below: Let $Y_i = (Y_1, Y_2, \dots, Y_N)$ for $i = 1, \dots, N$ be the independent observed variable, which follows a probability distribution that belongs to the exponential family of probability distributions, with mean $E(Y_i) = \mu_i$. The linear predictor \hat{Y}_i is based on the predictor variables $v_{1i}, v_{2i}, \dots, v_{Ti}$ for variables $t = 1, 2, \dots, T$, which are denoted by:

$$\hat{Y}_i = \beta_0 + \beta_1 v_{1i} + \dots + \beta_T v_{Ti} + \varepsilon_i \quad \text{where } i = 1, \dots, N \quad (2.10)$$

The regression model can be re-written in equation 2.11:

$$\hat{Y}_i = \beta_0 + \sum_{t=1}^T \beta_t v_{ti} + \varepsilon_i \quad \text{where } i = 1, \dots, N \text{ and } t=1, \dots, T \quad (2.11)$$

where β_0 is the intercept, β_i is the coefficient of variable v_{ti} for $i = 1, 2, \dots, N$ and $t = 1, 2, \dots, T$, and ε_i is the error term which is normally distributed with mean zero and variance one.

2.3.3 Logistic and Binomial regression model

Binomial and Logistic regression models are non-linear regression models, within the generalised linear regression family, when the observed variable (i.e. the response variable) has two possible outcomes (i) 0 or 1, or (ii) true or false. Therefore, the outcome can be represented by a binary or logistic indicator variable, and the model can be used to predict the chance of an event happening. It is particularly useful for modelling an individual patient's disease result (outcomes), for which the test result can be either positive or negative.

The observed variable Y_i is a Bernoulli random variable; this follows the Binomial distribution with two parameters (sample size ζ_i and probability π_i):

$$Y_i \sim \text{Binomial}(\zeta_i, \pi_i) \quad (2.12)$$

Y_i is considered as a Bernoulli random variable with the following probability statement:

$$\begin{aligned} P(Y_i = 1) &= \pi_i \\ P(Y_i = 0) &= 1 - \pi_i \end{aligned} \quad (2.13)$$

where $Y_i = 1$ represents true or positive, and $Y_i = 0$ represents false or negative outcomes. The linear regression model can be defined as:

$$\begin{aligned} \hat{Y}_i &= \beta_0 + \beta_1 v_{1i} + \beta_2 v_{2i} + \dots + \beta_T v_{Ti} + \varepsilon_i & \text{where } \hat{Y}_i &= 0, 1 \\ \hat{Y}_i &= \beta_0 + \sum_{t=1}^T \beta_t v_{ti} + \varepsilon_i & \text{where } i &= 1, \dots, N \text{ and } t=1, \dots, T \end{aligned} \quad (2.14)$$

2.3.4 Poisson regression model

Poisson regression is a non-linear regression model, again part of the generalised linear regression family. The outcome is discrete; for example, the number of positive incidence cases and, therefore, it is commonly used to model disease cases. Usually, the outcomes $Y_i = 0, 1, 2, \dots, N$, which are a set of observed counts arising from a Poisson process (i.e., the data Y_1, Y_2, \dots, Y_N in regions 1, 2, ..., N are mutually independent Poisson random variables). In addition, the population counts for each region are assumed to be fixed (i.e. non-random variables), denoted by p_1, p_2, \dots, p_N (McCullagh and Nelder, 1952). The observed variable is assumed to follow a Poisson distribution with mean μ_i for $i = 1, 2, \dots, N$.

$$Y_i \sim \text{Poisson}(\mu_i) \quad \text{where } i = 1, 2, \dots, N \quad (2.15)$$

The regression model is given in the following format:

$$\log(\hat{Y}_i) = \log(\phi_i) + (\beta_0 + \beta_1 v_{1i} + \beta_2 v_{2i} + \dots + \beta_T v_{Ti} + \varepsilon_i) \quad (2.16)$$

where \log is the link function of the Poisson regression model and ϕ_i is the offset, which is a variable measuring the units of exposure in region i , for example, number of expected cases in region i .

2.3.5 Bayesian hierarchical models

The Bayesian approach as applied to health studies has been defined as “the explicit use of external evidence in the design, monitoring, analysis, interpretation and reporting of a health technology assessment” (Spiegelhalter *et al.*, 2004). The Bayesian approach is an efficient way to estimate models and predict uncertainty with the given available data and prior distributions. Based on the Bayesian framework the posterior distribution covers the possible range of uncertainty of the estimated parameters. The observed Y_i , unobserved data z and unknown parameters θ can be described in terms of probability statements. The probability statements are conditional on the observed values of Y_i , written as:

$$p(\theta | Y_i, z) \tag{2.17}$$

The idea is to estimate θ conditional on Y_i and z . The prior distribution $p(\theta)$ is a term expressing the uncertainty of the unknown parameter θ prior to analysis, whereas the posterior distribution $p(\theta | Y_i, z)$ is a term expressing the uncertainty of θ after taking account of the data (Gelman *et al.*, 2003). The posterior is defined below:

$$p(\theta | Y_i, z) = \frac{p(\theta)p(Y_i, z | \theta)}{\int p(\theta)p(Y_i, z | \theta)d\theta} \propto p(\theta)p(Y_i, z | \theta)$$

$$p(\theta | Y_i, z) \propto p(Y_i, z | \theta)p(\theta) \tag{2.18}$$

where $p(Y_i, z | \theta)$ represents the observed data, e.g. disease incidence at region i . A range of non-spatial and spatial model structures have been used to estimate the posterior mean of the parameters of interest. A Bayesian regression method is applied in this study to explore the relationship between the observed disease cases and the explanatory variables (a series of social status, family structure variables and measures of deprivation). The results provide a summary of complex relations between variables. The posterior samples were drawn from Gibbs sampling based on Markov chain Monte Carlo (MCMC) methods, until the chain and posterior converge to a stationary distribution (Lawson *et al.*, 2003).

In this thesis, two types of Bayesian models are defined: non-spatial and spatial. The details of each model are discussed below.

2.3.5.1 Non-spatial model

Spatial information is not specified in the non-spatial model. Every region is assumed to be homogeneous. No neighbourhood information is given in the model, such that the model provides only global and average information as results. It is also assumed that no effects are contributed from the direct neighbourhood; each of the regions is completely independent. Such a model is useful only if the underlying model has no spatial variation, because the non-

spatial model gives only the average of the estimated parameters, there is no difference over space. The non-spatial Bayesian model is defined below:

$$Y_i \sim \text{Poisson}(\phi_i \theta_i)$$

$$\log \theta_i = \log \phi_i + \beta_o + \beta_1 v_{1i} + \dots \beta_T v_{Ti} + \delta_i \quad (2.22)$$

where ϕ_i is the offset (e.g. expected cases), β_o is the intercept, β_t is the coefficient of variable v_t and finally, δ_i is the unstructured heterogeneity (i.e. random effect).

2.3.5.2 Spatial model

The conditionally autoregressive (CAR) model was introduced by Besag *et al.*, (1991). Where spatial information is given in the model (e.g. direct neighbourhood), such information can be used to predict the underlying parameter of interest, such as incidence rate $\hat{\theta}_i$. In the CAR model, spatial correlation is included; therefore, the estimated parameters in region i depend on the neighbours j . There are two model settings, BYM and MIX models, as described below.

(i) *Besag, York and Mollie's (BYM) spatial model*

In the BYM model, area-specific random effects are included, which decompose into two components. The first component is δ_i the uncorrelated heterogeneity; this is the part measuring unstructured variation between areas. The second component is γ_i ; this is the component that models the structured variation in space (i.e. clustering component or correlated heterogeneity). Both δ_i and γ_i need to have a specified prior distribution. The model is defined as:

$$\log \theta_i = \log(\phi_i) + \beta_o + \sum_{t=1}^T \beta_t v_{ti} + \alpha_i \quad (2.19)$$

$$\alpha_i = \delta_i + \gamma_i \quad (2.20)$$

(ii) Bayesian Mixed (MIX) spatial model

A special type of spatial mixture model (MIX) was introduced by Lawson and Clark (2002). This spatial mixture model allows both smoothness and discontinuities and admits different forms of spatial variation. The MIX spatial model has four components. One of them is δ_i ; unstructured heterogeneity that measures the over-dispersion in an individual region. The other two are γ_i , representing the spatial correlation component and ϕ_i the component which models spatial correlation. The final component is λ_i ; it models discrete jumps. If all the $\lambda_i=1$, the MIX model converges back to the BYM model, if all the $\lambda_i=0$ the model is called pure jump (Lawson *et al.*, 2002; 2003).

The model is defined as below:

$$\log \theta_i = \log \phi_i + \beta_0 + \sum_{i=1}^T \beta_i v_{ii} + \delta_i + \lambda_i \gamma_i + (1 - \lambda_i) \phi_i \quad (2.21)$$

All the components v_i , u_i , ϕ_i and λ_i need to have specified prior distributions.

2.3.5.3 Bayesian model measurement

For GLMs the p -value can be used to measure the “goodness” of fit of the models, but for the Bayesian model it is different. Some kind of model measurement is needed to compare candidate models. In most cases concerning the use of the Bayesian model, the deviance information criterion (DIC) is used to measure how well the model is fitted, and the penalty value (pD) is used to measure how complex the model is (Spiegelhalter *et al.*, 2002 Gelman *et al.*, 2003). Both DIC and pD can be used to compare the candidate models; the smaller the DIC and pD the better the model fit. For full details of calculations and explanations please refer to Spiegelhalter *et al.*, (2002) and Gelman *et al.*, (2003). The summary of DIC and pD are given below in equations 2.22 and 2.23:

$$pD = \bar{D} - D(\bar{\theta}) \quad (2.22)$$

$$\begin{aligned} DIC &= pD + \bar{D} \\ &= \bar{D} - D(\bar{\theta}) + \bar{D} \\ &= 2\bar{D} - D(\bar{\theta}) \end{aligned} \quad (2.23)$$

where \bar{D} is the expectation measuring how well the model is fitted to the data. If \bar{D} is large it means that the model is fitted poorly to the data. It is defined as:

$$\begin{aligned}\bar{D} &= E(D(y, \theta) | y) \\ &= \frac{1}{L} \sum_{l=1}^L D(y, \theta^l)\end{aligned}\tag{2.24}$$

Finally, $D(y, \theta)$ is deviance and $D(\hat{\theta})$ evaluates the average of samples θ , which is defined as:

$$\begin{aligned}D(\hat{\theta}) &= D(y, \hat{\theta}(y)) \\ &= -2 \log p(y | \hat{\theta})\end{aligned}\tag{2.25}$$

2.3.6 Geographically weighted regression (GWR) modelling

Generally, in global models geographical variation in the relations is ignored and the process is assumed to be stationary. Geographically weighted regression (GWR) modelling is a type of local regression model. It is a well established technique that can be used to examine spatial variation in relations (i.e., local analysis rather than global analysis) and explore spatial patterns in parameters when spatial variation in parameters (non-stationary) is allowed. Information on local variation in parameters can lead to greater understanding of the relations between the target and explanatory variables.

2.3.6.1 Model Structure

In reality, some relations may vary over space, (i.e. a non-stationary model is required to describe this variation). The spatial variation in relations is ignored in a global regression model. Such local behaviour can be captured through the GWR model. If the non-stationary model is appropriate the results from the global model can provide misleading interpretations, and the spatial variation is only reflected in the residual map from the global model. The true underlying geographical pattern is hidden. The use of GWR allows to takes account of the

spatial variation if it does exist, and model the relations using a non-stationary process. The GWR model is defined below:

$$Y_i(x_i, y_i) = \beta_0(x_i, y_i) + \sum_{t=1}^T \beta_t(x_i, y_i) v_t(x_i, y_i) + \varepsilon(x_i, y_i) \quad (2.26)$$

where (x_i, y_i) represents the coordinates of location i , $Y_i(x_i, y_i)$ is the observed variable of location i , $\beta_0(x_i, y_i)$ is the intercept for location i , and $\beta_t(x_i, y_i)$ is the coefficient of variables t at location i . $\beta_0(x_i, y_i)$ and $\beta_t(x_i, y_i)$ are assumed as a continuous function in the regression model, rather than as constant and fixed in the global model. In GWR, $\varepsilon_i(x_i, y_i)$ represents the error term and it is assumed to follow a normal distribution with mean zero and variance σ^2 .

$\hat{Y}_i(x_i, y_i)$, $\hat{\beta}_0(x_i, y_i)$ and $\hat{\beta}_t(x_i, y_i)$ can be estimated from the maximum likelihood approach which is equivalent to using the least squares from the global model. However, within GWR, the parameters $\hat{\beta}_0(x, y)$ and $\hat{\beta}_t(x_i, y_i)$ are estimated through a local likelihood approach and estimated through an iterative process, until the estimates of the parameters converge. The estimation for location i is affected by the surrounding locations j . The amount of effect from location j is determined by the weighting scheme. For example, if regression point (x_i, y_i) is closer to data point (x_1, y_1) than data point (x_2, y_2) , then (x_1, y_1) causes more effect and more contribution to the estimation to (x_i, y_i) than location (x_2, y_2) (see Figure 2.1). The amount of effect depends on the choice of weighting scheme and the distance parameter d_{ij} between regression point i and the neighbouring (data) point j . The closer the data point to the regression point i , the greater the contribution to the estimation of the regression point i than the data point which is further away from the regression point i . For full details of the calculation and estimation of GWR please refer to Fotheringham *et al.*, (1998; 2002).

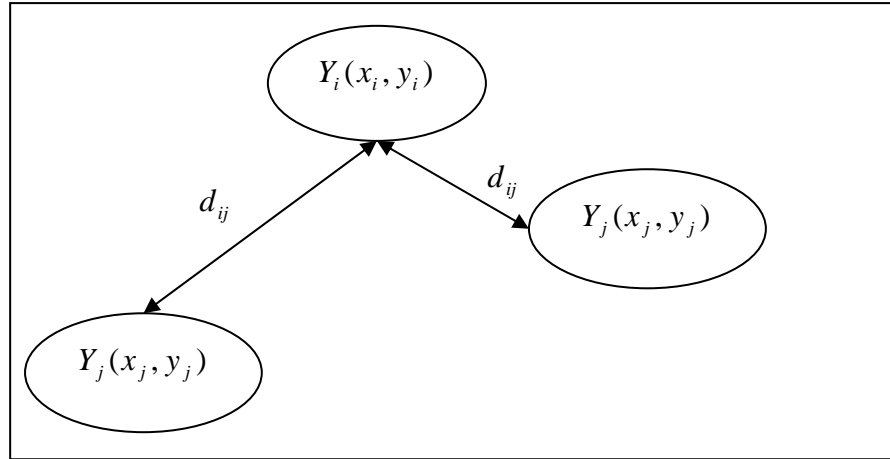


Figure 2.1 Regression point i and data point j in GWR model. The closer the data point j to regression point i , the greater the contribution to the regression point i from j .

2.3.6.2 Weighted Function: Kernel

Parameter estimation depends on the choice of and type of weighting function and kernel, where estimation is highly related to the kernel. Typically, there are two types of weighting functions being used in GWR, which are (i) Gaussian kernel (fixed kernel) and (ii) bi-square kernel (adaptive kernel); the weighting function is conditional on the size of kernel. In theory, the weight should decrease gradually as the distance between i and j increases, converging to zero.

(i) The Gaussian kernel function is defined as follows (Fotheringham *et al.*, 2002):

$$w_{ij} = \exp\left(-\frac{1}{2} \frac{d_{ij}^2}{d^2}\right) \quad (2.27)$$

Where d_{ij} is the distance between point j and regression point i and d is the bandwidth. The closer a point j to regression point i , the larger the weight given. The weight of j will be changed when the location of regression point i is changed.

(ii) The bi-square kernel utilises an adaptive method to calculate the weight w_{ij} . The size of the adaptive kernel may vary, but it covers the same number of data points in each of the kernels. The weighting function w_{ij} (where i is the regression point and j is the data point for $i = 1, \dots, N$ and $j = 1, \dots, N$) determines the weight for each data point. Each data point receives a different weight to the adaptive weight, which depends on the distance between i and j . Any points outside the kernel receive zero weight to the regression point and closer to the regression point receive more weight than those further away from the regression point. The amount of weight is determined by the bandwidth d . The function is given as below (Fotheringham *et al.*, 2002);

$$w_{ij} = \begin{cases} [1 - (\frac{d_{ij}}{d})^2]^2 & d_{ij} < d \\ 0 & otherwise \end{cases} \quad (2.29)$$

where d_{ij} is the distance parameter between regression point i and data point j . The choice of bandwidth is very important in terms of estimation of the parameters and the amount of smoothing of the parameters. In theory, a larger bandwidth can cause over smoothing (i.e. convergence to the global model) and a smaller bandwidth can also cause under smoothing (i.e. large spatial variation). Therefore, an optimal size of bandwidth is essential. The methods used to select the optimal bandwidth are to use some model measurements (e.g. cross-validation error, Akaike information criterion (AIC) and (or) correct Akaike information criterion (AICc)). The model with a certain kernel size which has the smallest measurement (e.g. AICc), is the optimal model with optimal kernel size.

2.3.6.3 Model measurement

The coefficients vary continuously over space, therefore, it is almost impossible to achieve completely unbiased estimation. Models with very few data can cause larger variation in local parameter estimation. Therefore the estimation becomes less reliable. On the other hand, a model with large number of data points can provide more reliable local parameter estimations. However, such models may

contain a large amount of bias as the distances between regression point i and data points j increase. Thus, it is important to get a balance between the bias and variance of the parameters in estimation. A trade-off is needed between the bias and variability of parameters and can be considered by introducing model selection indicators. There are many indicators available, such as DIC and BIC. For GWR, it is common to use the AIC to measure how well the model is fitted with certain explanatory variables at a given bandwidth size.

AIC was developed by Akaike in 1971 to assess the performance of estimated statistical models. The AIC of the model with bandwidth d is given as:

$$AIC(d) = D(d) + 2K(d) \quad (2.30)$$

Where D represents the deviance, K represents the effective number of parameters in the regression model, and d is the bandwidth in the kernel. The model with the smallest AIC value represents the model with the optimal bandwidth. Such a method is called minimum AIC estimator (MAICE). In practice, if the difference in AIC between two models is less than or equal to two, there is no significant difference between the two models, in which case both models are accepted as the best fitted model with optimal bandwidth.

AIC can reflect biased measurement (Akaike, 1974; Sugiura, 1978) when there are too many parameters and too few sample points (data). In order to avoid biased estimation from AIC, Sugiura (1978) derived a second order variant of AIC which is called c-AIC, and Hurvich and Tsai (1989) incorporated a small sample bias adjustment which led to a criterion called AICc defined below, where N is the total number of regions:

$$\begin{aligned} AICc(d) &= D(d) + 2K(d) + 2 \frac{K(d)(K(d) - 1)}{N - K(d) - 1} \\ &= AIC(d) + 2 \frac{K(d)(K(d) + 1)}{N - K(d) - 1} \end{aligned} \quad (2.31)$$

The other bandwidth selection criterion that can be used in GWR is called the Bayesian information criterion (BIC), and the calculation is given below:

$$BIC = -2\log(L) + K \log_e(N) \quad (2.32)$$

Where L is denoted as the model likelihood. BIC was derived from Bayesian theory, where each of the discrete number of candidate models have equal prior probabilities; the prior distributions on the model parameters. Again, the model with the smallest BIC is the better fitted model compared to the other candidate models. Both AICc and BIC can be used as a measurement method to compare candidate models; the best fitted model can be identified through comparing the AICc or BIC values.

2.3.7 Multilevel regression models

The term ‘multilevel’ refers to a nested membership relationship among units in a system (Centre for Multilevel Modelling, 2008). Multilevel modelling techniques allow the combination of different levels of information to explain the relations between observed variables and explanatory variables, given that some variables are measured at a lower level and some at a higher level. It is an appropriate tool for modelling data with complex hierarchical structures. It allows and helps users to fit a model when the target or observed outcome and the explanatory variables do not appear in the same level but the lower level information does nested within the higher level. Nowadays, it has become more common in studies of diseases, to understand how the diseases respond when different level information is available; for example, higher level socio-economic variables may be associated with an observed disease at an individual level. So it is possible to compare the different variables between individuals and within groups (Centre for Multilevel Modelling, 2008). The correlation between observed and explanatory variables might be different from individual to individual. For a normal single level model, such structural information is ignored, and this may provide misleading results.

2.3.7.1 Model structure

For multilevel regression models, there are few possible model structure (i) simple hierarchical model, (ii) random intercepts model, (iii) random coefficients model, (iv) Bayesian approach, and (v) conditionally autoregressive (CAR) model. Each of such model structure will be described in the following sections. The basic model structure is showed below.

$$\begin{aligned}
 y_{cd} &\sim \text{Binomial}(n_{cd}, \pi_{cd}) \\
 \log \text{it}(\pi_{cd}) &= \log\left(\frac{\pi_{cd}}{1 - \pi_{cd}}\right) \\
 &= \beta_0 + \beta_1 v_{1cd} + \beta_2 v_{2d} + u_d + e_{cd} \\
 \pi_{cd} &= \frac{1}{1 + \exp(-(\beta_0 + \beta_1 v_{1cd} + \beta_2 v_{2d} + u_d + e_{cd}))} \tag{2.33}
 \end{aligned}$$

Where y_{cd} is the observed variable, β_0 is the intercept, β_t is the coefficients for $t = 1, 2, \dots, T$, u_d is the random effect component at level d (higher level residuals), and e_{cd} is the random effect component at c level (lower level residuals). Where level c is lower than level d (e.g. individual c nested within region d). The first few terms in the model are assumed as the fixed part in the model and both u_d and e_{cd} represent the random part in the model. The random effect components are Normally distributed with mean zero and variance σ_u^2 and σ_e^2 .

$$u_d \sim \text{Normal}(0, \sigma_u^2) \qquad e_{cd} \sim \text{Normal}(0, \sigma_e^2)$$

(i) Simple hierarchical model

For the simple hierarchical model, all the parameters are assumed as constant and only showed the overall value for the coefficients, the model structure showed below,

$$\log \text{it}(\pi_{cd}) = \beta_0 + \beta_1 v_{1cd} + \beta_2 v_{2d} \tag{2.34}$$

(ii) Random intercept model

Within the simple hierarchical model the intercept is fixed as a constant term over space. However, adding the random effect term (random intercept) in the model, which allows to account for the probability of y_{cd} occur in different d , β_{0d} vary across the space at the higher level d , the model is defined below,

$$\log it(\pi_{cd}) = \beta_{0d} + \beta_1 v_{1cd} + \beta_2 v_{2d} \quad (2.35)$$

$$\beta_{0d} = \beta_0 + u_{0d}$$

Where β_{0d} the random intercept has two components which are the fixed term β_0 and the random effect u_{0d} it is a higher level specified component (level d) it is Normally distributed with mean zero and variance σ_{0d}^2 in equation ().

(iii) Random coefficients model

The coefficients can be vary , which allows to account for the difference in the lower level which nested within the higher level and also vary across the higher level. For example, the random coefficient account for the difference between explanatory variables (e.g. high socio-economic condition or low socio-economic condition) within the same higher level d (e.g. Output Areas) and to vary across in the higher study areas (Output Areas), the model structure is defined below,

$$\log it(\pi_{cd}) = \beta_{0d} + \beta_{1d} v_{cd} + \beta_2 v_d \quad (2.36)$$

$$\beta_{0d} = \beta_0 + u_{0d}$$

$$\beta_{1d} = \beta_1 + u_{1d}$$

(iv) Bayesian framework

Bayesian theory can be incorporate within multilevel modelling, which combining the prior information into the model, so that each unknown parameters resulting with a posterior distribution represent the possible uncertainly range for that parameter. Therefore, each of the parameters in model (equations 2.34, 2.35 and 2.36) will be given with a specified prior distribution simple to section 2.3.5.

(v) CAR model

The model in (i) to (iv) have not included with spatial information, thus each study area is assumed independent to other study areas. However, the neighbours normally quite similar to each other and there is certain affect from each other. Therefore, the spatial information is very helpful in understanding human behaviours. Conditionally autoregressive (CAR) model is commonly applied within multilevel regression model. When the areas are next to each other, which will received a weighted value one, otherwise the weighted value equal to zero. Such idea suggested the direct neighbour's information is taken into account in the prediction.

2.3.7.2 Parameters estimation process

There are two main approach of fitting multilevel model, which are (i) likelihood-based and (ii) Bayesian approach. For the likelihood-based included iterative generalized least squares estimation (IGLS) and restricted iterative generalized least squares (RIGLS), the estimation from both methods are procedure from an iterative process. However, when a model is based on Bayesian framework an alterative approached Monte Carlo Markov Chain (MCMC) can be used to estimate the unknown parameters, where Gibbs sampling will be applied in MCMC.

2.3.7.3 Model comparison

Chi-square goodness of fit test is commonly used to compare between the model, when maximum likelihood estimation is used in multilevel model, likelihood ration test can be used otherwise pD and DIC values can be used to measure how good it the model fitted and to compare between models. Details about pD and DIC please reference to section 2.3.5.3.

2.3.8 Available software for regression modelling

Since computing technology is improving rapidly, there are many software packages available for data analysis, and for constructing the regression model. Four software packages were used in this study, which are listed below:

- (i) S-plus
- (ii) WinBUGS,
- (iii) GWR 3.0
- (iv) MLwiN

S-plus was used to fit generalised linear regression models with various distributions, all the Bayesian models were fitted using WinBUGS, the local regression model was fitted by using GWR 3.0 and finally, MLwiN was used to fit multilevel regression models for the Chlamydia study.

2.4 Classification and regression trees (CART) analysis

Classification and regression tree (CART) is a classification method, which was developed by Breimen *et al.*, (1984). CART is a data analysis tool for predicting the dependent variable based on categorical predictor variables. It works by splitting the observed variable (e.g. patients) into groups, by using a binary trees method which often provides an illuminating view of the data. CART has been applied in many health studies (Chiogna *et al.*, 1996; Harper and Winslett, 2006). Particularly, for health studies, it might be interesting to identify the risk groups of patients and their common characteristics. An example is shown in Figure 2.2:

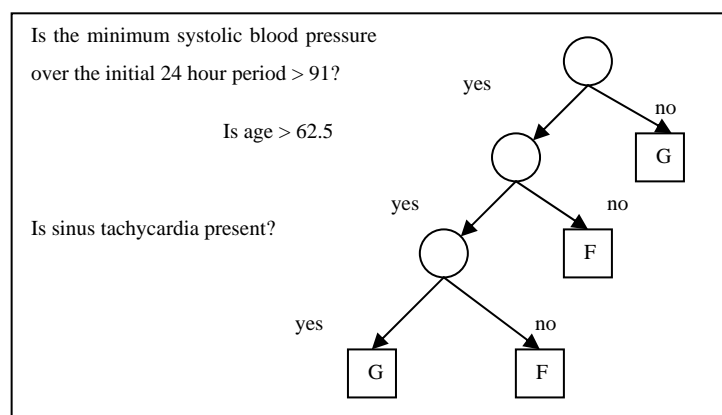


Figure 2.2 CART analysis of cardiac patients (Breimen *et al.*, 1984). In the tree, F represents low risk group heart attack patients and G represents high risk group heart attack patients.

Before starting to construct a regression tree it is necessary to define the predicted variable (i.e. the variable of interest). Each junction of the tree is called a node, the node at the end of a branch is called the terminal node. The parent node is the node being split further; thus, child nodes come from the parent node, and each parent node has two child nodes. If a variable is ordinal, variance is used to measure the purity in the group; if a variable is categorical then deviance is used to measure the purity in the group. An algorithm is used to split the data into sub-populations (child nodes); the sup-populations need to have higher purity (decreasing variance or deviance) than the parent node (Breimen *et al.*, 1984). Therefore, each of the variables (nodes) will have a best binary split given that it produces a node with the smallest variance or deviance. Users need to predefine a set of stopping rules to stop branching and stop expanding the tree. Some examples are listed below. In addition to the “stop rules”, stop branching occurs when the variance or deviance has not significantly reduced. It is necessary to define a set of stop rules to stop expanding the tree. Here are some examples of stop rules:

- (i) stop when nodes contain a certain number of cases,
- (ii) stop when reduction of variance is below a certain threshold,
- (iii) stop when a maximum number of terminal nodes has been produced.

2.4.1 CART algorithm

Variance is the main element in the CART, which is used to split the data into groups (or sub-groups) according to the best independent variable with reduced total variance. Firstly, each independent variable value is required to calculate the sum of the independent variable in that group ($\sum x$), the sum of square value ($\sum x^2$) and the number of items of data in that group (N). Secondly, sort the values of the independent variable into increasing order of the mean value of the dependent variable. Thirdly, to design where to split the independent variable, look for the sorted mean that produces the minimum variance. It is important to split the data based on the best independent variable in order to reduce the total variance calculated as following:

$$\frac{\sum_{\text{all groups}} [\sum x^2 - \frac{(\sum x)^2}{N}]}{\text{Total observations}} \quad (2.34)$$

Finally, choose a suitable sub-group of the data as the current group, and repeat the same process until either the data are split into groups and the size is less than the minimum number (stop rule (ii) in section 2.4) or the reduction in variance obtained by a split of the data is less than the minimum value (stop rule (iii) in section 2.4).

2.5 Decision theory

Everyone needs to make a decision when there is more than one choice; however, how can one make the best choice with the best outcome? It is a difficult question to answer. Decision analysis can be applied to predict the best choice with the best possible outcome. Some mathematical models can be used to make decisions with certain objective functions with specific risk (e.g. positive cervical cancer test result). The true future outcome cannot be predicted with full certainty, but it is possible to estimate the outcome with enough accuracy based on the collected data or distribution. It is particularly useful for studying screening systems, as it allows an analysis of which screening option provides the best outcome (Spiegelhalter *et al.*, 2004).

2.5.1 Decision tree model structure

The Decision tree can be applied based on decision theory; the tree can be used to determine the optimal outcome based on the given risk. Such a model enables a decision maker to break down a complex problem into several smaller problems. Decision theory has been applied in many health-care studies to evaluate health options. Below is one of the examples (Figure 2.3) from Ashby and Smith (2000) and Spiegelhalter *et al.*, (2004). It shows how to make a decision based on the estimated risk from collected data and also the prior distribution, which allows

one to calculate the cost of the treatment, so that the policy makers can make decisions about the patients' treatments.

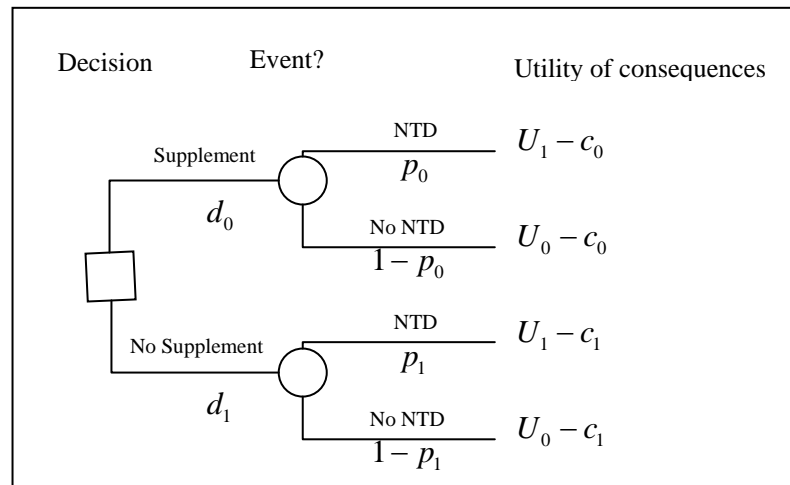


Figure 2.3 Decision tree for folic acid supplementation decision. The square node represents where to make a decision and the circular nodes represent events. The values or outcomes at the end of the tree are called utility of consequences. Where d_0 and d_1 represent the decision of taking a supplement or not, p_0 and p_1 represent the probabilities of having an neural tube defect (NTD) based on the earlier decision (taking supplement or not), U_0 and U_1 represent the utilities of taking NTD or not and finally, c_0 and c_1 represent the cost.

The decision can be made based on the cost of the outcome (consequences). Whichever gives the best cost for the treatment might be considered as the optimal decision. The same theory can be used in evaluating cervical cancer and Chlamydia screening systems to evaluate the optimal decision, which can increase the benefits for the patients and best use of the resource.

2.6 Simulation model

Simulation is a technique that facilitates learning about a disease process while observing the patients' or diseases' behaviour from a real system. The collected data can be used to design the system; and by asking some "what if" questions it is possible to identify any bottlenecks within the system and establish what can be

done to improve the real system and achieve the best possible outcomes (Winston, 1994; Taha, 1997; Hillier and Lieberman, 2001).

Particularly, for the NHS healthcare system, it might be interesting to know about each individual patient passing through the system, and that would provide a better opportunity to view the system, and to identify any areas within the current system that need improving. In addition, it is possible to forecast the future demand, which allows the NHS to prepare to provide better quality health services and better estimate resources. In terms of the patients' waiting time, it could also benefit the patients by reducing waiting times, and as a result, the system would operate in a more effective way. Finally, the patients' life quality (QoL) might be improved through such a study.

Simulation is used in this thesis to study both cervical cancer and Chlamydia screening systems in order to attain a better understanding of how the screening systems run and what activities could be undertaken to improve the detection rates. Two possible simulation frameworks are considered – the Markov and Semi-Markov models. The choice of simulation framework depends on the type of information available, as explained below.

2.6.1 Markov and Semi-Markov models

In some cases, it is of interest to know how a random variable (RV) changes in a given process. In particular, it could be of interest in a specific disease process. In a situation where a doctor wants to predict the probability of a patient moving from one state to another, a stochastic process model is particularly helpful to answer such a question. In particular, a type of stochastic process Markov chain has been applied in many areas, such as in marketing and finance, and especially in healthcare services. Markov chains have the special property that probabilities involving how the process will evolve in the future depend only on the present state of the process, and so are independent of events in the past. For a disease process, the parameter space T is countable (e.g. for cervical cancer these could be healthy, CIN 1, CIN 2, CIN 3, cancer and death).

Consider a continuous time chain $\{X(t)\}_{t \in [0, \infty)}$. It takes a value in a discrete state space S (e.g. pre-cervical cancer stages CIN 1, CIN 2 and CIN 3, cancer, death). The probability of moving from one state to another is called a transition probability ψ_{ij} and the time spent in state i before moving to state j is called the holding time h_{ij} . The reason a chain is said to be a semi-Markov chain is that each transition depends on the previous state or states, which means all transition probabilities are denoted as $\psi_{ij} \equiv \Pr(X_n = j \mid X_{n-1} = i)$. It exactly describes the pre-cervical cancer disease process, as each disease state is highly dependent on the previous state. The Semi-Markov model does not have the memory-less property (i.e. the current state is independent of past states) and it can be fitted with any distribution (Minh, 2000).

As simulation model can be used to simulate each individual patient's flow through the cervical cancer or Chlamydia infection process (e.g. how long a patient may remain in a state and how many patients stay in a state). It allows some understanding of the disease process and allows us to ask some “what if” questions: for example, what happens if a screening programme becomes available with a set of policies etc.. It allows us to examine and understand the screening programme without putting anyone (e.g. patients) at risk.

Chapter 3 Data

3.1 Introduction

For the study of diseases, there are two types of data available, *aggregated* and *individual* data, and it is more common to have aggregated data than individual data, due to confidentiality restrictions. As a result, data are often aggregated as summary counts rather than being provided as point-level data on individuals. Two diseases were studied in this thesis: the cervical cancer study used aggregated data and the Chlamydia study used individual data. The cervical cancer incidence data were drawn from 2004 statistics, and the Chlamydia data from 1999-2000. The Townsend indicator, social grade IV and V, and family structure (e.g. marital status and lone parents etc.) variables were downloaded from the UK census 2001 at different levels. The shape files for mapping were downloaded from Edina UKBorders.

3.2 Cervical cancer data

3.2.1 Cervical cancer national data 2004

The cervical cancer count data were provided by the Association of Public Health Observatories (APHO), which represents the nine Public Health Observatories (PHO) in England (Table 3.1). In total, 7179 cervical cancer cases (i.e. current cases in 2004, including new diagnosed cases) and 2391 deaths were recorded in 2004. The data were represented at district and unitary authority levels of the Cancer Registries in England. The total female population per age group was determined from the 2001 UK Census (between the ages of 0 to 4, 5 to 9, ..., 85 and over). These data were used to calculate the number of expected cases per region. The incidence rates were downloaded from the cancer research UK website (Cancer Research UK, 2005).

Table 3.1. Public Health Observatories (PHO) in England

Public Health Observatory (PHO)	Number of districts/unitary authorities in PHO
1. South West	45
2. South of England	67
3. London	33
4. East of England	48
5. East Midlands	40
6. West Midlands	21
7. North West	23
8. Yorkshire and Humber	43
9. North East	34

3.2.2 Townsend indicators, social grade and family structure national data at district or unitary authority level in 2001

The second dataset included in the cervical cancer study was composed of the Townsend indicators, social grade and family structure data for 2001, which were downloaded from the UK Census of 2001. Since the census is carried out once every ten years the closest matched year to 2004 was 2001. All variables were included in the analyses, which are listed in Table 3.2.

Table 3.2 Summary of explanatory variables used as indicators in the regression analysis.

Type of variables	Variables	Description	Table from UK census 2001
Townsend index score	(i) Unemployment	Employed population	KS009a Economic activity: all persons (from the key statistics)
	(ii) Households not owned	All households not owned by the tenant	KS018 Tenure (from the key statistics)
	(iii) Car ownership	All households with no cars/vans	KS017 Cars or vans: all households (from the key statistics)
	(iv) Over-crowded housing	Over one person per bedroom	UV 058 Person per room (from the census area statistics univariate tables)
Family structure	(v) Female marital status (for cervical cancer study)	(i) Proportion of single (officially single): single (never married) + divorced + widowed (ii) Proportion of married (officially married): married + remarried + separated (but still legally married)	ST002 Age by sex and marital status
	(vi) Marital status (for Chlamydia study)	(i) Proportion of single (officially single): single (never married) + divorced + widowed (ii) Proportion of married (officially married): married + remarried + separated (but still legally married)	KS004 Marital status
	(vii) Households with lone parent	(i) All lone parents (ii) Female lone parents	KS022 Lone parent households with dependent children
Social grade (proportion)	(viii) Proportion of Social grade IV + V	Grade VI: semi-skilled and unskilled manual workers Grade V: on state benefit, unemployed, lowest grade workers	UV050 Approximated social grade VI and V (low socio-grade)

Variables (i) to (iv) were used to calculate the Townsend index and other variables were used individually in the regression models. Variables (vi) to (viii) were expressed as a proportion of the total population.

At the beginning of this study the data were subjected to a Chi-square goodness of fit test, which showed that the data approximately followed a Poisson distribution. The national incidence rate per age group was provided by Cancer Research UK (Cancer Research UK, 2005). The incidence rate was used to calculate the number of expected cases.

3.2.3 Individual colposcopy clinical data from Portsmouth St Mary's Hospital 1998-2006

Colposcopy individual test results were provided by St Mary's Hospital, Portsmouth, between 1998 and 2006; these included some of the individual patients' information: (i) date of birth, (ii) date of smear, (iii) date of colposcopy, (iv) smear tests result and (v) colposcopy test results. Therefore, the ages of the patients and the screening intervals can be estimated based on dates of birth and date of tests. The patient's name was removed before the data became available for this PhD study. This was done for ethical issues and to protect the patient's privacy.

3.3 Chlamydia data**3.3.1 Chlamydia individual clinical data 1999 – 2001**

The individual patients' data were provided by St Mary's Hospital Portsmouth, which came from the second phase of an opportunistic screening trial study. Data were collected between October 1999 and September 2000, around the Portsmouth area. In total, 25553 tests consisted of 17342 patients, which included repeat tests to check that the infection had been cured. All the patients were allocated with a new ID number when they had repeat tests. The results and patients' data were provided in Excel format with matching patient IDs. The patients' ID numbers were used to link the tests and patients' records together as shown in Figure 3.1. The patient's records had patient name fields, but such information was removed before the data became available for research study to ensure patient anonymity.

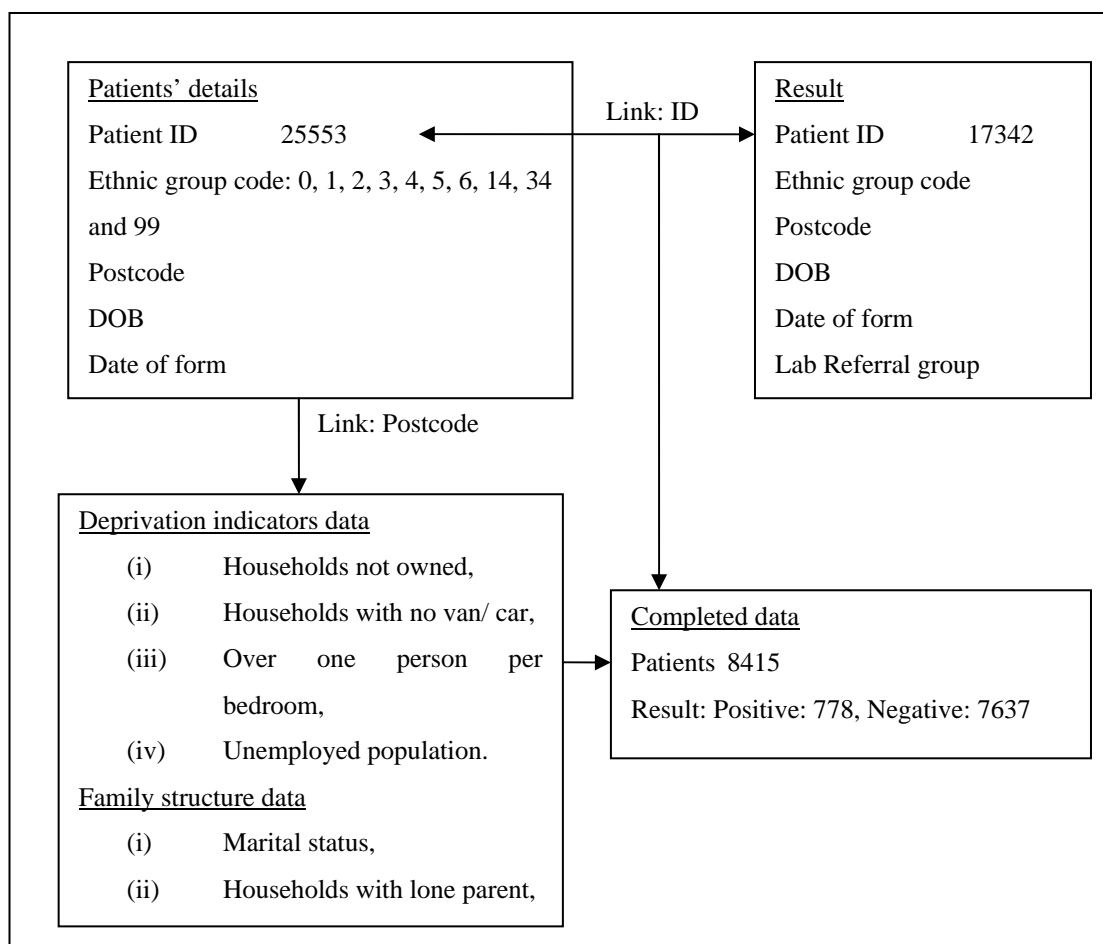


Figure 3.1 Linking the patient's details, results and explanatory data together.

Every time when a patient comes for a Chlamydia test, the patient receives a new patient's ID. However, a patient who has taken two Chlamydia tests and who has two different IDs, will have matching ID information attached in both of the patients' details and result records. Thus, the matching ID allows linking the same patient's history together in a very easy way. The postcode information is available from each individual patient's records. Therefore, that can be used to identify the location and it can be used to create a link to join the individual patient and the aggregated socio-economic data together.

The majority of the data were collected from the female population between the ages of 16 to 24 (Pimenta *et al.*, 2003 a;b). A list of personal information was available, including variables such as (i) patients' ID, (ii) gender, (iii) date of birth, (iv) date of form of the test, (v) the estimated age based on (iii) and (iv); (vi) ethnic group, (vii) postcode, and (viii) results, which are either positive or negative. Most importantly, the postcode information provided the spatial

information, which would allow the researcher to apply any spatial modelling based on that information. For full details of the trial study of the data please refer to Pimenta *et al.*, (2003 a;b)

3.3.2 Townsend indicators, social grade and family structure data at Output Area level from the Census 2001

A Postcode to Output Area (OA) look-up table was used to link the postcode and Output Area together. In addition to the non-available individual socio-economic data, the lowest available level was used in this study, which was the Output Area level. Townsend indicators, social grade and family structure data were collected through the UK census 2001; this was the closest year to the second phase of the Chlamydia screening trial study and all variables were collected at Output Area (OA) level. In total, there are 1365 Output Areas that were included in the trial study (Table 3.3).

Table 3.3 Chlamydia study summary at unitary authority level.

Location	No. of Output Area	No. of test	No. of positive test	Total population	Positive rate per 10000 people
Southampton	15	14	2	217,460	0.09197
Eastleigh	18	21	1	116,176	0.08608
Winchester	10	129	7	107,303	0.6524
East Hampshire	111	336	24	109,354	2.1947
Fareham	344	1,817	151	107,965	13.986
Gosport	246	1,502	160	76,381	20.9476
Portsmouth	621	4,595	433	186,717	23.1902
Total	1,365	8,414	778	921,356	8.4441

3.3.3 Townsend indicators, social grade and family structure data at Census area statistics (CAS) ward data

The purpose of having different levels of data is to examine the disease pattern and relation between Chlamydia incidence cases and deprivation indicator, social status and family structure factors at different spatial levels. An Output Area-to-CAS ward look-up table was used to link the Output Areas and wards together. All variables were collected through the UK census 2001; the Townsend scores at this level are available from the Census Dissemination Unit based on the 2001

census data (Census Dissemination Unit, 2006) and variables were listed in Table 3.2.

3.4 Data problem

It is common to encounter three data issues (i) aggregated data, (ii) individual data, and (iii) missing data. In particular, for the study of disease, most of the data are available at aggregated levels because of confidentiality restrictions on patients' personal data. Therefore, it is unlikely to achieve point level data, mostly data are only available at aggregated level, and it only provides an average picture of the study region. Individual data are often not available. Missing data is another common issue in studies of disease; it could be due to unobserved data, or when the observed data are too small; for example, for the cervical cancer data, all the observed records between zero and five were closed to protect patients' confidentiality. It is possible to model diseases despite a certain proportion of missing data; the details of how to deal with missing data are shown in each of the analysis chapters four to seven.

3.5 Summary

Two sets of data relating to two different diseases at national and local levels were used to study the relationships, if any, with deprivation indicator, social status and family structure variables. The missing data were overcome by employing different methods; for full details refer to each of the analysis chapters (4-7).

Chapter 4 Model for early detection of cervical cancer through cervical cancer screening programme

4.1 Introduction

Cancer is one of the most common causes of death; however, it can be prevented by early detection through regular screening tests. Most patients remain in an asymptomatic state at the early stages of cancer. When a patient becomes aware of symptoms, it is likely to be at a more advanced stage of the disease. Pre-cancer stages and early cancer stages can be detected through screening (Jonse, 1997). An example: in the case of colorectal cancer it was found that a regular screening test is essential in preventing cancer and it makes a remarkable difference in survival rates (Harper and Winslett, 2006). There are a number of national screening programmes available for parts of the general population that satisfy some basic requirements for taking the tests (e.g. age). The cervical cancer screening programme is one example of a national screening programme that is free of charge for every female who falls within a certain age group (Table 1.3). Within the cervical cancer screening programme, the smear test is the major test for every female patient. Patients who have moderate dyskaryosis, severe dyskaryosis or suspected invasive cancer from their smear test results, will need to take a further test, called a colposcopy, to investigate the abnormality further. It has been argued that, in the past, too many unnecessary colposcopies were performed. From a management point of view, it is helpful to reduce the number of unnecessary coploscopies as this can reduce the amount of wasted resources. A study demonstrated the possibility of removing the low risk population from the cervical cancer screening programme, which is also a possible way of increasing the efficiency of the current screening programme (Sherlaw-Johnson *et al.*, 1999).

From the patients' point of view, the waiting time can be reduced if unnecessary coploscopies are avoided. Thus, those in greatest need would have the chance to

Chapter 4 Model for early detection of cervical cancer through cervical cancer screening programme

detect cancer earlier and receive the appropriate treatment earlier. This chapter focuses on evaluating screening options and simulating the natural pre-cervical cancer process through a decision tree model and simulation model. Risk grouping through Classification and Regression Tree (CART) analysis can identify the elements associated with the development of Cervical Intraepithelial Neoplasia (CINs) and cervical cancer. Risk grouping will help in targeting patients to increase the efficiency of screening services (e.g. different risk groups may have different screening intervals and even different screening tests) (Breiman *et al.*, 1984; Sherlaw-Johnson *et al.*, 1999). Types of diagnostic tests are also important to the level of preventing cancer and cost-effectiveness of the screening programmes (Jenkins *et al.*, 1996; Sherlaw-Johnson *et al.*, 1999; Sherlaw-Johnson and Phillips, 2004). Other research has demonstrated how the use of different types of diagnostic tests (e.g. LBC and HPV tests) may affect the chances of detecting abnormal cells and the number of colposcopies required (Jenkins *et al.*, 1996; Sherlaw-Johnson and Philips, 2004). Details of medical terms can be found in Appendix A.

4.2 Objectives

This chapter analyses a set of clinical data; the individual colposcopy data from the Portsmouth St Mary Hospital between 1998 to June 2006. From the available data, the specific objectives are (i) to analyse which type of patients had a higher chance of requiring colposcopy than others. Separating the population into risk groups is one possible strategy to increase the chance of detecting cervical cancer at the early pre-cancer states; (ii) to construct a decision tree model to evaluate the effectiveness of each screening option, and (iii) to fit a simulation model, which describes the natural history of the pre-cervical cancer process. This model can describe a patient's flow through pre-cervical cancer and cervical cancer in the screening system from a normal (healthy) state to a cancer state over a period of time. All the analysis in this Chapter is intended primarily as a demonstration of what is possible, the regression analyses (Chapter 5) was intended to demonstrate that risk is related to known explanatory variables.

Chapter 4 Model for early detection of cervical cancer through cervical cancer screening programme

When the associated risk factors were identified from the regression models in Chapter 5, the general population can be divided into risk group, according to the identified associated risk factors in Chapter 5. The probabilities and other required parameters can be added into the decision tree and simulation models in this Chapter to demonstrate the real use of decision tree and simulation models.

4.3 Current screening programme

In the UK, the national cervical cancer screening programme is free of charge for every woman from the age of 25 to 64 (Table 1.3). Cervical cancer screening began in Britain in the mid 1960s, but it became an official screening programme by 1998 (Patnick, 2004). It was estimated that screening services cost around £150 million per year, including treating pre-cancerous lesions annually, equating to £37.50 per woman screened (Patnick, 2004). Annually, around 3.5 million women attend for cervical cancer screening tests (Patnick, 2005).

Cervical cancer screening tests are used to detect abnormal cells within the cervix, but it is not a test for diagnosing cervical cancer. Details about the national cervical cancer screening policy are summarised from the national screening guidelines (Patnick, 2004), attached in Appendix B.

4.3.1 Potential problems

No test can provide 100% accurate and effective results. Therefore, a certain number of unnecessary colposcopies are performed every year. Such unnecessary colposcopies increase the waiting time and waiting lists for other patients who may then have to wait longer before they can take the colposcopy. The potential problems are explained in Figure 4.1.

Chapter 4 Model for early detection of cervical cancer through cervical cancer screening programme

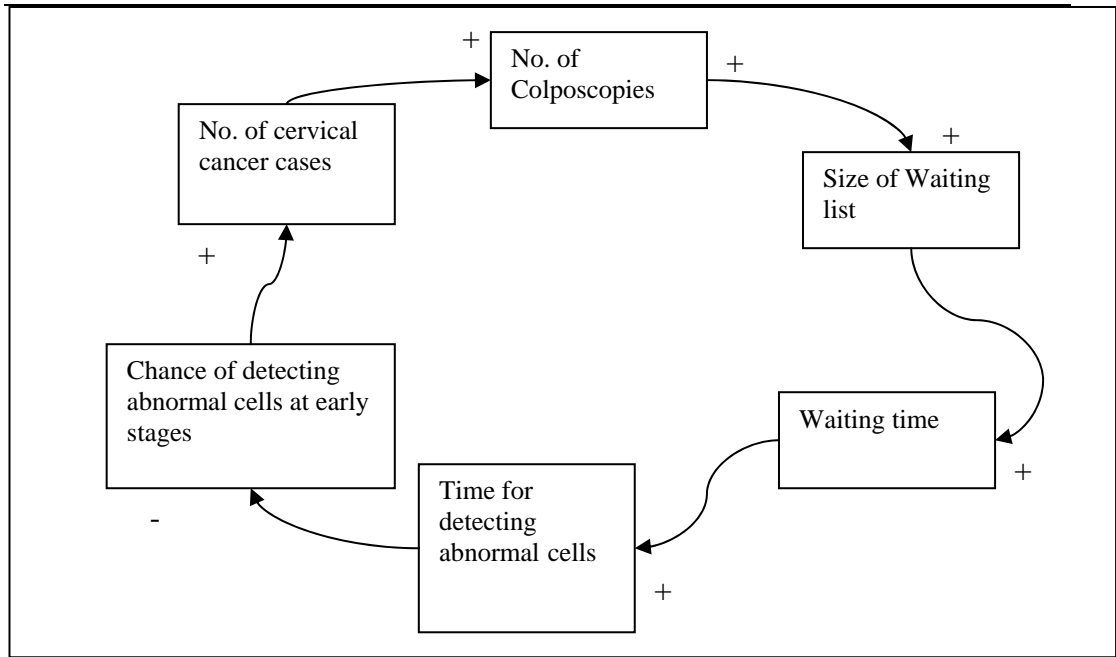


Figure 4.1. Problems associated with unnecessary colposcopies. The corresponding waiting list and waiting time increases at the same time as the number of colposcopies increases. Therefore, the time taken to detect abnormal cells increases and the chance of detecting abnormal cells at an early stage decreases. Finally the number of cervical cancer cases may increase.

4.4 Disease states

Disease states are used to measure patients' illnesses. Pre-cervical cancer and cervical cancer can be diagnosed at different states in their disease development. Carcinoma *in situ* (CIS) is a term used to describe atypical epithelia found in the cervix by measuring the full-thickness of the epithelium (Singer and Monaghan, 2000). The term was introduced by Broders (1932).

Pre-cervical cancer states are called CINs, a term introduced by Richart (1968). CINs are used to measure abnormalities within the cervix, and these are measured at three levels; CIN1, CIN2 and CIN3. Normally, disease states are expressed numerically (e.g CIN1, CIN2, etc). The earlier the disease is detected, the greater the chance of survival and the greater the benefit that may be achieved from treatment (Jonse, 1997). In this chapter, pre-cervical cancer disease states are considered.

Chapter 4 Model for early detection of cervical cancer through cervical cancer screening programme

Human Papilloma virus (HPV) is a type of virus. It is one of the common sexually transmitted infectious diseases (STIs). HPV can be detected through a colposcopy (Martin, 2000). For women with an abnormal cervical smear, the HPV DNA is found to be positive in a much higher percentage and is therefore a useful indicator of a high risk of developing cancer of the cervix. Some forms of HPV are especially related to cervical cancer incidence; for example, HPV 16 and HPV 18 (Goldie *et al.*, 2003; Arias-Pulido *et al.*, 2006). Therefore, the presence of HPV might be an indicator, which may be associated to the development of cervical cancer in the future. Details of medical terms are attached in Appendix A.

4.5 Data analysis

Patients are assumed as homogeneous under national healthcare policies, but individual patients differ from each other in a number of characteristics and physical conditions. Those characteristics include (i) sex, (ii) age, and (iii) socio-economic condition. Further characteristics focusing on physical condition include (iv) medical condition, (v) severity of illness, (vi) level of complications of disease states, (vii) speed of recovery, and (viii) response to medication. Patients with similar characteristics tend to have similar needs. Those patients can be considered as a group of patients requiring similar healthcare. In fact, these groups are typically heterogeneous and require more detailed modelling for classification; it is advantageous to be able to divide this heterogeneous group into smaller homogeneous (in terms of a particular measurement, e.g. patients' age) subgroups. Classification and Regression Tree Analysis (CART) is a data analysis tool to split patients into groups, by using a binary trees method which often provides an illuminating view of the data (Breiman *et al.*, 1984).

4.5.1 CART analysis and Colposcopy risk groupings

CART is commonly used in healthcare to divide the population into groups according to their risk; an example is maternity risk grouping to avoid complications and to increase benefits to both the medical teams and the mothers (Harper and Winslett, 2006). Another clinical study used different information

Chapter 4 Model for early detection of cervical cancer through cervical cancer screening programme

(i.e. collected data and expert opinion) to explore which is the best option for diagnosing possible heart disease in new born babies (Chiogna *et al.*, 1996). For the cervical cancer study the initial analysis attempted to classify patients into risk groups based on colposcopy results from Portsmouth clinical data. The results may help to predict which type of patients are likely to have positive colposcopy results as well as grouping patients into various groups according to their common characteristics.

Firstly, colposcopy results were defined as the predicted (dependent) variable. The list of independent variables included date of birth, age at smear test (equation 4.1), smear results, age at colposcopy (equation 4.2), screening interval (equation 4.3), and patients' history.

$$\text{Age of smear} = \text{Date of smear test} - \text{Date of birth} \quad (4.1)$$

$$\text{Age of colposcopy} = \text{Date of colposcopy} - \text{Date of birth} \quad (4.2)$$

$$\text{Screening interval} = \text{Date of colposcopy} - \text{Date of smear test} \quad (4.3)$$

The predicted variable was the characteristic predicted by the predictor (independent) variables. Thus, the predicted variables were assumed to be potentially related to the predictor variable. Once the predicted variable had been chosen, an algorithm was used to split the original population into sub-populations. The first node of the tree was the predicted variable. More details of the CART algorithm were discussed in Chapter two. Also within this study the CART algorithm was used slightly differently than was described in Chapter two. The difference was that each split was done manually. Since the relationship between Colposcopy and Human Papilloma Virus (HPV) was of interest, HPV absent and HPV present was used to divide individual patients into groups in CART analysis. The final CART tree and result summary are given in section 4.5.3.

4.5.2 CART analysis with PORT

The software package PORT was developed over a period of time under the direction of Dr Arjan Shanhani by various people and the funding for this work

Chapter 4 Model for early detection of cervical cancer through cervical cancer screening programme

came from a company called Practical Insights Ltd and the software can be purchased from the GeoData Institute, University of Southampton. The main purpose of the software is CART analysis. PORT is a data analysis programme which creates classifications and provides a variety of statistical information about the elements in the classes. Within the software users can choose the tree-based algorithm or manually split the data into groups. Therefore, the data can be classified into groups according to certain characters or factors that the users have chosen. PORT is a data analysis programme which creates classifications and provides a variety of statistical information about the elements in the classes. An example is classification of people in various risk groups for a particular condition. Another example is the classification of hospital patients from the point of view of the length of stay in the hospital.

4.5.3 CART Results

The classification results can be used to increase knowledge about the predicted variable and help to predict an individual patient's needs and resource utilisation. The healthcare needs and corresponding resources vary from group to group. For example, given an individual patient, who can be classified into a patient subgroup based on the past experience and data, and the type of healthcare needs. Therefore, the results can be predicted based on the CART results. A more suitable healthcare decision can be made based on which group the patient belongs to. The CART results are listed in Figure 4.2 and Table 4.1.

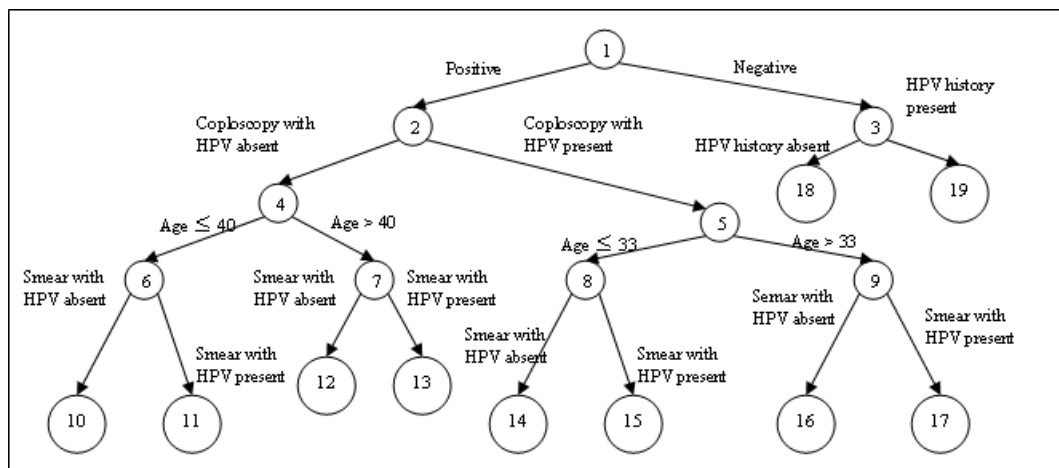


Figure 4.2. CART tree and number of risk groups for coploscopy results.

Chapter 4 Model for early detection of cervical cancer through cervical cancer screening programme

The coploscopy results (node 1) were divided into two groups, which are coploscopy positive (node 2) and coploscopy negative (node 3). Patients in node 2 were further divided into HPV absent or HPV present from the coploscopy results. A similar approach was applied to node 3; patients were divided into nodes 18 and 19, who had HPV history absent or HPV history present. Nodes 4 and 5 were divided further into nodes 6, 7, 8, 9 and there are 8 final nodes (nodes 10-17). So simply, it is possible to assume that nodes 10, 11, 12 and 13 represent one group, patients who had positive coploscopy and HPV absent, nodes 14, 15, 16 and 17 another group, who had positive coploscopy but HPV present, and nodes 18 and 19 are a final group who had negative colposcopy results.

The probability for each of the nodes was calculated as the percentage of patients within that group. Details of the calculation are listed (Table 4.1). For numerically independent variables (e.g. age of smear test), the minimum and maximum values are presented to indicate classification rules for forming these nodes. The final risk groupings were found as terminal nodes (nodes from which there are no further binary splits) and the terminal nodes are highlighted in Table 4.1.

Chapter 4 Model for early detection of cervical cancer through cervical cancer screening programme

Table 4.1 CART tree nodes summary for colposcopy results

Node	N	Field name	(%)	Minimum	Maximum	Deviance
1	1,639	All patients	---	---	---	2.93
2	1,534	Coploscopy positive	93.594% (1534/1639)	---	---	2.87
4	1,306	Coploscopy with non HPV	85.137% (1306/1534)	---	---	2.56
5	228	Coploscopy with HPV	14.863% (228/1534)	---	---	1.81
6	919	Coploscopy age \leq 40	70.368% (919/1306)	0	40	2.34
10	882	Smear with non HPV	95.974% (882/919)	---	---	2.33
11	35	Smear with HPV	3.808% (35/919)	---	---	1.81
7	387	Colposcopy age >40	29.632% (387/1306)	41	73	2.57
12	377	Smear with non HPV	97.416% (377/ 487)	---	---	2.57
13	10	Smear with HPV	2.584% (10/387)	---	---	1.70
8	174	Coploscopy age \leq 33	76.316% (174/228)	8	33	1.73
14	136	Smear with non HPV	78.161% (136/174)	---	---	1.59
15	38	Smear with HPV	21.893% (38/174)	---	---	1.79
9	54	Coploscopy age > 33	23.684% (54/228)	34	58	1.79
16	42	Smear with non HPV	77.778% (42/54)	---	---	1.77
17	12	Smear with HPV	22.222% (12/54)	---	---	1.36
3	105	Coploscopy negative	6.406% (105/1639)	---	---	0.05
18	104	History with non HPV	99.048% (104/105)	---	---	0.05
19	1	History with HPV	0.952% (1/105)	---	---	0.00

From the node tree (Figure 4.2) and node summary table (Table 4.1), it is apparent that overall 93.594% of patients had positive colposcopy results and only 6.406% of patients had negative results from 1998 to 2006 in Portsmouth. For the patients who had negative colposcopy results only 1% of the patients had a HPV history. Most patients had no HPV history (99%). Overall, 15% of patients who had positive colposcopy had HPV and 77% of them were below age 33. 85% of patients who had a positive colposcopy had no HPV and 70% of them were below age 40. Thus, over 66% (1,093/1,639) of positive colposcopy patients were under 40. This is not surprising as cervical cancer occurs more frequently in younger

Chapter 4 Model for early detection of cervical cancer through cervical cancer screening programme

women (e.g. 44% of cervical cancer patients were below age 45) (Moore-Higgs *et al.*, 2000).

The results show that age is one of the key factors associated with the colposcopy results. This is in accordance with previous studies which showed clearly that age and sexual behaviour are related to cervical cancer development (Moore-Higgs *et al.*, 2000 Bosch and Iftner, 2005). In addition, HPV was associated with the development of cervical cancer: HPV DNA was present in up to 99.7 percent of all cervical cancer cases, and infection with two common types of HPV (HPV 16 and HPV 18) accounts for more than 50 percent of all cervical cancer cases diagnosed each year (Goldie *et al.*, 2003).

The main interest is to understand how HPV presence or absence is related to positive colposcopy results, and to identify which types of patients are likely to have positive colposcopy results (e.g. HPV present or absent). Therefore, the colposcopy results were subdivided into a few subgroups. In practice, it is possible to group patients from nodes 10, 11, 12 and 13 into one group, (representing patients with positive colposcopy but HPV absent). Nodes 14, 15, 16 and 17 can be classified into a second group, (this group of patients had positive colposcopy and HPV present). Finally, nodes 18 and 19 can be classified into a third group (negative colposcopy and 99% of the patients had no HPV). From the CART results, patients from different groups may have different needs in terms of screening tests (e.g. screening interval) Different groups of patients may be assumed to have different probabilities of developing cervical cancer in the future and thus different diagnostic tests should be offered to different groups of patients. Such information can be utilised in decision tree and simulation models to estimate healthcare capacity and resources.

4.6 Decision tree model

Some studies demonstrated that the type of diagnostic tests (e.g. LBC, HPV) and the combinations of tests relate to the effectiveness and the cost of the cervical cancer screening programme (Jenkins *et al.*, 1996; Sherlaw-Johnson and Gallivan, 2000; Sherlaw-Johnson and Philips, 2004). In this section, a decision tree model was developed to evaluate screening options for the UK national cervical cancer screening programme. Details and definition of each test can be found from Chapter 1, section 1.4.2. Various options can be evaluated using a sequence of three decisions:

- (i) Current policy: no HPV test at all stages, thus, under the general decision tree model, the no HPV test option would be evaluated by using zero as the probability of a HPV test.
- (ii) HPV test for mild, moderate, severe dyskaryosis and severe dyskaryosis/ suspected invasive cancer results only.
- (iii) No colposcopy tests without HPV tests when the initial test results are abnormal.

The structure of the necessary decision tree link to current cervical cancer screening policy is displayed in Figure 4.3. A decision node, represented by a square, indicates that a decision needs to be made at that point in the process. A chance node, represented by a circle, indicates that a random event occurs at that point. There is a decision node for the i th regular smear test. Other decision nodes are for decision, HPV test or an urgent smear test. Before starting to evaluate the screening options some information is needed: the number of women and probabilities of each possible outcome. The percentage and cohort size used in this chapter is a dummy dataset, which is due to lack of available information, but it demonstrated the possible use of the decision tree and the potential benefit of evaluating screening options.

Chapter 4 Model for early detection of cervical cancer through cervical cancer screening programme

Having constructed the decision tree, it is now possible to analyse the screening options. Firstly, assume that there are 100,000 patients to start with. The decision tree model was developed by using Visual Basic Application (VBA) within Excel. The analysis processes follow the following steps.

1. Start at the left side of the decision tree and move right one column at a time. For each column, perform either step 2 or step 3 depending upon whether the nodes in that column are chance nodes or decision nodes.
2. For each chance node, calculate the expected values,

$$\text{Expected value} = \sum_i^n p_i x_i \quad (4.4)$$

Where p_i is the probability of x_i and n is the number of possible outcomes. The probability can be estimated from data, if data are available. The decision with the best outcome (i.e. minimum number of tests) can be accepted as the appropriate decision.

3. For each decision node, compare the expected values of its branches and choose the alternative whose branch has the largest or smallest expected values (it depends on the problem and, specifically whether the users want to obtain the maximum or minimum value).

4.6.1 Model for evaluating cervical cancer screening options

From the model (Figure 4.3), users can evaluate the number of smear tests, urgent smear tests, HPV tests and colposcopy tests needed for different screening options. Within the decision tree model, three risk groups were defined, (i) low risk, (ii) medium risk and (iii) high risk. The risk groups can be defined by the users. Three groups of patients (not an individual) were followed through the whole screening system. Therefore, patients within the same risk groups were assumed to be homogeneous. Everyone within the

Chapter 4 Model for early detection of cervical cancer through cervical cancer screening programme

same group was assumed to have the same chance of moving from one node to the others.

The decision tree structure (Figure 4.3), based on VBA, allows users to estimate the total number of smear tests, HPV tests and Colposcopy tests required for each group, and each option. The basic logic within the decision tree model is explained in Figure 4.4. The number of patients in each node (Figure 4.3) was equal to the percentage for that node multiplied by the number of patients in the previous node.

Chapter 4 Model for early detection of cervical cancer through cervical cancer screening programme

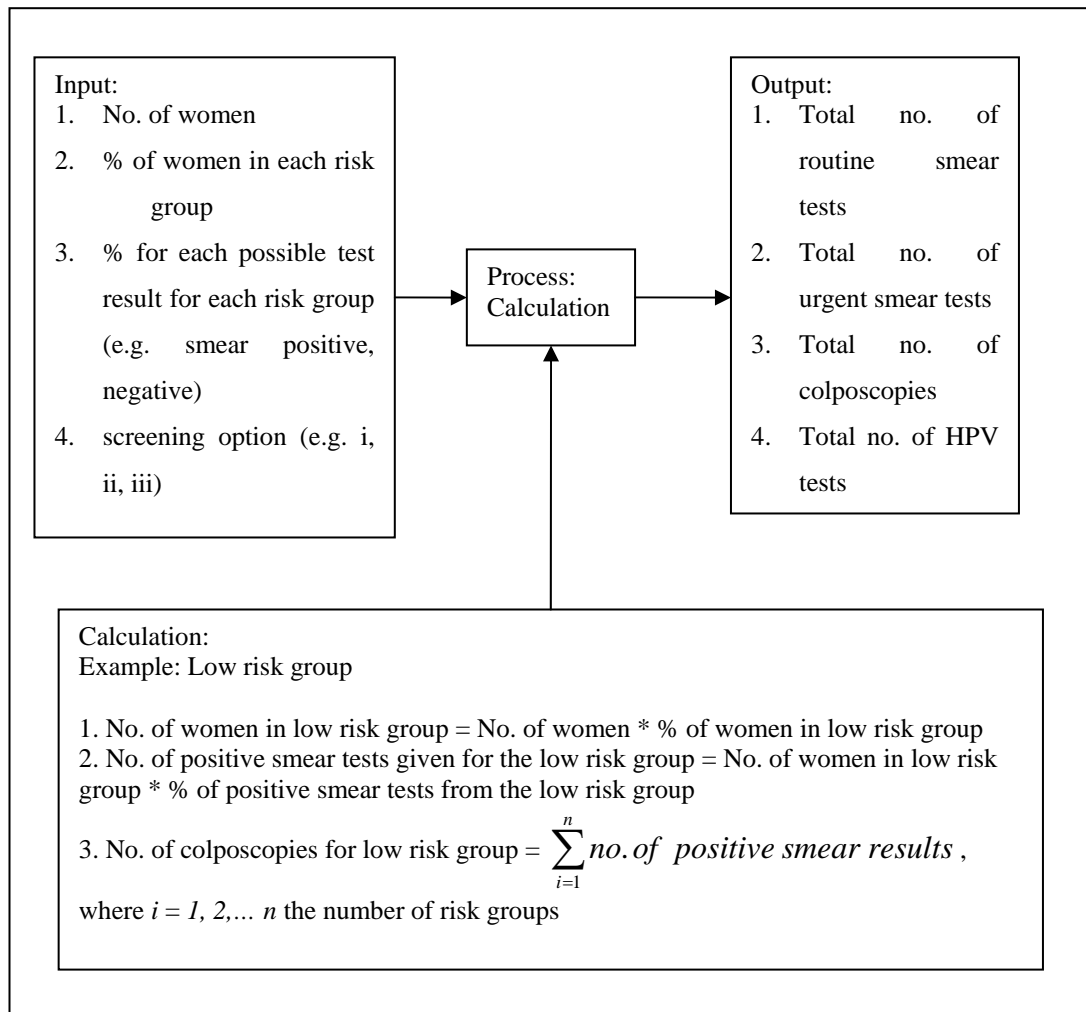


Figure 4.4 Schematic diagram showing the basic calculation within the decision tree model.

The user needs to enter the number of women (cohort size), the proportions belonging to each risk group from the total population, and the probabilities for each of the smear test results (positive, negative and inadequate); abnormal results (e.g. mild, moderate, severe dyskaryosis and cancer) and HPV test results. Details of the decision tree model within Excel are attached in Appendix C.

4.6.2 Decision tree results

Different events have different percentage values; for example, 80%, 15%, and 5% of the initial tests of the cohort of women of medium risk group will yield negative, abnormal, and inadequate results. Further, 50%, 35%, 10%

Chapter 4 Model for early detection of cervical cancer through cervical cancer screening programme

and 5% of the women who had abnormal results will have mild dyskaryosis, moderate dyskaryosis, severe dyskaryosis, or severe dyskaryosis/suspected invasive cancer. At any stage, the various percentages add up to 100%. In the decision tree model, “initial smear test” means the first smear test when a woman enters the national cervical cancer screening programme. Urgent smear test is used to represent a necessary smear test after the initial smear test. For example, if a woman has an inadequate smear at the initial test, and needs to take another smear test within a short period; this is classified as an urgent smear test. Table 4.2 shows the probabilities of each possible smear test result of various risk groups and the general population. The detailed calculations are attached in appendix C.

Table 4.2 Dummy dataset of risk groups, a set of dummy data were used to explain and demonstrate the use of decision tree model, due to the lack of available data.

Risk groups	Negative smear	Abnormal	Inadequate
Low	90%	5%	5%
Medium	80%	15%	5%
High	60%	35%	5%
General population	77.5%	17.5%	5%

The percentages used in this chapter represent a dummy dataset, since this information is not available. However, the dataset demonstrated the possible use of the decision tree and potential benefit for evaluating screening options. The decision tree model was used to evaluate a variety of options. Assume there is a cohort of 100,000 women, with 50%, 25% and 25% of the population from low, medium and high risk groups respectively with the general population percentage of 77.5%, 17.5 and 5% for negative, abnormal and inadequate results at the initial smear test. There were three options listed in section 4.6; current policy, HPV test and no HPV test. In the general decision tree model, the no HPV test option would be evaluated by using zero as the probability of a HPV test.

Chapter 4 Model for early detection of cervical cancer through cervical cancer screening programme

Results are shown in Table 4.3. The various options require different numbers of tests (e.g. smear, urgent smear, HPV and colposcopy). Option 1 (12,966) required the least number of tests and option 2 (132,681) required the largest number of tests in total. However, option 3 provided the largest percentage of positive colposcopies, where the percentage of positive colposcopy tests is equal to the total number of positive colposcopy tests divided by the total number of colposcopy tests. In terms of the efficiency of positive colposcopy tests, option 3 provides the largest percentage with the smallest number of colposcopy tests in total. Therefore, option 3 is the most effective option in this situation, given that we want to minimise the number of colposcopy tests. When the probabilities change the final option may change.

Table 4.3 Decision tree results for different screening options.

Summary	Initial smear test	Urgent smear test	HPV test	Colposcopy	Positive colposcopy	Total tests (smear, HPV and colposcopy)
Option 1	100,000	14,000	0	15,663	8,466 (54%)	129,663
Option 2	100,000	7,875	11,900	12,906	6,829 (52.9%)	132,681
Option 3	100,000	7,875	11,900	9,336	6,199 (66.4%)	129,111

It can be seen that, in this simple example, option 3 is the most effective, because it provides the largest percentage of positive colposcopy tests (66.4%).

The percentage information can be changed within the built-in interface, where users need to enter the percentage into the interface before starting to run the model.

4.7 Simulation model

Simulation is a widely used and effective tool for the analysis of complex systems, and also to manage the screening system. Examples of simulation application include the prevention of and treatment for diabetic retinopathy and HIV studies (Harper, 2002; 2003). Some simulation studies have demonstrated that human behaviour may affect the healthcare results and a simulation can capture and describe such factors as human behaviour in the model (Brailsford *et al.*, 2006). Other research has shown that the simulation model can provide a way of investigating the needs of NHS services at different geographical locations (Harper *et al.*, 2005). This is because of the complexities involved; it is not easy to provide an accurate model to capture the real situation. An alternative approach is to use simulation. Simulation models can also be used to forecast workforce and resource needed (Powell and Harper, 2004; Harper *et al.*, 2005)

Simulation may be defined as a technique that imitates the operation of a real world system as it evolves over time. A simulation model usually takes the form of a set of assumptions about the operation of the system, expressed as mathematical or logical relations between objects of interest in the system. In particular, when modelling a patient's pathway, similar patients can be grouped into sub-groups according to their behaviour, so that the focus is on individual patients passing through the healthcare system. The simulation process involves executing or running models through time on a computer, which generates representative samples of the measures of performance. In this respect, simulation may be seen as a sampling experiment on the real system, with the results being sample points.

In most simulation studies, users are concerned with the simulation of a system or a particular part of that system at a particular point in time. Thus, in order to model a system, users must understand the concept of a system. The patient-flows will be simulated over time, and each of the patients will be followed through over the whole simulation period. The simulation results allow us to understand the patient pathway and identify any problems

Chapter 4 Model for early detection of cervical cancer through cervical cancer screening programme

within the current system as well as estimate the capacity and resources required.

There are two common model frameworks, which are the Markov and Semi-Markov models. When time parameters are not available the Markov model is suitable to describe patients' activities; otherwise, the Semi-Markov model can be used. The Markov and Semi-Markov models are discussed in the next two sub-sections. Details of the simulation model are attached in Appendix D.

For the pre-cervical cancer process, the parameter space T is countable (e.g. healthy, CIN 1, CIN 2, CIN 3, cancer and death). Thus, for the pre-cervical cancer process the Markov chain with discrete parameter state is the most appropriate. In this section, a Markov chain model will be fitted to describe the pre-cervical cancer process. The simulation is shown in Figure 4.5 (Goldie *et al.*, 2003).

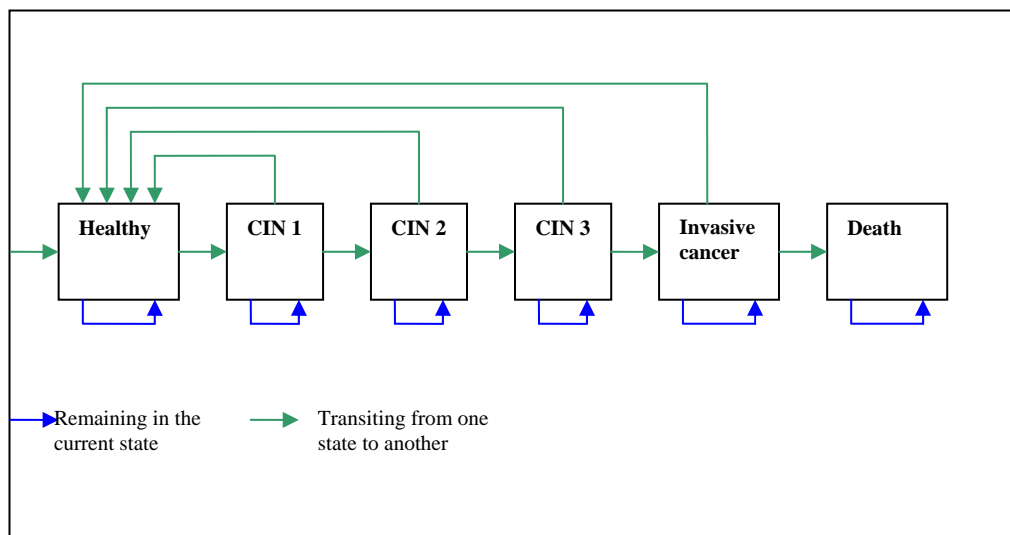


Figure 4.5. Cervical cancer disease model, describing how each healthy female could flow through the process (e.g. from health to CIN 1).

To model the pre-cervical cancer disease process, the Markov chain may not be an ideal choice, because of the memory-less property and the fixed distribution assumptions. The disease growth process does depend on the

Chapter 4 Model for early detection of cervical cancer through cervical cancer screening programme

previous states (more specifically, the present disease state depends on the previous disease state or states). Therefore, an alternative model needs to be introduced, called the semi-Markov chain. If the information on holding time is available for the above model (Figure 4.5), it is possible to construct a semi-Markov chain model. The model parameters and probabilities function applied in this section are a dummy dataset, which is due to lack of available information, but again it demonstrated the possible use of the simulation models.

4.7.1 Simulation results

The simulation model was developed in VBA within Excel; it demonstrated the possible use of the Semi-Markov model, when transition probabilities and holding time are available from clinical data. However, this information was not available at the time of writing; therefore, a set of dummy distributions were used in the simulation model.

Users are required to enter some information to start simulating a group of patients' flows through the disease system; e.g. simulation period in weeks, distribution for describing the chance of developing abnormal cells (e.g. Binomial), and distribution of length of stay in each state (e.g. Weibull). A set of simulated results is shown in Table 4.4.

Table 4.4 Summary of simulation results.

Information	Simulated Results
Simulation period	900 weeks
Simulated population	90,000
Number of patients with positive smear results	9,034
Length of stay in CIN 1	21-33 weeks
Length of stay in CIN 2	14-21 weeks
Length of stay in CIN 3	6-11 weeks

4.8 Summary

CART was used to divide the population into risk groups of developing cervical cancer; each individual person within the same group shares the same common characteristics (i.e. personal risk factors). Those characteristics determined the chance of developing cervical cancer in the patients' life time. Patients from different risk groups may have different probabilities of developing cervical cancer. By identifying the risk factors it helps in understanding which factors are related to cervical cancer.

Decision tree models help in evaluating screening options, when the policy makers face the situation of selecting the option which can maximize the returns (i.e. detecting more cases at the early pre-cervical cancer states) and minimize the number of tests. Such a method allows examination of which is the best option to increase the efficiency of the programme. Simulation allows understanding the natural disease process; it also allows to evaluate screening options if transition probabilities and holding times are available.

Therefore, all the above techniques demonstrated the possible way to identify the patient's common risk characters with different probabilities in developing cervical cancer in their life time. The combinations of both personal and national risk factors provide greater understanding of the associated risk factors of cervical cancer development.

Chapter 5 Cervical cancer regression study

5.1 Introduction

The cervical cancer screening programme in the UK was set up in 1988; the screening programme works effectively to reduce the risk of cervical cancer through detection of early pre-cancerous stages. In this context, it would be interesting to understand the relations between cervical cancer disease risk, and deprivation, social status and family structure factors. Knowledge of such relations may be of use in planning screening programmes to reduce risk and to target the necessary resources in order to increase the efficiency of the screening programme. Within the UK there is marked geographical variation in the cervical cancer incidence rate. This chapter demonstrates that individuals may have different risks of developing cervical cancer as a function of various personal (e.g. age, social status) and family structure variables and also the surrounding deprivation conditions. Certain geographically varying deprivation, social grade and family structure factors can provide valuable information about human behaviour and this behaviour may have implications for the likelihood of developing cervical cancer. The cervical cancer data were provided by the UK's Public Health Observatories (PHO). The deprivation, social grade and family structure variables were provided by the UK census 2001. Details of the data were given in Chapter 3.

The Poisson distribution is the most appropriate distribution to describe the underlying disease distribution for rare diseases (Richardson, 2003; 2004). Three types of regression model were applied in this chapter; the preliminary stage of this chapter used (i) generalised linear modelling to investigate the relationship between cervical cancer and indicator of social deprivation, social grade and family structure factors across England at the global level; (ii) a Bayesian hierarchical model was used to model the relationship between cervical cancer incidence and the same variables, and (iii) geographically weighted regression (GWR), was used to analyse the locally varying relationship between cervical cancer incidence, each variable was modelled as non-stationary.

Based on the results, it is possible to see, at least in principle, how the national screening programme could be made more efficient and effective by adapting deprivation indicator, social grade and family structure factors locally as a function of dividing the population into a number of risk groups according to the national and (or) personal common characteristics. For example, in Chapter 4, CART was used to divide patients into risk groups according to their personal characters. Combining both personal and national characters (i.e. risk factors) allows an understanding of potential risk factors, which are associated with cervical cancer development in a woman's life time.

The aim of this chapter was to use various regression models to examine the relationship between cervical cancer incidence and explanatory variables. The specific objectives of this study were (i) to explore geographical variation in cervical cancer incidence across England through various regression models at the district and unitary authority levels (where a district is smaller than a unitary authority) and (ii) attempt to explain the geographical variation through the explanatory variables.

5.2 Data

Two sets of data were included for analysis; cervical cancer count data for 2004 and explanatory data for 2001. The data are represented at district and unitary authority levels in England. A total of N regions was considered, where $N = 354$. For each region $i = 1, 2, \dots, N$ cells, let Y_i be the incidence (i.e., number of cervical cancer cases) and e_i be the expected cases for region i . However, any observed data between zero to five cases were closed to protect the patients. Therefore, such missing data were treated as truncated missing data between 0 to 5. The methods to overcome the missing data are discussed in detail in each of the regression sections. The expected number of incidence cases is around 21.56 cases per region. At the beginning of this study the data were subjected to a Chi-square goodness of fit test, which showed that the data approximately followed a Poisson distribution.

The expected cases were calculated by using the normalized incidence rate r_j per age group from the Cancer Research UK data (Cancer Research UK 2005). This rate was normalized by multiplying by the ratio between total cervical cancer cases (the data used here) and new diagnosed cases (~2.4) the data used in the Cancer Research UK rates. The normalized rate was then multiplied by the female population p_{ji} within that age group in region i , where j is the age group. The female population per age group was determined from the 2001 UK Census of those aged between 0 and 85 and over. The expected cases were calculated as shown in equation 5.1.

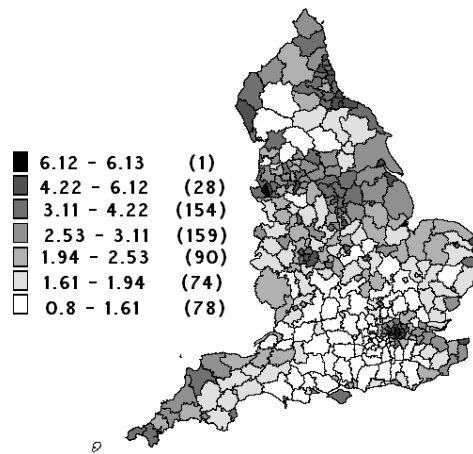
$$e_i = \sum_j r_j p_{ji} \quad \text{where } j=1, \dots, 85+, i = 1, \dots, 354 \quad (5.1)$$

where j is the age group (e.g. age 0 to 4, 5 to 9 ... 85 and over) and i is the number of study regions.

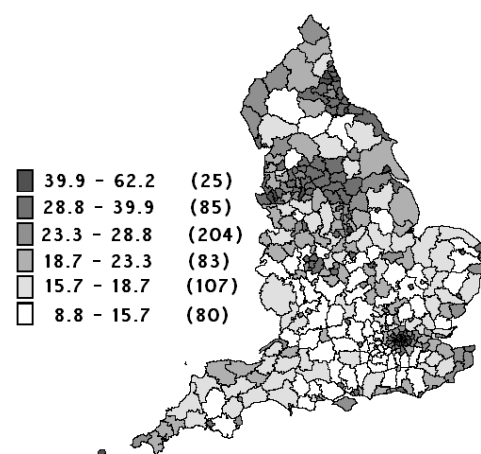
The Townsend index was constructed using four socio-economic variables: (i) unemployed population, (ii) household with no car/van, (iii) households not owned, and (iv) over-crowded housing (over one person sharing one room). The index was used as an individual variable in the regression models and other variables representing family structure and social grade were applied in the regression models. Details of data can be found in chapter 3. All explanatory variables are represented as proportions of the total population. The maps of each Townsend, social grade and family structure variables are displayed in Figure 5.1.

Figures 5.1a and 5.1b show greater percentages of unemployment and households without a car in the Midlands and the North of England. Figure 5.1c shows a less spatially structured pattern for percentage of households not owned while Figure 5.1d shows that some of the major cities (e.g. London, Manchester) have a greater density of people per room with potential overcrowding problems. Figure 5.1e shows a high percentage of lower social grade population in North and Midland areas. Figure 5.1f and 5.1g show more single population found in major cities and higher percentage of married population found in the rural areas. Figure 5.1h and 5.1i have very similar patterns, because a large proportion of lone parents are female lone parents.

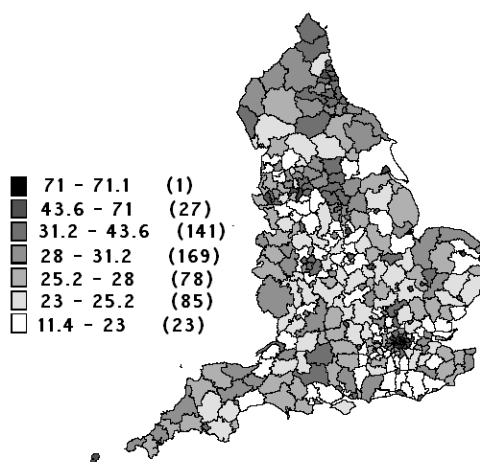
(a)



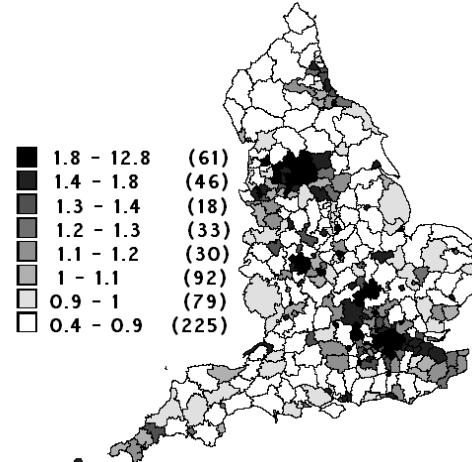
(b)



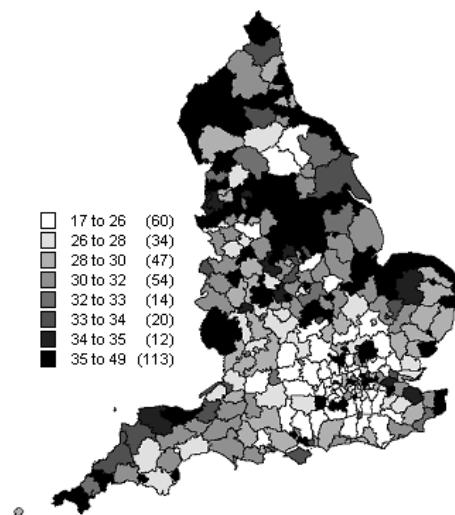
(c)



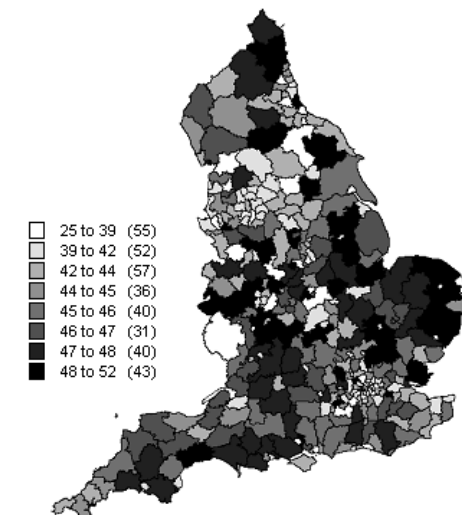
(d)



(e)



(f)



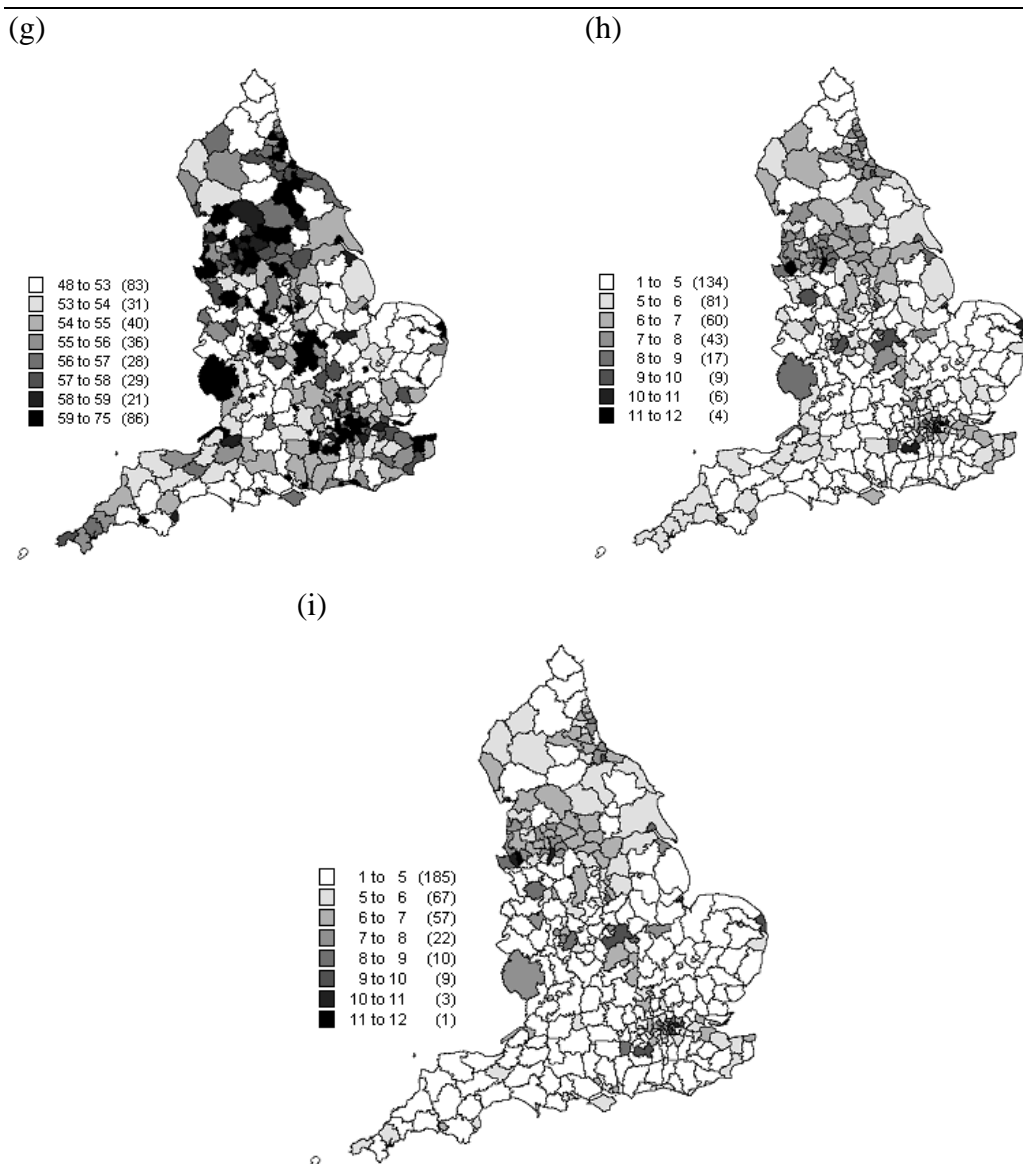


Figure 5.1. (a) percentage unemployed, (b) percentage of households with no car, (c) percentage of households not owned and (d) percentage of rooms occupied by more than one person, (e) percentage of socio-grade VI and V, (f) percentage of married female population, (g) percentage of single female population, (h) percentage of all lone parents households, and (i) percentage of female lone parents households.

5.3 Analysis

5.3.1 Exploratory analysis

Figure 5.2 show scatter plots between incidence and expected cases. While the relations between the observed and expected cases are clearly linear, there is a fair degree of scatter indicating that some variables other than expected cases or population may be affecting incidence. The two points at the extreme top-right of the plots represent Birmingham and Leeds.

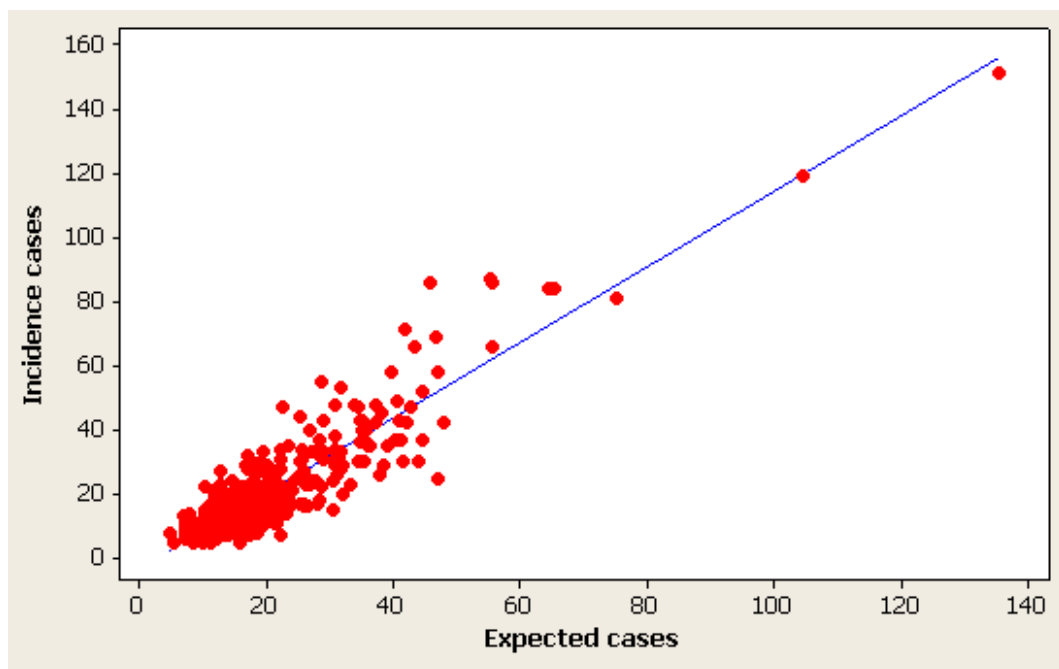


Figure 5.2 Incidence plotted against the expected cases for England 2004.

It is helpful to estimate the correlations between incidence and explanatory variables, which show how all variables are correlated together and whether the relationships are positive or negative. Table 5.1 and Figure 5.3 show the correlation values between incidence and explanatory variables. Most of the variables are positively correlated and proportion of married population is negatively correlated to incidence cases, apart from the Townsend index which is less correlated to other variables.

Chapter 5 Cervical cancer regression study

Table 5.1 Correlations between incidence cases and Townsend index score, low socio-grade and family structure proportion variables.

Correlations (p-values)	Incidence cases	Townsend index score	Proportion of female married population	Proportion of female single population	Proportion of all lone parent household	Proportion of female lone household
Townsend index score	0.386 (0.000)					
Proportion of female married population	-0.418 (0.000)	-0.841 (0.000)				
Proportion of female single population	0.418 (0.000)	0.841 (0.000)	-1.00 (0.000)			
Proportion of all lone parent household	0.523 (0.000)	0.639 (0.000)	-0.714 (0.000)	0.714 (0.000)		
Proportion of female lone household	0.523 (0.000)	0.646 (0.000)	-0.720 (0.000)	0.720 (0.000)	0.998 (0.000)	
Proportion of G4 + G5	0.401 (0.000)	0.331 (0.000)	-0.307 (0.000)	0.307 (0.000)	0.685 (0.000)	0.680 (0.000)

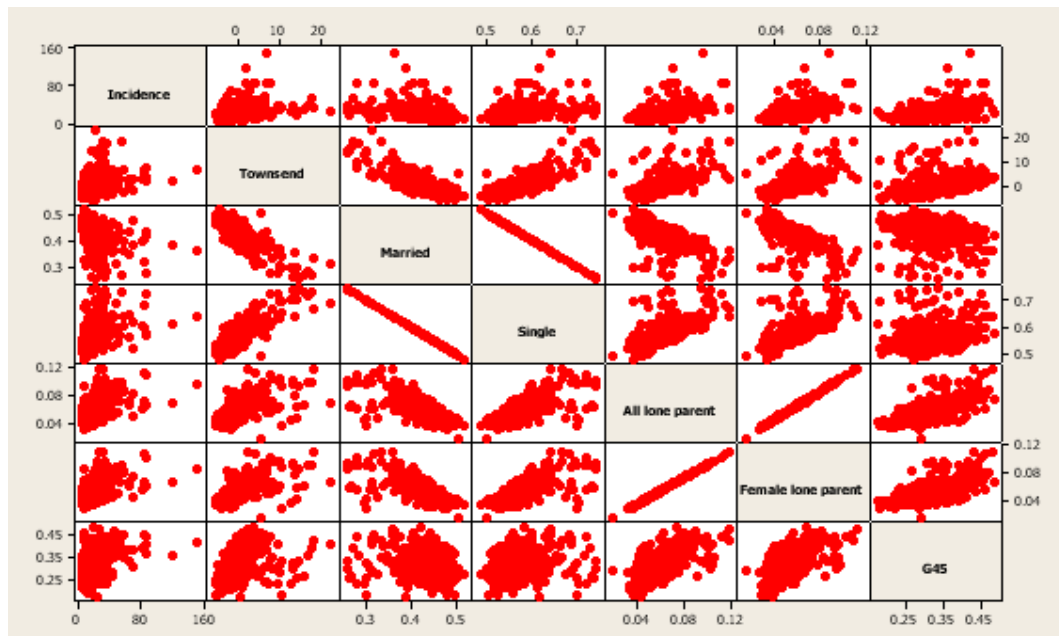
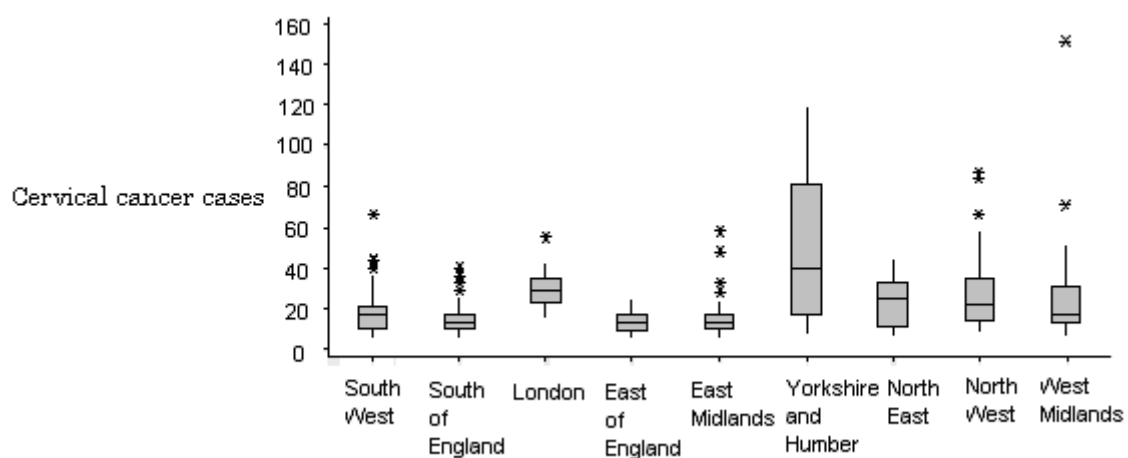


Figure 5.3 Matrix plot showing the relationships between variables.

The box-plots in Figure 5.4 show the variation within and between PHOs in incidence and mortality measured at the district or unitary authority level. The variables plotted are purely the number of incidence cases and death cases, with no information provided on the age structure or population size. Therefore, the plots need to be interpreted with care. The number of districts (or unitary authorities) per PHO is shown in Table 3.1. From Figure 5.4, it can be seen that the Yorkshire and Humber PHO had a large median and inter-quartile range compared with other PHOs.

(a)



(b)

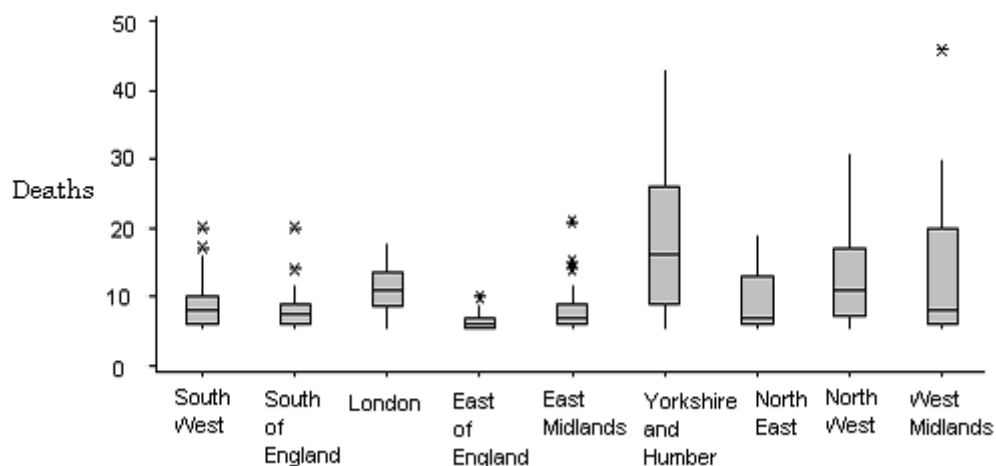


Figure 5.4. Box-plot of cervical cancer (a) incidence and (b) mortality per PHO, England 2004.

5.3.2 Townsend index

The Townsend index was applied again in this section and details of the calculation can be found from Chapter 2. Figure 5.5 shows the Townsend index for England in 2001. The Townsend index was used directly in all regression models as an individual explanatory variable,

However, there are some drawbacks of using Townsend index, as follows:

- (i) It is not possible to distinguish between individuals who are not able to buy a car or those who do not need a car. For example, those who live in a main city (e.g. London) may not need to have a car, since public transportation is more convenient,
- (ii) It is also not possible to distinguish between individuals who are not able to buy a property and those who do not want a property. It is more common for people who live in a main city to rent a house than buy a house.

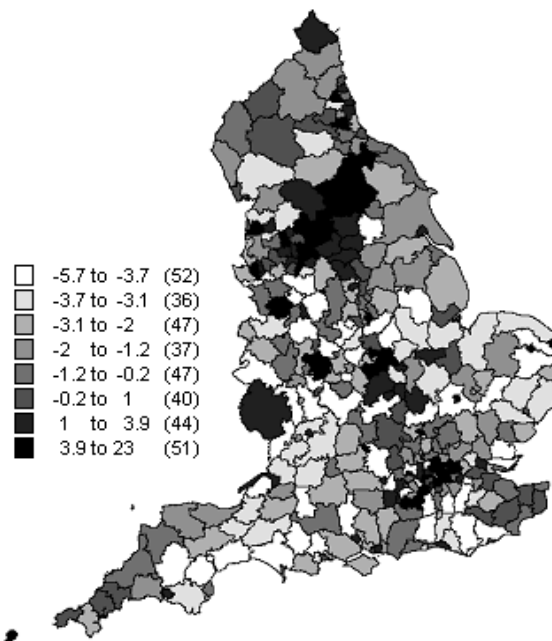


Figure 5.5. Townsend index, England 2001. North of England and Midlands have a larger Townsend index than the south. Southern areas in particular have a smaller Z. However, the areas near London have higher deprivation which is due to the main drawback of using a deprivation index.

5.4 Generalised linear regression model (GLM)

5.4.1 Truncated missing data

The 21 regions were truncated, which were treated as missing data (i.e. NA). In this section, all NA data were excluded from the GLM model. The more technical methods to overcome the truncation missing data are described in the next two sections of this chapter.

5.4.2 Methods

Various types of regression techniques are available to model the relations between variables. Generalised linear modelling (GLM) is a popular statistical modelling tool for exploring relationships between the observed (cervical cancer incidence) and explanatory variables (Gatrell and Bailey, 1996; Elliott *et al*, 2000). The outcomes from the models provide a summary of the complex geographical relations that may exist.

For the study of cervical cancer, it was assumed that the data (incidence) represent a set of observed counts arising from a Poisson process (i.e., the data Y_1, Y_2, \dots, Y_N in regions 1, 2, ..., N are mutually independent Poisson random variables). In addition, the population counts for each region are assumed fixed (i.e. non-random variables), denoted p_1, p_2, \dots, p_N and r_i denotes the assumed constant risk (McCullagh and Nelder., 1952). The relevant simple Poisson regression equation with offset variable is given (McCullagh and Nelder, 1952; Nakaya *et al.*, 2005).

$$Y_i \sim \text{Poisson}(\mu_i)$$

$$\hat{Y}_i = e_i \exp(\beta_0 + \sum_{t=1}^T \beta_t v_{it} + \varepsilon_i) \quad \text{For } i = 1, 2, \dots, N \quad (5.2)$$

where Y_i is the observed value at location i , e_i is the expected cases, β_0 is a constant term (i.e. intercept), β_t measures the relationship between the observed and explanatory variables, the v_{it} are the explanatory variables, and ε_i is the error term associated with location i , which is independent and normally distributed with mean of 0 and variance σ_i^2 .

The Chi-square test provides a method to test the level of significance between the observed cases and the Townsend index and explanatory variables. Hypotheses may be set up such as to test the association between the observed and explanatory variables. The test statistic is given below:

$$\chi^2 = \sum_{i=1}^N \frac{(Y_i - \hat{Y}_i)^2}{\hat{Y}_i} \quad (5.3)$$

Where \hat{Y}_i is the estimated cases per region, $i = 1, 2, \dots, N$.

The observed variable was:

1. Incidence, NA was excluded

The explanatory variables were:

1. Townsend index score
2. Proportion of female married population
3. Proportion of female single population
4. Proportion of all households with a lone parent
5. Proportion of households with a female lone parent
6. Proportion of Social grade IV and V population

Each individual explanatory variable was used in GLMs, to predict observed variable. The most significant variable was added into each model at each step, until the p -value was larger than the significance level (i.e. 0.05), in which case, the last variable was removed and the final model accepted.

5.4.3 Generalised linear modelling

Incidence cases and observed indirectly standardised incidence ratio (SIR) (with excluded missing data) were mapped in Figure 5.6, the cervical cancer incidence cases and SIR vary with location. For example, East Lindsey, North East Lincolnshire, Hartlepool, Sedgefield and Barrow-in-Furness had the largest cervical cancer incidence in 2004. Townsend index and other explanatory variables were investigated as explanatory variables within a GLM framework.

Firstly, the explanatory variables were treated in isolation, entered as a single explanatory variable into the GLM. Ten GLM models involving six explanatory variables were fitted (Table 5.3). Secondly, incidence was fitted with a single GLM as described in Table 5.3. The results suggest that all single variables were significantly related to the observed variables.

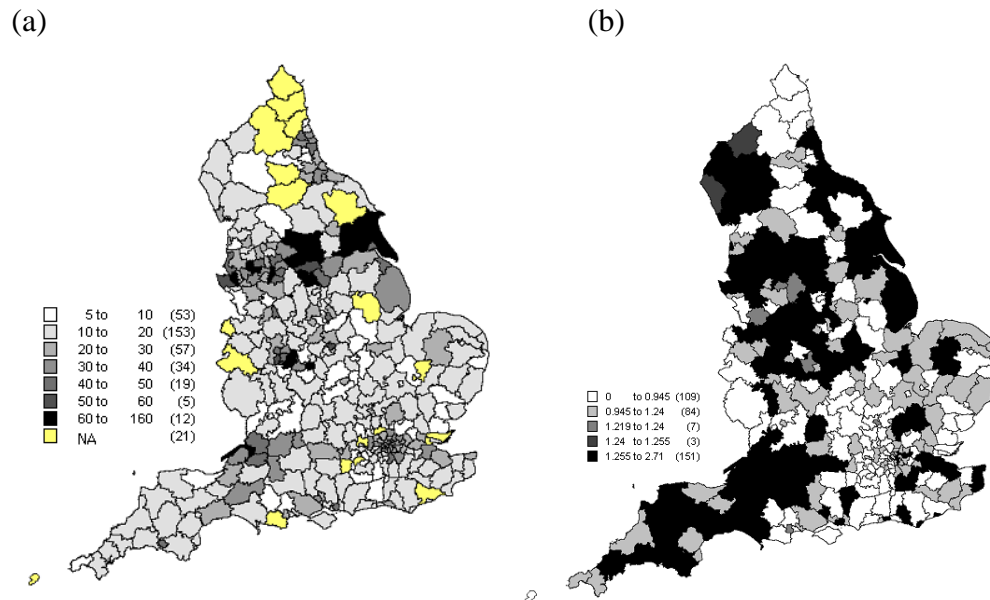


Figure 5.6. (a) Observed cervical cancer cases per region. The regions highlighted in yellow represent regions with missing data, (b) observed SIR per region. A few places have large incidence (e.g. West Devon, West Wiltshire, Sedgfield)

Chapter 5 Cervical cancer regression study

Table 5.2 (a) Summary of GLM (b) the intercept and coefficient values for each explanatory variable, where the regions with missing data were excluded from the modelling.

(a)

Model	Variables	Std.Error	t-value	P-value
1	Townsend index score	0.002285586	2.717489	0.006983208
2	Proportion of female married population	0.21514951	-3.152812	0.001723092
3	Proportion of female single population	0.2151495	3.152812	0.001723092
4	Proportion of household with lone parent	0.61764346	10.248450	0
5	Proportion of households with female lone parent	0.64785344	10.035571	0
6	Proportion of low social grade (IV+V) population	0.17947345	15.29755	0
7	Proportion of low social grade (IV+V) population + Townsend index score	0.19102757	15.135210	0
		0.00254778	-2.263933	0.02246866
8	Proportion of household with female lone parent + Townsend index score	0.877438931	10.914176	0
		0.003206611	-5.110045	2.015673e-007
9	Townsend index score + proportion of female married population	0.004338032	0.07847839	0.0069832
		0.407034440	-1.59990279	0.1101383
10	Townsend index score+ proportion of female single population	0.004338032	0.07847839	0.0069832
		0.407034440	1.59990279	0.1101383

(b)

Model	Variables	β_0	β_1	β_2
1	Townsend index score	0.233497811	0.006211054	
2	Proportion of female married population	0.5236252	-0.6783260	
3	Proportion of female single population	-0.1547008	0.6783260	
4	Proportion of household with lone parent	-0.1713738	6.3298879	
5	Proportion of households with female lone parent	-0.1424144	6.5015793	
6	Proportion of low social grade (IV+V) population	-0.6820043	2.7455034	
7	Proportion of low social grade (IV+V) population + Townsend index score	-0.722888707	2.891242526	-0.005768002
8	Proportion of household with female lone parent + Townsend index score	-0.30106881	9.57652309	-0.01638593
9	Townsend index score + proportion of female married population	0.5118985113	0.0003404418	-0.6512155378
10	Townsend index score + proportion of female single population	-0.1393170265	0.0003404418	0.6512155378

The final fitted models were models 1 – 6 in Table 5.2 and Figures 5.7 to 5.12. Models 7 and 8 in Table 5.2 were not accepted as the best fitted models, because some of the coefficients changed from positive to negative. For example in model 7 the coefficient of Townsend index score changed to negative. However, prior expectation would be that the relation with Townsend index to be positive. Further, the relationship between observed cases and Townsend index is positively related in Table 5.1. For these reasons models 7 and 8 were not accepted as the best fitted model. For models 9 and 10 the P-value are larger than 0.05 (5%). Therefore, the models were not accepted. Figures 5.7 to 5.12 show diagnostic statistics for each of the significant GLM models and the maps highlight that the North of England, Midlands and West of England had larger SIR from the fitted models. Some regions with larger Townsend index scores (greater deprivation) also had higher predicted SIR.



Figure 5.7. (a) Predicted SIR from model 1 in Table 5.2, (b) incidence plotted against fitted value and (c) normal plot for GLM predicting incidence when the missing data are excluded from the modelling.

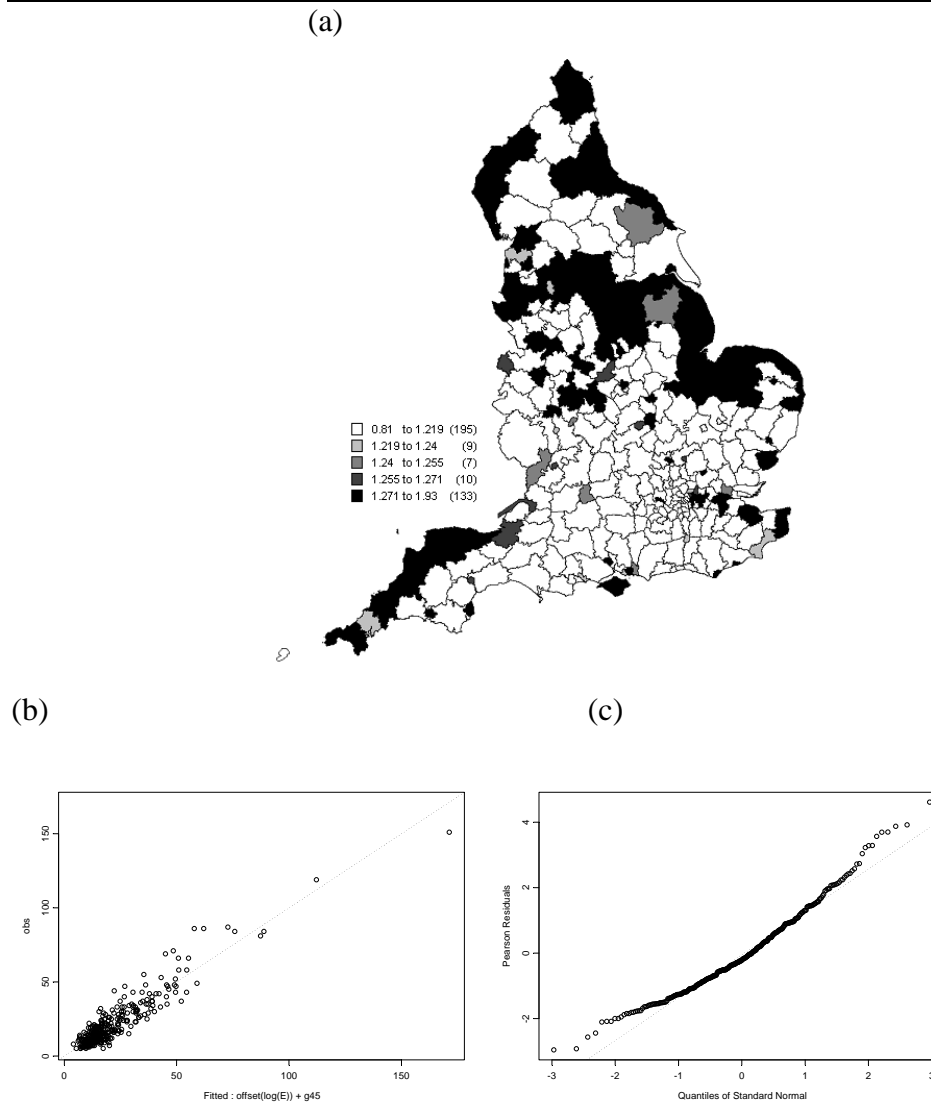


Figure 5.8. (a) Prediction SIR from model 2 in Table 5.2, (b) incidence plotted against fitted value and (c) normal plot for GLM predicting incidence when the missing data are excluded from the modelling.



Figure 5.9. (a) Prediction SIR from model 3 in Table 5.2, (b) incidence plotted against fitted value and (c) normal plot for GLM predicting incidence when the missing data are excluded from the modelling.



Figure 5.10. (a) Prediction SIR from model 4 in Table 5.2, (b) incidence plotted against fitted value and (c) normal plot for GLM predicting incidence when the missing data are excluded from the modelling.

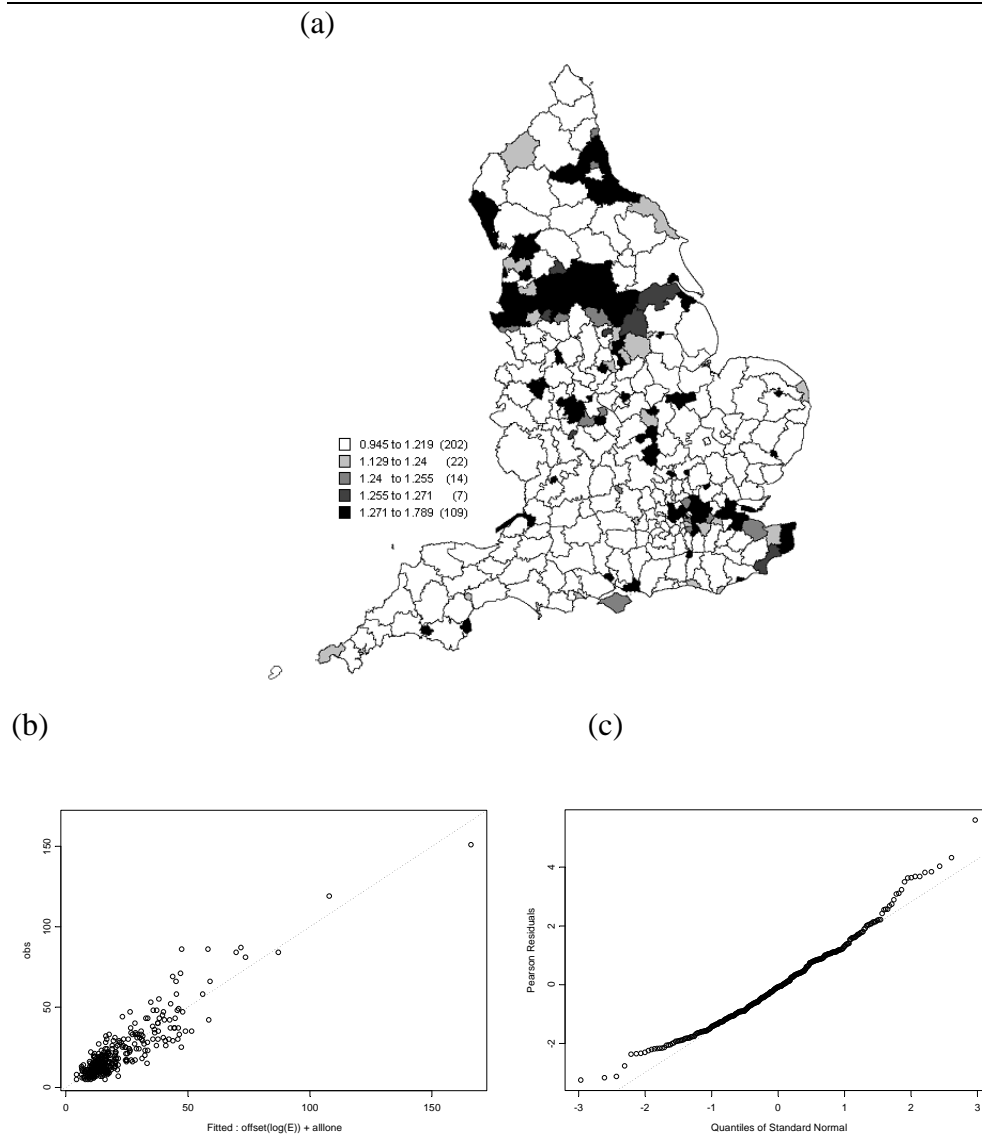


Figure 5.11. (a) Prediction SIR from model 5 in Table 5.2, (b) incidence plotted against fitted value and (c) normal plot for GLM predicting incidence when the missing data are excluded from the modelling.

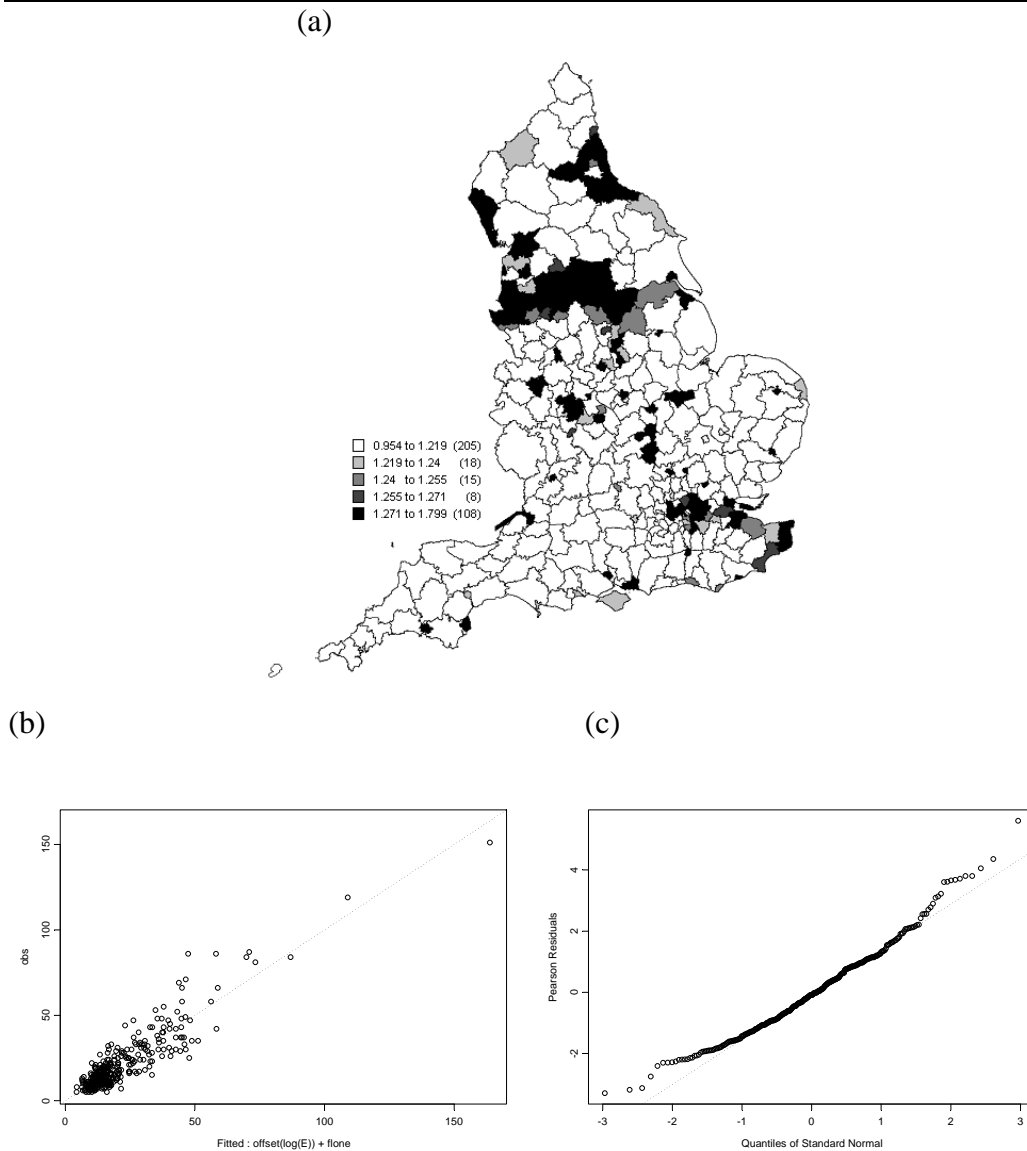


Figure 5.12 (a) Prediction SIR from model 6 in Table 5.2, (b) incidence plotted against fitted value and (c) normal plot for GLM predicting incidence when the missing data are excluded from the modelling.

GLM results showed that at the global level there is a significant relation between cervical cancer, Townsend index, proportion of social grade IV and V, proportion of single female, married female population and proportion of household with lone parent. The regions with higher Townsend index, larger proportion of population at low social grade, larger population of single female population, and also larger proportion of lone parent is likely have higher incidence rate. However, for the proportion of female married population is the other way around, when the proportion of female married population increase the incidence rate decrease. Such summary showed the chance of developing cervical cancer related personal marital status It would be more interesting to investigate further if

spatial information was taken into account, thus, in the next section the Bayesian approach is used to predict the incidence rate when the uncertainties are taken care of using prior distributions. Spatial information will be adopted in the regression models; therefore, it allows the spatial pattern and the spatial variation in cervical cancer incidence to be studied. As the literature showed deprivation is highly related to chance of developing disease (Rees *et al.*, 2002).

5.5 Bayesian regression model

Bayesian modelling has been demonstrated to be useful for analysing many types of complex epidemiological and biomedical data; examples include it being used to explore the relation between deprivation and socio-economic status (Abellan *et al.*, 2007) and mortality with income (Nakaya and Dorling, 2005), and estimating the disease relative risk (Richardson *et al.*, 2004). In this section, the framework of Bayesian hierarchical modelling was used to model the relationship between cervical cancer incidence rate and a range of deprivation, social grade and family structure factors across England. Three types of non-spatial and spatial models were used in the Bayesian framework. (i) non-spatial, (ii) BYM spatial, and (iii) MIX spatial models (e.g., the influence of the prior structure specified and the amount of smoothing of the risks actually performed). Interesting questions related to the performance of the models were investigated and discussed in the discussion Chapter 9.

5.5.1 Truncated missing data

On average, there are 21.56 incidence cases per study region. The data were represented at the district and unitary authority levels of the Cancer Registries in England. Twenty one regions had observed cases between zero to five cases. For confidentiality reasons, these 21 ($\approx 6\%$) observed data were closed. It is common to have closed or missing data in disease studies. In practice, it is possible to exclude or remove those 21 disclosed observed data from analysis; however, it could cause an amount of information to be discarded. It reduces the prediction power and the ability to detect the association or relation between the observed and the explanatory variables (Lunn *et al.*, 2006). All these undisclosed data were

treated as missing data and were computed through the Bayesian modelling by borrowing information from the first order neighbours. Prediction for these 21 regions was restricted to between zero and five cases.

5.5.2 Methods

A Bayesian regression method was applied in this study to explore the relationship between the observed cervical cancer incidence rate and the explanatory variables (a series of deprivation, social status and family structure variables). The final models provide a summary of the complex geographical relations.

A hierarchical framework has been used in many spatial disease studies, especially for those diseases that have smaller counts in small study regions. Incidence cases and incidence rate were estimated through a Bayesian approach. The model is defined below:

$$Y_i \sim \text{Poisson}(\mu_i), \quad \text{where } i = 1, 2, \dots, N \quad (5.4)$$

5.5.3 Model definition

The study area, England, was defined as D , which was split into 354 study regions d_i , so $\{d_i \in D\}$ for $i = 1, 2, \dots, 354$. With the Poisson assumption Y_i is assumed to be independently distributed with mean μ_i , indirectly standardised incidence ratio (SIR) θ_i , and expected cases e_i . The model is defined below:

$$\begin{aligned} Y_i &\sim \text{Poisson}(\mu_i) \\ \mu_i &= e_i \theta_i \\ \hat{\theta}_i &= \frac{\mu_i}{e_i} \end{aligned} \quad (5.5)$$

The closed or unobserved data were treated as a censoring problem, re-computed or re-estimated by borrowing information from a prior distribution $p(\theta)$, and also the neighbouring data. Lower and upper limits were specified for the closed data.

5.5.4 Model assumptions

In the spatial model, incidence Y_i is assumed to be conditionally independently distributed with standardised incidence ratio θ_i , which is also a positive random variable. Townsend index, low social status and the family structure variables were taken into account in the models. Each of the variables was tested individually in the analysis. The model was defined below:

$$Y_i | \theta_i \sim \text{Poisson}(e_i \theta_i) \quad (5.6)$$

It is common to use expected cases e_i to be the offset; Two types of model were applied in this study; (i) non-spatial, and (ii) spatial conditionally auto-regressive (CAR) models. The structure and the advantages of these two types of models are now discussed briefly.

5.5.5 Non-spatial model

Spatial information is not specified in the non-spatial model. Every region is assumed to be the same (i.e. no difference in terms of location). No neighbourhood information is available in the model, such that the model provides only global and average information to the users. It is also assumed that no effects are contributed by the direct neighbourhood; each of the regions is completely independent. Such a model is useful only if the underlying model has no spatial variation, because the non-spatial model only gives average estimated parameters, which do not vary over space. The non-spatial model was defined below:

$$\begin{aligned} \log \mu_i &= \log e_i + \beta_0 + \sum_t \beta_t v_{ti} + \delta_i \\ \theta_i &= \exp(\beta_0 + \sum_t \beta_t v_{ti} + \delta_i) \end{aligned} \quad (5.7)$$

Where β_0 is the intercept of the regression model, β_t is the coefficient of the explanatory variable v_{ti} and δ_i is the unstructured heterogeneity (random effect). Figure 5.13 shows the linkage between variables and parameters.

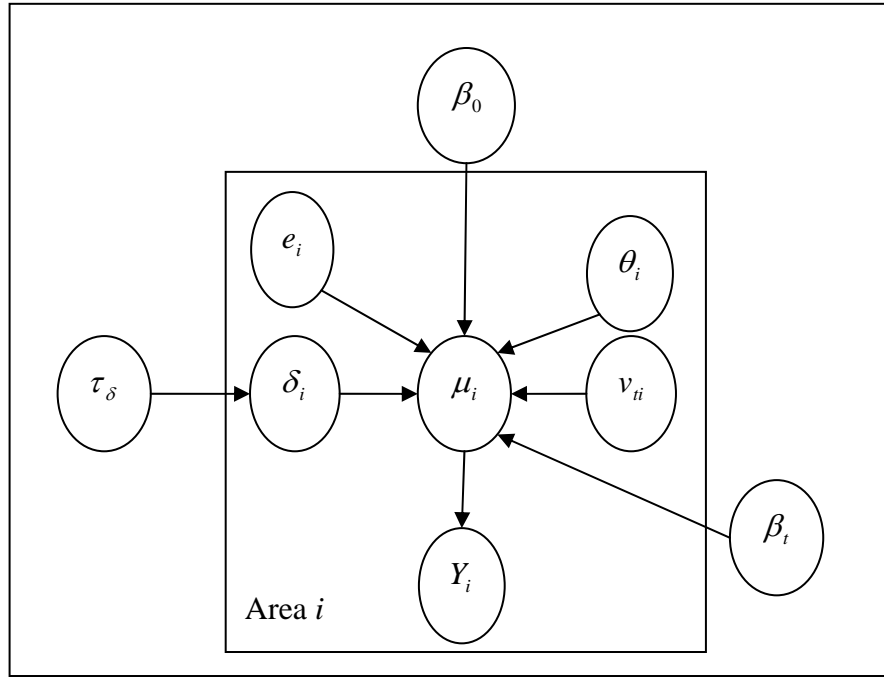


Figure 5.13 The link between variables and parameters in the non-spatial model.

5.5.6 Spatial model

Spatial models provide an underlying model with given observed spatial data, which can be used to predict the underlying incidence rate with the available spatial information. The conditionally autoregressive (CAR) model was introduced by Besag *et al*, 1991. CAR models are commonly used by statisticians and epidemiologists and it is facilitated within WinBUGS. Information from direct neighbours is taken into account. The weighting function w_{ij} of all the direct neighbours has an equal weight of one (Lawson *et al.*, 2003):

$$w_{ij} = \begin{cases} 1, & j \text{ is direct neighbour of } i \\ 0, & \text{otherwise} \end{cases} \quad (5.8)$$

Where w_{ij} is the weighting function for measuring the association between regions i and j . There are two common CAR model settings, which are BYM and MIX as described above.

(i) BYM spatial model

In the BYM model, $\hat{\theta}_{BYMi}$ was defined as the indirectly standardised incidence ratio.(SIR). Instead of modelling directly the correlation between log incidence

standardised ratio $\log \hat{\theta}_{BYMi}$ as an independent function and several covariates, an alternative approach is to model the conditional independence between incidence rates in different areas. Area-specific random effects are decomposed into two components. The first component is δ_i the uncorrelated heterogeneity; this is the part measuring unstructured variation between areas. The second is γ_i ; this is the component that models the structured variation in space (i.e. clustering component or correlated heterogeneity). The model is defined as below and the model structure is described in Figure 5.14:

$$\log \mu_i = \log e_i + \beta_0 + \sum_t \beta_t v_{ti} + \alpha_i \quad (5.9)$$

$$\alpha_i = \delta_i + \gamma_i \quad (5.10)$$

$$\hat{\theta}_i = \exp(\beta_0 + \sum_t \beta_t v_{ti} + \alpha_i) \quad (5.11)$$

The prior distributions of uncorrelated and correlated heterogeneity δ_i and γ_i are specified below:

$$\delta_i \sim Normal(0, \tau_\delta) \quad (5.12)$$

$$\gamma_i \sim Normal(0, \tau_\gamma) \quad (5.13)$$

For the clustering component, a spatial correlation structure is used. Estimation of the posterior mean of incidence standardised ratio $\hat{\theta}_{BYMi}$ in any area depends on neighbouring areas.

The parameters τ_δ and τ_γ control the variability of clustering component δ_i and correlated heterogeneity γ_i . The prior distribution of these two parameters can be specified, and the Gamma distribution is the most common choice as suggested by Bernardinelli *et al.*, (1995).

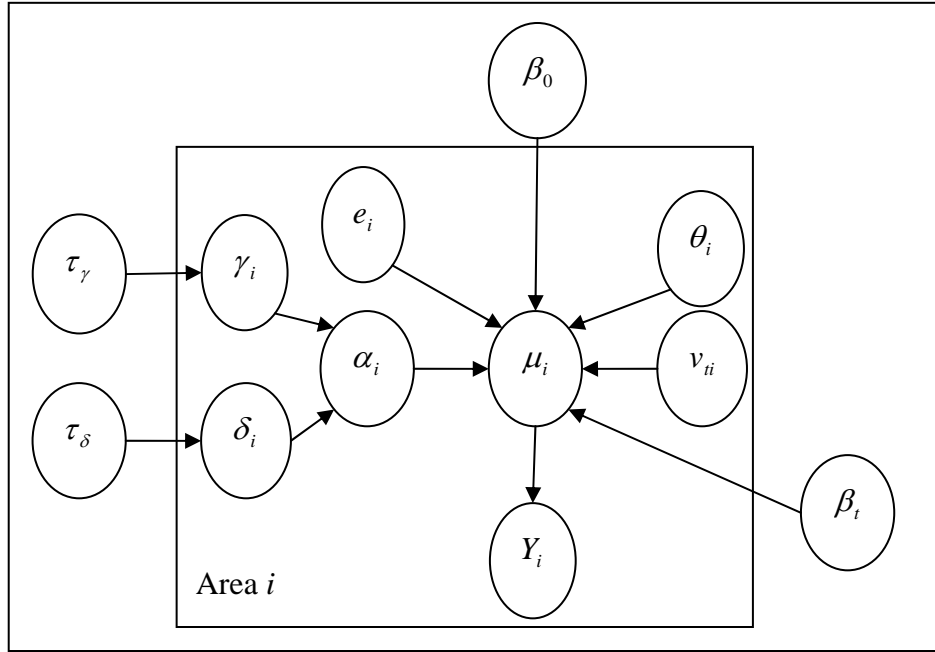


Figure 5.14 The link between variables and parameters in the BYM model.

(ii) MIX spatial model

The BYM method is used to analyse the incidence rate in small areas and it is considered as a smoothing technique. However, such models could over smooth the spatial incidence rate where large discontinuities exist in the risk surfaces. Therefore, in both theory and practice, it is important to maintain the discontinuities and smoothness within the model and maps. Thus, a special type of spatial mixture model (MIX) was introduced by Lawson and Clark (2002). This spatial mixture model allows both smoothness and discontinuities and admits different forms of spatial variation.

Within the BYM model, the non-spatial model has one random effect component and the spatial model has two random effect components. The MIX spatial model has four components. One of them is δ_i ; unstructured heterogeneity that measures the over-dispersion in an individual region. This is assumed to be a fixed component. Two of them are γ_i and φ_i the mixing components. These two components represent different aspects of spatial correlation. Figure 5.15 describes the link between the parameters. The final one λ_i models discrete jumps. If all the $\lambda_i=1$, the MIX model is the same as the standard BYM model. If all zeros, the model is called pure jump (Lawson *et al.*, 2002; 2003).

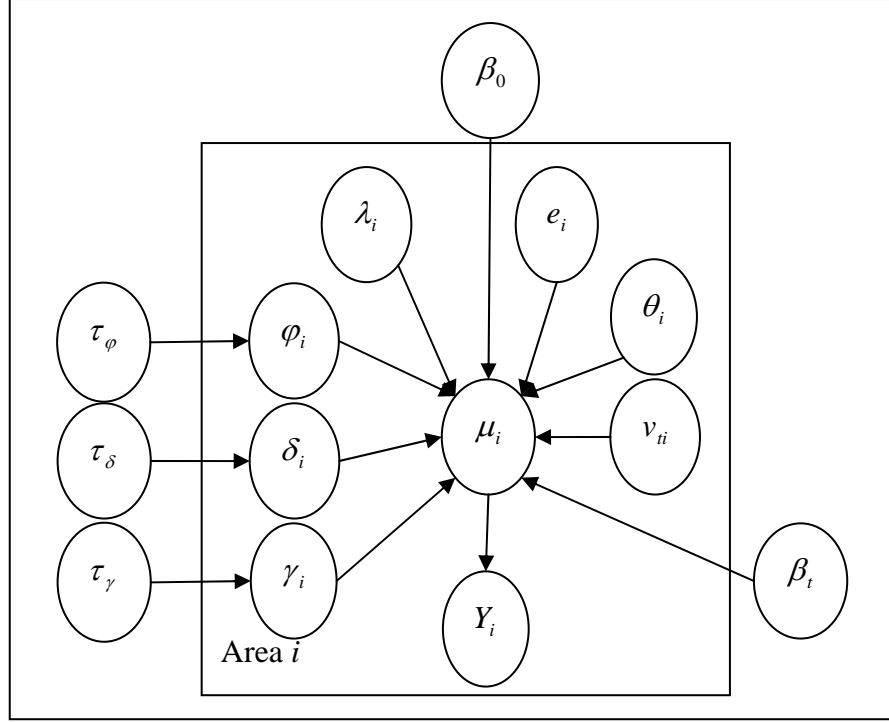


Figure 5.15 The link between variables and parameters in the MIX model.

The model is defined as below:

$$\log \mu_i = \log e_i + \beta_0 + \sum_t \beta_t v_{ti} + \delta_i + \lambda_i \gamma_i + (1 - \lambda_i) \varphi_i \quad (5.14)$$

$$\hat{\theta}_{MIXi} = \exp(\beta_0 + \sum_t \beta_t v_{ti} + \delta_i + \lambda_i \gamma_i + (1 - \lambda_i) \varphi_i) \quad (5.15)$$

The jump γ_i is given a prior i . It examines the total absolute difference between neighbours:

$$\gamma_i \sim \text{Beta}(0.5, 0.5) \quad (5.16)$$

5.5.7 Analysis and results

In this section, three Bayesian hierarchical models ((i) non-spatial, (ii) BYM spatial, and (iii) MIX spatial models) were applied to explore the relation between cervical cancer SIR and deprivation, social grade and family structure variables at the district or unitary authority level in England in 2004. All the Bayesian models are listed in Table 5.3 – Table 5.5 and mapped in Figure 5.16-5.18.

Table 5.3 Summary of non-spatial Bayesian Poisson regression models.

(a)

Model	Variables	β_0 (95% CI)	β_1 (95% CI)	β_2 (95% CI)
1	Townsend index score	0.1636 (0.1234, 0.2025)	0.01053 (0.002544, 0.01852)	
2	Proportion of female single population	-0.3838 (-0.8746, -0.01081)	0.9622 (0.3185, 1.813)	
3	Proportion of female married population	0.6039 (0.2765, 0.9457)	-1.023 (-1.823, -0.2605)	
4	Proportion of household with lone parent	-0.3003 (-0.4389, -0.1673)	7.724 (5.672, 9.869)	
5	Proportion of household with female lone parent	-0.2644 (-0.3935, -0.1391)	7.916 (5.776, 10.15)	
6	Proportion of low social grade (IV+V) population	0.1527 (0.1143, 0.1897)	0.08666 (0.05574, 0.1177)	
7	Proportion of household with female lone parent + proportion of low social grade (IV+V) population	-0.1993 (-0.3521, -0.04974)	6.61 (3.935, 9.324)	0.03217 (-0.004757, 0.06884)
8	Proportion of household with female lone parent + Townsend index score	-0.4194 (-0.5871, -0.2586)	10.9 (8.034, 13.86)	-0.01569 (-0.02589, -0.005774)
9	Proportion of household with female lone parent proportion + proportion of female single population	0.729 (0.3631, 1.219)	12.81 (9.965, 16.13)	-2.195 (-3.254, -1.433)
10	Proportion of households with female lone parent + proportion of female married population	-1.42 (-1.963, -0.8573)	12.63 (9.716, 15.59)	2.108 (1.092, 3.068)

(b)

Model	Models (variables)	pD	DIC
1	Townsend index score	103.3577023	974.70254708
2	Proportion of female single population	103.32035773	975.5729926
3	Proportion of female married population	102.95395444	976.78772316
4	Proportion of household with lone parent	93.23612062	969.9301876
5	Proportion of household with female lone parent	91.12125211	967.19954957
6	Proportion of low social grade (IV+V) population	97.56668097	973.20375348
7	Proportion of household with female lone parent + proportion of low social grade (IV+V) population	95.12799991	967.12499188
8	Proportion of household with female lone parent + Townsend index score	93.12824428	967.83376986
9	Proportion of household with female lone parent proportion + proportion of female single population	90.51250534	967.09715974
10	Proportion of households with female lone parent + proportion of female married population	89.52749918	963.62040584

Chapter 5 Cervical cancer regression study

Table 5.4 Summary of BYM CAR Bayesian Poisson regression models.

(a)

Model	Variables	β_0 (95% CI)	β_1 (95% CI)	β_2 (95% CI)
1	Townsend index score	0.1458 (0.1104, 0.1807)	0.02306 (0.01241, 0.03363)	
2	Proportion of female single population	-0.7927 (-1.257, -0.199)	1.654 (0.6269, 2.457)	
3	Proportion of female married population	0.8681 (0.4578, 1.215)	-1.671 (-2.476, -0.7157)	
4	Proportion of household with lone parent	-0.2371 (-0.3978, -0.08078)	6.488 (3.996, 9.063)	
5	Proportion of household with female lone parent	-0.227 (-0.3788, -0.08095)	7.005 (4.414, 9.7)	
6	Proportion of low social grade (IV+V) population	0.1447 (0.1086, 0.1804)	0.05525 (0.01942, 0.09107)	
7	Townsend index score + proportion of female married population	0.2723 (-0.5423, 0.9069)	0.02054 (0.003082, 0.04032)	-0.2943 (-1.771, 1.593)
8	Townsend index score + proportion of households with female lone parent	-0.1082 (-0.3146, 0.09071)	0.01021 (-0.004598, 0.02493)	4.769 (1.094, 8.565)
9	Townsend index score + Proportion of low social grade (IV+V) population	0.1434 (0.107, 0.1796)	0.01997 (0.008008, 0.03197)	0.02319 (-0.01761, 0.06476)

(b)

Model	Models (variables)	pD	DIC
1	Townsend index score	106.02894151	957.99103804
2	Proportion of female single population	106.80700759	960.27818055
3	Proportion of female married population	107.46463116	959.94719854
4	Proportion of household with lone parent	108.81552651	964.28449211
5	Proportion of household with female lone parent	106.90209847	959.53606315
6	Proportion of G4 + G5 population	109.7265438	96341250304
7	Townsend index score + proportion of female married population	108.74620904	962.51777259
8	Townsend index score + proportion of households with female lone parent	105.54561402	955.67957353
9	Townsend index score + Proportion of G4 + G5 population	105.14123698	958.24877874

Chapter 5 Cervical cancer regression study

Table 5.5 Summary of MIX CAR Bayesian Poisson regression models.

(a)

Model	Models (variables)	β_0 (95% CI)	β_1 (95% CI)	β_2 (95% CI)
1	Townsend index score	0.162 (0.1312, 0.1918)	0.02194 (0.01357, 0.03045)	
2	Proportion of female single population	-0.7628 (-0.8929, -0.6475)	1.628 (1.418, 1.832)	
3	Proportion of female married population	0.8551 (0.5591, 1.112)	-1.604 (-2.208, -0.9156)	
4	Proportion of household with lone parent	-0.2443 (-0.3771, -0.1133)	6.891 (4.814, 8.989)	
5	Proportion of household with female lone parent	-0.234 (-0.3476, -0.1182)	7.469 (5.49, 9.452)	
6	Proportion of low social grade (IV+V) population	0.1606 (0.1293, 0.1918)	0.04438 (0.01856, 0.07247)	
7	Townsend index score + proportion of female married population	0.3893 (-0.07695, 0.8459)	0.01695 (0.003931, 0.03085)	-0.5275 (-1.578, 0.5523)
8	Townsend index score + proportion of household with female lone parent	-0.1792 (-0.3246, -0.02895)	0.00484 (-0.005615, 0.01577)	6.451 (3.751, 9.057)
9	Townsend index score + proportion of low social grade (V+V) population	0.1594 (0.1289, 0.1898)	0.01948 (0.01006, 0.02885)	0.01481 (-0.01227, 0.04337)

(b)

Model	Variables	pD	DIC
1	Townsend index score	74.48339393	947.00286159
2	Proportion of female single population	75.56180522	946.91556045
3	Proportion of female married population	76.78314502	947.73158404
4	Proportion of household with lone parent	70.3796227	953.56697841
5	Proportion of household with female lone parent	72.92979225	948.69604725
6	Proportion of low social grade (IV+V) population	77.21551043	957.22739088
7	Townsend index score + proportion of female married population	75.6384955	943.07269939
8	Townsend index score + proportion of household with female lone parent	76.41373642	951.03644554
9	Townsend index score + proportion of low social grade (V+V) population	81.58074762	949.74908547

The best fitted models were:

(i) Non-spatial model:

$$\hat{\theta}_i = \exp(\beta_0 + \beta_1 v_{All\ lonei} + \delta_i)$$

$$\hat{\theta}_i = \exp(\beta_0 + \beta_1 v_{Female\ lonei} + \delta_i)$$

$$\hat{\theta}_i = \exp(\beta_0 + \beta_1 v_{Female\ lonei} + \beta_2 v_{G45i} + \delta_i)$$

(i) BYM CAR model:

$$\hat{\theta}_i = \exp(\beta_0 + \beta_1 v_{Townsendi} + \alpha_i)$$

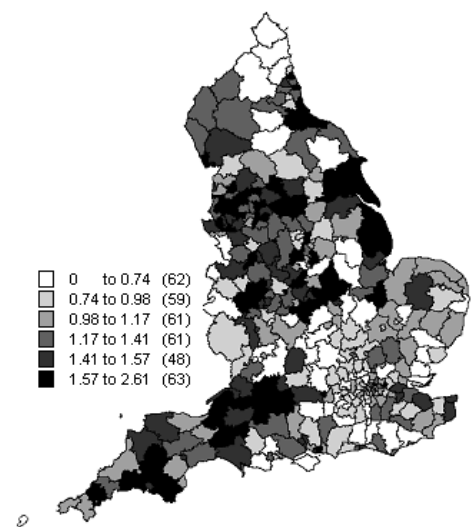
$$\hat{\theta}_i = \exp(\beta_0 + \beta_1 v_{Townsendi} + \beta_2 v_{Female\ lonei} + \alpha_i)$$

$$\hat{\theta}_i = \exp(\beta_0 + \beta_1 v_{Townsendi} + \beta_2 v_{G45i} + \alpha_i)$$

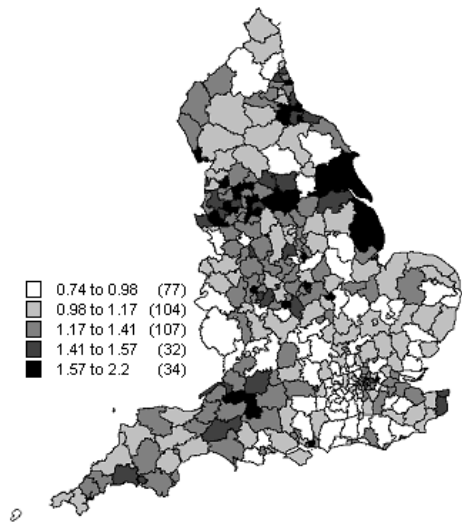
(ii) MIX CAR model:

$$\hat{\theta}_i = \exp(\beta_0 + \beta_1 v_{Townsendi} + \beta_2 v_{Marriedi} + \delta_i + \gamma_i u_i + (1 - \gamma_i) \varphi_i)$$

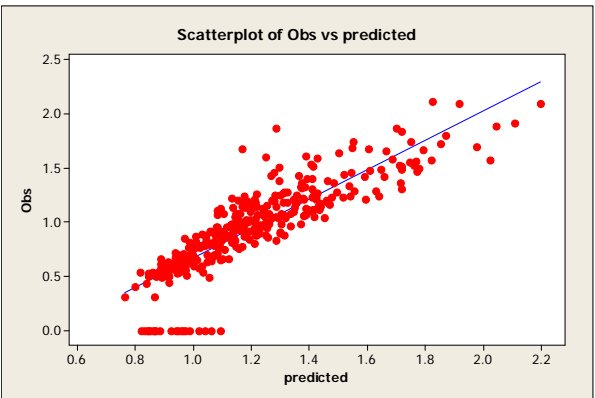
(a)



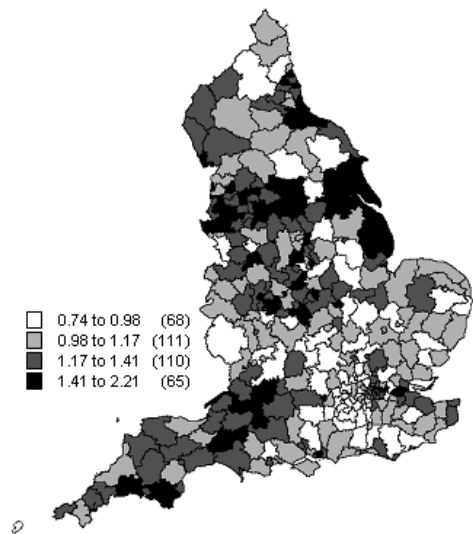
(b)



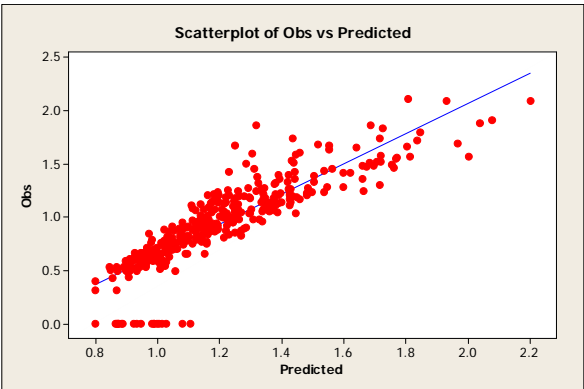
(c)



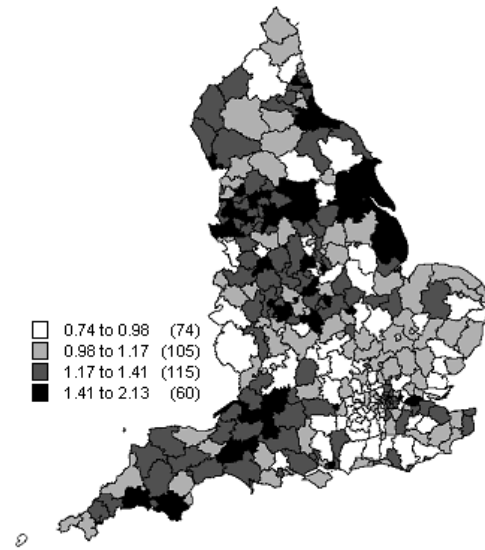
(d)



(e)



(f)



(g)

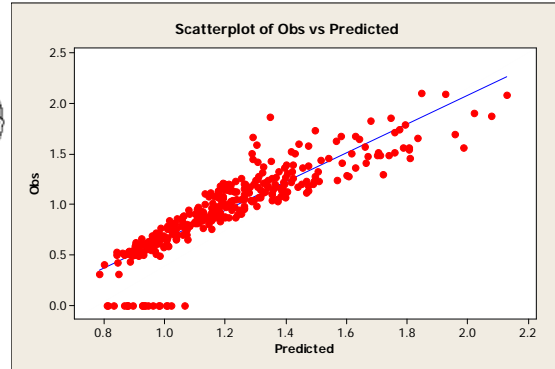
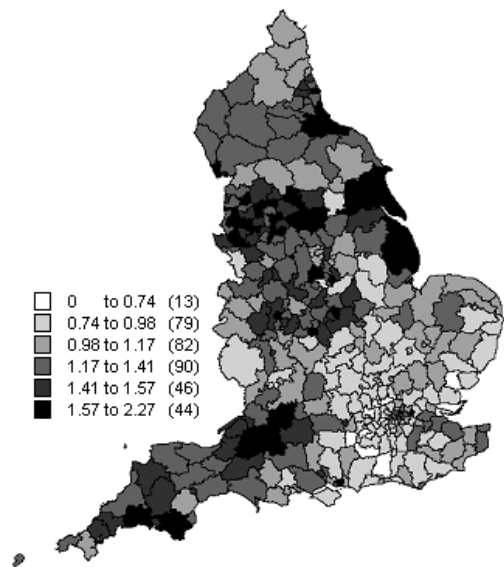
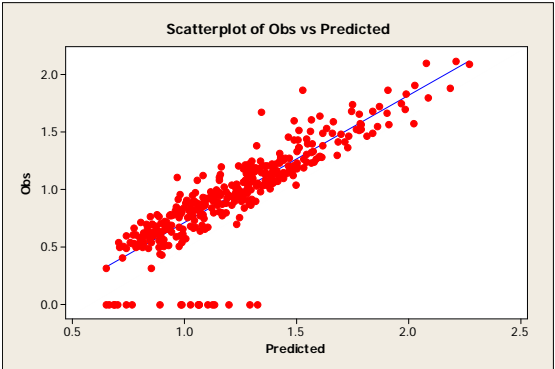


Figure 5.16 The estimated mean indirectly standardised incidence ratio $\hat{\theta}_i$ for different non-spatial models; (a) raw ratio $\hat{\theta}_i$, (b) $\hat{\theta}_i$ from the non-spatial model with proportion of household with lone parent, (c) SIR plot against fitted value, (d) $\hat{\theta}_i$ from the non-spatial model with proportion of household with female lone parent, (e) SIR plot against fitted value, (f) $\hat{\theta}_i$ from the non-spatial model with proportion of household with female lone parent and lower social grade population.

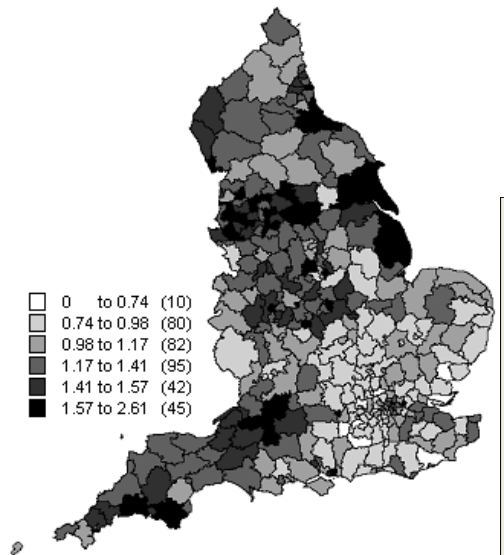
(a)



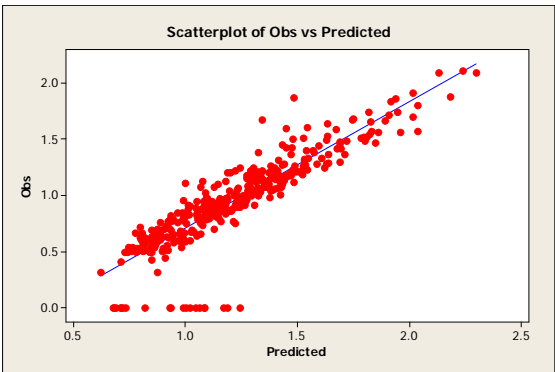
(b)



(c)



(d)



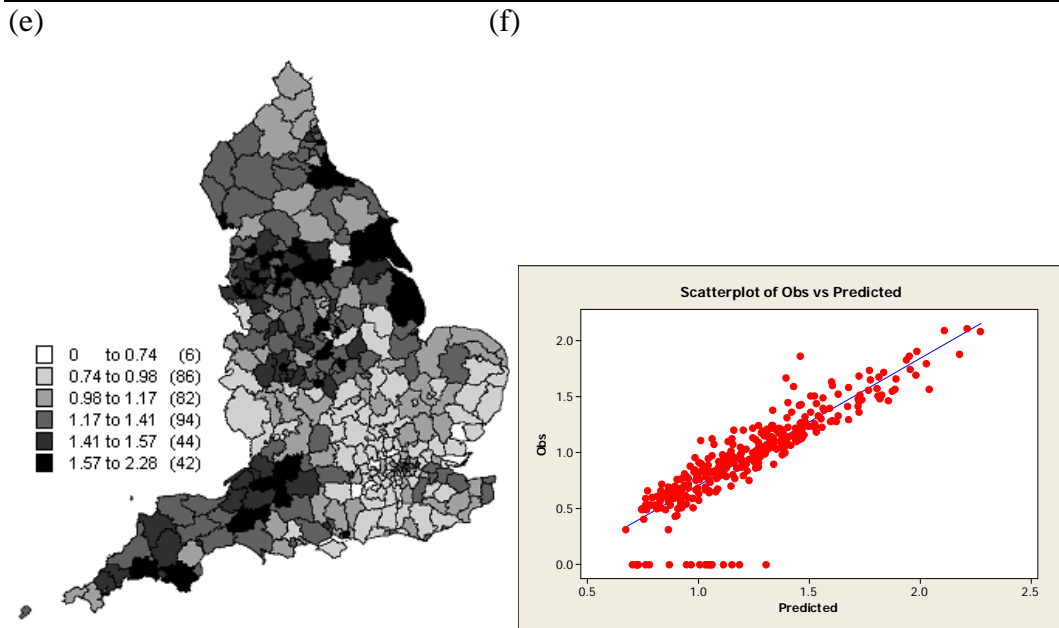


Figure 5.17 The estimated mean indirectly standardised incidence ratio $\hat{\theta}_i$ for different BYM CAR models; (a) $\hat{\theta}_i$ from the BYM model with Townsend index, (b) SIR plot against fitted value, (c) $\hat{\theta}_i$ from the BYM model with Townsend index and proportion of household with female lone parent, (d) SIR plot against fitted value, (e) $\hat{\theta}_i$ from the BYM model with Townsend index and proportion of low social grade population and (f) SIR plot against fitted value.

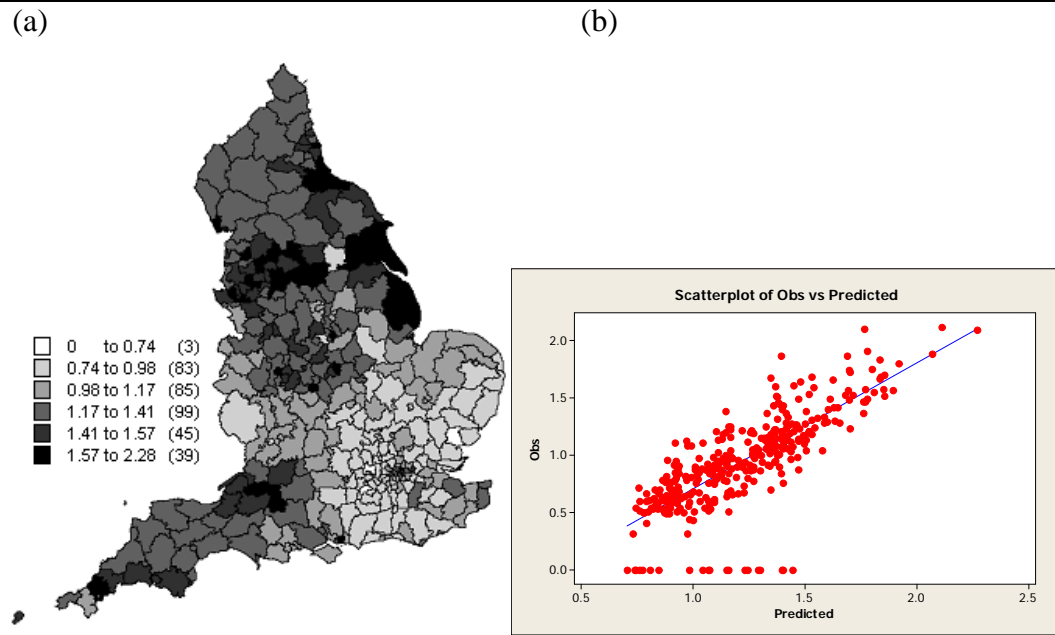


Figure 5.18 The estimated mean indirectly standardised incidence ratio $\hat{\theta}_i$ for different MIX CAR models; (a) $\hat{\theta}_i$ from the MIX model with Townsend index and proportion of female married population, (b) SIR plot against fitted value.

Townsend index, proportion of household with lone parent, proportion of population with low social grade (IV+V) and proportion of female married population are the most significant variables associated to cervical cancer SIR in all three types of Bayesian regression models. In simple terms, the analysis suggests that on average, as the Townsend index, and proportion of household with lone parent, and proportion of population with low social grade increase cervical cancer incidence increases, but when the proportion of female married populations increase the standardised incidence ratio (SIR) decreases.

Based on prior knowledge, the marital status do affect the incidence rate, for example if a woman is single status the chance for her to have more than one sexual partner is higher than a married woman, the marital status factor does reflect some possible information about personal sexual behaviours, which may be related to the chance of cervical cancer development. From the prior knowledge, it is expected regions with high deprivation and with a large population of single females or female lone parents may have a higher incidence

rate, where the chance of developing cervical cancer is related to female reproductive history and personal sexual behaviour.

The values of β_0 , β_1 and β_2 are shown in Table 5.3a, 5.4a, and 5.5a and the model measurements in Table 5.3b, 5.4b, 5.5b. The posterior mean of SIR for all three models is shown in Figure 5.16, 5.17 and 5.18. Clearly, all final fitted models have a very similar posterior mean, but Figure 5.18 has a clear spatial pattern compared to the raw data Figure 5.16a. This arises because the spatial models have the advantage of spatial smoothing. By comparing the pD and DIC values between models, the spatial MIX models have smaller pD and DIC values than the BYM spatial and non-spatial models.

Three types of Bayesian regression models were used to examine the relations between target and explanatory variables. The link between parameters is shown in Figures 5.13-15. The Bayesian regression models highlighted that the Townsend index and some of the family structure variables (e.g. female lone parent and married population) are significantly related to cervical cancer development, which is very similar to the GLM results in the section 5.4. In the next section of this chapter, the geographically weighted Poisson regression model is used to study the same set of data. However, other methods are provided to deal with the missing data issue and an opportunity is provided to model the variables as non-stationary.

5.6 Geographically weighted Poisson regression model (GWPR)

Traditionally, global models have been applied in disease studies (Best *et al.*, 2000). In such studies, it is assumed that the disease can be modelled adequately by a geographically invariant process. Simply, the disease model and, specifically, its parameters are spatially stationary or spatially constant. Possible spatial variation in the parameters in such models is neglected. Such missing information can be very important and also meaningful in disease analysis. Since some of the relations between the target variable and some explanatory factors may not be

spatially constant, the covariance may vary spatially (Nakaya and Dorling, 2005). In these circumstances, local regression modelling is a viable alternative approach for exploring the relationship between the target and explanatory variables at the local level, taking into account spatial variation in parameters and allowing the covariance to vary locally based on geographical context.

Many studies show that ill health issues are related to socio-economic status (e.g. Green and Richardson, 2002; Jarup *et al.*, 2002; Lawson and Clark, 2002; Richardson and Best, 2003; Richardson *et al.*, 2004). Other studies show that such relations may also vary between regions and such variation should be taken into account in modelling (Pascutto *et al.*, 2000).

GWR is well developed for different modelling frameworks (e.g., Gaussian and Poisson). For example, Gaussian GWR was applied to long-term limiting illness in the Northeast of England, and the results showed complex regional variation in regression parameters (Fotheringham *et al.*, 1998). Technically, geographically weighted Poisson regression (GWPR) can be applied where proportions are predicted (e.g. in many disease studies).

The global regression models described early in section 5.4.3 showed significant relationships between cervical cancer incidence and Townsend index, proportion of low social grade population, proportion of single population and proportion of married population (i.e. representing the family structure and also reflecting some of the information about the possible sexual behaviour of the general population and households with lone parent). This section demonstrates a statistical tool, geographically weighted Poisson regression (GWPR), for analysing the relation between cervical cancer disease incidence and the same set of explanatory factors in England. A kernel was used as a spatial weighting function to estimate spatial variation in the Poisson regression parameters. Local parameters were estimated which describe the spatial variation in the relationships between incidence and deprivation, social grade and family structure characteristics. Within the GWPR model, the spatial data were modelled as the result of non-stationary processes over space.

5.6.1 Truncated missing data

One way of dealing with this truncation is to estimate the basic true mean from the data, including the missing data, accepting the estimated mean to be true; then a random number can be drawn for each of the regions based on the estimated true mean. So in practice, for each region, we drew a large number 6000000 of random values based on the true mean, and the first 100 random numbers between zero and five were used to replace the missing data. The set of 100 samples is sufficiently large enough to represent the probabilities of the missing data. Finally, 100 sets of 21 data, were imputed in place of the missing data from the original dataset and applied the GWPR models 100 times.

5.6.2 Prediction mean

100 different sets of realisations were drawn to replace the missing data for the 21 regions. Thus, the GWPR models were run 100 times and the average predictions were estimated from the outcomes from the 100 GWPR models. The average prediction was represented by $E(Y_i)$ and defined below:

$$E(\hat{Y}_i) = \frac{\sum_{n=1}^{100} \hat{Y}_i^n}{100} \quad (5.17)$$

Where n is the number of GWPR models; in this case, n is equal to 100, \hat{Y}_i^n is the prediction outcome from the n th model from region i , where $i = 1, 2, \dots, N$ and $E(\hat{Y}_i)$ is the overall average prediction from the 100 GWPR models. It is also important to examine the variation from the 100 GWPR predictions. Such information explains how much variation there is within the prediction when the missing data vary between zero to five cases.

5.6.3 Prediction variance

It is useful to calculate the overall variance for the 100 predictions; it shows the overall variation resulting as a function of the uncertainty due to the truncation of the distribution. The variance provides information on prediction uncertainty and

parameter estimation uncertainty; therefore, it allows for comparison of the variation within the prediction between the observed and missing data. It summarises the prediction variation due to the truncation. The variance is denoted as $\text{var}(\hat{Y}_i)$, which is the overall variance out of the n model predictions, and was calculated as:

$$\text{var}(\hat{Y}_i) = \frac{\sum_{n=1}^{100} (\hat{Y}_i^n - E(\hat{Y}_i))^2}{99} \quad (5.18)$$

5.6.4 Analysis and results

When GWR was first developed, the Gaussian model was commonly used in disease studies (Fotheringham *et al.*, 1998; Nakaya *et al.*, 2005). For disease counted cases and rare diseases with small numbers of cases, the Poisson model is a more appropriate framework to describe the underlying distribution (Lawson *et al.*, 2003; Gelman *et al.*, 2003). Many disease analysis studies for small areas applied the Poisson model to describe the disease distribution (e.g. Green and Richardson, 2002; Lawson and Clark, 2002; 2003).

5.6.5 Variable definition

Thus, estimated incidence ratio (SIR) $\hat{\theta}_i$ is equal to:

$$\hat{\theta}_i = \frac{\mu_i}{e_i} \quad (5.19)$$

Therefore, a GWPR cervical cancer incidence model can be defined as below:

$$\begin{aligned} \hat{Y}_i &\sim \text{Poisson}(e_i \exp(\sum_{t=0}^T \beta_{ti} v_t)) \\ \log \hat{Y}_i(x_i, y_i) &= \log e_i (\beta_0(x_i, y_i) + \beta_1(x_i, y_i) v_{1i} + \dots + \beta_t(x_i, y_i) v_{ti}) \\ \hat{Y}_i(x_i, y_i) &= e_i \exp(\beta_0(x_i, y_i) + \beta_1(x_i, y_i) v_{1i} + \dots + \beta_t(x_i, y_i) v_{ti}) \end{aligned}$$

Where the link function for the Poisson model is log and $\log e_i$ represents the expected cases for exposure to cervical cancer, also called the “offset”, which is a

measurement unit of exposure for region i . Most disease studies based on the Poisson distribution framework used the expected cases e_i as the offset. Both global Poisson and GWPR models were preformed, and summarised in the next two sections.

5.6.6 Global Poisson regression model

To examine possible determinants of the geographical patterns in the cervical cancer incidence, a traditional global Poisson regression model with offset expected cases was fitted, which is based on the demographic composition of each region. Deprivation, social grade and family structure variables were tested; the variables were described in Chapter 3. All explanatory variables were significant to the observed incidence. For full details of measurements of all 100 candidate models please refer to Appendix E. Table 5.6 shows the results from one of the random models with different variables. The final fitted global Poisson regression model is defined as:

$$\hat{Y}_i = e_i \exp(-0.718 + 2.832v_{G45i}) \quad (5.20)$$

Where v_{G45i} represents the proportion of low social grade population, which is significant at the global level. Equation 5.20 is one of the 100 models.

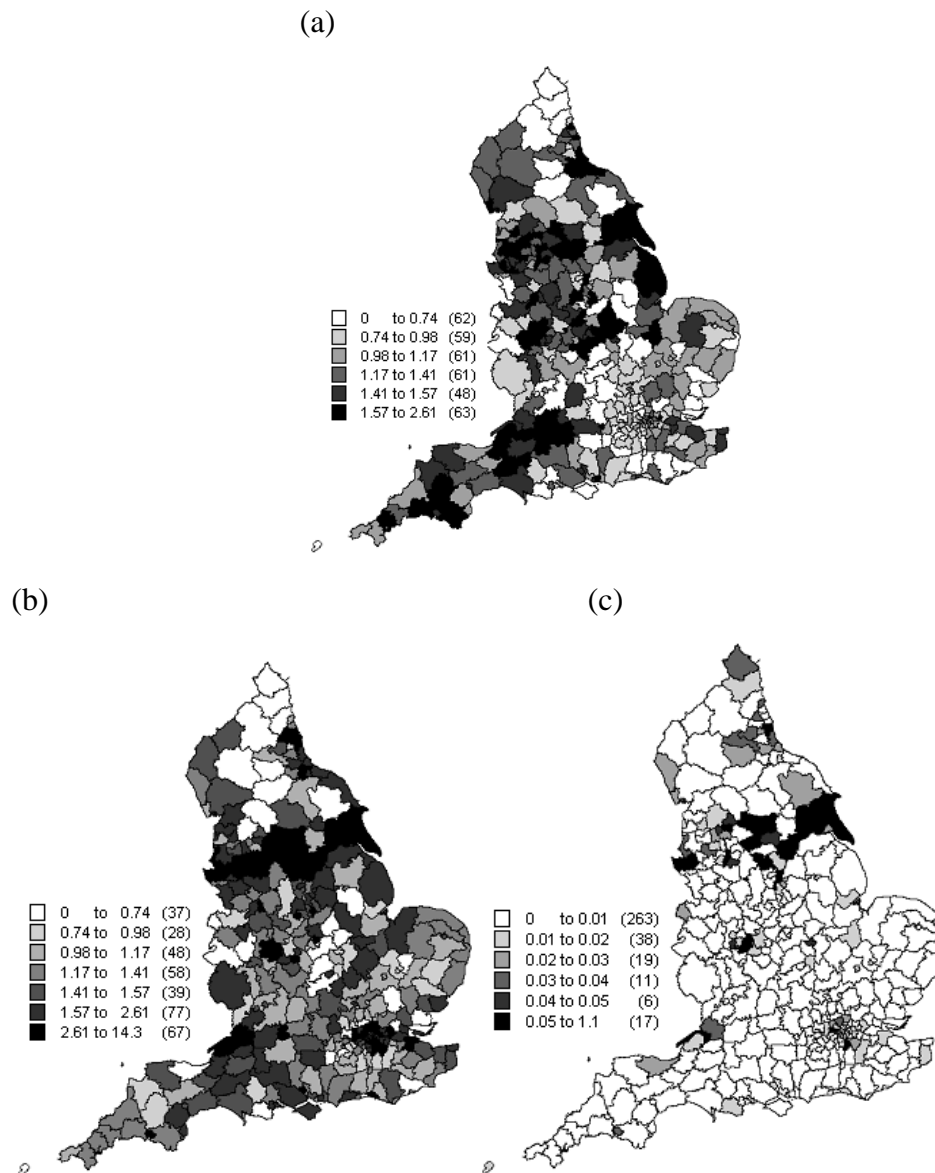
5.6.7 GWPR results

The GWPR method applied within this study was described in Chapter 2. The overall prediction means and variances of $\hat{Y}(x_i, y_i)$ are displayed in Figures 5.19 and estimated parameters mapped in Figure 5.20. The final fitted model is given as (model 6 in Table 5.6);

$$\hat{Y}_i(x_i, y_i) = e_i \exp(\beta_0(x_i, y_i) + \beta_1(x_i, y_i)v_{G45i}(x_i, y_i)) \quad (5.21)$$

The proportion of low social grade population is significant to incidence rate. The estimated predicted mean, variance, and residual values are displayed in Figure 5.19. The overall estimated $\hat{\beta}_0(x_i, y_i)$ and $\hat{\beta}_1(x_i, y_i)$ out of the 100 samples are

displayed in Figure 5.20. As summarised in this study, the use of an adaptive weighted function and the optimal bandwidth were selected based on the smallest AICc in Table 5.6; therefore the optimal kernel size was 91 regions. It is clear that the contribution of the explanatory variables varies over space (Figure 5.20). Results of the 100 models can be found in Appendix E.



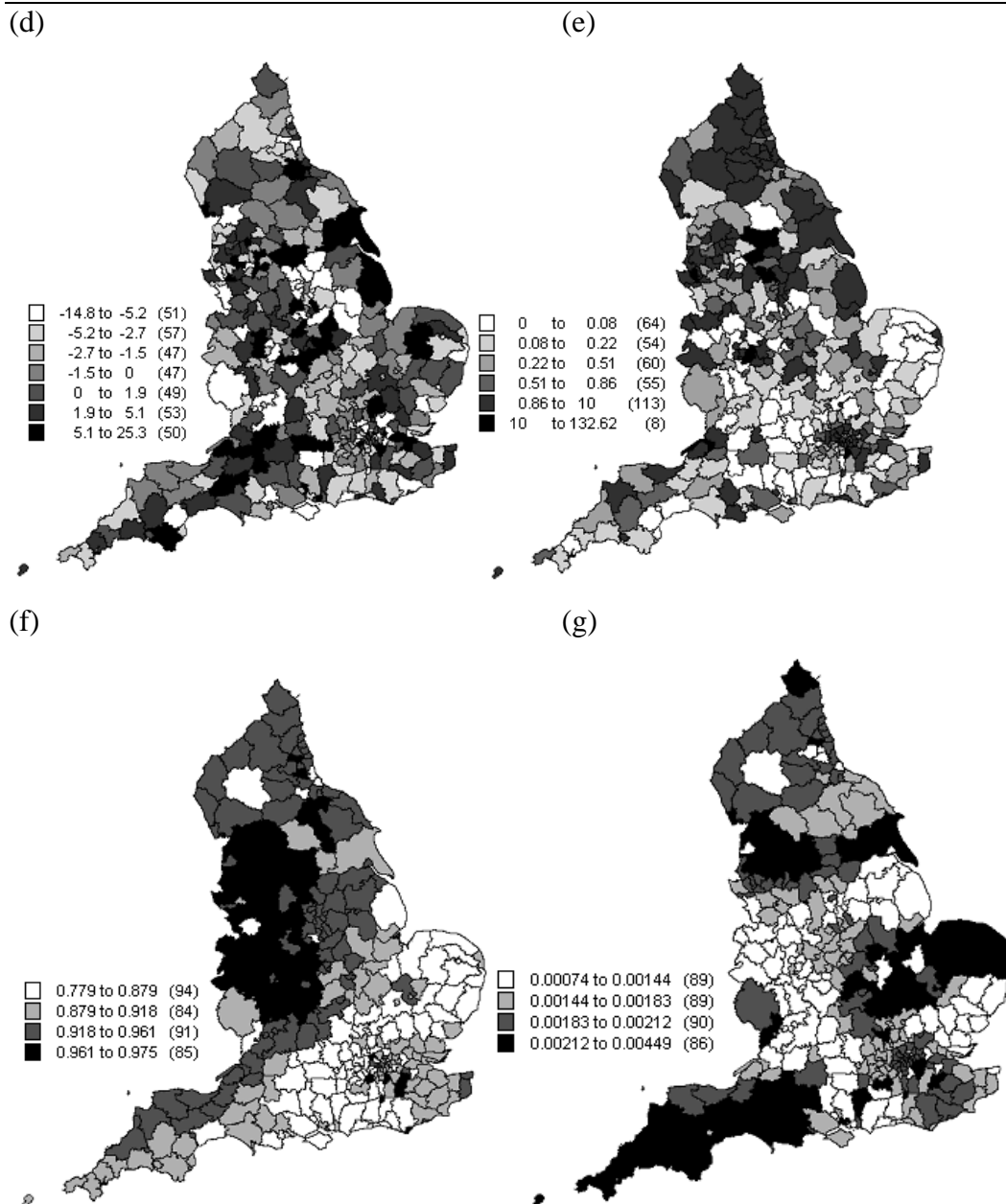


Figure 5.19. (a) Raw SIR, (b) the prediction mean of SIR out of the 100 models, (c) the variance of the predictions of SIR, (d) the mean of residual values between the observed and predicted cases and (e) the residual variance, (f) the mean of R-square value and (g) variance of R-square value.

Figure 5.19 a and b have a similar pattern and the variance of the 100 models in 5.19c showed very small variation between the 100 models. Residual value (d) showed spatial correlation and also small variation in (e). The R-square values of local models (f) are generally high between 78 to 98% and again the variation of the R-square value from the 100 model is relatively small in (g).

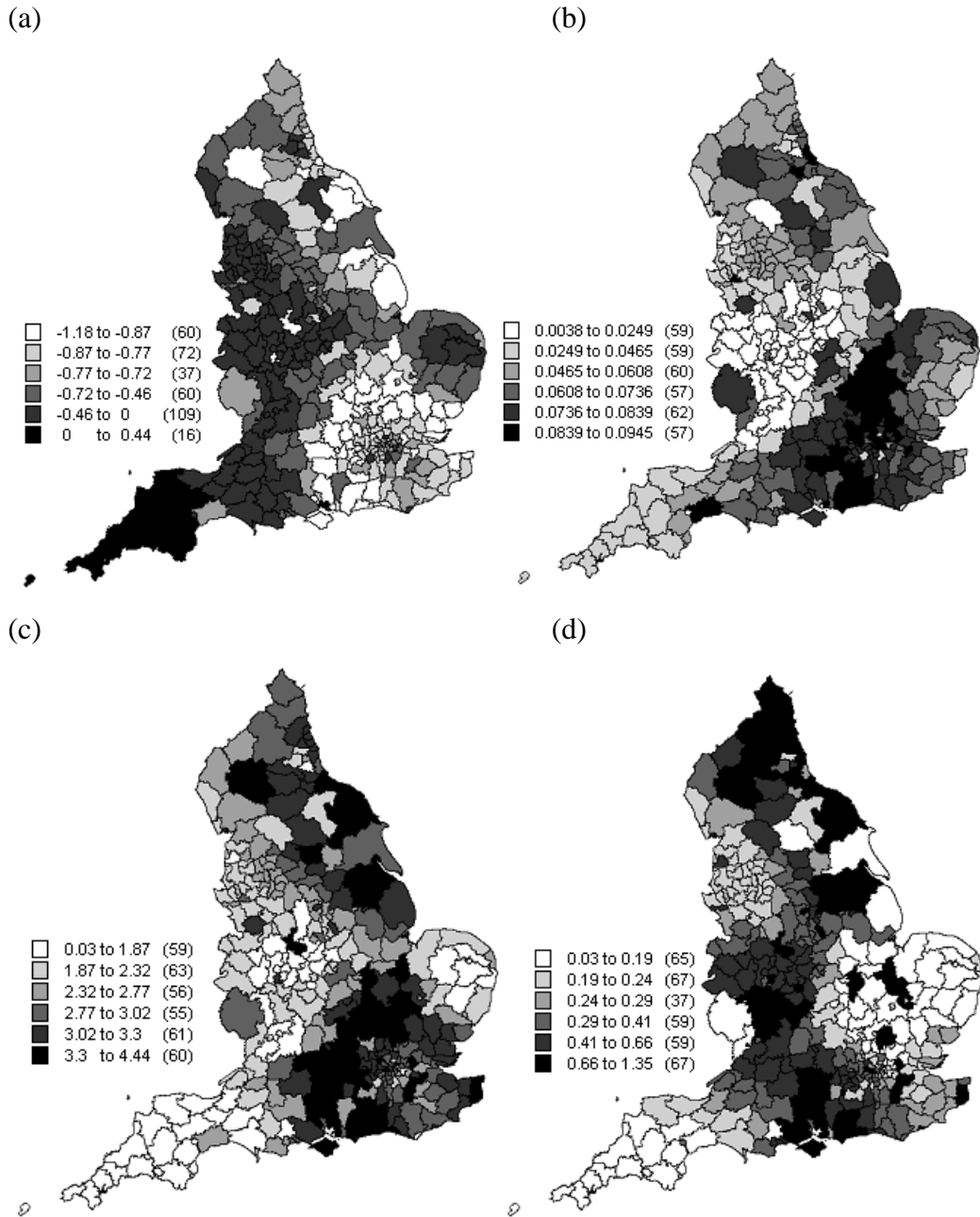


Figure 5.20. Estimated parameters (a) mean $\hat{\beta}_{oi}$, (b) variance $\hat{\beta}_{oi}$, (c) mean $\hat{\beta}_{li}$, (d) variance $\hat{\beta}_{li}$ from model 6 in Table 5.6.

Figure 5.20 shows the spatial variation of the local parameter estimates from model 6 (covariate is proportion of low social grade population). The local estimates of $\hat{\beta}_{li}$ are positive in all areas (Figure 5.20c). Areas with high estimated coefficient are the South, Midlands, and North of England. The West of England (e.g. Cornwall) has a lower incidence rate and also low estimated coefficient. It could be related to the population structure; the proportion of elderly population is higher in the West of England than the South of England. It is clearly shown that

Chapter 5 Cervical cancer regression study

the contribution of the explanatory variable proportion of low social status in the final fitted model, and the relationship varies over space. Therefore, global models are not suitable to describe behaviour of the explanatory variables.

As described in section 5.5.7, the models were compared using the local AICc. The smallest AICc values were assumed to provide the best fitted model out of the candidate models. From Table 5.6, it can be concluded that the best fitted model is model 6, which is the GWPR model with the smallest AICc and kernel size with 91 regions.

Table 5.6 Summary statistics of model comparisons.

Model	Variables	Kernel	AICc (global)	BIC (global)	AICc (local)	BIC (local)
1	Townsend index score	91	853.023703	860.728109	640.3845	709.916097
2	Female single proportion	91	849.382834	857.08724	651.31749	725.374232
3	Female married proportion	91	849.382834	857.08724	651.31749	725.374232
4	All lone parent proportion	91	750.371283	858.075689	594.871272	666.851433
5	Female lone parent proportion	91	754.678029	762.32435	597.276846	669.281713
6	G4 + G5 proportion	91	612.968809	620.673215	539.322475	610.348925
7	G4 +G5 + Female lone parent proportion	91	614.875115	626.414434	539.797014	641.997312
8	G4 + G5 + Townsend index score	91	612.221974	623.761293	538.485982	637.542841
9	G4 + G5 + Female married proportion	91	613.471141	625.010511	539.672056	642.525934
10	G4 + G5 + All lone parent proportion	91	614.963647	626.502967	539.207776	641.453094

Kernel size is very important to the prediction. In theory, the kernel size should vary between different models; however, unexpectedly all the models in Table 5.6 have the same kernel size. It could be due to the fact that all the variables exhibit similar spatial patterns, such that all models had similar bandwidth values.

Figure 5.21 shows the optimal size of bandwidth is 91 regions. The size is relatively large, which may be due to the sample size being small in most of the regions; therefore, it is necessary to have a larger kernel to cover sufficient data in order to predict reliably. Different sizes of kernel were tested; the optimal

bandwidth provides the smallest AICc value. When the AICc converges at a certain size of kernel, that is the optimal kernel size.

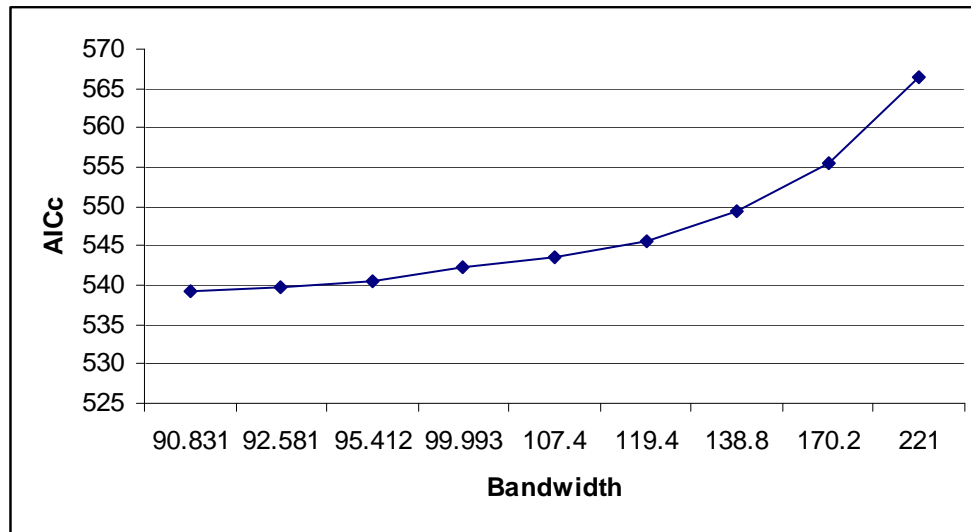


Figure 5.21 Kernel bandwidth and the corresponding AICc values.

It is interesting to examine which exploratory variables are fitted adequately by stationary parameters, and which variables required a non-stationary process. If the inter-quartile range at local level is greater than twice the standard error at global level, this would indicate that variable requires a non-stationary process (Table 5.7).

Table 5.7 Test for indicating non-stationary variables.

Parameters	2* S.E. (Global)	Inter-quartile range (Local)	Stationary or non-stationary variable
Intercept	0.062	0.458604	Non-stationary
Proportion of low social grade (IV+V) population	0.178	1.045168	Non-stationary

Table 5.7 shows that the explanatory variable (low social status population) has a larger inter-quartile range than twice the global standard error. This indicates that cervical cancer incidence (or incidence rate) is related to the proportion of low social grade population and is better fitted by a non-stationary model in GWPR rather than using a global regression model. Thus, the local model provides a more meaningful model of the relationships between incidence rate and

proportion of low social status. This allows the non-stationary process within models to increase the prediction power, and as a result, lead to a better understanding of the relationship between incidence cases (and/or incidence rate) and low social grade population. Figure 5.20 shows that the proportion of low social status population does affect the incidence rate spatially over England. The contribution is greater in the south and north of England than the west, which might be related to the population structure, (high percentage of elderly in west compared to the rest of England).

The Bayesian CAR and GWPR models showed there was spatial pattern. Some of the regions had higher incidence rates e.g. Midlands. Thus, a national fixed screening programme may not be suitable and work effectively over the whole country.

5.7 What does this mean for the design of screening programmes?

Based on the regression results in this chapter, some of the personal variables (e.g. being single or married) and the social status level are associated with cervical cancer incidence. Therefore, it should be possible to divide the general female population into risk groups according to (i) their deprivation condition, social status and family structure variables or (ii) geographical regions. Each risk group can have a different screening policy and screening interval. By dividing the population into groups, it allows use of the decision tree and simulation models in Chapter 4 to evaluate the screening options such that each group may be assigned a different screening interval given different probabilities of developing cervical cancer.

In reality, the health department (e.g. NHS) has only limited resources. Therefore, the use of decision tree and simulation models based on the regression results would allow policy makers to evaluate each potential policy based on the risk categories, and to evaluate which is the best combination of screening policies given limited resources. The optimal policy can provide the best use of resources and increase the efficiency of the overall screening programme compared to the current national fixed programme.

5.7.1 CART analysis based on regression results

There is a missing link between chapters 4 and 5. Therefore, a possible method to fill in the gap between chapters, is to create a simple decision tree model based on the results from the regression models from chapter 5 as follows;

1. Use regression to predict incidence
2. Divide the population into groups according to their deprivation condition characteristics (e.g., high deprivation index score vs. low deprivation index score etc.)
3. Create a decision tree model for the population based on the overall incidence rate.
4. Create a new decision tree model for the groups based on the set of incidence rates
5. Explore the effect of different screening intervals for the whole population (it would depend highly on the availability of transition probabilities)
6. Explore the effect of different sets of screening intervals for the groups (it would depend highly on the availability of transition probabilities)

In reality, the transition probabilities and transition times for different screening intervals for the general population (or for groups) is not available. Only if continuous clinical data (or follow up data) are available (i.e. follow each individual patient for many years) can the transition probabilities be estimated. Otherwise, it is not possible to compare the screening options with various screening intervals because of the lack of information about transition probabilities (or knowledge).

The transition probability describes how a patient might be expected to develop cervical cancer from a particular group. It is possible to estimate the incidence rate per group, but it is not possible to know the time interval and the related probabilities over which a patient might be expected to develop cervical cancer. But the incidence rate per group may be enough for a simple analysis.

The original pre-cervical cancer and cancer disease process should follow the structure in Figure 5.22. However, the transition probabilities and transition time parameters are required to be estimated from any possible continuous clinical

data, such data are rarely available. Thus, the following section demonstrated a possible simple analysis with risk grouping.

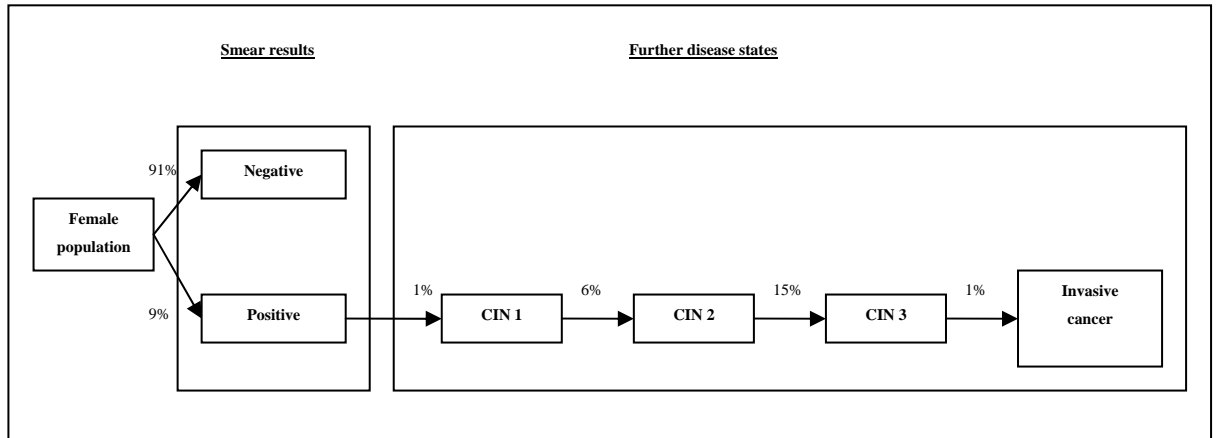


Figure 5.22 Pre-cervical cancer and cervical cancer disease process, showing summarised probabilities from Jenkins., *et al* 1996 and the positive and negative probabilities were recommended by Dr Harindra from Portsmouth St Mary hospital.

The overall idea is that if an individual is tested, that will reveal a realisation of the disease (or not) based on the incidence rate per group and can treat the disease for that individual. This tells the policy makers in the long run the number of cases that they have revealed (removed) and the number that they have missed. This simple analysis may be enough to reveal that a selective screening policy would be more efficient than a global one. Figure 5.23 demonstrates the use of decision tree model to estimate the number of expected cervical cancer cases from a general female population, Table 5.8 showed the result based on the model in Figure 5.24. The probabilities in Figure 5.24 were estimated from the national cervical cancer data.

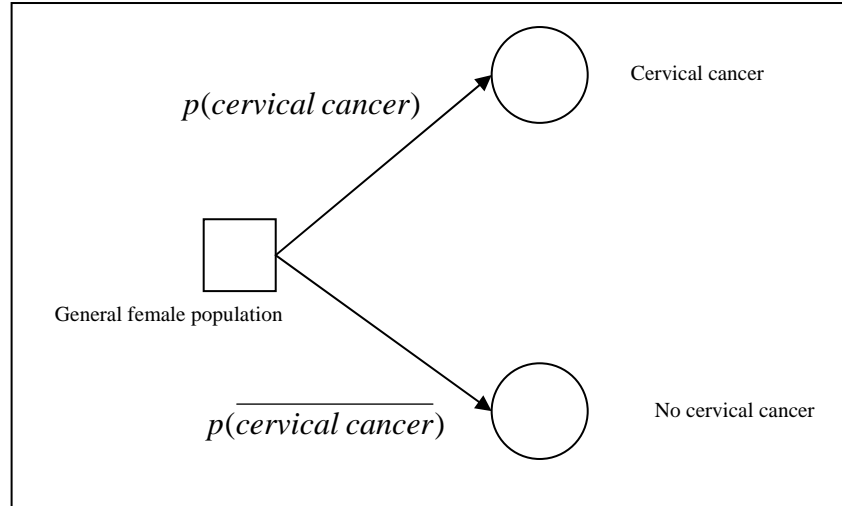


Figure 5.23 Decision tree model.

In Figure 5.23 Two probabilities are needed, which can be estimated from the national cervical cancer data. The calculation is listed below;

$$p(cervical\ cancer) = \frac{\sum_{i=1} cervical\ cancer\ incidence}{\sum_{i=1} at-risk\ population} \quad (5.22)$$

$$\overline{p(cervical\ cancer)} = 1 - p(cervical\ cancer) \quad (5.23)$$

Where $p(cervical\ cancer)$ represents the probability of cervical cancer presence, and $\overline{p(cervical\ cancer)}$ represents the probability of cervical cancer absence. Figure 5.23 shows both cervical cancer probabilities, which were estimated from the national cervical cancer data.

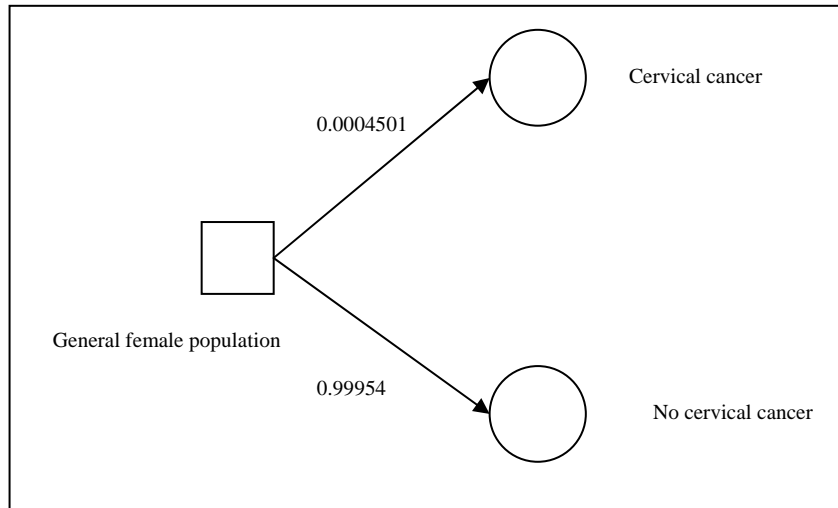


Figure 5.24 Decision tree model based on the observed national cervical cancer data.

Assume there are 20,000,000 females at risk. The positive and negative probabilities were listed in Figure 5.24. The expected number of positive cervical cancer cases and negative cervical cancer cases can be calculated in the following two equations (5.24 and 5.24). Table 5.8 shows the results.

$$\text{positive cervical cancer cases} = \text{total at risk population} \times p(\text{cervical cancer}) \quad (5.24)$$

$$\text{negative cervical cancer cases} = \text{total at risk population} \times \overline{p(\text{cervical cancer})} \quad (5.25)$$

Table 5.8 Decision tree outcome based on the information from Figure 5.24.

Group	Total patients
General female population	20,000,000
Cervical cancer cases	9,020
No cervical cancer	19,990,980

Figure 5.25 demonstrates the use of a CART tree, which is a potential method of splitting the population into risk groups according to certain characteristics.

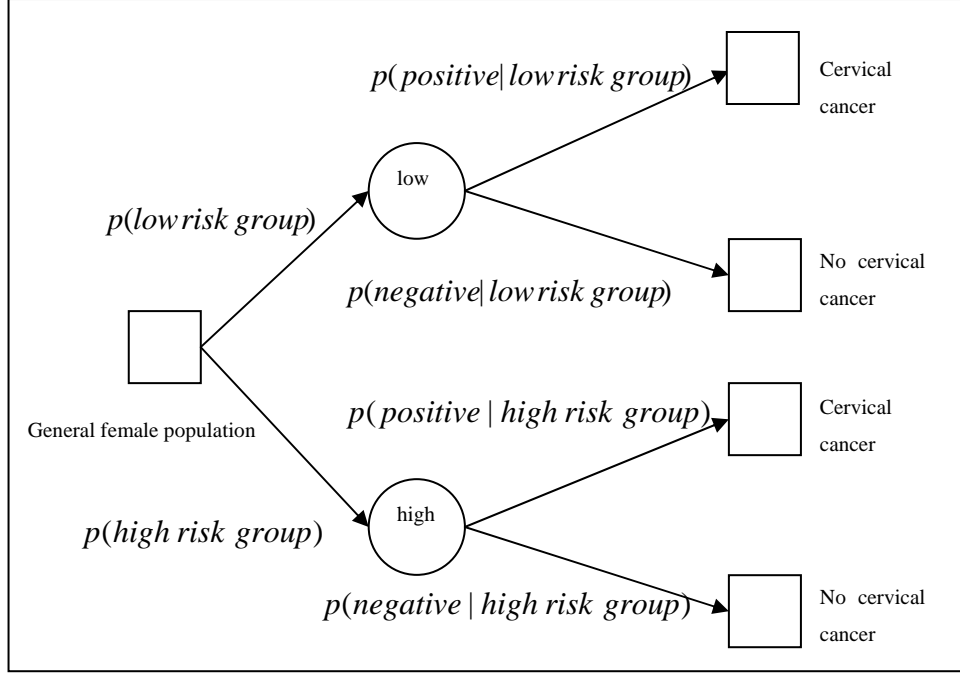


Figure 5.25 Decision tree with the best split from CART, that allows splitting the general female population into two risk groups (i.e. high or low risk),

The national cervical cancer count data can be used to estimate the probabilities between states and groups. Those probabilities are needed for decision tree models. The probabilities in Figure 5.25 can be calculated as below,

$$p(\text{low risk group}) = \frac{\sum_{i=1}^{\text{regions at low risk}}}{\text{Total regions}} \quad (5.26)$$

$$p(\text{high risk group}) = \frac{\sum_{i=1}^{\text{regions at high risk}}}{\text{Total regions}} \quad (5.27)$$

$$p(\text{positive}|\text{lowrisk group}) = \frac{\sum_{i=1}^{\text{positive cases in lowrisk regions}}}{\sum_{i=1}^{\text{at risk population in low risk regions}}} \quad (5.28)$$

$$p(\text{negative}|\text{lowrisk group}) = \frac{\sum_{i=1}^{\text{negative cases in low risk regions}}}{\sum_{i=1}^{\text{at risk population in low risk regions}}} \quad (5.29)$$

$$p(\text{positive} | \text{high risk group}) = \frac{\sum_{i=1}^{\text{positive cases in high risk regions}}}{\sum_{i=1}^{\text{at risk population in high risk regions}}} \quad (5.30)$$

$$p(\text{negative} | \text{high risk group}) = \frac{\sum_{i=1} \text{negative cases in high risk regions}}{\sum_{i=1} \text{at risk population in high risk regions}} \quad (5.31)$$

In this section, CART tree was used to split the population into two risk groups with the best split. The whole population can be divided into two groups according to the Townsend index score (i.e. high or low). Low risk regions represent a low Townsend index score (low deprivation regions) (i.e. below the best split point) and high risk regions represent a high Townsend index score (i.e. above the best split point). For details about the best split in CART please refer to Chapter 2.

5.7.2 Risk grouping

CART techniques can be used to divide the whole population into groups; Chapter 2 explained the CART theory and decision tree theory and section 5.7.1 demonstrated how to split the population into groups according to the deprivation indicator. Such techniques show that the chance of developing cervical cancer varies between groups, which means that some groups with certain characters (e.g. low deprivation condition) may have higher chance of developing cervical cancer in their life time compared to other risk groups.

5.7.2.1 CART based on the Townsend index score

Townsend index score was identified as a significant factor to the cervical cancer development. Therefore, Townsend index was chosen as a factor to divide the population into risk groups. By applying CART methods the population was divided into two risk groups according to the Townsend index score (e.g. high or low deprivation). The probabilities of developing cervical cancer were estimated per group.

Data: cervical cancer data (observed national counts)

Missing data: excluded from analysis (in total there are 354 regions, 21 with NA counts)

Average (overall) incidence rate: 0.00045

Predicted variable: incidence rate per 10000 women per region

Independent variable: Townsend index score

Best split: <1.45

The best split is Townsend index equal to 1.45, which suggests that any regions with an index less than 1.45 are identified as the low risk group and any regions with index above 1.45 are identified as the high risk group. Figure 5.26 shows the decision tree model. The probabilities were estimated from the national cervical cancer count data. Table 5.9 shows the number of expected cervical cancer cases per group.

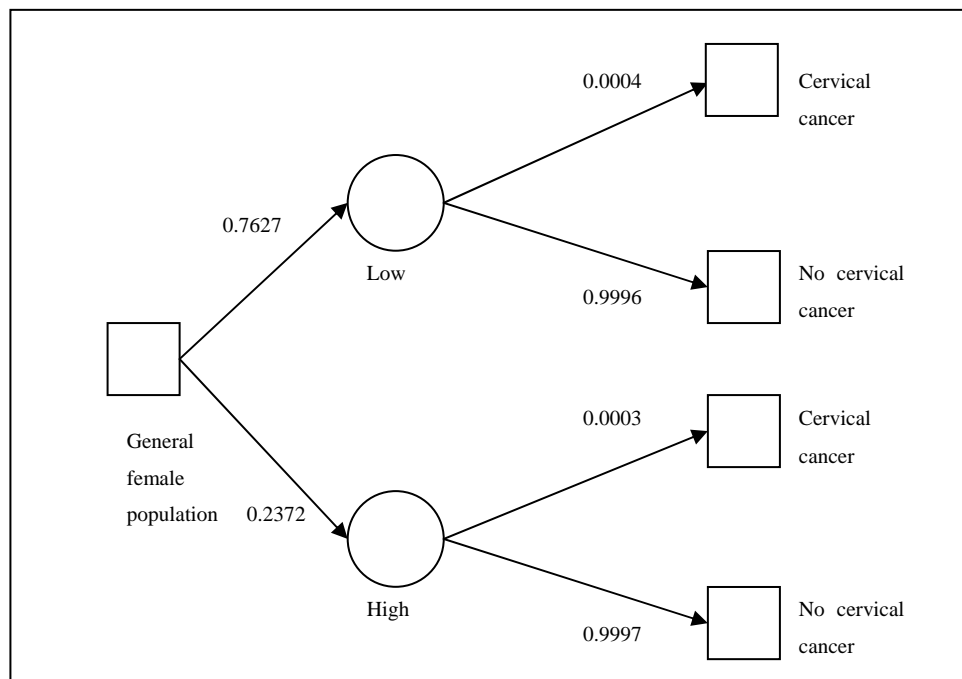


Figure 5.26 Decision tree model with two risk groups. The low risk group has a low Townsend index score and the high risk group has a high Townsend index score. The probabilities were estimated from the observed national data.

Table 5.9 Potential decision tree (Figure 5.26), based on the national observed data. Population was split into two risk groups.

Group	Total patients
General female population	20,000,000
Low risk group	15,254,000
Cervical cancer	6,102
No cervical cancer	15,247,898
High risk group	4,746,000
Cervical cancer	1,424
No cervical cancer	4,744,576
Total cervical cancer	7,579

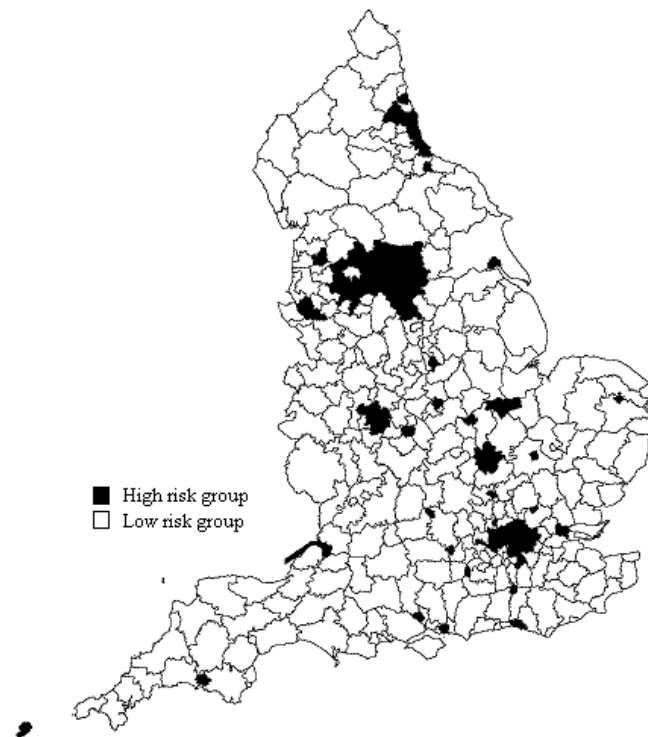


Figure 5.27 Risk grouping by the Townsend index score. the black areas represent the high risk group and the white areas represent the low risk group.

5.8 Summary

In this chapter, various regression models were used to explore the relations between cervical cancer incidence and Townsend index as well as other variables. The results summarised that the Townsend index, personal marital status and

social grade were significant to cervical cancer incidence. Particularly, the GWPR results showed that the significant variable proportion of low social grade population was fitted best by non-stationary models, which means the contribution from the same variable might vary over space. The non-stationary process can increase the prediction power and increase understanding of the relationship between cervical cancer incidence and explanatory variables over space. If the non-stationary assumption and spatial information were ignored it could lead to misleading explanation of the results. It demonstrated that not everyone has the same chance of developing cervical cancer; some patients with certain characteristics (e.g. high or low social status) have a different chance of developing cervical cancer. Therefore, the lower risk patients may not need to be screened as often as the high risk patients. The maps (Figures 5.27) showed the risk grouping according to the Townsend index. In practice, it is possible to screen the high risk population in the black areas in the map more often than the white areas. This could reduce the amount of wasted resources. The exact resources can be added into the high risk groups. That may help to detect the higher risk patients at any early pre-cervical cancer states.

However, something is missing. That is the time parameters and transition probabilities. If the transition time and probabilities were available between states, it would allow estimation of number of patients in each state for a long run. The result in section 5.7.2.1 allows evaluation of the effectiveness of the screening options and it also allows comparison of different options.

Both Chapter 4 and this chapter demonstrated the use of different mathematical models, which allows identifying the personal and other factors that are associated with cervical cancer incidence (e.g. age). The simple analysis in section 5.7 showed that the chance of developing cervical cancer does vary between risk groups. Surely, including both personal and general associated risk factors may increase the screening efficiency. The efficiency of the national cervical cancer screening policy at local and (or) national levels is discussed in Chapter 8.

Chapter 6 Chlamydia study

6.1 Introduction

On average there is a 10% positive rate of Chlamydia within the entire UK population most commonly under the age of 25 (Primarolo, 2006). The numbers of new Chlamydia cases has increased annually from the mid-1990s; the confirmed cases rose by 5% between 2004 and 2005. It has become the most common sexually transmitted infectious disease (STD/ SDI) in the UK and it is also one of the sexual health issues causing most concern worldwide. The UK Department of Health carried out an opportunistic screening trial study in 10 centres to collect data between 1999 to 2000, which will be used to inform an extension of the screening programme to the whole country in the future as part of the National Chlamydia Screening Programme (NCSP). Portsmouth was one of the chosen sites for the opportunistic screening trial of Chlamydia. The target population was estimated at 30,000 (Evenden *et al.*, 2006). If it were necessary to set up an official national screening programme as has been done for other diseases such as cervical cancer, it would be interesting to consider what type of screening system should be adopted. If a fixed screening programme with fixed screening interval was implemented (e.g. cancer screening programme) to cover the whole population, this may increase work pressures and overstretch healthcare resources.

This chapter demonstrates (i) how to model individual patients' age, which will be used in the regression model in chapter 7, (ii) how to use the decision tree model to examine what is the benefit of setting up a screening programme and what is the best possible screening option, and (iii) how to use simulation to understand the screening systems and to predict the necessary resources. All the techniques used in this chapter are mainly used to demonstrate the analysis that would be possible if data were available. Real data were used where available, but where data were missing, (e.g. the transition time and some transition probabilities) dummy datasets were used to demonstrate the use of the methods.

The associated risk factors can be identified through the regression models in Chapter 7, those risk factors can be used to divide population into groups, according to the identified associated risk factors in Chapter 7. The probabilities and other required parameters of each risk group can be added into the decision tree and simulation models in this Chapter to demonstrate the potential use of decision tree and simulation models to evaluate the screening options, thus, the optional screening policy can be achieved through such a process.

6.2 Data

The individual Chlamydia data collected between 1999 and 2000 in Portsmouth were used in this chapter and some of the probabilities from NHS reports (Primarolo, 2006) are used in the decision tree model and simulation models (Pimenta *et al.*, 2003 a, b; Health Protection Agency, 2006). For details of the Chlamydia data please refer to chapter 3.

6.3 Age

It is common to have aggregated data in most disease studies; for example in Chapters 4 and 5, national cervical cancer data were used. Information on an individual patient's age is not usually available, but it provides rich information for understanding how a disease relates to patients within the given age range, and which age groups have a higher chance of developing certain types of disease. Such information helps to identify the high risk populations, to target the high risk populations and to run the screening programme in a more efficient way.

6.3.1 Age category

It is common to model the individual patient's age as a categorical variable (Clayton and Hills, 1993; Evenden *et al.*, 2006), that is, split the patients into two groups or a few groups based on the risk (e.g. low or high risk). In the Chlamydia study, the age group 16 to 24 was classified as high risk group; the rest of the ages were classified as the low risk group. However, this strategy causes some loss of information. In practice, it is possible to include the individual age in regression,

but it causes complications in modelling (Jackson *et al.*, 2008); it is not easy to model the age (Jackson *et al.*, 2006). It is helpful to estimate the age distribution, which is a possible method to overcome the use of age categories (lost information).

6.3.2 Age distribution

Kolmogorov-Smirnov is a test used to examine the underlying distribution, which is similar to Chi-square goodness of fit test. The main application of this test was to test the age distribution. Firstly, the positive risk z_i and the normalised positive risk w_i are defined below. The reason to normalise the risk z_i is because the risk z_i is always assumed the same (equal risk rate) with different data size; therefore, $\sum_{i=12}^{41} w_i = 1$ for $i = 12, 13, \dots, 41$ in total 28 age classes were recorded from ages 12 to 41, there were no cases recorded between ages 0 to 11 due to the sample collection. The basic distribution $F(x)$ is tested in the following section (DeGroot and Schervish, 2002).

$$z_i = \frac{x_i}{y_i} \quad (6.1)$$

$$w_i = \frac{z_i}{\sum_{i=12}^{41} z_i} = \frac{z_i}{\sum_{i=12}^{41} \frac{x_i}{y_i}} \quad (6.2)$$

Where y_i is the number of tests, x_i is the number of positive incidences (positive cases), i is the patient's age $i = 12, \dots, 41$; the test statistic is given below,

$$n^{1/2} D_n^* \quad (6.3)$$

Where $D_n^* = \sup_{-\infty < x < \infty} |F_n(x) - F^*(x)|$, D_n^* is the maximum difference between the same distribution function $F_n(x)$, and the hypothesised distribution function $F^*(x)$. Therefore, it is necessary to estimate the sample mean $E(\overline{x_i})$ and sample

variance $\hat{\sigma}_{xi}$ given sample size n equal to 28, since the recorded ages started from age 12 to 41, so in total there were 28 age classes.

In this section, the age distribution is tested; the potential distributions were considered (i) Normal or (ii) Log-Normal distribution. The Kolmogorov-Smirnov test is set up as following,

The hypothesis test is given below,

H_0 : the basic distribution is Normal (or LogNormal) distribution, $F(x) = F^*(x)$

H_1 : the basic distribution is not Normal(or LogNormal) distribution, $F(x) \neq F^*(x)$

Level of significance: 95%

p -value: 1.36

Table 6.1 Kolmogrov-Smirnov test results from both Normal and Log-Normal distributions.

	Test 1	Test 2
Distribution	Normal	Log-Normal
$E(\bar{x}_i)$	6.259	2.941
$\hat{\sigma}_{xi}$	6.259	0.298
D_n^*	0.118	0.0972
$n^{1/2}D_n^*$	0.635921684	0.523484
P-value	1.36	1.36

Since both $n^{1/2}D_n^*$ values (0.636 and 0.523) are less than 1.36, H_0 is accepted.

The basic distribution can be Normal and Log-Normal. Using a QQ-plot to compare the fitted distribution (Figure 6.1), showed that both Normal and Log-Normal distributions were fitted well, but it does not show which is better. Log-Normal had a smaller $n^{1/2}D_n^*$ value than Normal. Therefore, it is possible to accept a Log-Normal distribution as the underlying distribution. The final fitted observed and expected distributions are shown in Figures 6.2 and 6.3.

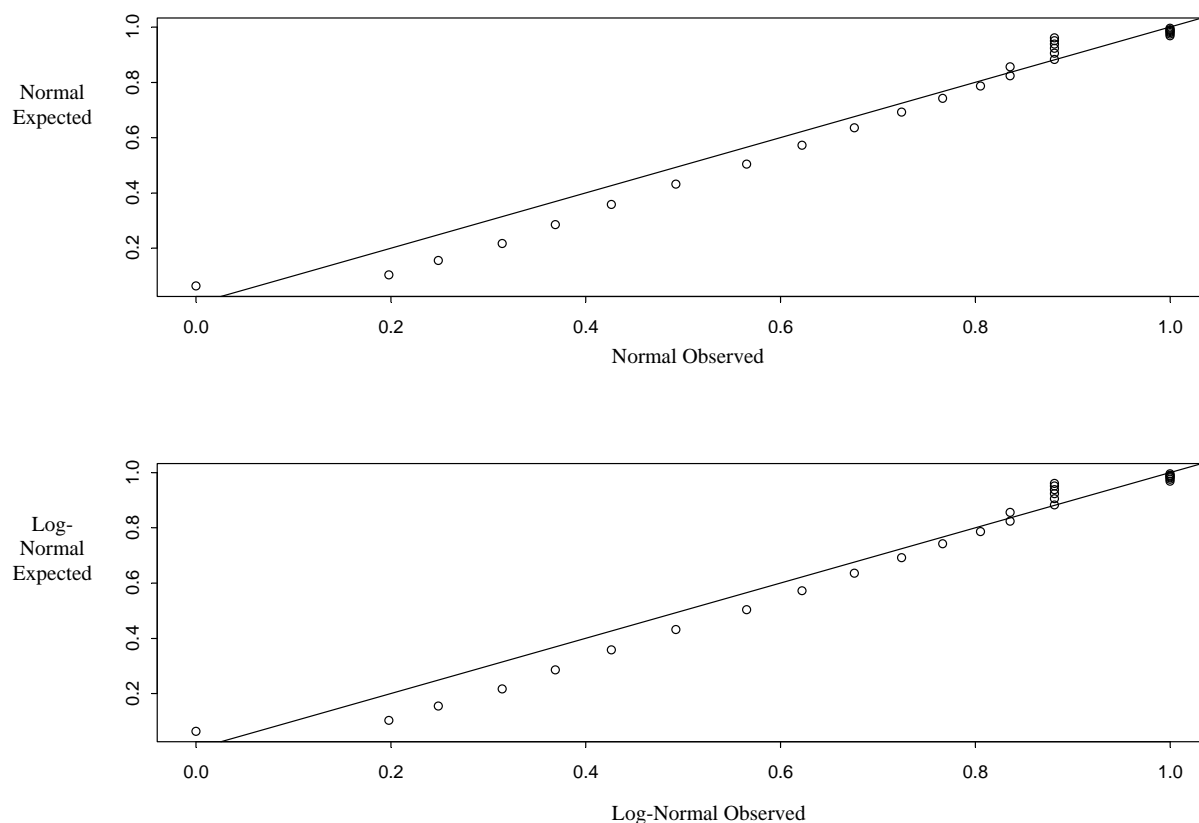


Figure 6.1 Q-Q plot, to compare the goodness of fit between the observed risk and expected risk with two distributions. The Log Normal distribution is fitted better than the Normal distribution, thus, the basic distribution is Log-Normal.

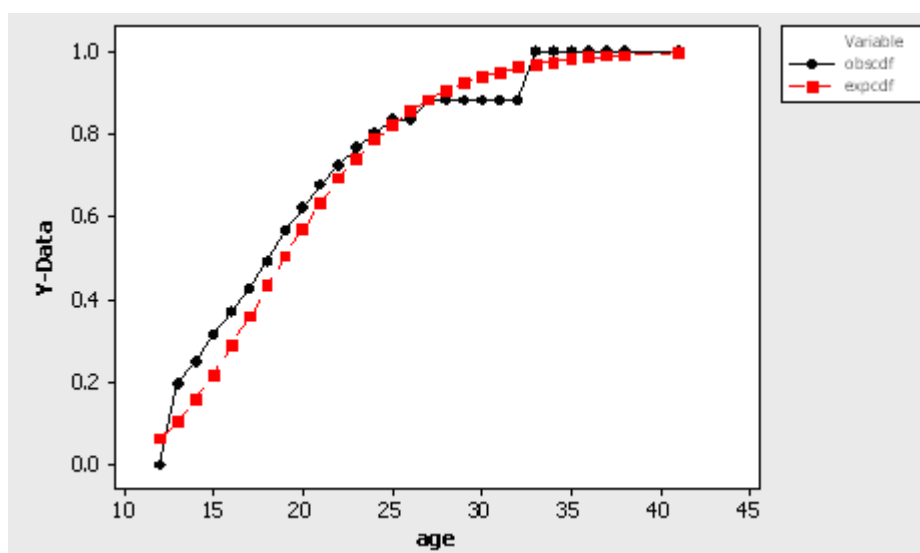


Figure 6.2 The cumulative distribution function (CDF) curve of observed and expected distributions. The line with dots represents the observed CDF and the line with squares represents the experimental CDF.

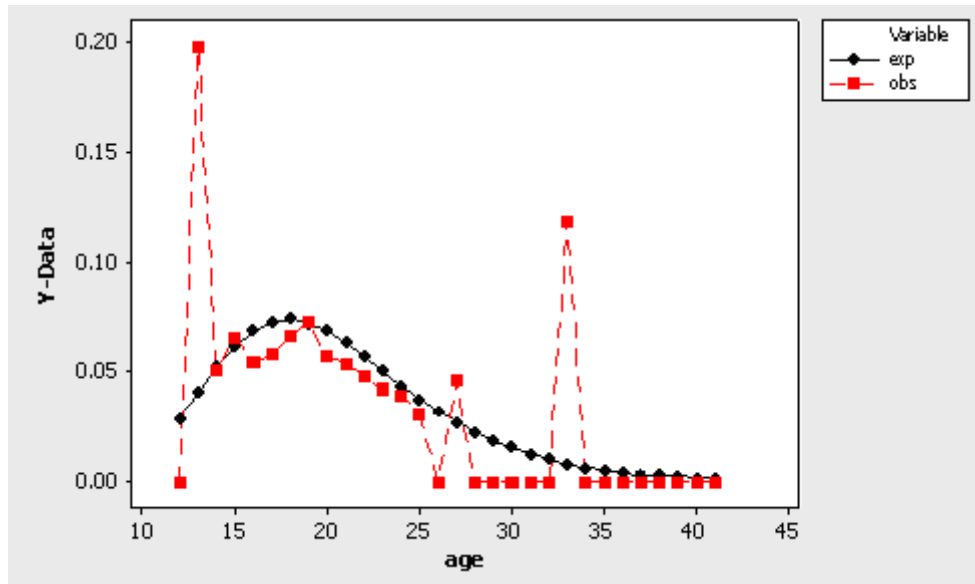


Figure 6.3 The observed and expected probability per age classes, the line with squares represents the observed probability, the line with dots represents the experimental distribution.

In theory, age and sex are treated as extra information, which contribute to the risk of developing disease at the individual level. However, such data can over-complicated the model in practice. Thus, it was suggested to estimate the age by indirect standardization by Jackson *et al.*, (2006; 2008) and the best choice for modelling age in a Poisson model is the *logit* function. Therefore, in practice it is required to define the baseline risk of disease per age. Thus for the Chlamydia study, the expected probabilities approximately followed the Log-Normal distribution, but that only represents the expected probabilities per age. However, the risk of developing Chlamydia is the parameter of interest in the model. The expected probabilities from the Log-Normal distribution are not used in the regression model directly (in Chapter 7). Rather, the expected probabilities from the Log-Normal distribution were transformed into the *logit* scale and finally multiplied by the scale factor 1.6843 (Table 6.2 and Figure 6.4). The final outcome represents the risk of developing Chlamydia infection rather than the probability of developing Chlamydia infection per age.

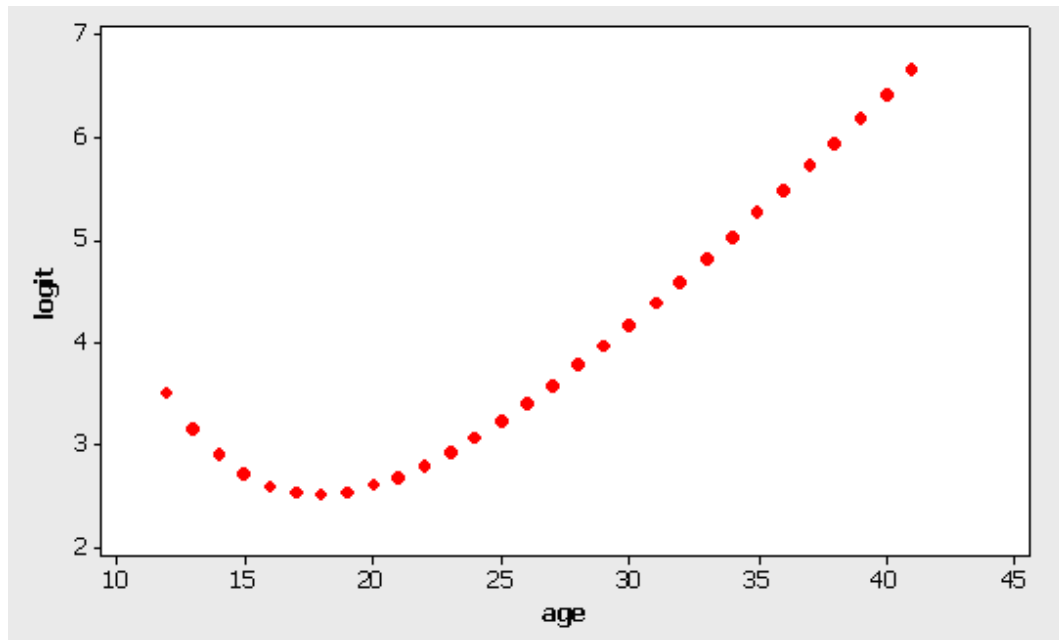


Figure 6.4 *Logit* of the expected risk per age group

6.4 Decision tree model

Decision theory became more popular in evaluating healthcare options, particularly with in the NHS, where it has been applied in many NHS health studies; for example, assessing cost-effectiveness of chest physiotherapy, screening and also treatments (Claxton *et al.*, 2004) and there is some evaluation of examples of healthcare technologies (Claxton, 1996; Claxton *et al.*, 1999; 2000; Claxton and Thompson, 2001; Sculpher *et al.*, 1997) Decision theory was used to evaluate the screening options; a decision tree model was constructed in this section based on the Chlamydia disease states described in Figure 6.5 and the simple version of disease states in Figure 6.6. Based on the tree structure, it allows us to evaluate each of the possible screening options, which might provide better opportunities to detect more positive Chlamydia cases. The Chlamydia tree structure in Figure 6.7 is based on Figure 6.6.

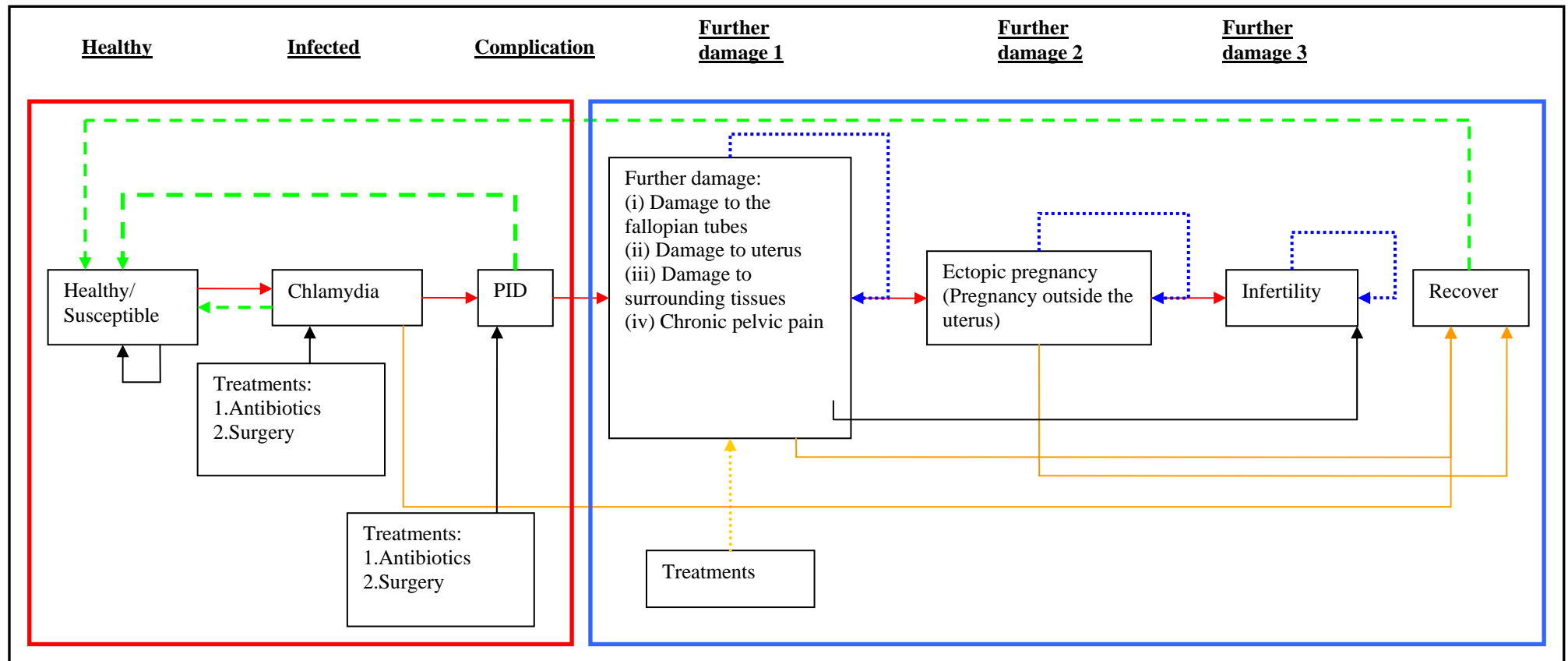


Figure 6.5 Chlamydia disease states.

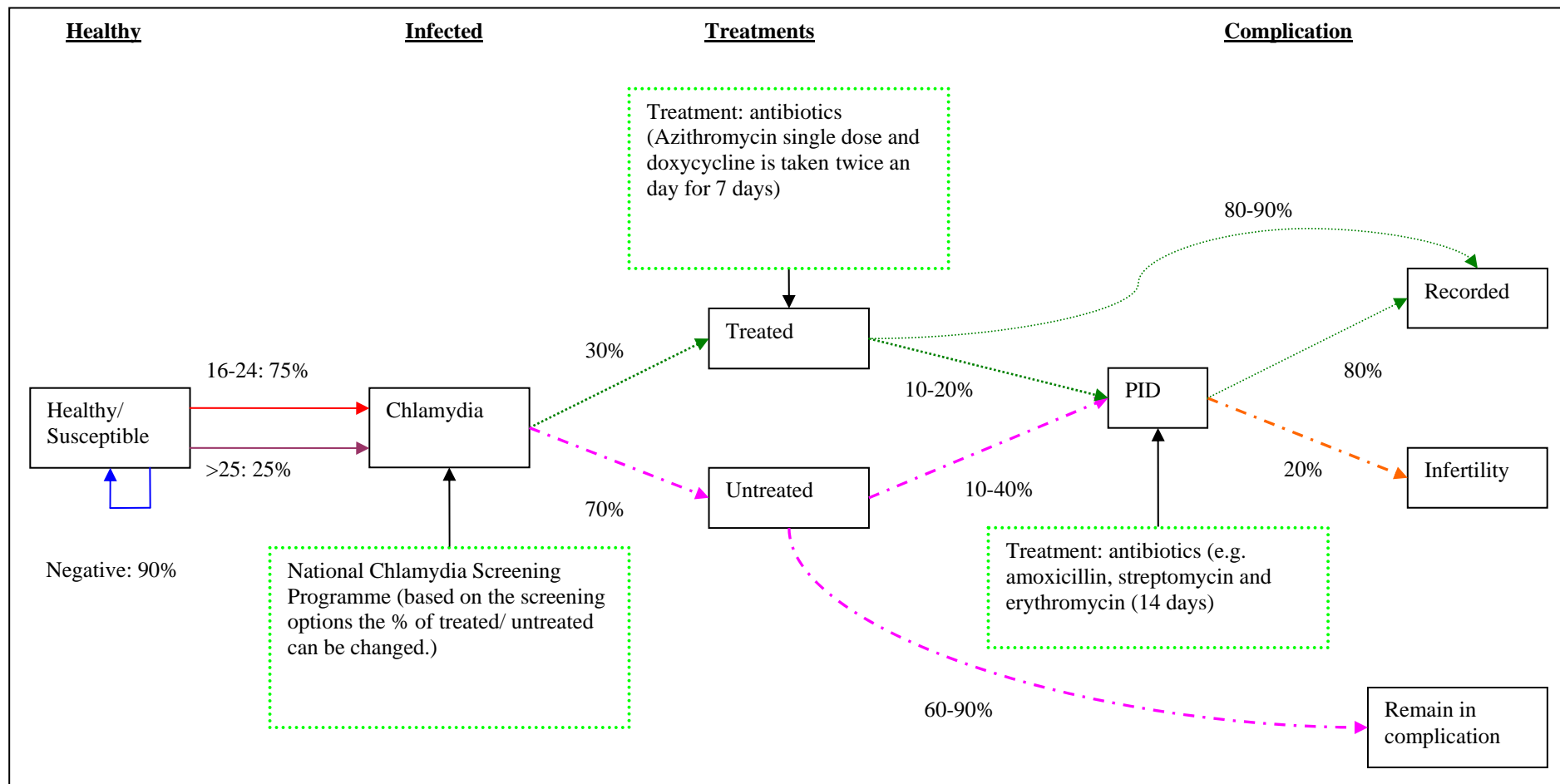


Figure 6.6 Simple version of Chlamydia disease states.

6.4.1 Decision tree structure

A decision tree model was constructed to evaluate the best option for Chlamydia screening, which would achieve the best payout (i.e. more detected cases and a reduction in the number of cases of PID, complications and infertility). The model structure is shown in Figure 6.7; the circle represents the points at which decisions should be made and the square box represents the final output from that route. The aim is to make a decision which can maximise the payout, which is to decrease the number of undetected cases through screening.

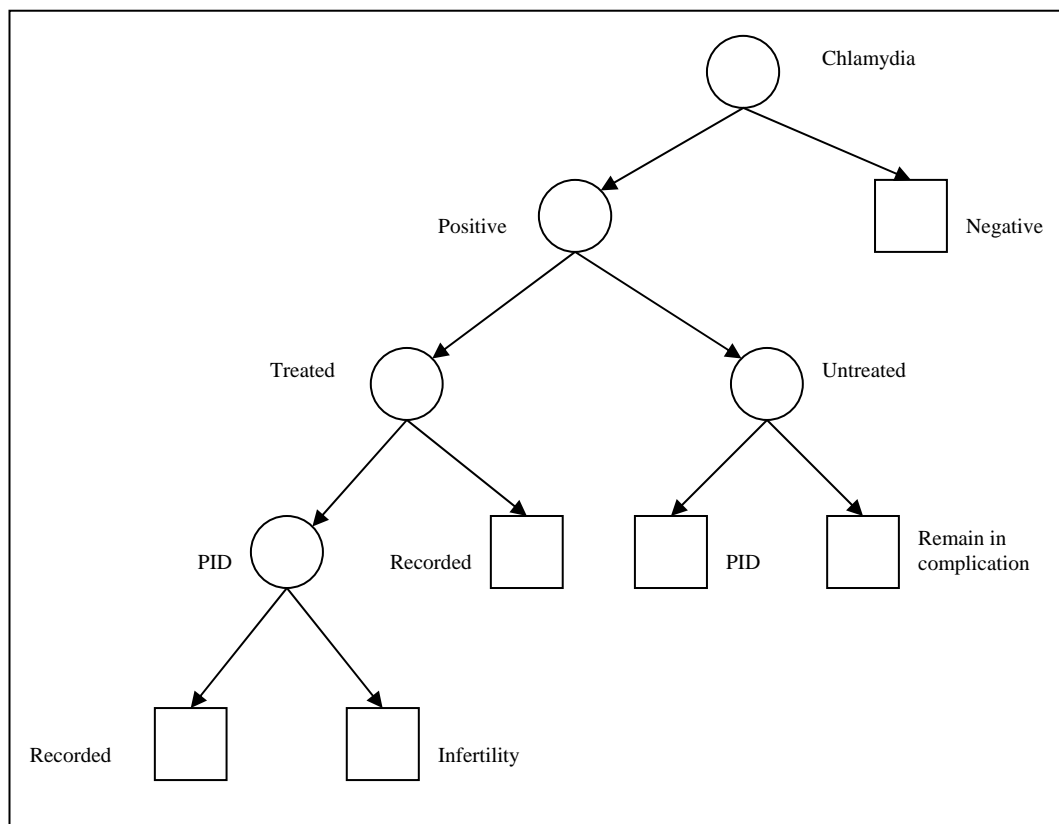


Figure 6.7 Chlamydia decision tree model.

An example was utilised to explain the use of decision theory, to identify the best option and to identify what might be possible to improve with the given best screening option. The result of the example is shown in Table 6.2. The results in Table 6.2 show that if policy makers only concentrate on improving the detection rate for the positive number of Chlamydia tests, this is not enough. In practice, it is necessary to adopt a combination of improving screening to avoid undetected cases and also to increase the quality of treatment services to take care of those patients who remain in undetected states (i.e. patients transmit to further

damage stages e.g. infertility), otherwise the number of cases of infertility would increase despite the screening becoming more effective and the number of undetected cases decreasing. The details of the decision tree model are given in Appendix F. Based on the lack of probability information, the probabilities used in this section are dummy data, which can demonstrate the use of a decision tree model to evaluate screening options.

Table 6.2 Examples of screening options for Chlamydia

Option	Treated %	Untreated %	Recorded	PID	Infertility	Remain in complication
No screening option	0.3	0.7	2910	1750	90	5250
Option 1	0.5	0.5	4850	1250	150	3750
Option 2	0.7	0.3	6790	750	210	2250
Option 3	0.8	0.2	7760	500	240	1500

The probabilities in Figure 6.6 and Table 6.2 were collected from a list of reports from the NHS and HPA (Health Protection Agency, 2006; Primarolo, 2006). In Table 6.2 there were four options, the first option is no screening, which reflects the current situation, when no official screening is available for anyone at regular periods. Options 1 to 3 represent the official screening becoming available with different levels of efficiency of screening policies. For example, Option 1 had less efficiency compared to Options 2 and 3. Therefore, more patients had PID and remained in complication, whereas Option 3 was more effective and it detected more Chlamydia patients, so fewer patients had PID and remained in complication. The treatment efficiency is needed to increase to match with the screening efficiency; otherwise more patients will become infertile. The use of the decision tree model helps in evaluating the screening options, it clearly shows that more patients with Chlamydia are recorded if there is a screening system available; however, it is noticeable that the number of infertility patients will increase if the treatment for PID does not improve when Chlamydia screening becomes available. Therefore, the decision tree model allows the identification of possible bottle necks. On the other hand it is necessary to improve the treatment for the subsequent illness.

In section 6.5, a simulation model is used to demonstrate what is possible in terms of improving the screening system and predicting the possible demand in the future.

6.5 Simulation

Simulation modelling allows those concerned to simulate disease behaviours and the patients' response to treatments; therefore, decision makers can create a simulation model to model and analyse a life threatening condition when applying the new system. Two key characteristics are needed to describe the patients' behaviours and response to treatments. Such characteristics can be described by the transition probabilities and holding times, which describe how likely a patient is to move to further disease states and how long it takes a patient to transfer to other states from his or her current state. Both transition probabilities and time parameters are described in the next two sections. Parameters and probabilities in the simulation model are dummy datasets due to lack of information.

6.5.1 Simulation model structure

A simulation model describes a disease system; it describes the process of each individual transforming through the system, which has a finite number of disease states. The period of staying in a disease state is called Length of Stay (LoS) and the chances or probabilities of transforming to other states (Figure 6.8) are described by the holding time t_{ij} and transition probabilities p_{ij} .

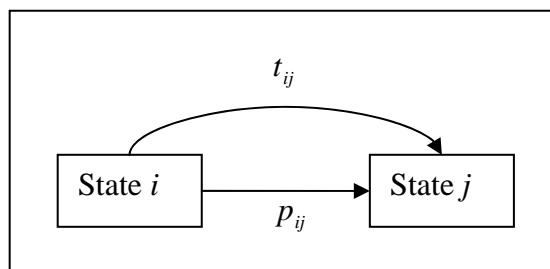


Figure 6.8. Transition time and probability from state i to state j

To simulate the passage through the system, data such as holding time and transition probabilities within each state are required. Two types of models are

suitable for the model structure, Markov and Semi-Markov models, and use of one or the other depends on what types of information are available. For the Markov model the transition probabilities are required, but not the holding times, whereas the Semi-Markov model requires both transition probabilities and times. A variety of distributions can be used - for example, Weibull, Normal and Log-Normal. The Weibull distribution is commonly used to describe the holding time if data are not available (Harper, 2002).

Patients may stay healthy for a certain time before moving to the next state (e.g. Chlamydia positive, PID etc). Such a study allows understanding of the natural disease process and it is also possible to ask some “what if” questions, which help to evaluate and improve the screening options.

The simulation model was developed by using Visual Basic for Applications (VBA) within Excel and details were attached in Appendix G. Users are allowed to change the probabilities between stages and the parameters of the transition distribution. The length of the simulation period and numbers of patients to start with are decided by the users, and the simulated period is measured by month. In theory, it allows changes to the types of the distribution to be made. Each individual patient was followed through the tree (Figure 6.9 and 6.10) and the results were stored in an Excel worksheet. An example was used to demonstrate how to use this simulation model and the results. A uniform distribution was used to describe the transition probability and the holding time was described by the Weibull distribution, which is one of the common choices when the transition time distribution is unknown, or it is not possible to estimate from the data, or there are no data available. The transition probabilities shown in Figure 6.9 were collected from a list of reports and literature (Health Protection Agency, 2006; Primarolo, 2006). There is no national screening programme available in England. Therefore, the results were compared between (i) no screening programme is available and (ii) screening programme is available.

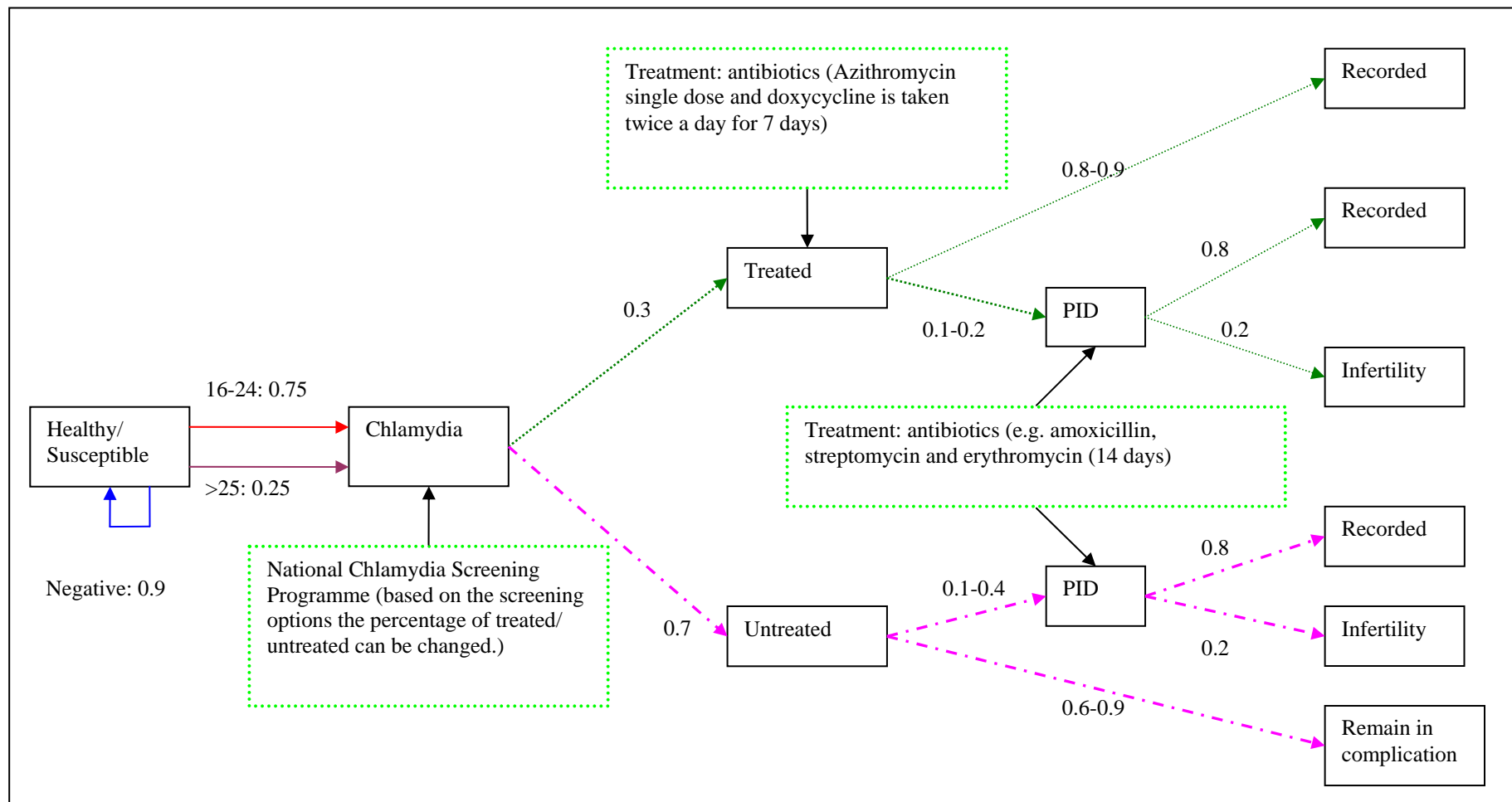


Figure 6.9. Chlamydia disease system

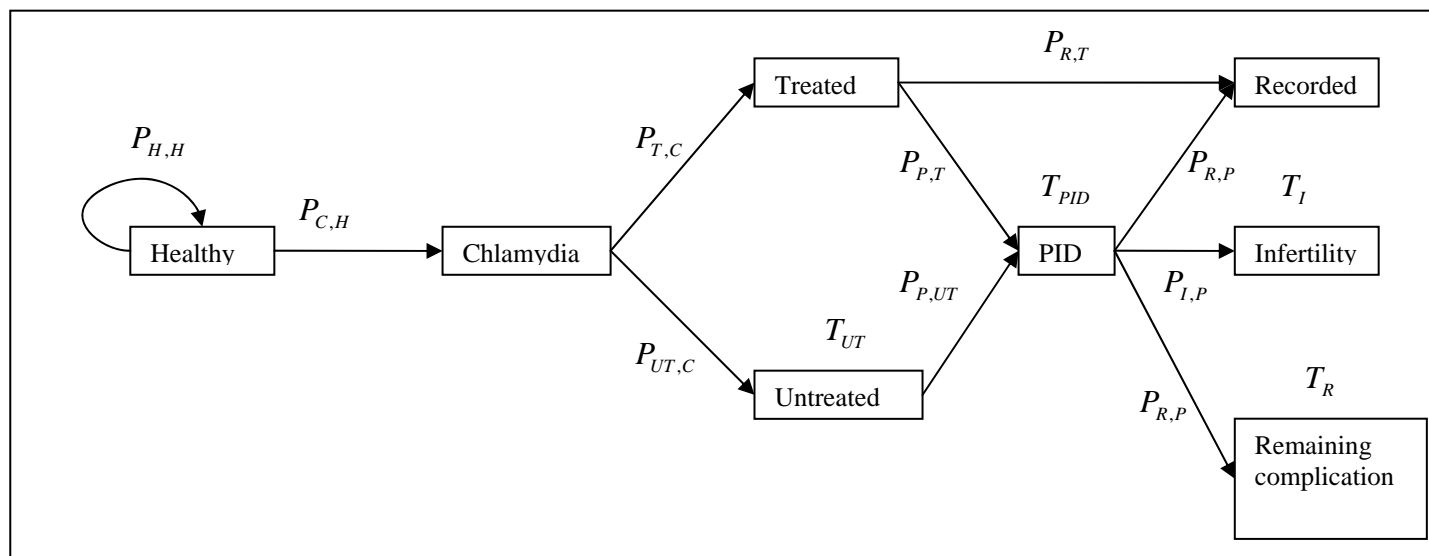


Figure 6.10 Chlamydia simulation model.

6.5.2 Simulation results

A set of dummy holding time and transition distribution and summary transition probabilities (Health Protection Agency, 2006; Primarolo, 2006) were adopted to demonstrate the use of simulation in a Chlamydia screening programme, and when individual patient data become available, the distribution and parameters can be estimated from the available data. Users are allowed to change the distribution within the built-in model to adapt it to the purpose required.

To start simulation, users need to enter the simulation period in months (i.e. start and end dates are needed and the simulation period can be estimated from the given dates), and the number of patients go into the simulation system per month as listed below:

Table 6.3 Simulation results from a simple Chlamydia simulation model

Information	Result
Transition distribution between states	Uniform
Holding time between states	Weibull
Simulation period (months)	10
No of patients per month	1,000
Final positive patients	1,040
No. of patients in complication state	449
No. of patients in PID state (some of them will be recorded after receiving treatment)	330
No. of patients in infertility state	64
No. of patients recorded at the final state (from PID or complication after receiving treatment)	527

The simulation with dummy datasets was used to demonstrate what might be possible. It summarised the natural Chlamydia process, if the required data (e.g. transition time) are available. Table 6.3 showed an example of how many patients were in each state. Users can change the distributions and full details of the simulation model can be found from Appendix G.

6.6 Summary

The results from decision tree and simulation models show a possible method to evaluate the screening option and to understand the Chlamydia disease process when required data become available. In particular, Section 6.3 demonstrated how to model the individual patients' age, which had been commonly modelled as categorical (e.g. low or high risk categories) (Clayton and Hills, 1993; Jackson *et al.*, 2006; 2008), causes a loss of information. Ideally, keeping the age of individual patients may increase the prediction power, and at least will not lose any information. Thus, modelling age distribution is a possible way forward. Such distributional information will be used in Chapter 7 in some of the regression models.

Chapter 7 Chlamydia regression models

7.1 Introduction

Nationally, the Chlamydia infection rate is 1 in 10 (Primarolo, 2006), and 70% of female and 50% of male patients remain asymptomatic at the early disease stages (Health Protection Agency, 2006); however, the consequences of undiagnosed and untreated Chlamydia can lead to complications. A significant proportion of female patients (10-40%) develop Pelvic Inflammatory Disease (PID) (Health Protection Agency, 2006), the acute Chlamydia infection including salpingitis. If this condition remains untreated it can cause serious damage, such as chronic pain, ectopic pregnancy and even infertility. These complications and damage only occur in female patients. The male will commonly experience a urethral discharge from the penis, and further complications and damage include inflammation and fertility problems. There is approximately a one in two chance of a man experiencing impaired fertility or epididymitis (Health Protection Agency, 2006).

This chapter demonstrates the use of the generalised linear regression model, multilevel Bayesian regression model and geographically weighted regression (GWR) model to explore the relationship between positive Chlamydia results and the deprivation indicator (Townsend index), social status and family structure variables at different regional levels. The multilevel regression model is a well known type of statistical tool for exploring the relationship between target and explanatory variables when the target and explanatory variables are observed at different levels; (i.e. multilevel). The GWR allows local variation in relations (i.e. non-stationary model). Ignoring the spatial variation could cause misleading interpretations of the relations between target and explanatory variables. In order to identify the linkage between positive Chlamydia cases and deprivation condition, social status and family structure variables, which would provide a better understanding of Chlamydia and enable policy makers to view the current healthcare problems from a more informed position, more valuable information can be added into the screening policy.

The objectives of the study were (i) to explore the relationship between positive Chlamydia incidence and deprivation indicator, social grade and family structure factors, (ii) to identify risk factors that are associated with positive Chlamydia infection, (iii) to display the Chlamydia incidence pattern in the Portsmouth area between 1999 and 2000, (iv) to divide the population into sub-groups, and (v) to target the high risk sub groups and to suggest possible screening strategies. The area covered by this Chlamydia study is showing in Figure 7.1. However, some of the Output Areas (OAs) do not contain any observed data. Therefore, there are many disconnected Output Areas in the mapping. Thus, the study area was reconsidered in the data section.



Figure 7.1 Study area at CAS Ward level 2001, the basic shape file is available from Edina.



Figure 7.2 STD clinics in the Portsmouth area; A represents the St. Marys hospital, B represents the family planning clinic, C represents the X-Perience young persons centre, D represents the social services in Fareham, E represents the family planning in Gosport, F represents the Brune park youth centre, G represents the Lee on the Solent focus youth club.

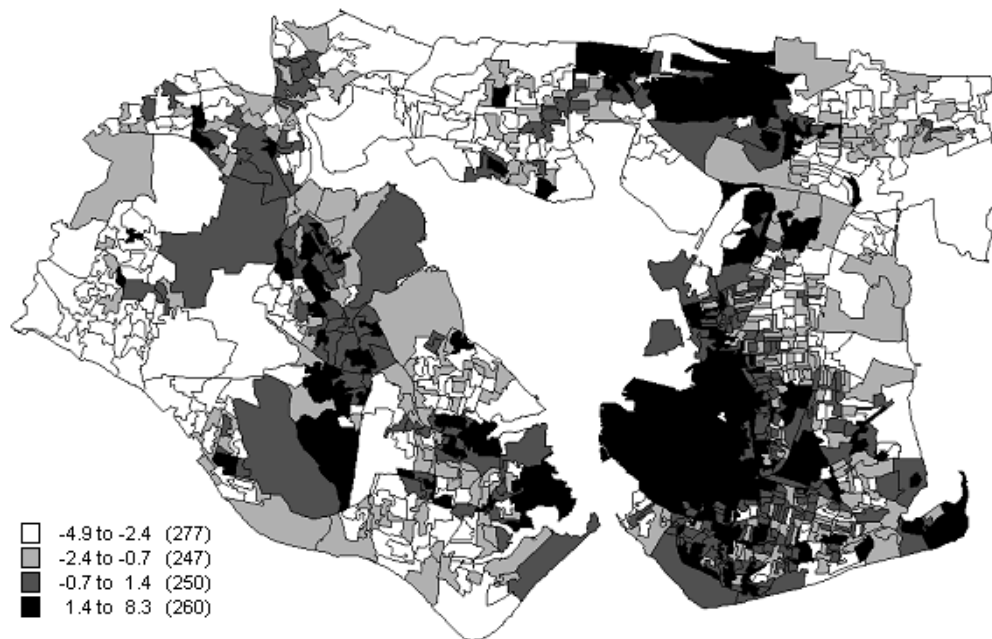
Patients are free to choose where to take their Chlamydia test. Therefore, it is possible to believe that where the test is taken depends highly on the location of the clinics. This explains why very few cases came from Fareham, Southampton etc., there are other clinics available and closer to Fareham and Southampton. For this reason, regions not close enough to the clinics and disconnected regions were not included in this Chlamydia study.

7.2 Data

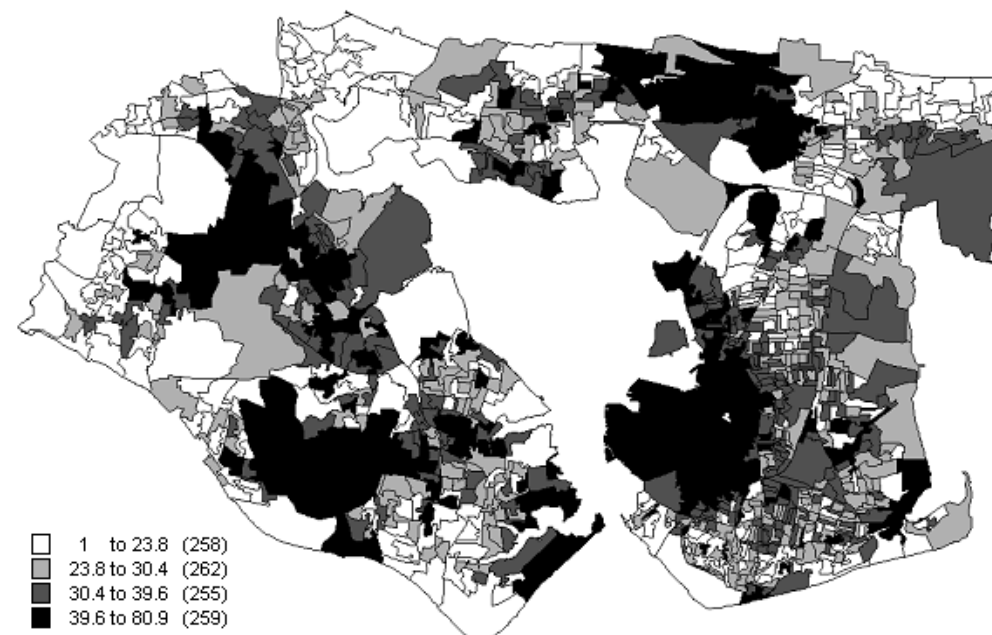
The data were described fully in Chapter 3; since every Output Area contains observed cases, the overall study area was reconsidered in this section as shown in Figure 7.2 and 7.3 to avoid disconnected areas. The individual data on the occurrence of Chlamydia in Portsmouth between 1999 and 2000 were used. The

individual deprivation indicator, social grade and family structure variables were not available; the lowest available level is Output Area (OA) and the second lowest level is CAS Ward levels. Therefore the chosen levels for the explanatory variables were OA and CAS Ward levels. The explanatory variables were downloaded from the UK censuses 2001. For full details of the data please refer to Chapter 3. The explanatory variables were mapped at OA and CAS Ward levels and displayed in Figures 7.3 to 7.4.

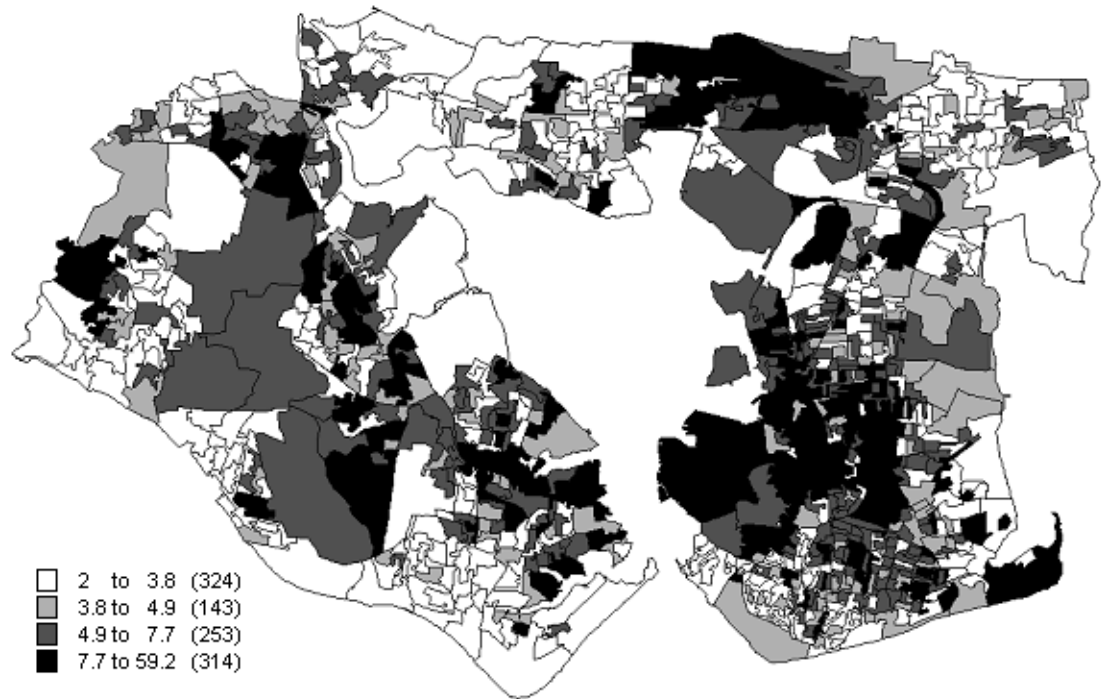
(a)



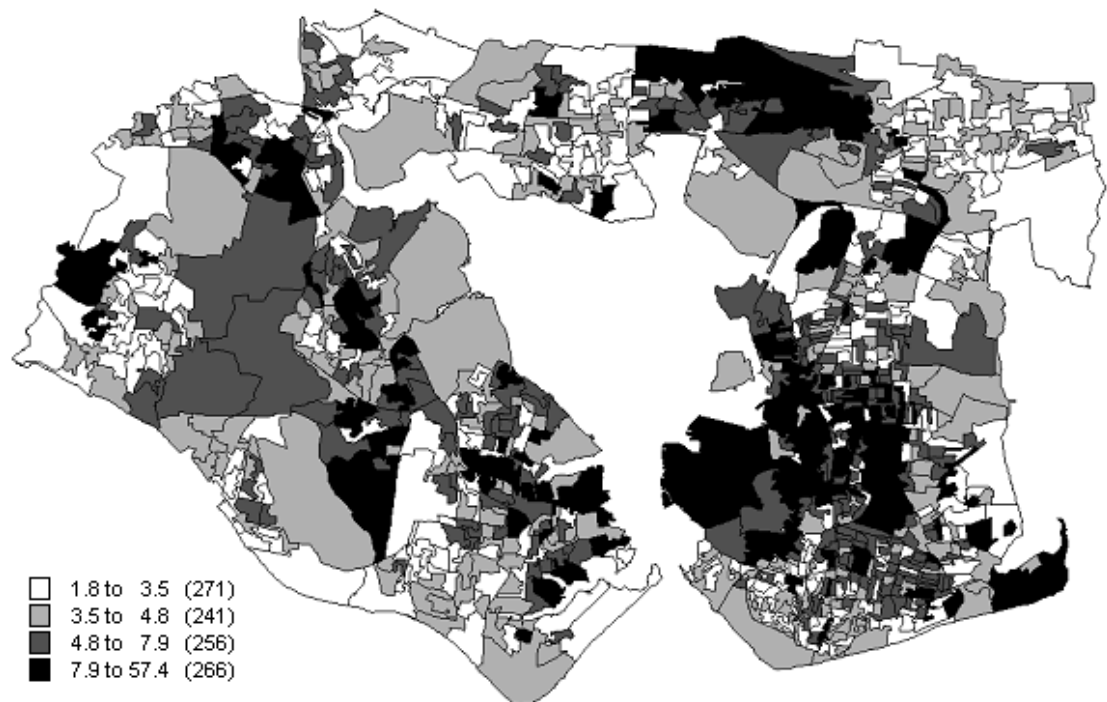
(b)



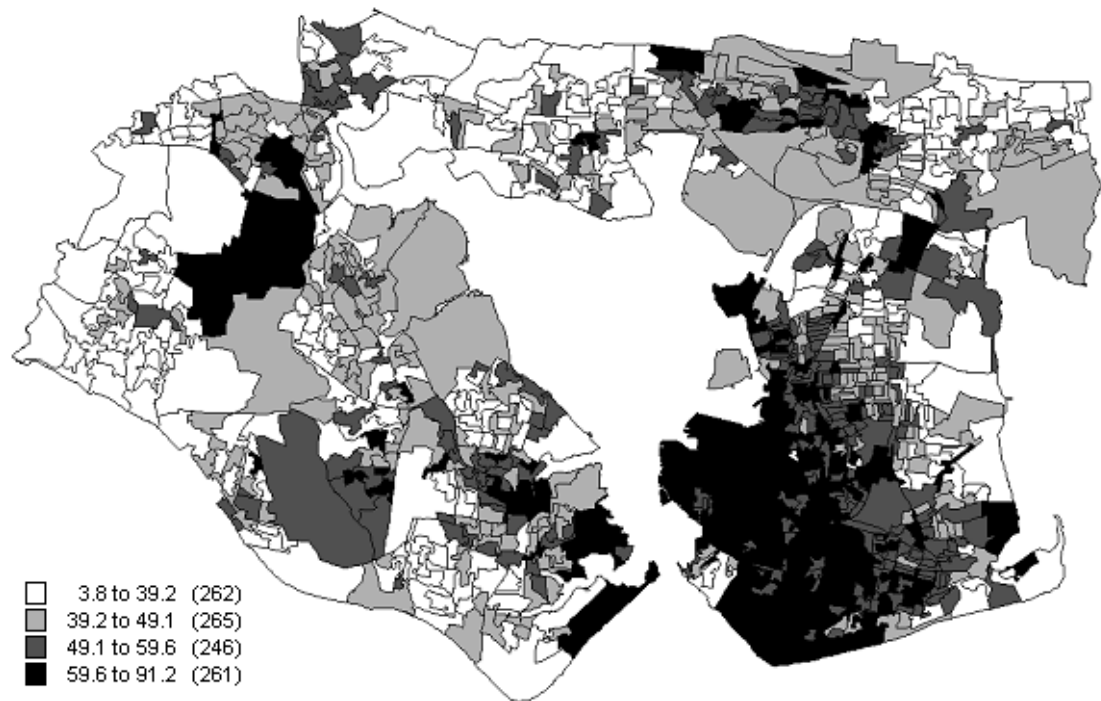
(c)



(d)



(e)



(f)

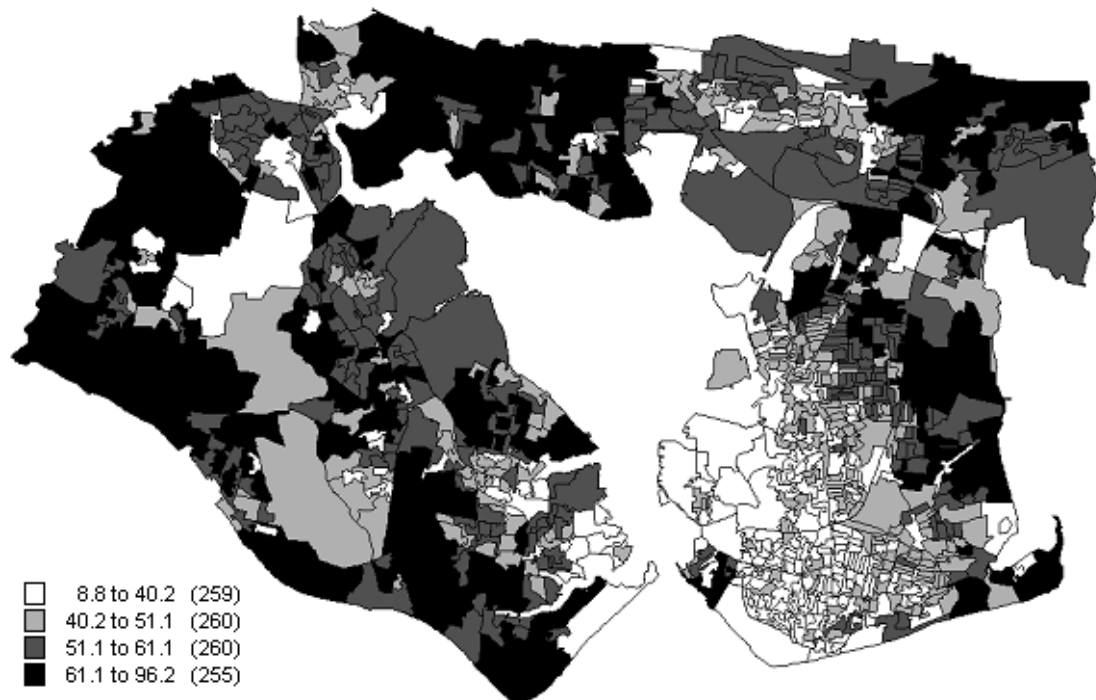


Figure 7.3 (a) Townsend index, (b) percentage of socio-grade IV+V, (c) households with all lone parents, (d) households with female lone parents, (e) single population and (f) married population at Output Area level.

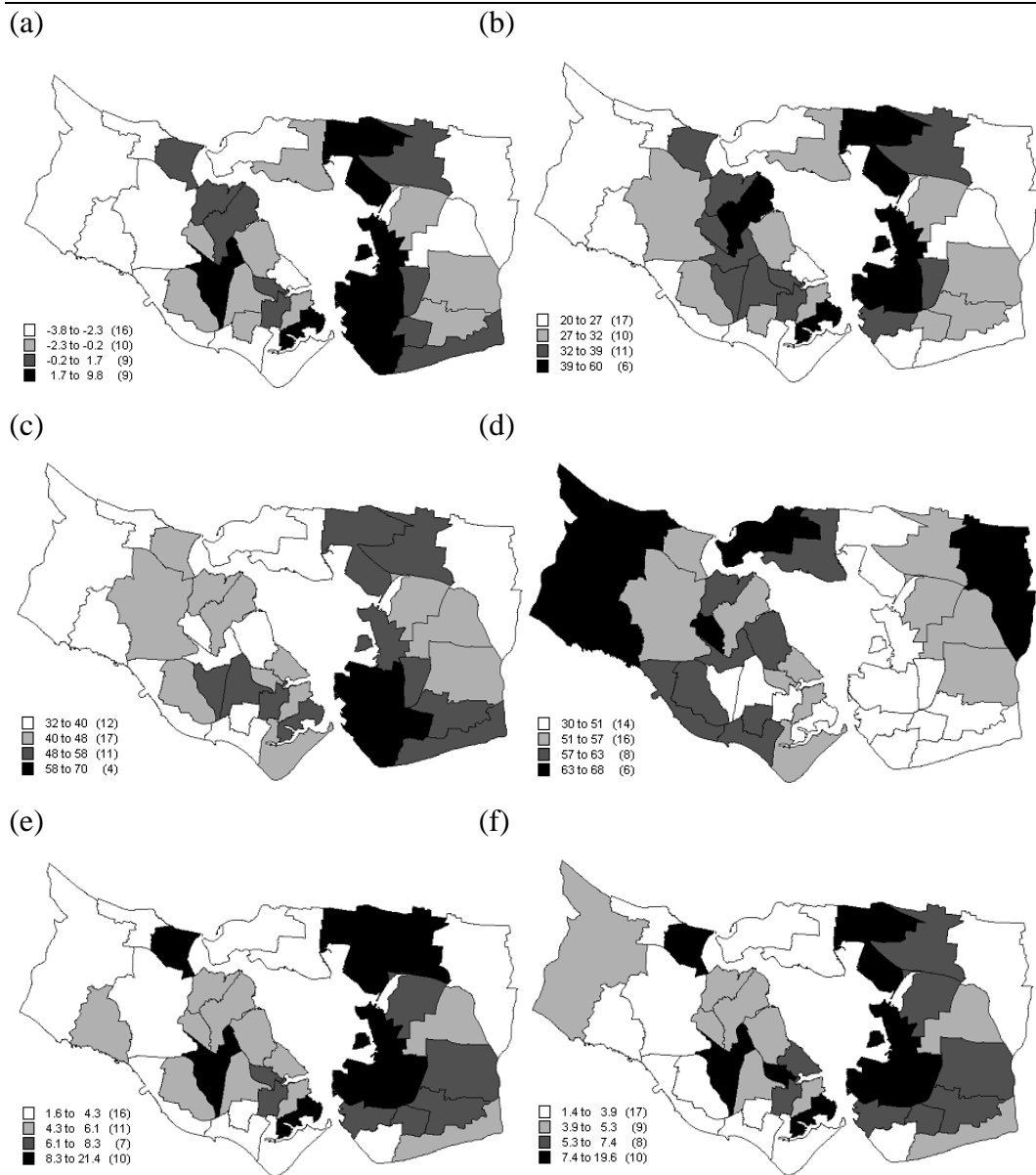


Figure 7.4 (a) Townsend index, (b) percentage of socio-grade IV+V, (c) households with all lone parents, (d) households with female lone parents, (e) single population and (f) married population at CAS ward level.

In addition, the missing test results (or incomplete records) and the repeated test records (i.e. a patient had a repeat test to check for the current stage) were removed from the analysis and modelling work, since interest is in the first test result only.

7.3 Exploratory analysis

7.3.1 Age

The age probabilities were calculated (equation 7.1); it is clearly shown that the peak age is around 20 (Figure 7.5). Most of the research showed that the peak age group is under age 25 (Pimenta *et al.*, 2003b; Health Protection Agency, 2006). A number of research studies and literature described the at-risk age as being under 25 years for both females and males; age is one of the important factors which can be used to target the high risk population (Pimenta *et al.*, 2003 a, b).

$$P(\text{positive result} | \text{age } x) = \frac{\text{No. of positive cases in age } x}{\sum_{k=1}^K \text{total positive cases}} \quad (7.1)$$

Where K is the total number of positive cases

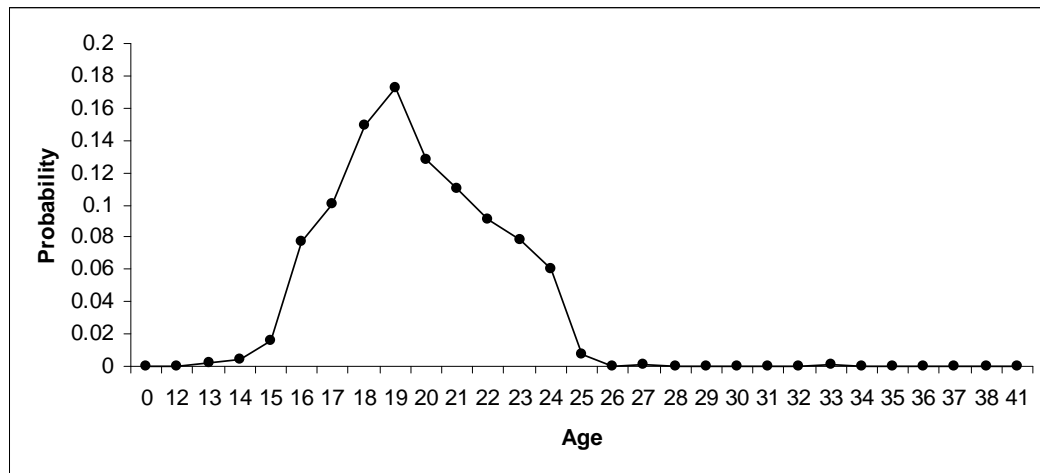


Figure 7.5 Probabilities of positive test results plotted against age.

7.3.2 Ethnic origin

Data on ethnic origin were available from the patient's information; Table 7.1 shows the number of positive cases per ethnic group. It is clearly shown that 93% of the positive cases came from the white ethnic origin group and the remaining 7% of the information came from the unknown group. The sample size from non-white ethnic groups was relatively small compared to the white ethnic group. Therefore, this information is not particularly useful for further analysis.

Table 7.1 Positive cases by ethnic origin group.

Ethnic Group	Ethnic Code	No. of Positive Cases
White	0	712
Black Caribbean	1	1
Black Other	2	2
Indian	3	1
Bangladeshi	5	1
Chinese	6	2
Black African	14	2
Other	34	5
Unknown	99	55

7.3.3 Location

Figures 7.6, 7.7 and 7.8 show the positive rate patterns in postcode sectors, in the CAS Wards and local authority. Table 7.2 shows the number of positive tests per CAS Ward. It seems that some of the postcode sectors and CAS Wards had more positive cases than the others; for example, SO14 and SO15 (but those postcode sectors had very few observed cases), PO2 and PO4 and an Output Area Charles Dickens (00MRMR) had much larger number of observed cases and also had relatively high positive rates. It is interesting to identify particular disease patterns, if any, and disease trends by location, and this will be commented on in the regression model section.

Table 7.2 Positive rate by CAS Ward.

CAS Ward	No. of positive	No. of test	positive rate
00MRMP	24	216	0.1111
00MRMQ	36	544	0.0662
00MRMR	51	382	0.1335
00MRMS	19	242	0.0785
00MRMT	26	275	0.0945
00MRMU	8	203	0.0394
00MRMW	33	354	0.0932
00MRMX	27	303	0.0891
00MRMY	22	257	0.0856
00MRMZ	26	292	0.0890
00MRNA	47	399	0.1178
00MRNB	30	257	0.1167
00MRNC	32	419	0.0764
00MRND	55	455	0.1209
24UEFT	13	150	0.0867
24UEFU	8	128	0.0625
24UEFW	7	91	0.0769
24UEFZ	17	168	0.1012
24UEGA	8	127	0.0630
24UEGC	3	90	0.0333
24UEGD	9	124	0.0726
24UFFL	7	86	0.0814
24UFFM	15	108	0.1389
24UFFN	5	74	0.0676
24UFFP	6	61	0.0984
24UFFQ	12	91	0.1319
24UFFR	10	102	0.0980
24UFFS	13	84	0.1548
24UFFT	5	82	0.0610
24UFFU	22	182	0.1209
24UFFW	11	92	0.1196
24UFFX	3	29	0.1034
24UFFY	2	34	0.0588
24UFFZ	14	124	0.1129
24UFGA	13	72	0.1806
24UFGB	3	67	0.0448
24UFGC	6	96	0.0625
24UFGD	13	118	0.1102

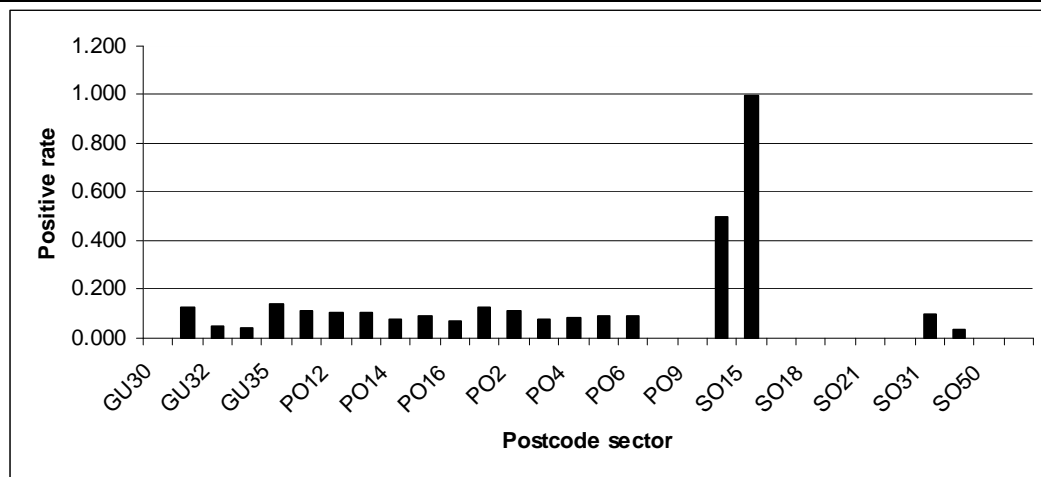


Figure 7.6 Positive rates per postcode sector.

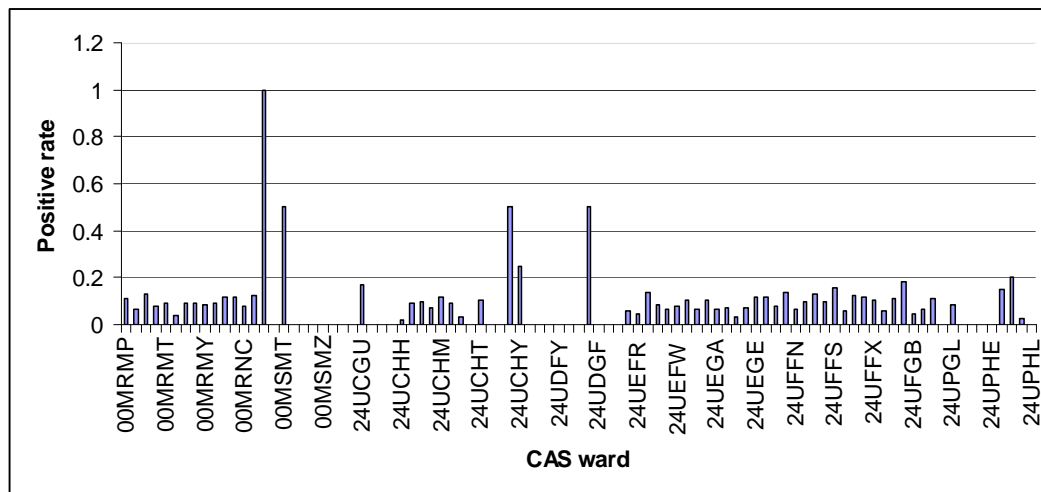


Figure 7.7 Positive rates per CAS Ward.

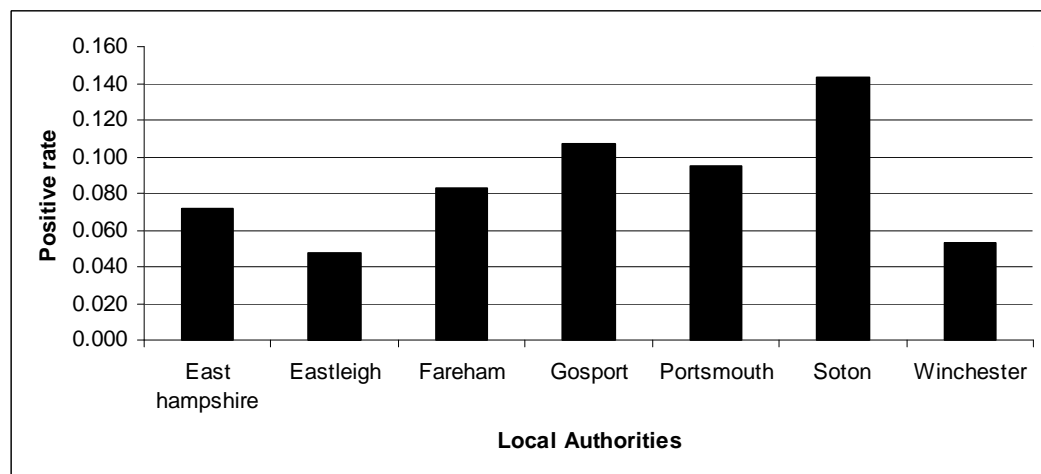


Figure 7.8 Positive rates per local authority.

7.3.4 Townsend index

Again the Townsend variables were used to measure the relative deprivation in Portsmouth (Townsend *et al.*, 1989). Townsend index variables were used to measure the deprivation at both Output Area and CAS Ward levels, as shown in Figure 7.3a and Figure 7.4a. The greater the Z value the greater the deprivation, which is depicted by the darker colour on the maps. The calculation of the Townsend index was explained in Chapter 2.

Figure 7.3a shows a random spatial pattern but Figure 7.4a shows the southern part of the study area, which is near the sea, has higher levels of deprivation than other areas.

7.4 Generalised linear regression model

For the study of Chlamydia, the data at Output Area level represents a set of observed counts $Y_i = (Y_1, Y_2, \dots, Y_N)$, where N is the total number of study regions ($N = 1030$ Output Areas as the final study areas), which arise from a Poisson process. The expected case was pre-defined. The regression equation is given by:

$$Y_i \sim \text{Poisson}(\mu_i)$$

$$\hat{Y}_i = e_i \exp(\beta_0 + \sum_{t=1}^T \beta_t v_{ti} + \varepsilon_i) \quad \text{For } i = 1, 2, \dots, N \text{ and } t = 1, 2, \dots, T \quad (7.2)$$

Where β_0 is the intercept and the β_t represent the coefficients of the explanatory variables v_{ti} and e_i is the expected cases in region i .

7.4.1 Generalised linear regression model at Output Area level

The explanatory variables were added into the GLM model as single explanatory variables. In total, there are five models with single explanatory variables; the coefficient value of each model is shown in Table 7.3 and other summary information is shown in Table 7.4.

Table 7.3 The intercept and coefficient value for each explanatory variable at Output Area level.

Model No	β_0	β_1	β_2	β_3
1.	-0.9928989	1.4789398		
2.	-0.7393471	3.3163026		
3.	-1.367185	1.675429		
4.	0.3082438	-1.6754278		
5.	-0.49734973	0.07931188		
6.	-1.07497365	0.04608768	1.10554479	
7.	0.03056997	0.04608775	-1.10554223	
8	-0.63490962	0.06509723	0.40669120	
9	-0.65679942	0.04761635	2.13824715	
10	-1.386411429	0.002529101	2.533898005	1.338719986
11	-0.047692691	0.002529202	2.533896734	-1.338717048

Table 7.4 Summary of GLM applied with explanatory variable at Output Area level.

Model No	Variables	Std.Error	t-value	P-value
1	Proportion of low social grade population (IV+V)	0.2738327	5.400888	1.366698e-007
2	Proportion of households with lone parents	0.47552288	6.974013	2.818639e-010
3	Proportion of single population	0.2598869	6.446763	1.410082e-010
4	Proportion of married population	0.2598869	-6.446757	1.410136e-010
5	Townsend index score	0.01258824	6.300476	7.189273e-010
6	Townsend index score + proportion of single population	0.01632178	2.823691	0.0000000007
		0.33245393	3.325408	0.0009316457
7	Townsend index score + proportion of married population	0.01632178	0.04608775	0.0000000007
		0.33245395	-1.10554223	0.0009316714
8	Townsend index score + proportion of low social grade (IV+V) population	0.01938381	0.06509723	0.0000000
		0.42073842	0.40669120	0.3337915
9	Townsend index score + proportion of households with lone parents	0.01596642	2.982280	0.000000001
		0.63345241	3.375545	0.001134184
10	Townsend index score + proportion of households with lone parents + proportion of single population	0.01987323	0.1272617	0.000000001
		0.63137369	4.0133095	0.001134184
		0.34359939	3.8961652	0.000102843
11	Townsend index score + proportion of households with lone parents + proportion of married population	0.01987323	0.1272668	0.000000001
		0.63137371	4.0133074	0.001134184
		0.34359939	-3.8961566	0.000102847

At the single variable level, all variables were significantly related to Chlamydia incidence (Table 7.4). The final fitted models are shown in Table 7.4. Models 10 and 11 were significantly related to the positive incidence of Chlamydia at Output Area level. In summary, Townsend index, proportion of households with lone parents and proportion of single population are related positively with Chlamydia incidence rate. When the proportion of married population increases the incidence rate decreases. The raw and predicted SIR from model 11 are mapped in Figure 7.9.

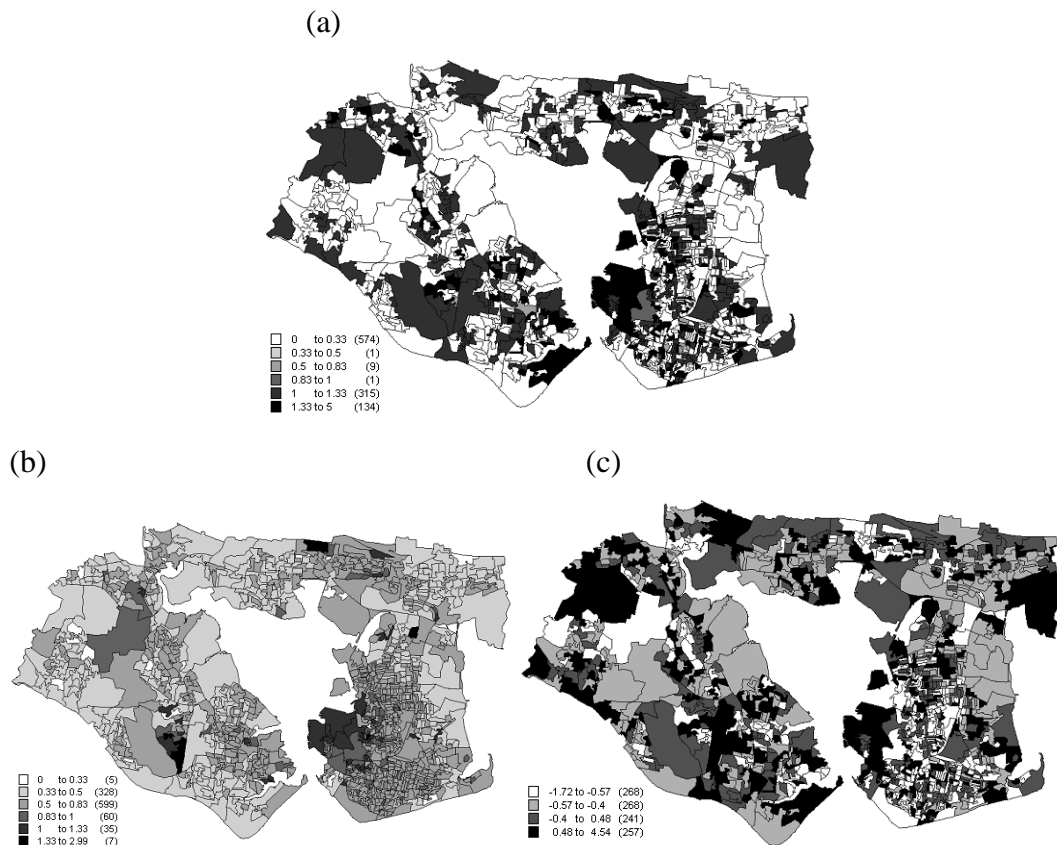


Figure 7.9 (a) Map of raw SIR, (b) estimated SIR from model 11 in Table 7.4, (c) Residual value between the observed and predicted cases per region.

7.4.2 Generalised linear regression model at CAS Ward level

In this section, the same model structure was applied at the CAS ward level. The coefficient value of each model is shown in Table 7.6 and other summary information is shown in Table 7.5.

Chapter 7 Chlamydia regression models

Table 7.5 Summary of GLM of each explanatory variable at Ward level.

Model	Variables	Std.Error	t-value	P-value
1	Townsend index score	0.01172566	3.638591	0.0003733398
2	Proportion of single population	0.3491374	1.685270	0.09161565
3	Proportion of married population	0.3491374	-1.685270	0.09161565
4	Proportion of households with lone parents	0.95495019	3.677735	0.0004041337
5	Proportion of low social grade (IV+V) population	0.4134746	4.311681	0.00003131768
6	Townsend index score	0.02408778	-0.1029605	0.00037334
	+ proportion of low social grade population	0.86808570	2.1441704	0.03041612
7	Townsend index score + proportion of households with lone parents	0.01670212	1.512871	0.0003733
		1.38446298	1.486249	0.1447018

Table 7.6 The intercept and coefficient value for each explanatory variable at Ward level.

Model	Variables	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
1	Townsend index score	-0.05024602	0.04266487	
2	Proportion of single population	-0.3140181	0.5883909	
3	Proportion of married population	0.2743728	-0.5883909	
4	Proportion of households with lone parents	-0.2633694	3.5120536	
5	Proportion of low social grade (G4+G5) population	-0.6040492	1.7827703	
6	Townsend index score + proportion of low social grade population	-0.627799980	-0.002480091	1.861323662
7	Townsend index score + proportion of households with lone parents	-0.18355514	0.02526815	2.05765660

At the ward level, only the Townsend index (model 1), proportion of households with lone parents (model 4), and proportion of low social grade population (model 5) were significant. Models 2, 3 and 7 in Table 7.5 showed large P-values (> 0.05). Therefore, those variables were not significant at Output Area level. Model 6 was not accepted as the best fitted model because the coefficient of Townsend index changed to negative, but the relation between incidence and Townsend

index is expressed to be positive. GLM summary and coefficients are shown in Tables 7.5 and 7.6. The observed SIR and predicted SIR from models 1, 4 and 5 are mapped in Figure 7.10.

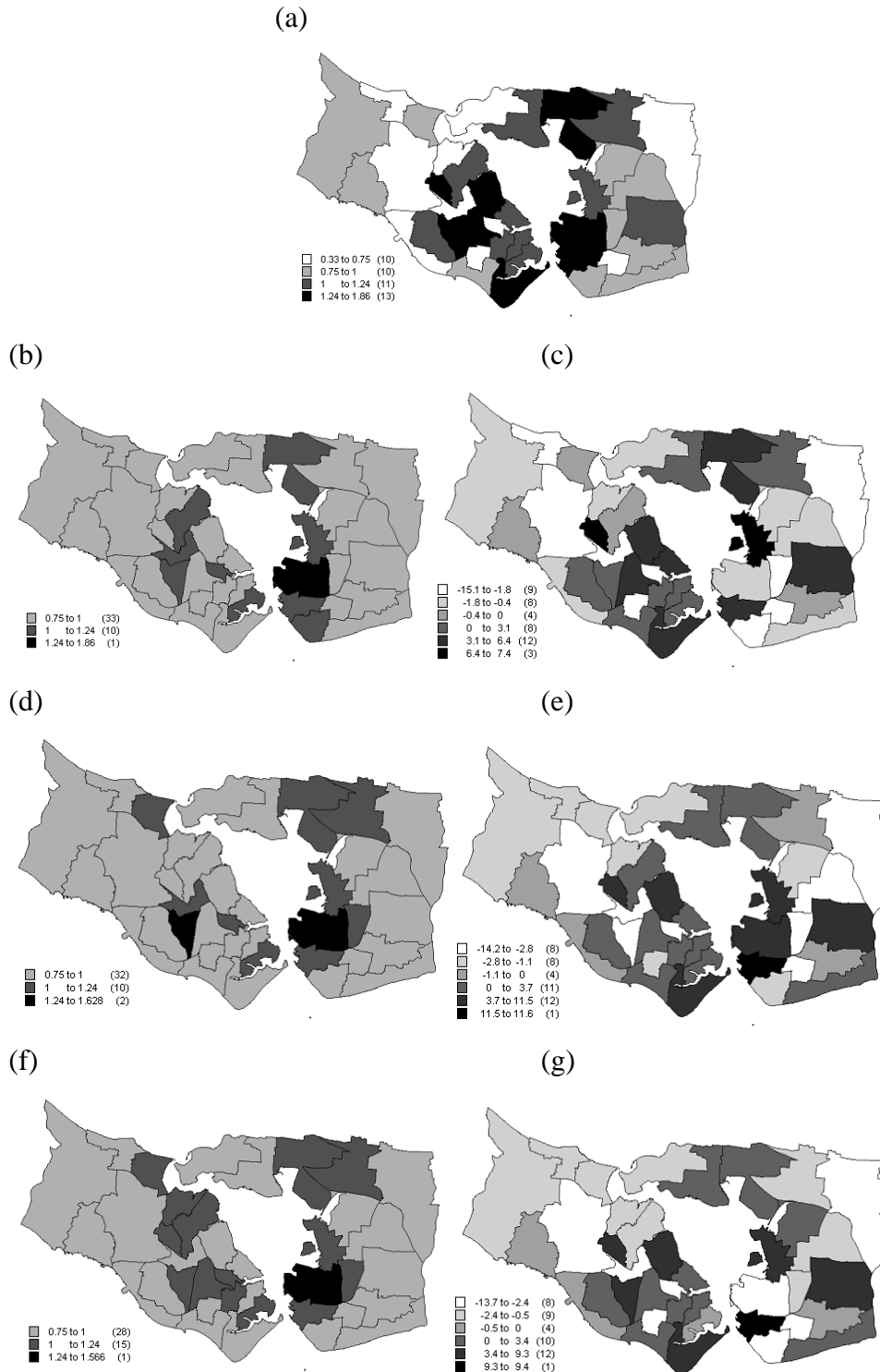


Figure 7.10 (a) raw SIR, (b) predicted SIR from model 1, (c) the residual from the observed and predicted cases from model 1, (d) predicted SIR from model 4, (e) residual from the observed and predicted cases from model 4, (f) predicted SIR from model 5 and (g) residual from the observed and predicted cases from model 5.

7.5 Multilevel regression model

A multilevel model was used to model the relation between individual Chlamydia results (for example, positive or negative) and explanatory variables at the Output Area level. When the size of the study region is small and the relative population is also small, the Binomial distribution is suitable to model the counted data (Rasbash *et al.*, 2005). Binomial multilevel regression can be used to model the outcome with positive or negative results, $Y_{ij} \sim \text{Binomial}(\zeta_{ij}, \pi_{ij})$ with a hierarchical model structure. The Chlamydia multilevel hierarchical Bayesian regression model was based on Bayesian estimation by using MCMC. So a set of prior distributions were given in the model.

For the Chlamydia study, individual test records were observed from the clinical study from Portsmouth 1999-2000. The multilevel model was used to model the test results (i.e. positive or negative) for each individual patient in the Portsmouth area within a one year period given that the explanatory variables are available at Output Area level; this is because social status, family structure as well as other information is not available from the individual. Therefore, each individual is nested in the Output Area level.

7.5.1 Multilevel regression model structure

A multilevel regression model is often used with the observed and explanatory variables appearing or measured at different levels. With particular relation to the Chlamydia study, the observed test results was recorded per patient but the deprivation, social status and family structure variables were not available at the individual level, but at national level only. Therefore, the multilevel regression model is a suitable choice to explore the relation between observed and explanatory variables when they appear in different levels. In particular, in this study patients (i) are nested within Output Area (j). The model structure is described in Table 7.7 and Figure 7.11.

Table 7.7 This table summarises the model structure. The detailed structure is displayed in Figure 7.11.

Levels	Levels	Description
2 (j)	Output Area	Explanatory variables are available at this level.
1 (i)	Patients	Test result: positive or negative and patients' personal information is available at this level.

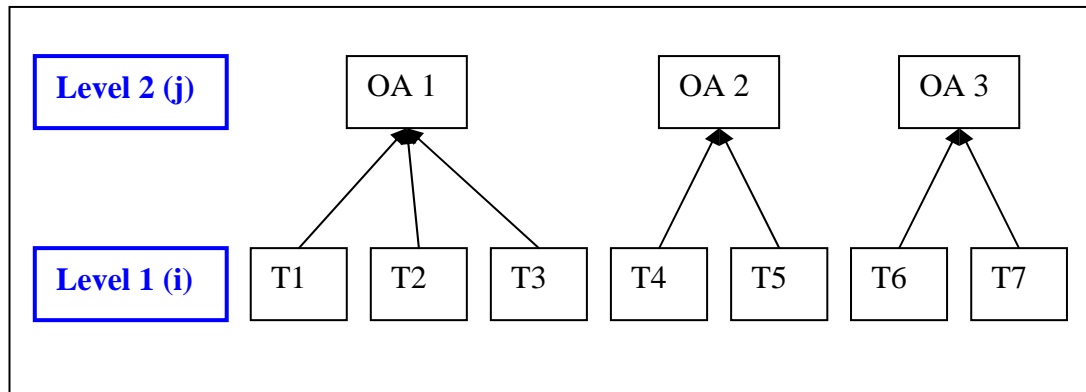


Figure 7.11 Unit diagrams, individual patients and each of the test results were measured at different times (months) within the study period.

In this study, the test results were analysed. Some of the variables within the model are described in Table 7.8.

Table 7.8 Data used in the model.

Variables	Description
Test result	Identifying code for each test (level 1 unit)
Patient	Identifying code for each
Postcode	Identifying code for each Postcode
Output Areas	Identifying code for each Output Area (level 2 unit)
Population (within Postcode)	Population at Postcode level, but age information is not available
Results	Positive, negative
Age	Patient's age at the time of test
Result time	Month at the time of test
Ethnic	0, 1, 3, 4, 5, 6, 14, 34, 99
Sex	F, M
Explanatory variables at Output Areas	(i) Townsend index score, (ii) Social status variables, (iii) Family structure variables.

7.5.2 Model

Let Y_{ij} be the positive result, π_{ij} be the chance of having a positive result, and ς_{ij} be the fixed variable at Output Area level, where i represents patient level (level 1) and j represents Output Area level (level 2). The at-risk population is defined as female population of age 16 to 24 (Pimenta *et al.*, 2003 a; b).

$Y_{ij} \sim \text{Binomial}(\varsigma_{ij}, \pi_{ij})$, the simple two level regression model is defined below:

$$\text{logit}(\pi_{ij}) = \beta_0 + \beta_1 v_{ij} + \beta_2 v_j \quad (7.4)$$

In practice, it is possible to have a (i) random intercept or (ii) random coefficient models it allows to measure the different between Output Areas.

(i) Random intercept model

$$\text{logit}(\pi_{ij}) = \beta_{0j} + \beta_1 v_{ij} + \beta_2 v_{2j}$$

$$\beta_{0j} = \beta_0 + u_{0j}$$

Where β_{0j} the random intercept has two components which are the fixed term β_0 and the random effect u_{0j} it is a higher level specified component (level j) it is Normally distributed with mean zero and variance σ_{0j}^2 details about the theory can be found in Chapter 2.

(ii) Random coefficients model

$$\text{logit}(\pi_{ij}) = \beta_{0j} + \beta_{1j} v_{ij} + \beta_2 v_j$$

$$\beta_{0j} = \beta_0 + u_{0j}$$

$$\beta_{1j} = \beta_1 + u_{1j}$$

The coefficients can be vary , which allows to account for the difference in the lower level which nested within the higher level and also vary across the higher level. For example, the random coefficient account for the difference between explanatory variables within the same higher level j (e.g. Output Areas) and to vary across in the higher study areas (Output Areas).

7.5.3 Analysis

In this section, two models ((i) global non-spatial model and (ii) CAR model) were used and the ages were modelled in two different methods (i) age categories and (ii) *logit* of the risk within the given age. For the age categories there are only two categories based on the age groups, which are the higher risk group (category 1) and the lower risk group (category 0). Different age ranges of the categories were examined; results are summarised in Tables 7.9-7.10, models 2 and 3. Age was also modelled as function of *logit* of risk per age (Figure 7.12): results are summarised in Tables 7.19-7.12. Finally, Table 7.11 to 7.12 showed the results based on simple two levels multilevel models and Table 7.13 and 7.14 showed the CAR model random intercept results. Details about the CAR model please refer to Chapter 2 in section 2.3.7.1.

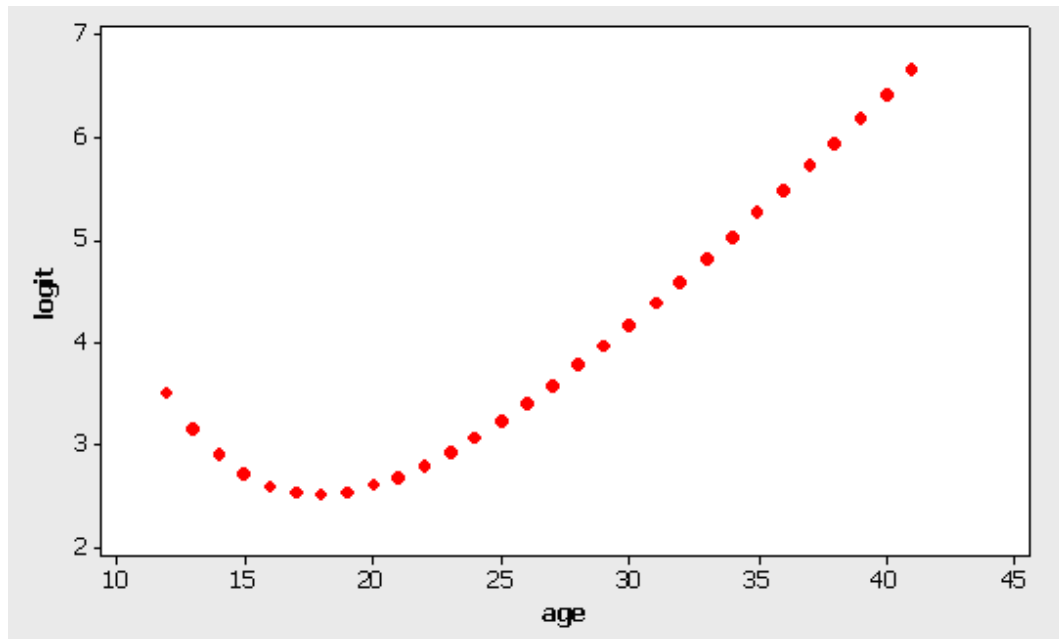


Figure 7.12 Age can be modelled as a function of *logit* of risk per age.

In this study all simple (global) hierarchical regression models, random intercept, random coefficients and CAR models were applied. Each model was run for 500000 times. However, only the simple hierarchical models are the best fitted models all other models were too complex and received higher penalty values but not obvious improvement in the model fitting. Therefore, of the simple hierarchical and CAR models results were displayed in the following. Details about the model structures please reference to Chapter 2.

Table 7.9 Global regression model with age categories, model 2: category 0 is the age group between 0 to 12 and 27 to 41; category 1 is between ages 13 to 26. Model 3: category 0 is the age group between 0 to 15 and 25 to 41; category 1 is between ages 16 to 24

No	Model
1	-2.245(0.041) <i>cons</i>
2	-2.553(0.44) <i>cons</i> +0.31(0.441) <i>age</i> _1 _{ij}
3	-2.302(0.191) <i>cons</i> +0.059(0.196) <i>age</i> _2 _{ij}
4	-2.263(0.042) <i>cons</i> -0.042(0.014) <i>Townsend</i> _j
5	-2.362(0.157) <i>cons</i> +0.214(0.277) <i>Single</i> _j
6	-2.150(0.128) <i>cons</i> -0.209(0.266) <i>Married</i> _j
7	-2.428(0.061) <i>cons</i> +2.213(0.529) <i>All lone parent</i> _j
8	-2.644(0.116) <i>cons</i> +1.156(0.309) <i>G4 + 5</i> _j

Table 7.10 Global regression model measurement based on the models in Table 7.9.

No	\bar{D}	$D(\bar{\theta})$	pD	DIC
1	4272	4270.97	1.03	4273.03
2	4272.86	4270.96	1.90	4274.76
3	4272.93	4270.94	1.99	4274.92
4	4263.98	4261.99	1.99	4265.97
5	4272.43	4270.41	2.02	4274.46
6	4272.35	4270.41	1.94	4274.29
7	4257.06	4255.09	1.97	4259.04
8	4259.98	4258.03	1.95	4261.93

Table 7.11 Global regression model with *logit* of risk per age.

No	Models
1	-2.245(0.041) <i>cons</i>
2	-1.216(0.348) <i>cons</i> -0.485(0.164) <i>log it</i> (<i>age</i> _{ij})
3	-1.380(0.369) <i>cons</i> -0.494(0.172) <i>log it</i> (<i>age</i> _{ij}) +2.214(0.54) <i>All lone parent</i> _j
4	-1.661(0.352) <i>cons</i> -0.460(0.156) <i>log it</i> (<i>age</i> _{ij}) +1.132(0.316) <i>G4 + 5</i> _j
5	-1.255(0.339) <i>cons</i> -0.489(0.160) <i>log it</i> (<i>age</i> _{ij}) +0.042(0.014) <i>Townsend</i> _j
6	-1.559(0.367) <i>cons</i> -0.485(0.160) <i>log it</i> (<i>age</i> _{ij}) +1.662(0.644) <i>All lone parent</i> _j +0.596(0.376) <i>G4 + 5</i> _j

Table 7.12 Global regression model measurement based on the models in Table 7.11.

No	\bar{D}	$D(\bar{\theta})$	pD	DIC
1	4272	4270.97	1.03	4273.03
2	4263.8	4261.76	2.04	4265.83
3	4248.97	4245.77	3.21	4252.18
4	4252.39	4249.43	2.96	4255.34
5	4255.86	4252.84	3.03	4258.89
6	4246.99	4242.97	4.03	4251.02

Table 7.13 CAR regression model with *logit* of risk per age.

No.	Model	Ω_u
1	-0.531(0.197) <i>logit</i> _ <i>age</i> _{ij} + <i>u</i> _{0j} <i>cons</i>	442.782(166.218)
2	0.173(0.12) <i>Townsend</i> _j + <i>u</i> _{0j} <i>cons</i>	451.594(156.302)
3	22.828(14.28) <i>Single</i> _j + <i>u</i> _{0j} <i>cons</i>	1638.514(1112.166)
4	0.392(1.673) <i>Married</i> _j + <i>u</i> _{0j} <i>cons</i>	432.200(149.404)
5	29.537(8.553) <i>All lone parent</i> _j + <i>u</i> _{0j} <i>cons</i>	878.224(372.696)
6	5.989(1.894) <i>G4 + 5</i> _j + <i>u</i> _{0j} <i>cons</i>	469.841(164.489)
7	-0.500(0.172) <i>logit</i> _ <i>age</i> _{ij} +27.377(6.851) <i>All lone parent</i> _j + <i>u</i> _{0j} <i>cons</i>	755.135(270.258)
8	-0.548(0.177) <i>logit</i> _ <i>age</i> _{ij} +5.94(1.985) <i>G4 + 5</i> _j + <i>u</i> _{0j} <i>cons</i>	481.839(169.511)
9	-0.46(0.168) <i>logit</i> _ <i>age</i> _{ij} +19.967(5.468) <i>All lone parent</i> _j +3.292(2.429) <i>G4 + 5</i> _j + <i>u</i> _{0j} <i>cons</i>	642.273(210.944)

Table 7.14 CAR regression model measurement based on the models in Table 7.13.

No.	\bar{D}	$D(\bar{\theta})$	pD	DIC
1	3632	3155.12	476.88	4108.88
2	3639.64	3164.92	474.73	4114.37
3	3609.62	3153.86	455.77	4065.39
4	3639.87	3164.64	475.23	4115.1
5	3616.76	3156	460.76	4077.52
6	3637.66	3163.89	473.77	4111.43
7	3612.11	3148.46	463.65	4075.77
8	3629.02	3154.34	474.68	4103.7
9	36170.6	3150.96	466.64	4084.24

The best fitted model is model 6 in Table 7.11 and 7.12. It suggests that the choice of age range within the age category did not have much effect in modelling. In

this section, the patients' age was modelled as (i) categorical variable and (ii) function of $\text{logit}(\text{risk}/\text{age}=x)$, to examine which is the best method to model age. It was required to compare the DIC values between Tables 7.10, 7.12 and 7.14. The results show that it is better to model the age as a function of $\text{logit}(\text{risk}/\text{age}=x)$, because such a model (model 6 in Table 7.11) had a smaller DIC value (4251.02) compared to the other models with age categorical information in Table 7.9 and 7.10.

Finally, to summarise the results from applying the global regression models, overall, the global model 6 in Table 7.11 and 7.12 with smallest DIC value in Table 7.12 (4251.02) showed a better fit than the CAR model in Table 7.13 and 7.14. The CAR regression model had larger pD (penalty) value, although the CAR model had smaller DIC value. The final fitted model reveal that when the patient's age decreases, the proportion of households with lone parents increases and the proportion of low social grade population increases then the Chlamydia positive rate is predicted to increase.

7.6 Geographically weighted regression model

The basic GWR theory was described and explained in detail in Chapter 2; a summary is given below.

Assuming that the underlying distribution is Poisson (e.g the underlying distribution is allowed to change according to the dataset), and μ_i is the mean value:

$$Y_i \sim \text{Poisson}(\mu_i) \quad (7.5)$$

The geographically weighted regression model is given below:

$$\hat{Y}_i = \beta_0(x_i, y_i) + \sum_t \beta_t(x_i, y_i)v_{it} \quad (7.6)$$

In the following analysis section, GWR models were constructed at different levels (i) individual, (ii) Output Area level and (iii) CAS Ward levels.

7.6.1 Output Area Level

In this section, all variables were measured at the same Output Area level. The optimal kernel size was 848 (model 4 and 5) regions out of 1030 regions in total (Table 7.15). The kernel size is relatively large, the original data had 90% zero positive cases; therefore it is necessary to have a larger kernel to include more information. Results are shown in Table 7.15. The observed cases and prediction parameters are mapped in Figures 7.13 and 7.14.

Offset variable: No. of expected positive cases at ward level

*No. of expected positive cases_i = Positive rate * Total test_i, i = 1, ..., N.*

$$\text{Positive rate} = \frac{\sum_{i=1}^N \text{Total positive cases}_i}{\sum_{i=1}^N \text{Total test}_i} \quad (7.7)$$

$$\begin{aligned} \text{Positiverate} &= \frac{665}{6972} \\ &= 0.095381526 (\sim 9\%) \end{aligned}$$

Table 7.15 GWPR results at Output Area levels.

Models	Variables	Kernel size	AICc (Global)	AICc (Local)
1	Townsend index score	848	1130.682391	416.927276
2	Proportion of married population	974	1127.501404	426.68265
3	Proportion of single population	974	1127.501328	426.68261
4	Proportion of households with lone parents	848	1128.85458	412.854978
5	Proportion of low social grade (IV+V) population	848	1140.881568	411.380884
6	Proportion of low social grade (IV+V) population + proportion of households with lone parents	869	1123.729741	415.899557
7	Townsend index score + proportional of households with lone parents	917	1122.099498	418.185964
8	Proportion of low social grade (IV+V) population + proportion of single population	939	1122.404625	425.477156
9	Proportion of low social grade (IV+V) population + Townsend index score	939	1122.404625	425.477156

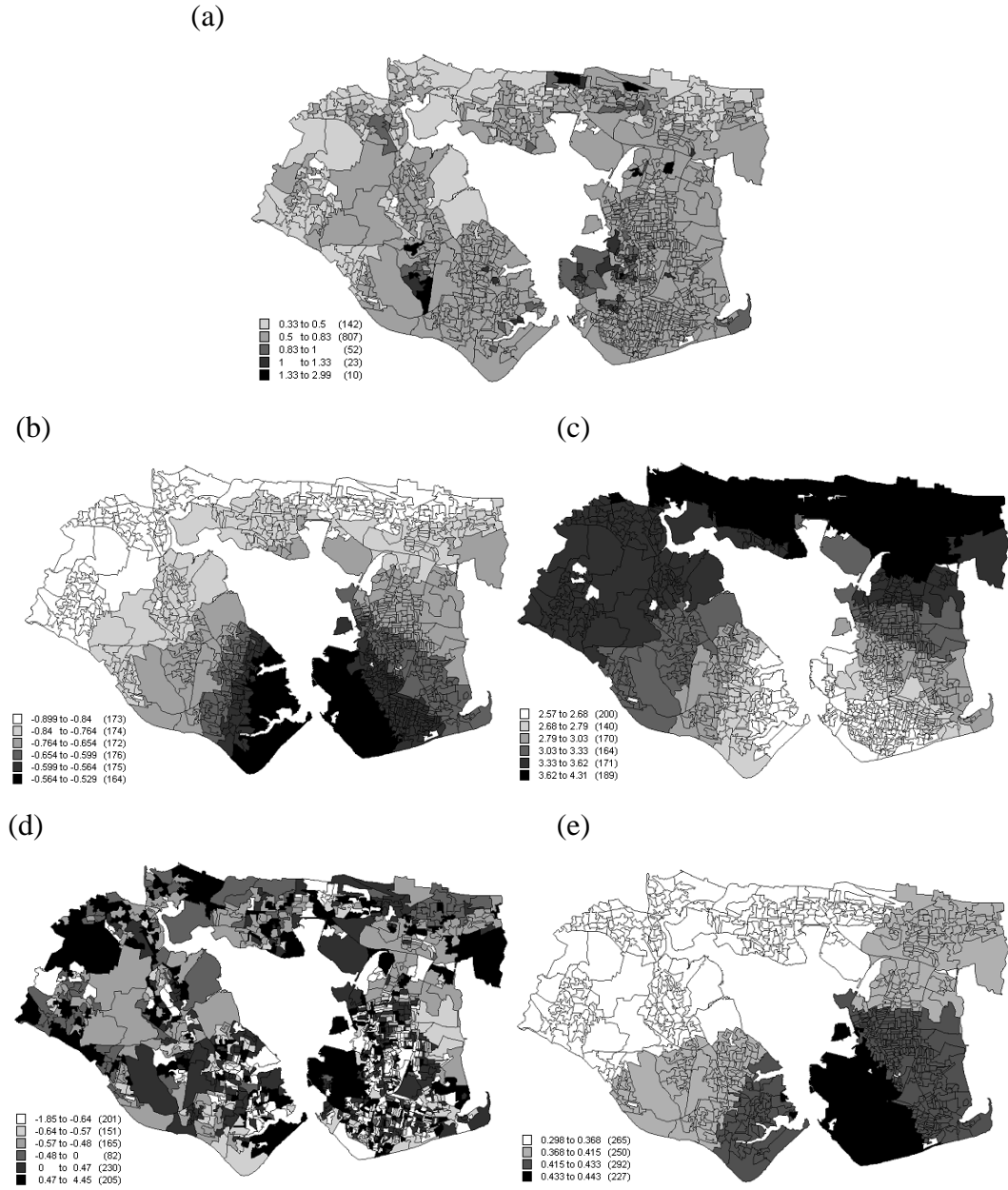


Figure 7.13 (a) Predicted SIR of model 4 in Table 7.15, (b) $\hat{\beta}_{0i}$, (c) $\hat{\beta}_{1i}$ with all lone parents proportion, (d) residual value between observed and predicted cases, and (e) R^2 value of model 4.

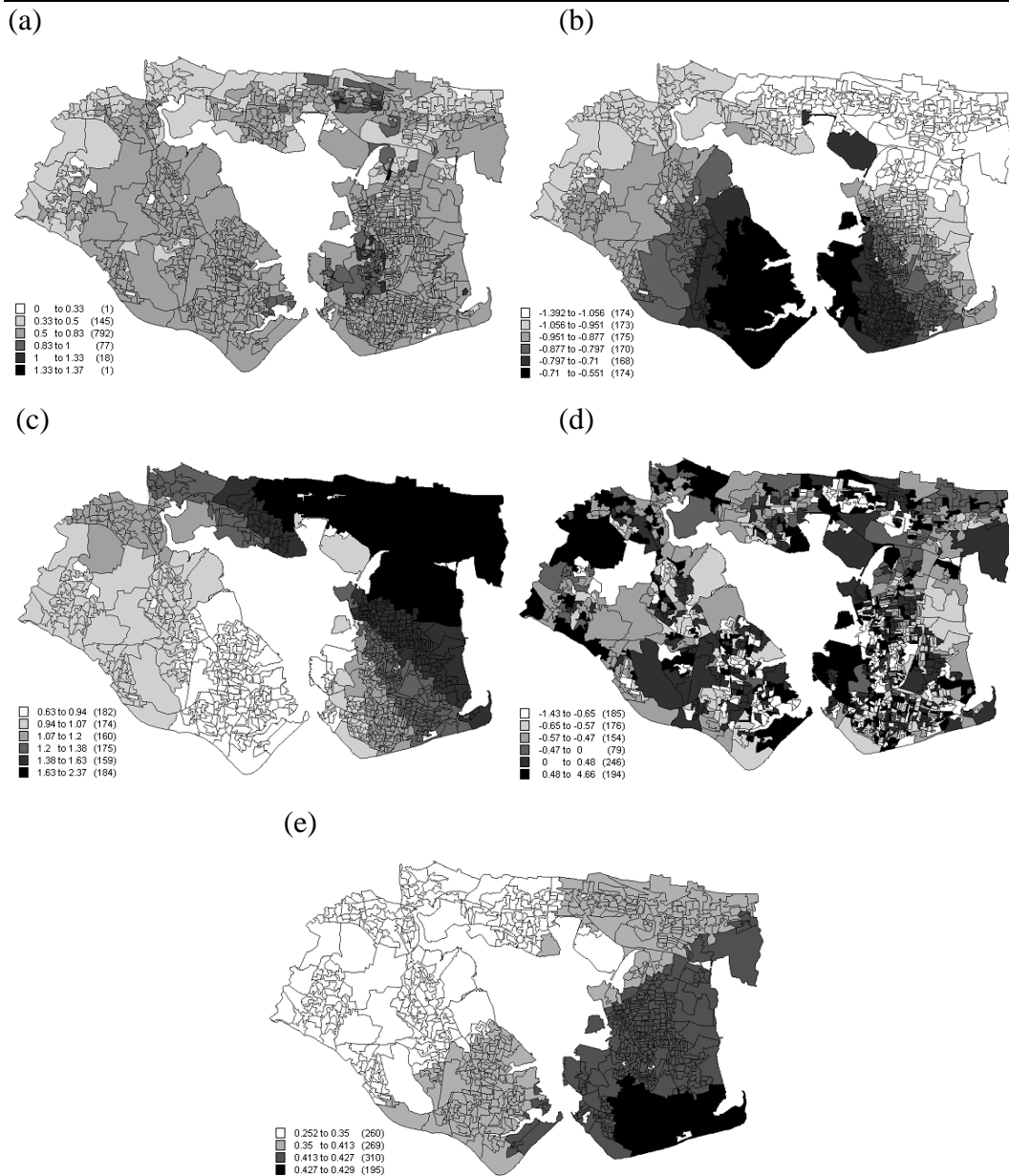


Figure 7.14 (a) Predicted SIR of model 5 in Table 7.15, (b) $\hat{\beta}_{0i}$, (c) $\hat{\beta}_{1i}$ with all lone parents proportion, (d) residual values between observed and predicted cases, and (e) R^2 value of model 5.

The proportion of low social status population and the proportion of households with lone parents are positively related to the incidence rate at Output Area level. It suggests that when the proportion of low social status population and/or the proportion of households with lone parents increase, the incidence rate increases. The results show that the models lack prediction power; the power was not enough to explain the observed variable. This may relate to the sampling issue, since the data were only collected between 1999 and 2000 around the Portsmouth area and for the female population between the ages of 16 to 24. The cases outside

this age range were not available for this study. Thus, it is possible to study the same data set a higher level, that is, examine the incidence pattern at a more aggregated level.

7.6.2 CAS Ward level

Output Area is the lowest available level in the UK census; the second lowest level is called CAS Ward. By using the Output Area-to-CAS Ward look-up table, which was provided by the UK census, it allowed linkage of Output Area and CAS Wards together. The look-up table lists which Output Areas were covered by which CAS Wards. In total there are 93 Wards under study. The same type of regression models were used, and the results are shown in Table 7.16 and Figure 7.15. The optimal kernel size was 36 regions from model 5.

Offset variable: No. of expected positive cases at ward level

*No. of expected positive cases_i = Positive rate * Total test_i, i = 1, ..., N.*

$$\text{Positive rate} = \frac{\sum_{i=1}^N \text{Total positive cases}_i}{\sum_{i=1}^N \text{Total test}_i} \quad (7.8)$$

$$\begin{aligned} \text{Positiverate} &= \frac{661}{6978} \\ &= 0.094726283 \text{ (~9\%)} \end{aligned}$$

The final fitted global model is model 5 in Table 7.16, which has the smaller AICc at both global and local level and the model is listed below,

$$\hat{Y}_i = e_i \exp(-0.604 + 1.783G45_i) \quad (7.9)$$

Where \hat{Y}_i is the predicted cervical cancer cases, e_i is the expected cases and $G45_i$ represents the proportion of low social grade (IV+V) population in region i . The local models with different explanatory variables are listed in Table 7.16.

Table 7.16 Summary results at CAS Ward level

Models	Variables	Kernel size	AICc (Global)	AICc (Local)
1	Townsend index score	36	48.525901	48.642314
2	Proportion of single population	31	58.341262	55.810529
3	Proportion of married population	31	58.341262	55.810529
4	Proportion of households with lone parents	33	48.674031	47.131103
5	Proportion of low social grade (IV+V) population	36	43.850892	44.591749
6	Proportion of low social grade (IV+V) population + proportion of households with lone parents	36	45.56487	47.283182
7	Proportion of low social grade (IV+V) population + Townsend index score	36	46.203309	48.697664
8	Proportion of low social grade (IV+V) population + proportion of single population	36	46.153311	47.784771

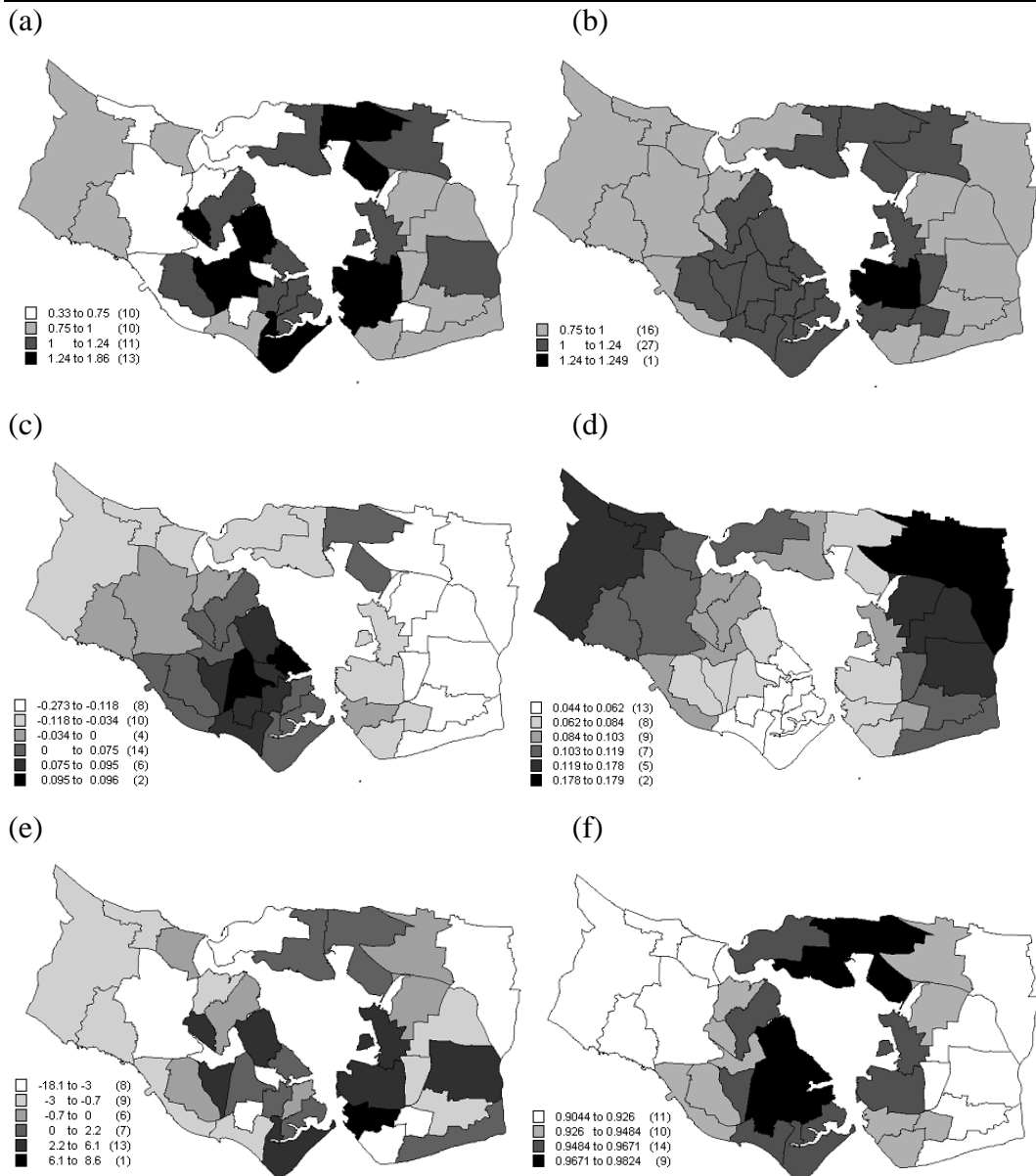


Figure 7.15 (a) Observed SIR, (b) predicted SIR from model 5, (c) $\hat{\beta}_{oi}$, (d) $\hat{\beta}_{li}$, (e) residual values between observed and predicted cases, and (f) R^2 values at CAS Ward level.

The results shown in Table 7.16 showed that there is only small difference between the global and local models since both of the AICc values are so similar (i.e. difference is less than 2). Therefore, in the CAS Ward level the global model is good enough to provide reliable estimation for parameters e.g. $\hat{\beta}_{li}$. It is useful to examine the global and local models by comparing the global standard error and the local inter-quartile range. If twice the global standard error is larger than the inter-quartile range then it could indicate that the variable is under a stationary process (Table 7.17). Therefore, the global model with proportion of low social

grade population is better than the local model of the Chlamydia incidence rate at CAS Ward level.

Table 7.17 Test for indicating stationary variables.

Parameters	2* S.E. (Global)	Inter-quartile range (Local)	Stationary or non-stationary variable
Intercept	0.296	0.201607	Stationary
Proportion of lower social grade (G4 + G5) population	0.826	0.430932	Stationary

At the individual level, the model showed that age is related to positive incidence of Chlamydia, at both of the Output Area and CAS Ward levels, the proportion of households with lone parents, proportion of low social status population are both significant to the positive incidence. The coefficients vary spatially at OA level, it suggests that the associated factors affecting or contributing to the positive incidence vary spatially. It provides information to understand how human behaviour and the surrounding deprivation condition or personal social and marital status might contribute to the chance of developing Chlamydia.

7.7 What does this mean for the design of screening programmes?

Based on the regression results in this chapter, age and some of the personal variables are associated with Chlamydia incidence at different levels. Therefore, potentially the general population can be divided into risk groups according to (i) the associated factors (e.g. family structure, deprivation and social grade) or (ii) geographical regions. Each risk group can have a different screening policy and screening interval. The advantage of dividing the population into groups and allocating different screening policies and screening intervals to risk groups, is that the screening programme can be made more efficient. Decision trees and simulation models in Chapter 6 can be used to evaluate the screening options. When health resources are limited, but the aim is to reduce the number of undetected Chlamydia cases optimal screening options are needed to run the programme in an efficient way.

Different risk groups may have different probabilities of developing chlamydia. By adding this information into the decision tree and simulation models, it allows us to answer some questions (e.g. what is the optimal combination of screening interval for each risk groups). At the same time, the limited resources can be well used and the number of undetected cases can be reduced through regular screening.

CART techniques can be used to split the population into risk groups according to their deprivation condition. The positive rate and negative rate per group can be estimated from the observed data at Output Area level. Decision tree model can be used to estimate the expected number of positive and negative cases per group, which showed it is a possible method to evaluate screening policy. The total study regions can be divided into two groups (i.e. low or high risk regions). In practice, the population within the high risk regions can be considered as the higher risk population, who may need to take Chlamydia screening test more often than the low risk population in the low risk regions. Such study demonstrated the real use of decision tree model and risk groups. It could prevent Chlamydia cases remaining undetected.

7.7.1 CART analysis

There is a missing link between chapters 6 and 7 which is similar to that for chapter 4 and 5. Therefore, a possible method to fill in the gap between Chapters, is to create a simple decision tree model based on the results from the Chlamydia GWPR regression models at Output Area level from Chapter 7,

1. Use regression to predict incidence
2. Divide the population into groups according to their deprivation condition or other characteristics (e.g., low Townsend index score vs. high Townsend index score)
3. Create a decision tree model for the population based on the overall incidence rate.
4. Create a new decision tree model for the groups based on the set of incidence rates

5. Explore the effect of different screening intervals for the whole population (it would depend highly on the availability of transition probabilities)
6. Explore the effect of different sets of screening intervals for the groups (it would depend highly on the availability of transition probabilities)

In fact, the transition probabilities for different screening intervals are not available. Only if continuous clinical data (or follow up data) are available (i.e. follow a group of patients for many years) can the transition probabilities be estimated. Otherwise, it is not possible to compare the screening options based on different screening intervals, because of lack of information about transition probabilities.

7.7.2. Risk grouping

CART techniques can be used to divide the population into groups. Each group is assumed to have different probabilities of developing the disease. Section 7.7.2.1 demonstrated how to split the population into groups according to the Townsend index score at OA level. Such techniques show that the chance of developing Chlamydia varies between groups, which means some groups with certain characters may have higher chance of developing Chlamydia in the patients' life time.

The Chlamydia screening process should follow the structure in Figure 7.16. However, the transition probabilities were required to be estimated from any possible continuous clinical data. Such data are rarely available.

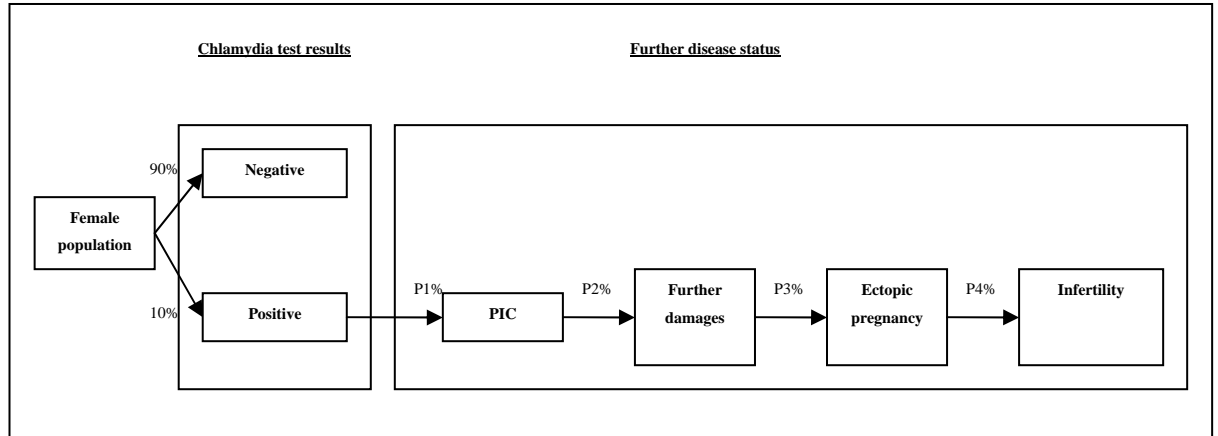


Figure 7.16 Chlamydia disease process, disease status with the transition probabilities (e.g. P_1 , P_2, \dots , $P_4\%$) and the positive and negative probabilities were provided from Primarolo (2006). Further damage status include damage to the fallopian tubes, damage to the uterus, damage to the surrounding tissues, and chronic pelvic pain.

7.7.2.1 CART based on the observed Chlamydia data

Transition probabilities and transition time are missing in the decision tree and simulation models in Chapter 6, which describe how a patient from a particular risk group might be expected to get the disease. It is possible to estimate the incidence rate per group, but the transition time parameter and the related transition probabilities over which a patient might be expected to get Chlamydia infection are not known. But the incidence rate per group is enough for a simple analysis.

The overall idea assumes that if an individual is tested, that reveals a realisation of the disease (or not) (based on the incidence rate per group) and the individual can be treated for the disease. This tells the policy makers in the long run the number of cases that they have revealed (removed) and the number that they have missed. This simple analysis may be enough to reveal that a selective screening policy would be more efficient than a national fixed policy. Figure 7.17 demonstrates the use of a decision tree model to estimate how many Chlamydia cases arise from a general female population. The probabilities in Figure 7.16 and 7.20 were estimated from the observed Chlamydia data and the expected Chlamydia cases are listed in Table 7.19.

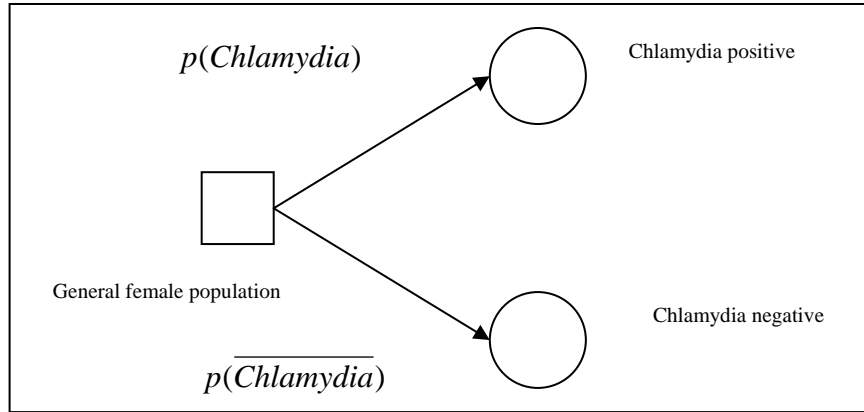


Figure 7.17 Decision tree model.

Assume there are 20000000 population at risk. The positive and negative probabilities are listed in Figure 7.18. The expected Chlamydia positive cases and expected Chlamydia negative cases can be calculated in the following two equations (7.7 and 7.8). Table 7.18 showed the results.

$$\text{Chlamydia positive cases} = \text{total at - risk population} \times p(\text{Chlamydia}) \quad (7.7)$$

$$\text{Chlamydia negative cases} = \text{total at - risk population} \times p(\overline{\text{Chlamydia}}) \quad (7.8)$$

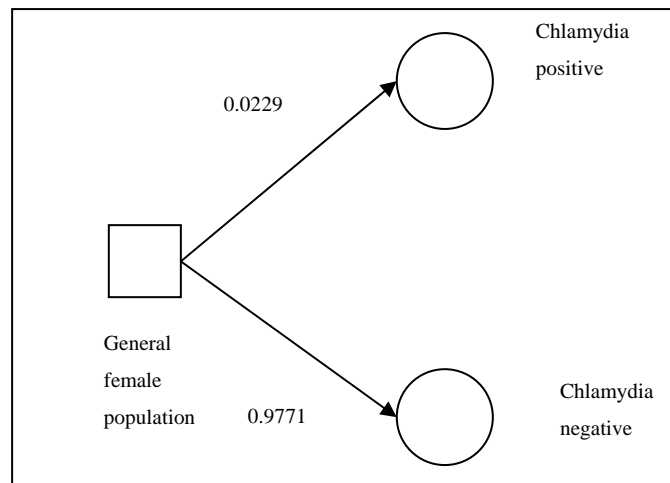


Figure 7.18 Decision tree model based on the Chlamydia data. Both the positive and negative probabilities were estimated from the available Chlamydia data at OA level described in Chapter 3.

Table 7.18 Decision tree outcome based on the information from Figure 7.18.

Group	Total patients
General female population	20000000
Chlamydia positive	458000
Chlamydia negative	19542000

Figure 7.19 demonstrates the use of the CART technique to split the population into risk groups according to their characters, people in the same group are assumed to have the same probability of developing Chlamydia in their life time. However, patients from different groups are assumed to have different probabilities of developing Chlamydia.

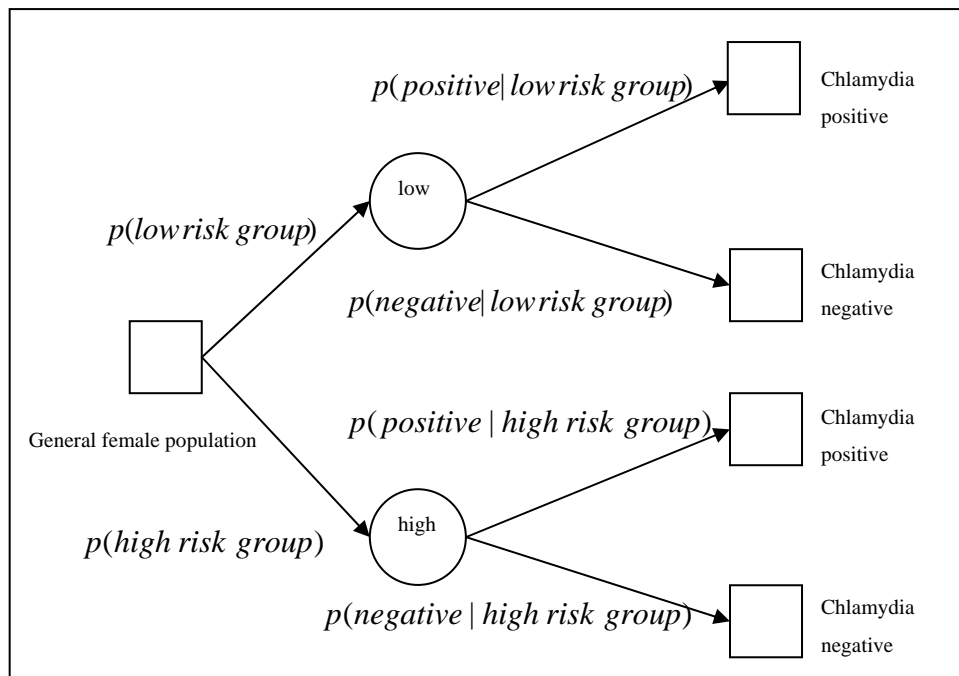


Figure 7.19 Potential decision tree with best split from CART analysis, which splits the general female population into two risk groups. The Chlamydia count data can be used to estimate the probabilities between status and groups.

In Figure 7.19 Six probabilities are needed, which can be calculated as below,

$$p(\text{low risk group}) = \frac{\sum_{i=1}^{\text{regions at low risk}}}{\text{Total regions}} \quad (7.9)$$

$$p(\text{high risk group}) = \frac{\sum_{i=1} \text{regions at high risk}}{\text{Total regions}} \quad (7.10)$$

$$p(\text{positive} | \text{low risk group}) = \frac{\sum_{i=1} \text{positive cases in low risk regions}}{\sum_{i=1} \text{at risk population in low risk regions}} \quad (7.11)$$

$$p(\text{negative} | \text{low risk group}) = \frac{\sum_{i=1} \text{negative cases in low risk regions}}{\sum_{i=1} \text{at risk population in low risk regions}} \quad (7.12)$$

$$p(\text{positive} | \text{high risk group}) = \frac{\sum_{i=1} \text{positive cases in high risk regions}}{\sum_{i=1} \text{at risk population in high risk regions}} \quad (7.13)$$

$$p(\text{negative} | \text{high risk group}) = \frac{\sum_{i=1} \text{negative cases in high risk regions}}{\sum_{i=1} \text{at risk population in high risk regions}} \quad (7.14)$$

Where low risk regions have lower deprivation condition (i.e. low Townsend index score) and the high risk regions have high deprivation condition. For details about the best split point please refer to Chapter 2.

7.7.2.2 CART based on the Townsend index score

In this section, the observed Chlamydia data and Townsend index were used to divide the population into groups, where the probabilities of developing Chlamydia vary between groups.

Data: Chlamydia data (observed individual counts) summarised at OA level

Average (overall) incidence rate: 0.02

Predicted variable: incidence rate per 100 women per region

Independent variable: Townsend index score

Best split: Townsend index score <0.5

The analysis shows that the best split is Townsend index equal to 0.5, Figure 7.20 shows the decision tree model, the low risk regions representing the regions with index score less than 0.5. Otherwise, regions are identified as high risk regions. Table 7.19 shows the expected number of Chlamydia cases per group.

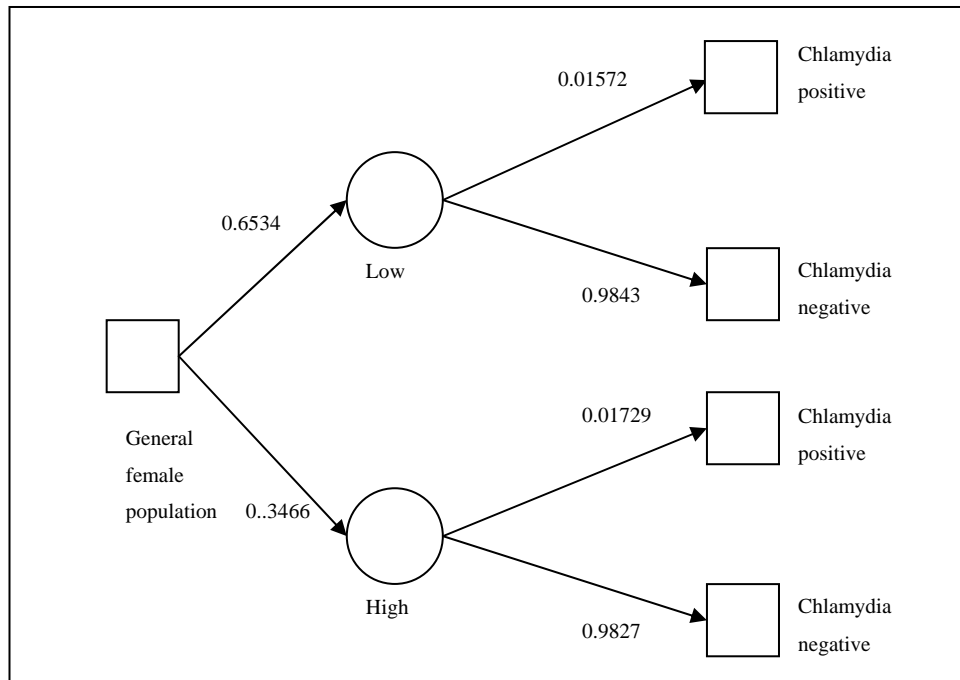


Figure 7.20 Decision tree model with two risk groups. The low risk group has a low Townsend index score and the high risk group has a high Townsend index score. The probabilities were estimated from the observed data.

Table 7.19 Potential decision tree, based on the observed data. Population was split into two risk groups.

Group	Total patients
General female population	20000000
Chlamydia positive	325254
Chlamydia negative	19674676

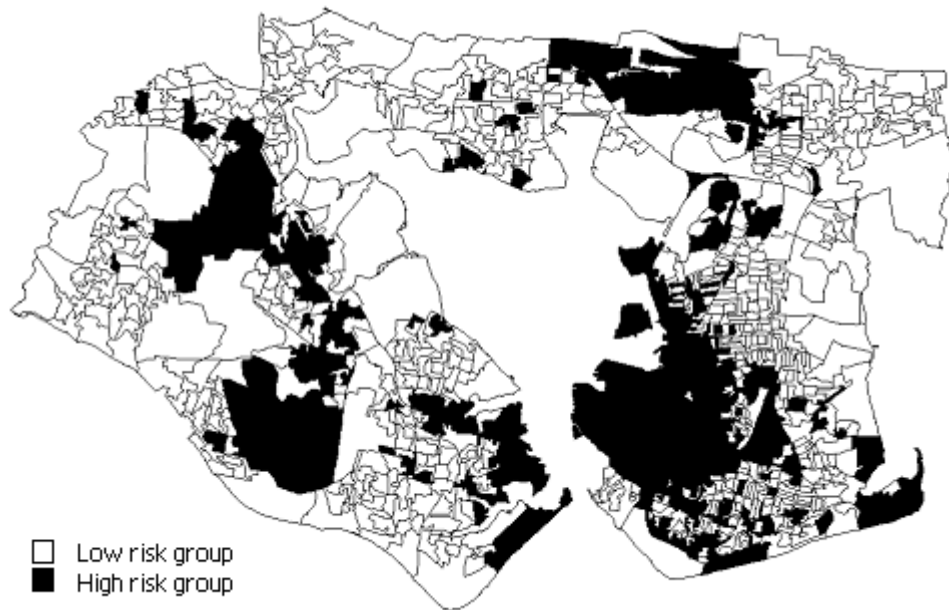


Figure 7.21 Risk grouping by the Townsend index score, the black represents the high risk group with high deprivation condition and the white represents the low risk group, with low deprivation condition.

7.8 Summary

The main task of this chapter was to explore the relationship between positive Chlamydia infections and some personal variables or deprivation condition, or social grade (status) at different regional scales. Obviously, different regional scales reflected different disease patterns. The lowest level (i.e. individual) did not show any interesting pattern. The results in section 7.5.3 showed the global model was better fitted than the CAR models. However, the prediction at individual level was inaccurate. This could be due to sampling issues, and will be discussed in detail in Chapter 8. At the CAS Ward level, the model showed a greater strength of prediction. The south of the study area had a higher Chlamydia incidence which could be due to a greater proportion of young population. The results from a various regression models showed there is a relationship between Chlamydia infection and low social grade (or status). The simple analysis showed that the chance of developing Chlamydia does vary between risk groups. The decision tree model demonstrated that not everyone has the same chance of developing Chlamydia. Some patients with certain characteristics (e.g. high or low socio-

economic status) have different chances of developing Chlamydia. Therefore, the lower risk patients may not need to be screened as often as the high risk patients. The maps (Figure 7.21) showed the risk grouping according to the Townsend index score. In practice, it is possible to screen the high risk population in the black areas in the maps more often than the white areas in the maps. Such a strategy can reduce the amount of wasted resources, the exact resources can be provided to the high risk groups, which may help to increase the detection rate and also reduce the number of undetected asymptomatic cases.

The observed data included only female patients between the ages of 16 to 24. Therefore, all the cases beyond such groups were excluded from the data, and it could cause a sampling issue. The sampling issue could explain why the predictions from GWPR were not close to the observed counts. Therefore, the expected positive Chlamydia cases were much smaller in Table 7.19, because the predicted positive cases were much smaller than the observed data due to the sampling issue. However, something is missing. That is the time parameters and transition probabilities. If the transition time and probabilities were available between status, it would allow estimation of the number of patients in each state for a long run. The result in section Table 7.19 allows to evaluate the effectiveness of the screening options and it also allows to compare different options.

Chapter 8 Discussion

8.1 Introduction

It is important to understand the aim of screening for cervical cancer and Chlamydia, which is to prevent the development of invasive cervical cancer by detecting and treating precancerous lesions at an early stage (i.e. before the abnormal cells change to cancer cells), and to avoid the undetected asymptomatic cases of Chlamydia, which in turn will avoid further complications or forward infected new cases. There are three stages: (i) prevention, (ii) treatment and (iii) follow-up. In practice, stages one and two can be treated as separate stages. If both screening programmes are available and useful at the first stage, then cervical cancer abnormalities can be detected within this stage and the second and third stages may not be necessary or may become less important in practice, whereas for Chlamydia cases can be detected early enough and treated, thus avoiding further damage and need for further treatments. Therefore, in terms of cost effectiveness, screening can prevent resources being overstretched and reduce unnecessary pressure on NHS services.

Consider the question: “Does location affect the likelihood of incidence, death or survival?” If the chances of developing cervical cancer and/or becoming infected by Chlamydia, and the level of response to treatments are associated with surrounding environmental variables and personal factors (e.g. age, diet, smoking etc.), then local variation in policy may be considered as an effective way of improving healthcare services at the national level (Rogerson *et al.*, 2006). If the risks of developing cervical cancer and Chlamydia vary geographically, local policies and local screening programmes could be introduced which might result in a reduction in the risk of developing cervical cancer and undetected Chlamydia cases.

8.2 Cervical cancer

The overall aims of the National Cervical Cancer Screening Programme are to reduce the risk of cervical cancer development in the female population. One of the objectives of this PhD study was to identify the associated risk factors which can in turn increase knowledge of cervical cancer, and it may also allow for the improvement of the effectiveness of the screening programmes. In this study, the relationships between the incidence of cervical cancer and, deprivation, social status, and family structure variables were investigated by applying regression models. It was shown that Townsend index, proportion of households with female lone parent, proportion of married female population and low social status are associated with the cervical cancer incidence. It follows that these known and widely available spatially varying factors could be taken into account when considering modifying healthcare services locally to increase the efficiency and effectiveness of screening programmes.

Survival has a subtly different interpretation to incidence and mortality. Survival rate may be related positively, at least in principle, to the effectiveness of the healthcare programme locally. That a factor such as proportion of married female population is negatively significant in GLMs, implies that personal marital status is related to the chance of developing cervical cancer in a woman's life time. Such a variable may reflect some information about personal sexual behaviour. For example, a married woman is less likely to have more than one sexual partner, and sexual behaviour is highly related to cervical cancer development.

The maps of the residuals from the fitted GWPR models illustrate clearly that the residuals are spatially autocorrelated. This points to the need for an alternative modelling approach (e.g., spatially auto-regressive modelling), but also highlights that these personal factors explain only a part of the variation in cervical cancer counts. Other variables are missing from the models. While speculative at this stage, it is interesting to argue that a missing variable may be current local spatial variation in the effectiveness of the healthcare programme. Of course, several alternative interpretations are possible.

Three main points of technical discussion are of interest; (i) the benefits of using different regression models, (ii) missing data, and (iii) cervical cancer mapping.

The important aspect of hierarchical models, which has been discussed widely in many studies, is that such models allow borrowing of strength between different data sets. Therefore, this leads to improved and more stable estimates of the parameters. With different model structures, the level of flexibility of modelling random effects is different. So the advantage of the Bayesian method is that it provides spatial smoothing that can reduce the chance of obtaining extremely high or low estimated rates that may falsely indicate disease clustering. Of course, it could also over-smooth in some cases, but a different model setting allows for different amounts of smoothing (e.g. BYM and MIX models: Jarup *et al.*, 2002). Therefore, it is also useful to compare the spatial pattern within a few different methods. Three Bayesian models were utilised to estimate incidence rate with different levels of flexibility of random effects. The BYM model produced a better fitted model.

The posterior mean of incidence rates exhibited spatial variation. In particular, the Midlands and the North of England had relatively high incidence rates. These high rates could be related to deprivation, personal marital status and social status. It is common to have missing or un-observed data. In the present study, it was necessary to apply an appropriate method to take care of the missing data. The advantage of the Bayesian model is that it allows for the borrowing of information from the available data, prior distribution and direct neighbours (if spatial information is available), so that the missing data can be handled appropriately and do not cause extreme estimation.

Nationally, the cervical cancer screening programme is free of charge for every woman between the ages of 25 to 64. However, there could be local variation in the efficiency and the accuracy of the test results. A patient's attributes, response to the test and the treatment are crucial to cervical cancer development, as the objective of the screening programme is to prevent cervical cancer development.

Both Bayesian and GWPR results indicated all variables are significant to the cervical cancer incidence rate, but the Townsend deprivation, proportion of lone parent population, proportion of low social status population and proportion of marital status population are the most important variables within the Bayesian regression models. Every human is unique, and as a result, a fixed screening

system might not be flexible enough to cover different requirements of different types of patients. The national screening policy uses age to divide target patients into groups (young or mature groups). Such a policy ignores the variation between patients and their backgrounds (e.g. family history, personal behaviours). It is possible to target patients in two ways: (i) divide the population into risk groups according to their age and socio status (e.g. low, medium and high), or (ii) divide the study area (England) into a few regions and give each region a different screening policy or screening interval.

From the GWPR results it is clear that the relationship between incidence rate and proportional of low social status population varies spatially. Specifically, the local parameters mapped in Figure 5.20 in Chapter 5 varied spatially. The residual map (Figure 5.19) seems to exhibit some autocorrelation, which suggests that social status, cannot explain spatially correlated variation in incidence rates completely. Some other possible variables may be missing from the model (e.g., sexual behaviour, personal HPV history, family history etc.).

There are two types of regression models of interest here; *local* and *global*. A global model describes only the average type of behaviour which is of limited use if behaviour varies over space. However, local models provide information about the spatial patterns of relationships. An advantage of using local spatial analysis is that it is possible to link the outputs of spatial techniques to the powerful visual display capabilities of geographical information systems. Practically, the global model gives only an average value for each estimated parameter across all study areas, which means it assumes that every study region within the study area is the same. This may not be true in some cases. The local model allows spatial variation in parameters. If no spatial variation is expected, then the global model may provide useful and meaningful outputs.

Within a geographical information systems environment, the GWPR method can be used as a visualisation and exploratory spatial analysis tool. If users are also interested in seeking those variables that can be described adequately by a stationary process and which are best fitted by non-stationary models, then it is possible to apply a mixed GWPR. A mixed GWPR is a semi-parametric GWPR model; it allows some variables to vary spatially and others to remain constant.

In terms of prediction, some of the regions face the situation of under-prediction or over-prediction. There are two important cases (i) those regions that are relatively large (i.e. the size of the cell is large), and (ii) those regions that include extreme cases. For the first case, when the regions are relatively large then the distance between the regions (point i and j) is larger, so that the accuracy of the prediction may be reduced by the distance. In the second case, the prediction can be underestimated or overestimated, because of the moderating influence of the neighbours.

Many studies (Richardson *et al.*, 2004) have suggested that poor health outcomes often appear in the most deprived areas. So it seems that some relationships between health outcomes and low social status variables exist in the present study. The strength of the relationship between health and low social status should be a concern to all governments that espouse ideals of equality.

In the GWPR analysis, one hundred random numbers were drawn from a Poisson distribution to replace the missing data. Although this simple method was used to deal with the missing data in this study, the results showed very limited variance and the predictions for those areas with missing data were very similar, which suggests that the results are not affected greatly by replacing the missing data. Further research is needed to explore other possible methods to solve the missing data problem. For example, it is very common to have missing data in clinical trials. Lavori *et al.*, (1995) demonstrated an approximate Bayesian bootstrap method to solve the truncation problem at unit-level. Lai and Ying (1994) also demonstrated the general principle of constructing M-estimators of regression parameters to deal with censored data.

In general, it is possible to assume that in a particular population a group of women may behave differently compared to others, which may be related to the given geographical location, Townsend deprivation condition, and personal conditions (e.g. age). In practice, healthcare planners can identify those related factors to divide a general population into varying risk groups (e.g., low, medium and high), since different risk groups may have different chances of developing cervical cancer in their lifetimes. As well as in terms of considering the screening options, it is possible to have more than three options: for example, it is possible

to consider whether or not to provide HPV tests for the low risk cohort, instead only providing the HPV test for those women from medium and high risk cohorts. Such an approach can increase the accuracy of estimating better screening options for different risk groups. A clinical study in the UK demonstrated that up to 25% of the available resources can be saved if the low risk group women (i.e. from the age of 50 upwards) withdraw from the national screening programme (Sherlaw-Johnson *et al.*, 1999).

Implementation of a locally adaptive screening policy is likely to be extremely difficult to achieve in practice for a variety of reasons. A more feasible alternative is to divide the population into spatially varying risk groups, each of which may have a different screening frequency. This may increase the chances of detecting abnormal cells at a timely stage such that certain activities could be applied to control disease incidence. However, it is acknowledged that it is never an easy task to improve NHS services, since the healthcare system is extremely complex.

In addition to the complexity of the screening programme system, it is possible to consider other details within the screening programme, e.g. diagnostic tests. The use of diagnostic tests is highly related to the accuracy of detection rate; for example, some studies showed that by accepting the HPV test in the cervical cancer screening programme, the detection rate at which the pre-cancer states are identified can be increased (Sherlaw-Johnson *et al.*, 1997; Sherlaw-Johnson and Gallivan, 2000). Another study shows the use of LBC and the combination of Liquid Based Cytology (LBC) and HPV tests in the screening programme in the UK; however, the combination of both tests does not increase the cost-effectiveness of the screening programme, which causes a higher chance of an inappropriate colposcopy being administered during a patient's life time (Sherlaw-Johnson and Philips, 2004). HPV and LBC tests may be considered for introduction into the screening programme and by 2000 both tests were piloted in the screening programme (Patnick, 2008). Finally, the HPV vaccine is available, and some clinical studies in Italy and Germany showed that the use of the vaccine significantly reduces the incidence of cervical cancer, and it might be considered as a means of achieving increased cost-effectiveness in screening programmes (Ferko *et al.*, 2007; Hammerschmidt *et al.*, 2007).

8.3 Chlamydia

The aim of Chlamydia screening is to (i) prevent undetected positive Chlamydia cases, (ii) provide treatments to positive patients, and (iii) to prevent and avoid further complications or forward transmission. Chlamydia screening provides an opportunity to identify more asymptomatic cases. Treatments can be provided at the right timing for patients and, thus, result in the prevention of forward transmission. Thus, in practice, it is necessary to have an effectively operating system. At the moment, there is not an official screening period: if a person suspects they might be infected with Chlamydia, then they can be tested, unlike the cervical cancer screening programme. It might be that there is a need for a recommended screening period particularly for high risk patients who could get regular and more timely check-ups. This is one possible way to reduce the number of undetected cases and to prevent forward transmission.

One of the key elements is to identify the high risk group from the general population, which is one of the objectives within this PhD study. Various regression models were used to study the relationship between Chlamydia incidence cases and deprivation, social status, lone parents and marital status variables.

The U.S. Preventive Services Task Force (USPSTF) reported that women, pregnant women, and men should be screened, especially if aged under the age of 25, which carries a higher risk than other age groups. The USPSTF listed the risk factors which feature in the personal histories of those with Chlamydia and other STIs, including number of sexual partners and sexual behaviours (e.g. use of condom) (USPSTF, 2007; Meyers *et al.*, 2008). Thus, an official screening programme should screen those high risk populations regularly. If and when the Chlamydia screening programmes become available for high risk groups this could be a way of alleviating overstretched NHS resources.

No matter which type of regression models is used, multilevel, aggregated level, global and local level models showed that there is a relationship between incidence of disease and age of the patients and some of the deprivation, social status and family structure variables: particularly larger proportion of households with lone parent, and larger proportion of low social grade population may

increase Chlamydia incidence rates, at OA level, and the proportion of low social grade was also positively related to positive Chlamydia cases at CAS ward level. However, the predicted patterns at other lower level (e.g. individual levels) were not similar to the observed patterns. One possible reason for the inaccurate prediction is missing data (e.g. on sexual behaviour). Some of the personal information may not exist such as numbers of sexual partners, or STI history, as well as the explanatory variables not being available at the personal level. Such information would increase the prediction power.

A second possible reason is a sampling problem, since the data were collected between 1999 and 2000 and for those female patients between ages 16 to 24; which means the proportion above the age of 24 was not recorded. Those unobserved data could lead to a major loss of data, which may lead to inaccurate prediction power.

A third reason is a typical data issue, which may cause ecological bias. Individual level survey data have insufficient power to study small-area variations in health. The best way to overcome such an issue concerning data is to combine both individual and aggregated data in modelling; such a technique is called “Hierarchical related regression for combining aggregate and individual data” (Jackson *et al.*, 2006; 2008). However, it is not possible to collect any possible variables at the personal level for the Chlamydia study. Thus, it is not possible to apply the hierarchical related regression model in the Chlamydia study.

The incidence pattern shown in Chapter 7 appears random and also the deprivation, low social status and family structure variable patterns do not appear similar to the incidence pattern at OA level, which explains why the prediction is inaccurate, possibly related to the second and third reasons above. In principle, if it is possible to divide a population into groups according to the related risk at a certain scale (e.g. CAS Wards), for example, high and low risk groups, each of the groups at CAS ward level may have different probabilities of developing Chlamydia or becoming infected. Then each of the risk groups can attend screening tests at different, but regular intervals; such as once every six months for the high risk group at ward level. This could be a possible way to improve the chances to detect more cases, as it becomes possible to assume that a particular

population group may have a higher chance of infection and/ or of transmitting Chlamydia than the others at a certain scale (e.g. Ward level). If risk varies spatially, then a population can be divided into spatially varying risk groups, where each of the risk groups in the different locations should have different screening intervals. However, this may not be possible in practice, as this will cause over-complications for general practice. Therefore, a possible solution might be segmentation, to divide England into broad geographical parts based on some variables (e.g. deprivation condition, social status and also family structure), where each part can have a different screening interval. Most important is to screen those members of the high risk groups. Since the high risk groups might be relatively small, it may be easier and more feasible to set up a screening programme to incorporate the entire high risk population. However, it is not possible to screen the entire population including both low risk and high risk population, as this would overstretch the NHS. Particularly, the overall incidence rate is 10%, but 75% of the incidence cases come from the age group between the ages of 16 to 24 (Health Protection Agency, 2006).

8.4 Summary

Mostly people understand how personal ill-health conditions, family history and personal sexual behaviour are directly associated with the development of some long-term and infectious (i.e. short-term) diseases. However, the various types of regression models indicated that some of the social status, deprivation and marital status variables were significantly correlated with cervical cancer and Chlamydia broadly; this suggests that deprivation condition, social and marital status variables are associated with human health. Such information may be useful for policy makers. For example, it might be helpful to divide a population into groups according to their personal deprivation condition and social status or the socio-economic status of the region within which they live. Alternatively, it might be possible to consider segmentation based on deprivation condition to divide England into a few parts according to other possible variables (e.g. social status) geographically. Each of the groups might then be allocated a different screening policy; e.g. a different screening test or screening interval. From a financial point of view, this may save resources or make better use of limited resources. From the

patient's point of view it may increase the chances of detecting and preventing long-term disease.

Chapter 9 Conclusion

9.1 Introduction

In reality, people are not homogeneous; every human is unique. Individual characteristics and behaviours contribute to subsequent disease experience, some of the main factors being age, sex and genetic make-up; but also lifestyle factors and social background. Social and lifestyle factors are important, some of which are strongly associated with cervical cancer development (e.g. the female and their partners' sexual activity, smoking, diet and job etc.) and some of which may be associated with Chlamydia (e.g. sexual behaviour, age, socio-grade, culture and sexual partners' behaviour). The preliminary results presented here suggest that cervical cancer screening intervals and frequency should vary with deprivation condition, social status and family structure factors which themselves can be mapped from the UK 2001 Census of Population.

Generally, regression models are used to explore the relations between observed health outcomes and explanatory variables. By using GLMs, a simple picture is provided of the linkage between the observed and explanatory variables. By using Bayesian methods, the posterior distributions describe the uncertainty of the parameters, and the local regression can provide more information on the local variation in relationships.

9.2 Cervical cancer

The GLM results showed that all of the explanatory variables (Townsend index, low social status and marital status) were associated with cervical cancer incidence. The Bayesian results showed that Townsend index score, proportion of households with lone parent, and also proportion of low social grade population were positively associated with cervical cancer incidence rate. It highlighted that the Midlands and the West of England had greater incidence rates than the rest of England. It is important to understand that such associated variables may be

insufficient alone to explain the geographical patterns evident in incidence rates since the significant variables are associated to incidence, but not the main cause of the development of cancer. The main causes of cervical cancer can be attributed to family history and patients' sexual behaviours as well as their partners' sexual behaviour.

In addition to human behaviour (e.g. family history, personal history and personal sexual behaviour), individuals may have different chances of developing cervical cancer and also show different responses to the pre-cancerous treatments. The detection rate of cervical cancer as well as the mortality and survival rates can vary spatially across the UK. Nevertheless, identifying the associated factors does provide information that may be useful in planning screening programmes.

Traditionally, global regression models have been used to explore the relationships between health outcomes and explanatory variables. However, such techniques do not take account of spatial variations in the relationships. This study demonstrated the use of GWPR to examine the relations between cervical cancer incidence rates and Townsend index score, proportion of households with lone parents, and also proportion of low social grade population across England, and also demonstrated how the use of local modelling provides a great deal more information for health analysis than traditional global modelling. However, the GWPR indicated that proportion of low social status population was the most significant variable correlated with cervical cancer broadly. This suggests that social status is associated to human health. The relation between low social status and the incidence rate varies spatially: it has more contribution in the south and north of England than west of England, which may related to the population structure (e.g. larger proportion of elderly population). Such information may be useful for policy makers. It is possible to target patients in two ways: (i) divide the population into risk groups according to their age and social status (e.g. low, medium and high), or (ii) divide the study area (England) into a few regions and give each region a different screening policy or screening interval. Each of the groups might then be allocated a different screening policy, (e.g. a different screening test or screening interval). From the financial point of view this may save resources or make better use of limited resources. From the patient's point of view it may increase the chances of detecting and preventing long-term disease.

Two healthcare studies from the NHS (Herbert and Smith, 2007) showed that changes in policy regarding only age and frequency make poor use of resources (Raffle., 2004). The results from this study demonstrated that incidence rates vary spatially across England. So the ideal way to improve the use of resources may not be to adopt a fixed screening interval for different age groups, but to consider a more adaptive programme for different parts of England or alternatively, divide the population into different deprivation condition, social state and family structure risk groups. A NHS study from Herbert and Smith (2007) showed that the number of cases with Cervical Intra-epithelial Neoplasia (CIN3) has increased for women between age 20-24 because of trends in sexual behaviours, where increasing numbers of young people become more active sexually when they are still in their mid-teens (Herbert and Smith, 2007). The change in sexual behaviour arises in part because of socio-economic changes from time to time and from place to place. If that is true, then recognising the associated factors may be useful for long-term prevention. For example, it is possible to improve sex education in local schools by teaching mid-teen pupils about protective sex. Such an approach can be valuable as an intervention factor for long-term prevention.

Further research is required to investigate the link between more specific probabilities of transition between pre-cancerous and cancerous status (e.g., CIN1, CIN2, and CIN3).

9.3 Chlamydia

GLM, Bayesian multilevel regression and GWR regression models were used in the Chlamydia study, and the results showed that some of the variables (e.g. deprivation condition, low social status and marital status variables) were associated with the development of Chlamydia and or chance of being infected at different spatial levels. However, the variable maps and the observed incidence maps did not show similar patterns except at Ward level and, thus, the predictions were inaccurate. This might be related to sampling problems, and in addition, the individual data collected from the surveys usually cause problems such as lack of prediction power. Similarly, aggregated data might also cause problems, such as loss of information. The ideal way to overcome such bias issues is to combine

both individual and aggregated data in the model. In reality, it is not possible to collect any variables at an individual level, and the total number of Chlamydia cases may not be available at the national level either. Without such data the predictive power cannot be increased easily. It is important to understand that the deprivation condition, social status and family structure variables may not be sufficient to explain the geographical patterns evident in the Chlamydia incidence rate. This is most likely because the chance of developing or becoming infected by Chlamydia is highly related to personal sexual behaviour, the number of sexual partners, and the patients' partners' sexual behaviour.

Individuals may have very different behaviours and, thus, varying chances of developing Chlamydia; however, information on such behaviour (e.g. family history, personal history and personal sexual behaviour) is not available. In terms of the disease modelling, the responses to treatments also vary from person to person; thus the chance of developing Chlamydia and the level of suffering can differ from one individual to the next. However, the detection rate of Chlamydia could be quite different across England. Nevertheless, identifying the associated factors does provide information that may be useful in planning programmes.

Further research is required to investigate the link between more specific probabilities of transition between Chlamydia and other further damaging stages (e.g., PID) and socio-economic variables at the local scale.

9.4 Summary

A better understanding of disease development, disease spatial patterns and the contributing factors (i.e. causes and associated factors) are very important in healthcare planning. This research demonstrated the use of various mathematical models to examine the relationship between cervical cancer incidence and deprivation condition, social status and family structure variables and Chlamydia infection.

Cervical cancer is related to two sets of factors (i) personal and (ii) socio-economic factors. In Chapter 4, the CART results showed that the age and if a

patient had positive HPV history, who is likely to require a colposcopy in her life time. In Chapter 5, the regression results identified that cervical cancer incidence is related to age, deprivation condition (Townsend index score), social status (low status) and family structure (single or married female population) at the national level.

Chlamydia infection is inaccurately related to age and deprivation condition, social status and family structure variables at individual level, but it is positively related to the proportion of low social grade population at CAS ward level.

To operate a screening system, it is necessary to estimate and allocate resources efficiently; however, it is the most difficult task faced by the healthcare industries and limited resources can easily be overstretched. Where relationships (i.e. between disease incidence and deprivation condition, low social status and family structure variables) exist they can be used to inform the optimisation of screening programmes.

Appendix A

Medical Terms Explanation

Cancer “It arises from the abnormal and uncontrolled division of cells that then invade and destroy the surrounding tissues” (Martin, 2000).

Cervical cancer “Cancer of the neck (cervix) of the uterus” (Martin, 2000).

Chlamydia “Chlamydia is a genus of virus-like bacteria that cause disease in man and birds” (Martin, 2000).

Colposcopy Colposcopy is a type of diagnosis test to diagnose cervical precancer, which is available in NHS cervical cancer screening programme (Singer and Monaghan, 2000).

Carcinoma in situ (CIS) “Cancer that arises in epithelium, the tissue that lines that skin and internal organs of the body. It may occur in any tissue containing epithelial cells. In many cases that site of origin of the tumour may be identified by the nature of the cells it contains” (Martin 2000). CIS is a term used to measure the thickness of the epithelium that is covered by undifferentiated neoplastic cells (Singer and Monaghan, 2000).

Cervical Intraepithelial Neoplasia (CIN) CIN is a term measuring the amount of cells changed to abnormal cells in the cervix; the abnormal cells may become the invasive cervical cancer cells. It can be measured in three grade levels (e.g. CIN 1, CIN 2 and CIN 3) according to the amount of abnormal cells (Martin, 2000).

Human Papilloma Virus (HPV) “HPV is a member of the papovavirus group that causes warts, including genital warts. There are over 50 strains of HPV: certain strains are considered to be causative factors in the development of anal and genital cancers, especially cervical cancer, but additional

factors are necessary before the cells become malignant. HPV is one of the most common sexually transmitted infections.” (Martin, 2000).

Pelvic Inflammation Disease (PID) “PID is an acute or chronic condition in which the uterus, Fallopian tubes, and ovaries are infected. The inflammation is the result of infection spreading from an adjacent infected organ or ascending from the vagina; it may also result from a blood-borne infection, such as tuberculosis” (Martin, 2000).

Ectopic pregnancy “The pregnancy development of a fetus at a site other than in the uterus. The may happen if the fertilized egg cell remains in the ovary to the uterus (the Fallopian tube) or if it lodges in the free abdominal cavity” (Martin, 2000).

Infertility “Inability in a woman to conceive or in a man to induce conception” (Martin, 2000).

Appendix B

Cervical Cancer National Screening Guidelines

Grade	Explanation	Action
NEGATIVE	No abnormalities detected	Routine recall after three to five years
ABNORMAL	Cellular appearances which cannot be described as normal	Refer for colposcopy after one borderline change or three abnormal tests at any grade in a ten year period
Borderline changes	Endocervical cell changes	Refer for colposcopy after one test is reported as borderline
	Squamous cell changes	Refer for colposcopy after three tests in a series are reported as borderline
Mild dyskaryosis	Cellular appearances consistent with CIN 1	Ideally refer for colposcopy but it remains acceptable to recommend a repeat test after one test reported as mild dyskaryosis. If two tests are reported as mild dyskaryosis refer for colposcopy
Moderate dyskaryosis	Cellular appearances consistent with CIN 2	Refer for colposcopy
Severe dyskaryosis	Cellular appearances consistent with CIN 3	Refer for colposcopy
Suspected invasive cancer	Possibility of invasive cancer	Refer for colposcopy. Women should be seen urgently within two weeks of referral
INADEQUATE	The test cannot be interpreted. It may be too thick or too thin, obscured by inflammatory cells, blood, incorrectly labelled or does not contain the right type of cells	Repeat the test. Refer for colposcopy after three consecutive inadequate samples

Appendix C

Model for Evaluating Cervical Cancer Screening Options

A decision tree model for the cervical cancer screening programme is illustrated in Figure C1. The model was designed to compare three pre-cervical cancer screening options, which were (i) current screening programme, (ii) provide HPV test to patients, who had mild dyskaryosis, moderate, severe dyskaryosis and severe dyskaryosis/ suspected invasive cancer result and (iii) provide colposcopy to patients, only if patients had HPV positive.

The aims of comparing these options were to evaluate the effectiveness of each option and find the best option, which can reduce the number of unnecessary colposcopy tests. The healthcare providers can consider changing the screening policies, having HPV tests instead of repeat smear test and not having colposcopy immediately.

Just before starting to evaluate each option, the user needs to select an option first, and then click the “Confirm” button and finally the “Evaluation” button, if the user wants to clear all the information in the main screen, then press the “Clean” button.

UserForm5

Option1 | Result |

Population

Enter population size

Risk Group Probability

	Negative	Mild	Moderate	Severe	Cancer	Inadequate
LRG	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
MRG	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
HRG	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>

Options

Definitions

Graph summary

Options

Evaluation

Cancel

Exit

Select an option → ☐ Option 1 *Current Policies*

☐ Option 2 *Further Smeat Test for the Patient, who has Moderate dyskaryotic from Low Risk Group.*

☐ Option 3 *Further Smeat Test for the Patient, who has Moderate dyskaryotic from both Low and Medium Risk Groups.*

Confirm the option and the information

Start evaluating for the selected option

Figure C1. The main screen for the decision tree model.

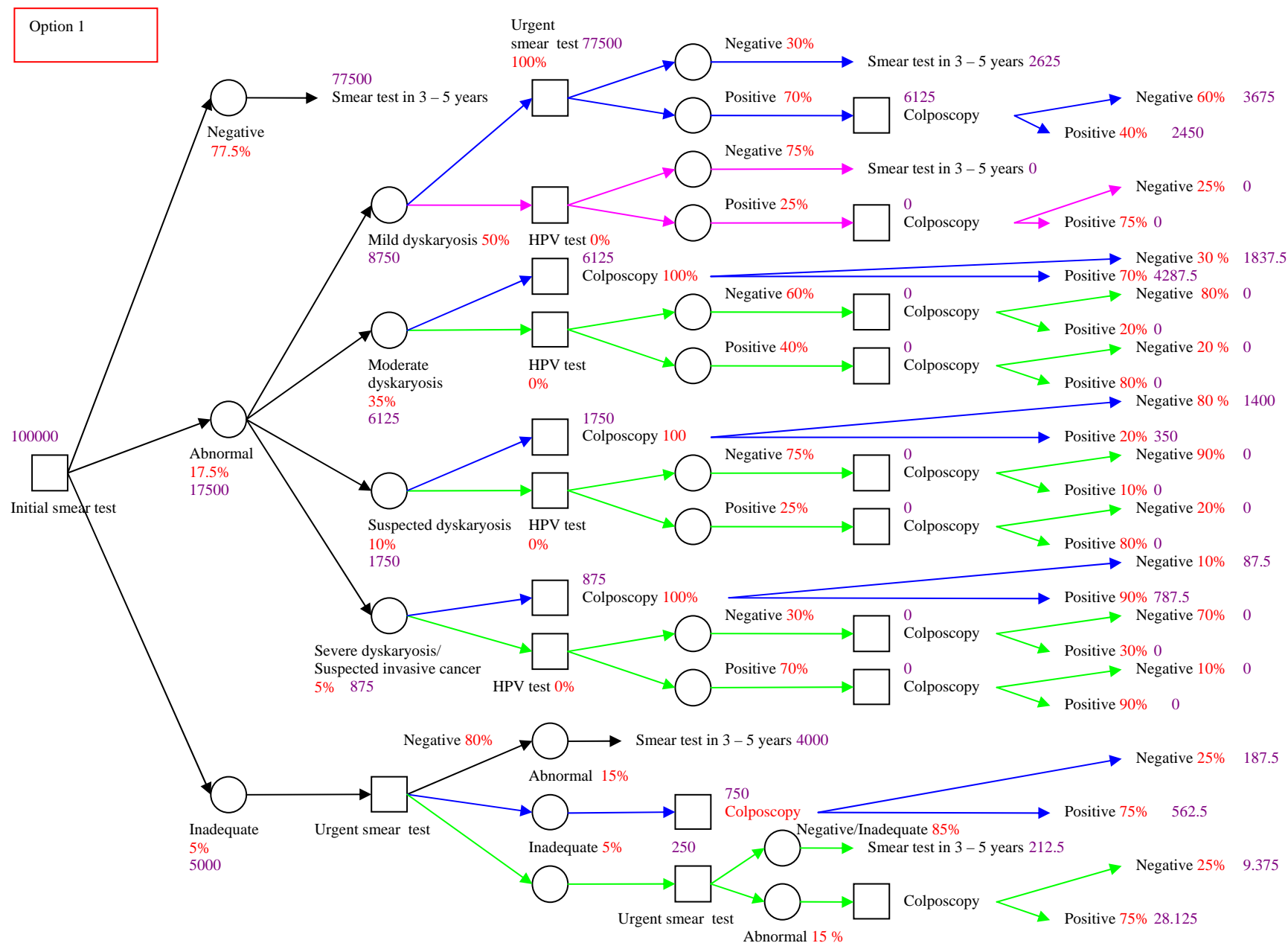
The summary results are displayed in Figure C2. If any needed information was missing in Figure C1, an error message appears to remind the users which particular information is needed.

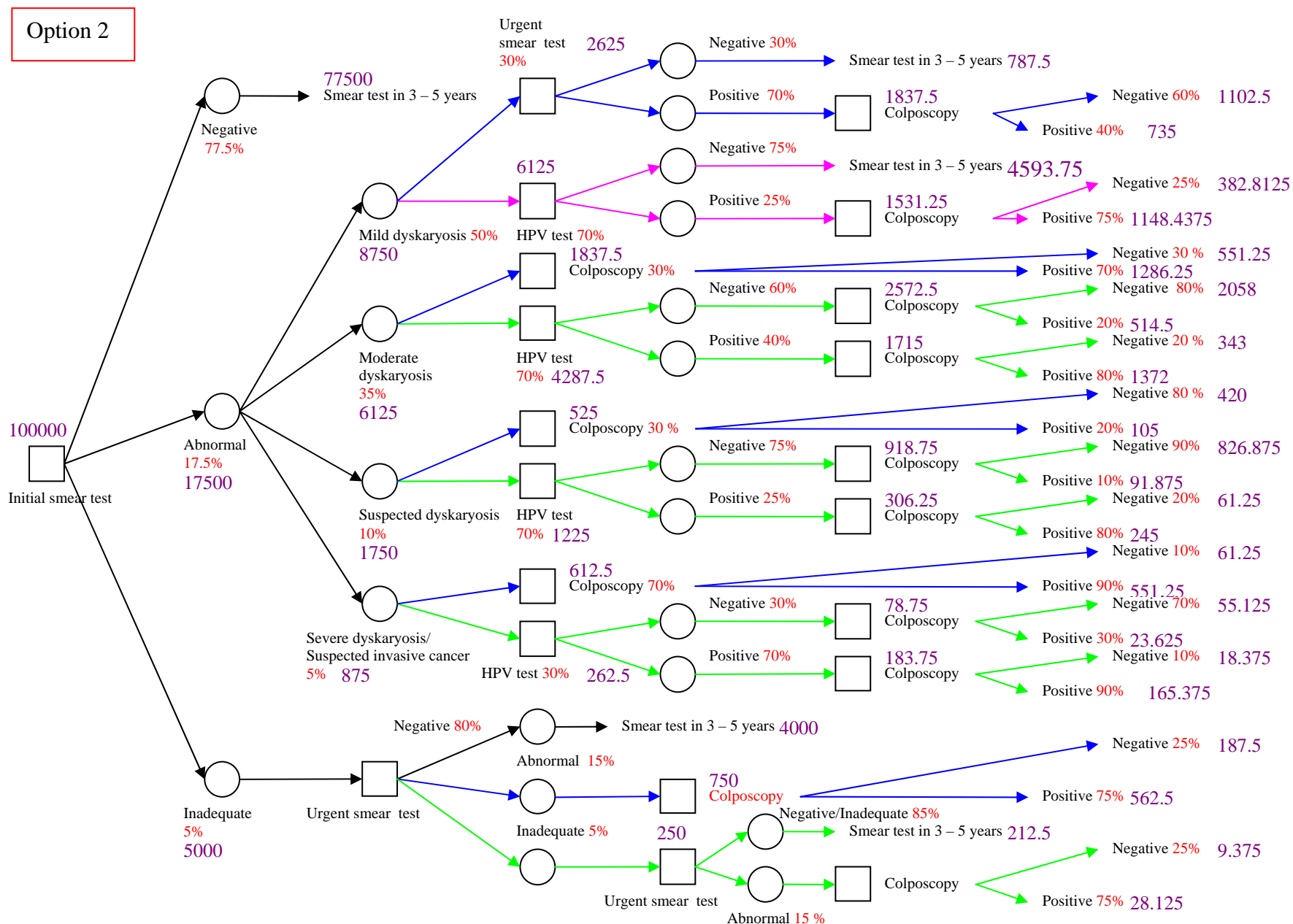
UserForm5

Option1 | Result |

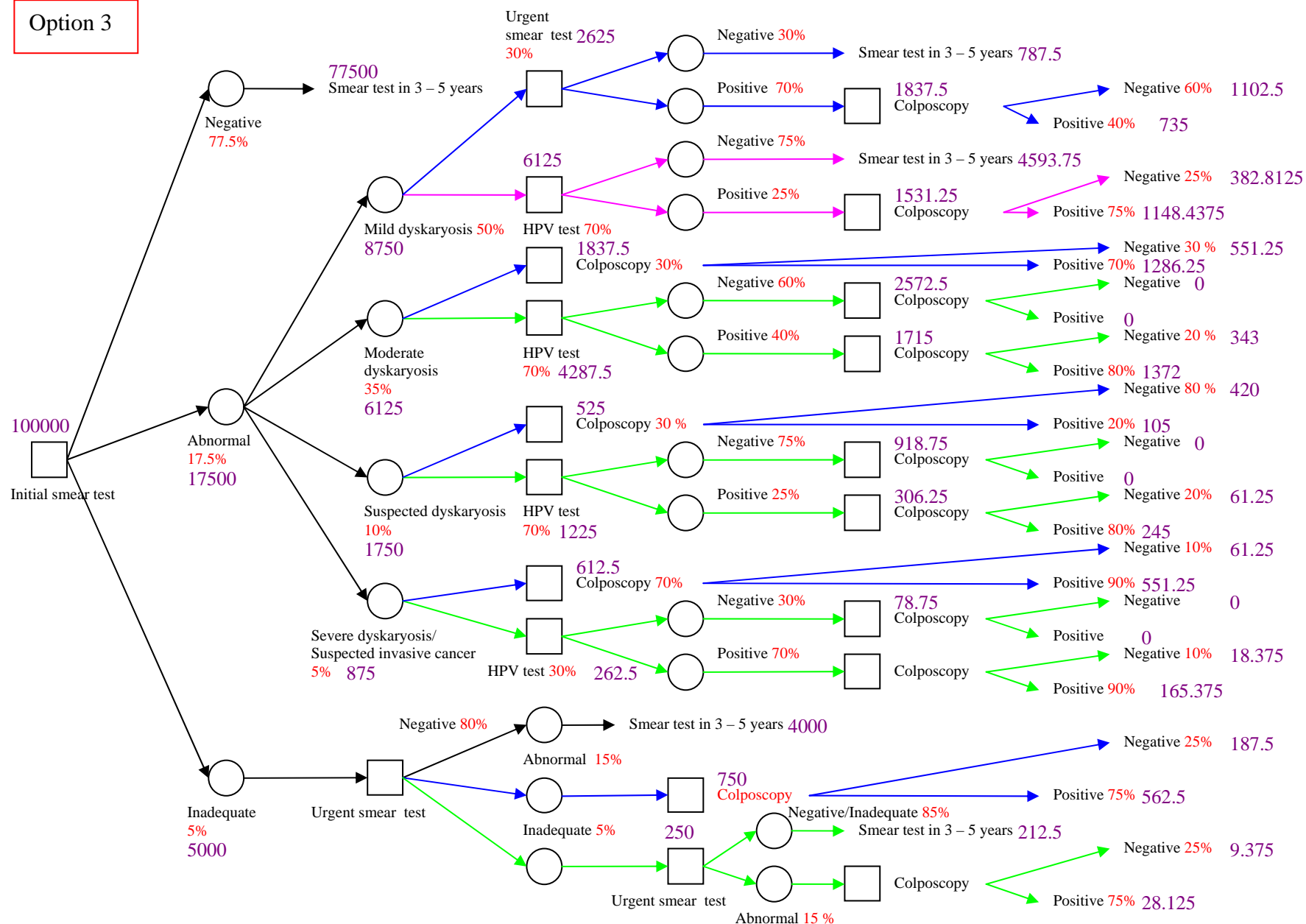
Total smear tests	<input type="text"/>	Total mild dyskaryotic	<input type="text"/>
Total routine smear tests	<input type="text"/>	Total moderate dyskaryotic	<input type="text"/>
Total urgent smear tests	<input type="text"/>	Total severe	<input type="text"/>
Total colposcopy	<input type="text"/>	Total invasive cancer	<input type="text"/>
Total negative tests	<input type="text"/>	Total inadequate	<input type="text"/>

Figure C2. Result summary for the selected screening option.





Option 3



Summary	Initial smear test	Urgent smear test	Total colposcopy	Total positive colposcopy	Total HPV test	Total tests (smear, colposcopy and HPV)
Option 1	100000	14000	15662.5	8465.625 (54%)	0	12966.25
Option 2	100000	7875	12906.25	6828.9375 (52.9%)	11900	119775
Option 3	100000	7875	9336.25	6198.9375 (66.4%)	11900	129111.25

Therefore option 3 is the most effective, because it provides the largest percentage of positive colposcopy (66.4%).

Appendix D

Cervical Cancer Simulation Model

A simulation model was developed to simulate the pre-cervical cancer disease process when data are available for estimating the simulation parameters e.g. distributions parameters. However, such data are not available, thus dummy distributions and parameters were used to demonstrate the use of the simulation model. The simulation model requires a random number generator, which is called “ITVCMath.dll”. Please follow the following instructions in order to link the generator and the model together.

Instruction

Step 1: Go to Tool within Excel

Setp 2: Select Macro from Tool

Setp 3: Select Visual Basic Editor

Setp 4: Go to Tool within the Visual Basic Editor

Setp 5: Go to references

Setp 6: Browse the random number generator and click “OK”

Total number of positive smear patients per month were produced by clicking the “Binomial” button in Figure D1.

Month (i)	No. of patient per month				
1	9				
2	9				
3	16				
4	9				
5	8				
6	7				
7	9				
8	9				

Binomial

Figure D1 Step one to produce simulated patients, who had positive smear result.

The lengths of times of stay in each of the disease states were estimated by the Weibull distribution. The user is allowed to change the distribution and the parameters inside the code. An example is shown in Figure D2. A summary of total patients in each state is given in Figure D3 per month.

Month (i)	No. of Individual	Normal	CIN 1	CIN 2	CIN3			
1	1	33	29	21	8			
1	2	36	24	20	10		Random	
1	3	36	27	20	10			
1	4	34	29	19	10			
1	5	34	31	21	11			
1	6	35	31	19	9			
1	7	34	26	20	10			
1	8	33	28	18	10			

Figure D2 An example showing the times of stay in each possible disease state for each individual patient.

Time		Stage			
Week (i)		Normal	CIN 1	CIN 2	CIN 3
Month 1	9	0	0	0	0
Month 2	20	0	0	0	0
Month 3	32	0	0	0	0
Month 4	45	0	0	0	0
Month 5	50	0	0	0	0
Month 6	56	0	0	0	0
Month 7	67	0	0	0	0
Month 8	77	0	0	0	0
Month 9	87	0	0	0	0
Month 10	99	0	0	0	0
Month 11	110	0	0	0	0
Month 12	120	0	0	0	0
Month 13	126	0	0	0	0
Month 14	133	0	0	0	0
Month 15	148	0	0	0	0
Month 16	154	0	0	0	0

Figure D3 Total patients in each state per month.

Appendix E

Cervical cancer GWPR results

Model	Variables	Kernel size	AICc (global)	BIC (global)	AICc (local)	BIC (local)
Random 1	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	612.968809	620.673215	539.322475	610.348925
Random 2	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	610.313424	618.017830	536.064037	607.092674
Random 3	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	619.199152	626.903558	544.085911	615.101333
Random 4	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	604.149583	611.853989	530.546542	601.566662
Random 5	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	605.868535	613.572941	532.851371	603.883733
Random 6	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	600.748651	608.453057	528.786566	599.814494
Random 7	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	604.846257	612.550663	531.596209	602.616981
Random 8	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	610.850499	618.554905	537.414116	608.441280
Random 9	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	604.146856	611.851262	531.225499	602.248834
Random 10	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	608.709612	616.414018	535.416812	606.431451
Random 11	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	597.383411	605.087817	524.371359	595.393448
Random 12	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	602.653705	610.358111	529.750664	600.768160
Random 13	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	741.235997	748.940403	585.806422	657.800429
Random 14	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	609.471271	617.175677	535.833222	606.860776
Random 15	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	608.248842	615.953248	534.331422	605.349351
Random 16	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	608.816634	616.521040	535.524496	606.541333
Random 17	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	612.846426	620.550832	539.018499	610.052060
Random 18	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	600.259876	607.964282	526.309510	597.333755
Random 19	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	604.301925	612.006331	530.885972	601.911188
Random 20	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	606.614449	614.318855	533.076753	604.095948
Random 21	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	605.922535	613.626941	533.345661	604.368320
Random 22	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	609.650209	617.354615	536.328890	607.347887
Random 23	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	610.594769	618.299175	537.798756	608.823638
Random 24	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	610.093808	617.798214	537.263698	608.278534
Random 25	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	614.182364	621.886770	540.076933	611.094551
Random 26	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	607.638178	615.342584	533.908497	604.936805
Random 27	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	603.108900	610.813306	530.272566	601.294785
Random 28	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	614.339132	622.043538	540.154719	611.173890
Random 29	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	612.652290	620.356696	538.290571	609.312235
Random 30	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	605.900160	613.604566	533.095988	604.119520
Random 31	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	600.332028	608.036435	527.029383	598.060495
Random 32	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	612.416391	620.120797	538.370861	609.387097
Random 33	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	601.903725	609.608131	529.182309	600.208612
Random 34	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	596.394304	604.098710	523.614381	594.645127
Random 35	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	604.303897	612.008303	531.252234	602.275586

Appendix E

Random 36	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	598.294507	605.998913	526.515442	597.541765
Random 37	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	602.315367	610.019773	527.825936	598.853810
Random 38	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	609.290608	616.995014	535.762959	606.776912
Random 39	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	598.216931	605.921337	525.067194	596.093567
Random 40	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	607.871200	615.575606	533.923007	604.944259
Random 41	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	608.355362	616.059768	534.759057	605.796040
Random 42	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	615.813697	623.518103	541.428419	612.453023
Random 43	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	2947.626977	2955.331383	2445.934700	2518.356819
Random 44	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	609.052703	616.757109	535.343087	606.375374
Random 45	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	610.505383	618.209789	536.896840	607.911584
Random 46	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	608.438809	616.143215	535.152370	606.174284
Random 47	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	604.622725	612.327131	531.500426	602.519560
Random 48	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	600.720939	608.425345	526.957152	597.978528
Random 49	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	609.886674	617.591080	536.202910	607.231862
Random 50	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	606.023040	613.727446	532.412195	603.432068
Random 51	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	603.213627	610.918033	530.507036	601.529117
Random 52	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	611.032407	618.736813	537.409128	608.430500
Random 53	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	601.660757	609.365163	528.045529	599.068785
Random 54	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	605.258893	612.963299	532.114011	603.136507
Random 55	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	602.458332	610.162738	529.739668	600.762311
Random 56	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	608.168092	615.872498	534.789154	605.809419
Random 57	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	602.729019	610.433425	529.899757	600.923639
Random 58	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	845.798522	853.502928	633.764914	703.295557
Random 59	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	612.796409	620.500815	538.702946	609.721710
Random 60	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	605.750367	613.454773	532.474824	603.496660
Random 61	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	605.249045	612.953451	532.159185	603.181024
Random 62	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	610.502165	618.206571	537.039513	608.059681
Random 63	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	596.081241	603.785647	522.142957	593.182254
Random 64	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	605.554574	613.258980	532.165617	603.196504
Random 65	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	603.894797	611.599203	530.459547	601.489820
Random 66	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	604.625235	612.329641	531.146424	602.173575
Random 67	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	609.188967	616.893373	535.509324	606.532238
Random 68	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	604.273669	611.978075	530.921657	601.951211
Random 69	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	609.312593	617.016999	535.761283	606.789585
Random 70	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	603.143825	610.848231	529.920201	600.952429
Random 71	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	604.710004	612.414410	532.020146	603.043395
Random 72	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	603.168681	610.873087	530.048085	601.076981
Random 73	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	606.700860	614.405266	533.744670	604.759334
Random 74	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	610.354973	618.059379	536.903836	607.923550

Appendix E

Random 75	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	602.428887	610.133293	529.578966	600.603398
Random 76	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	606.978669	614.683075	532.963213	603.997184
Random 77	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	610.571738	618.276144	536.299174	607.328080
Random 78	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	607.038241	614.742647	533.334869	604.362105
Random 79	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	606.388341	614.092747	533.163571	604.184198
Random 80	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	596.286813	603.991219	524.605675	595.621671
Random 81	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	603.990405	611.694811	529.796006	600.816695
Random 82	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	600.453838	608.158244	528.000680	599.028160
Random 83	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	606.314463	614.018869	532.686381	603.720598
Random 84	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	604.574409	612.278815	531.153201	602.181261
Random 85	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	607.277628	614.982034	534.162393	605.176753
Random 86	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	605.569972	613.274378	532.054953	603.084787
Random 87	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	607.381718	615.086124	534.939131	605.954667
Random 88	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	609.526113	617.230519	535.872546	606.892748
Random 89	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	602.025641	609.730047	528.519946	599.545851
Random 90	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	604.388393	612.092799	530.713100	601.734608
Random 91	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	608.479036	616.183442	534.825439	605.846241
Random 92	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	609.688141	617.392547	536.086960	607.110063
Random 93	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	606.072333	613.776739	532.669251	603.687401
Random 94	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	605.393509	613.097915	531.737609	602.751588
Random 95	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	600.221389	607.925795	527.671037	598.698375
Random 96	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	608.808338	616.512744	534.693288	605.718945
Random 97	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	610.172743	617.877149	536.710503	607.740176
Random 98	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	603.362964	611.067370	530.611640	601.625018
Random 99	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	610.748860	618.453266	537.073551	608.093716
Random 100	$e_i \exp(\beta_{0i} + \beta_{1i} G45_i)$	91	604.425023	612.129429	531.453545	602.482296

Appendix F

Chlamydia Decision Tree Model for Evaluating Screening Options

Decision tree model for evaluating Chlamydia screening options in Chapter 6 is shown in Figure E1. Users need to enter some information in order to carry out the calculation e.g. number of women to start with, and the probabilities of each event. Grouping idea is demonstrated in decision tree model, there are two groups of population (i) between ages 16-25 and (ii) over age 25. After entering all the information in Figure E1, users need to click the “Calculate” button in Figure E2 to start calculate. And the results are summarized in Figure E2.

Information			
No. Women		10000	
Screening Option		no	
Prob	Treated	0.3	
	Untreated	0.7	
Risk groups		Prob	
Low risk group: age > 25		0.25	
High risk group: 16-24		0.75	
Chlamydia Treatment		Upper Prob	Lower Prob
PID Treated		0.1	0.2
PID Untreated		0.1	0.4
PID Treatment		Prob	
Recorded		0.8	
Infertility		0.2	

Users need to confirm the screening option at this state.

Figure E1. Information page for the decision tree model. Users need to enter the number of women to start with, and the probabilities that associated to each events.

Results			
States	Low-risk group	High risk group	Total
Healthy (Chlamydia -ve)			9000
Chlamydia +ve	250	750	1000
Treated	75	225	300
Untreated	175	525	700
PID	55	165	220
Recorded	107.75	323.25	431
Infertility	11	33	44
Remain in complication	131.25	393.75	525

Click on this button to start.

↓

Calculate

Figure E2. Results summary of decision tree model.

A set of probabilities from NHS reports and HPA website were used, but the probabilities are allowed to change in order to compare the efficiency between different options (HPA, 2006; Primarolo, 2006).

Appendix G

Chlamydia Simulation Model

The simulation model in Chapter 7 is discussed in details in this Appendix. To start simulation users need to enter some information like the start and the end date of the simulation period. The simulation period is measured in months and the number of simulated average patients per month are needed. All information needed is shown in Figure F1.

The simulation model is required to attach with a random number generator, which is called “ITVCMath.dll”. Please follow the following instructions in order to link the generator and the model together.

Instruction

Step 1: Go to Tool within Excel

Setp 2: Select Macro from Tool

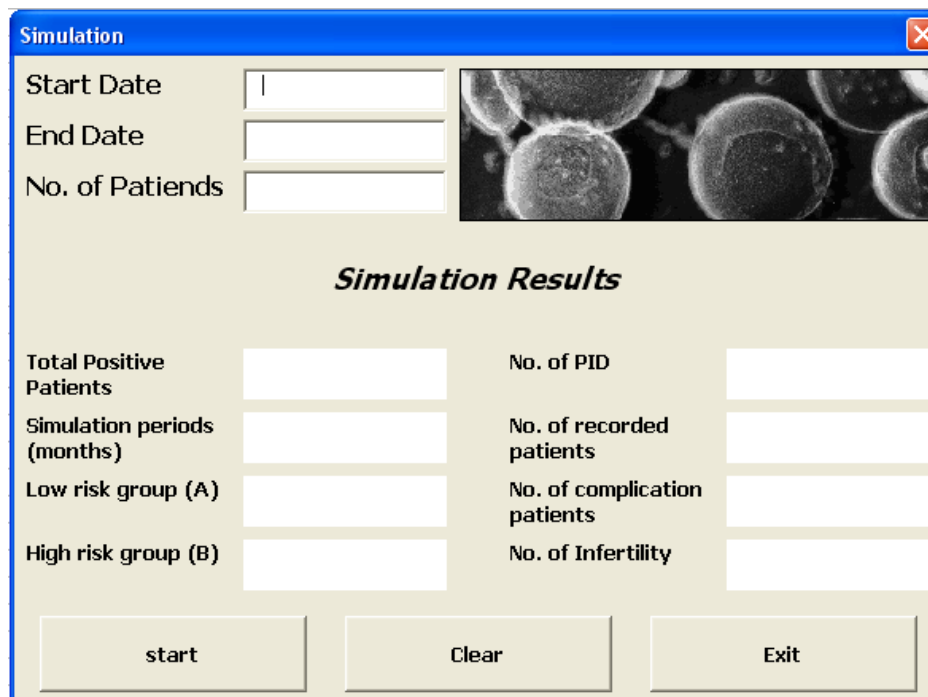
Setp 3: Select Visual Basic Editor

Setp 4: Go to Tool within the Visual Basic Editor

Setp 5: Go to references

Setp 6: Browse the random number generator and click “OK”

Each of the patients is simulated and flows through the simulation model, which is similar to the real life process and the simulated results are summarized in Figure F1 below.



The image shows a software window titled "Simulation" with a standard Windows-style title bar (blue with a close button). The window is divided into several sections. At the top left, there are three input fields labeled "Start Date", "End Date", and "No. of Patiends" (note the typo). To the right of these fields is a small rectangular image showing several spherical cells, possibly representing oocytes or sperm. Below the input fields, the section is titled "Simulation Results" in a bold, italicized font. Under this title, there are two columns of input fields. The left column contains "Total Positive Patients", "Simulation periods (months)", "Low risk group (A)", and "High risk group (B)". The right column contains "No. of PID", "No. of recorded patients", "No. of complication patients", and "No. of Infertility". At the bottom of the window, there are three buttons: "start", "Clear", and "Exit".

Simulation Results	
Total Positive Patients	No. of PID
Simulation periods (months)	No. of recorded patients
Low risk group (A)	No. of complication patients
High risk group (B)	No. of Infertility

start Clear Exit

Figure F1 Simulation user interface, details of information is needed for starting the simulation process and results will be displayed in the results section.

Presentations

Both of the cervical cancer and Chlamydia applications have been presented at a number of international conferences. The details of conferences at which the author presented the research are listed below.

Cheng, M.Y.E., Atkinson, P.M., Shahani, A.K. (2006) Investigating the relations between cervical cancer geographical variation in socio-economic deprivation [oral presentation]. International conference in GIS and public health, Hong Kong, 27-29 June 2006.

Cheng, M.Y.E., Atkinson, P.M., Shahani, A.K. (2007) Geographically weighted Poisson regression of the relation between cervical cancer and socio-economic deprivation in the UK [poster]. *Research show cases*, University of Southampton, 22 March 2007.

Cheng, M.Y.E., Atkinson, P.M., Shahani, A.K. (2007) Bayesian hierarchical modelling of the relation between cervical cancer and socio-economic deprivation in the UK [oral presentation]. *The International Conference of the Royal Statistical Society*, York, 16-20 July 2007.

Cheng, M.Y.E., Atkinson, P.M., Shahani, A.K. (2007) Geographically weighted Poisson regression of the relation between cervical cancer and socio-economic deprivation in the UK [oral presentation]. *The 7th International workshop of Geographical Information System (IWGIS)*, Beijing, 14-15 September 2007.

References

Abellan, J.J., Fecht, D., Best, N., Richardson, S., Briggs, D.J. (2007) Bayesian analysis of the multivariate geographical distribution of the socio-economic environment in England. *Environmetrics*, 18, 745-758.

Akaike, H. (1974) New look at statistical-model identification. *IEEE Transactions on automatic control*, 19, 716-72.

American Academy of Family Physicians (2008), Smear process.

<http://familydoctor.org/online/famdocen/home/women/reproductive/gynecologic/138.html>

Arias-Pulido, H., Peyton, C.L., Joste, N.E., Vargas, H., Wheeler, C.M. (2006) Human papillomavirus type 16 integration in cervical carcinoma in situ and in invasive cervical cancer. *Journal of clinical microbiology*, 44, 1755 – 1762.

Ashby, D. and Smith, A.F.M. (2000) Evidence-based medicine as Bayesian decision-marking. *Statistics in medicine*, 19, 3291-3305.

Bernardinelli, L., Clayton, D., Montomoli, C. (1995) Bayesian estimates of disease maps: how important are priors. *Statistics in medicine*, 14, 2411-2431.

Besag, J., York, J., Mollie, A. (1991) Bayesian image restoration with applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43, 1-20.

Best, N., Ickstandt, K., Wolpert, R.L. (2000) Spatial Poisson regression for health and exposure data measured at disparate resolutions. *Journal of the American statistical association*, 95, 1076-1088.

Bosch, F.X. and Iftner, T. (2005) *The aetiology of cervical cancer*. Publication no. 22. Sheffield, National Health Service Cervical Screening Programme, Department of Health publications.

Brailsford, S.C., Sykes, J., Harper, P.R. (2006) Incorporating human behaviour in healthcare simulation models. In proceeding of the *Winter Simulation Conference*, 466-472.

Breiman, L., Friedman, J.H., Stone, C.J., Olshen, R.A. (1984) *Classification and regression trees*. Chapman & Hall/ CRC.

Broders, A.C. (1932) Carcinoma in situ contrasted with benign penetrating epithelium. *Journal of the American medical association*, 99, 1670-4.

Cancer Research UK (2005), Cervical cancer incidence rate.
<http://info.cancerresearchuk.org/cancerstats/types/cervix/incidence/>

Cancer Research UK (2008), Cervical cancer vaccine.
<http://www.cancerhelp.org.uk/help/default.asp?page=16024>

Census Dissemination Unit (2006), National deprivation indicators, 2001.
<http://cdu.census.ac.uk/related/deprivation.htm>

Centre for Multilevel modelling (2008), multilevel modelling.
<http://www.cmm.bristol.ac.uk/MLwiN/index.shtml>

Chiogna, M., Spiegelhalter, D.J., Franklin, R.C.G., Bull, K. (1996) An empirical comparison of expert-derived and data-derived classification trees. *Statistics in medicine*, 15, 157-169.

Claxton, K., Posnett, J. (1996) An economic approach to clinical trial design and research priority-setting. *Health economics*, 5, 513-24.

Claxton, K. (1999) The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *Journal of health economics*, 18, 341-364.

Claxton, K., Lacey, L.F., Walker, S. (2000) Selecting treatments: a decision theoretic approach. *Journal of the royal statistical society series A*, 163, 211-225.

Claxton, K. and Thompson, K. (2001) A dynamic programming approach to efficient design of clinical trial. *Journal of health economics*, 20, 797-822.

Claxton, K., Ginnelly, L., Sculpher, M., Philips, Z., Palmer, S. (2004) *A pilot study on the use of decision theory and value of information analysis as part of the NHS health technology assessment programme*. Publication no. 31. Health technology assessment, 8.

<http://www.nchta.org/pdfexecs/summ831.pdf>

Clayton, D., Hills, M. (1993) *Statistical models in epidemiology*. New York, Oxford University press.

Cooper, G. M. (1993) *The cancer book*. Boston, Jones and Bartlett.

DeGroot, M. H., Schervish, M. J. (2002) *Probability and statistics (3rd edition)*. Boston, Addison Wesley.

Elliott, P., Wakefield, J.C., Best, N., Briggs, D.J. (2000) *Spatial epidemiology: methods and application*. Oxford, Oxford University press.

Evenden, D., Harper, P.R., Brailsford, S.C., Harindra, V., (2006) Improving the cost-effectiveness of Chlamydia screening with targeted screening strategies. *Journal of the operational research society*, 57, 1400-1412.

Fenwick, E., Claxton, K., Sculpher, M. (2001) Representing uncertainty: the role of cost-effectiveness acceptability curves. *Health economics*, 10, 779-89.

Ferko, N., Debicki, D., Barnfi, F., Marocco, A., Mantovani, L.G. (2007) Estimating the long-term health and economic impact of a prophylactic cervical cancer vaccine on the burden of cervical disease in Italy. *Value in health*, 10, A440-A441.

Fotheringham, A.S., Charlton, M.E., Brunsdon, C. (1998) Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. *Environment and planning A*, 30, 1905-1927.

Fotheringham, A.S., Brunsdon, C., Charlton, M.E. (2002) *Geographically weighted regression: the analysis of the spatially varying relationships*. Chichester, John Wiley & Sons.

Gatrell, A.C. and Bailey, T.C. (1996) Interactive spatial data analysis in medical geography. *Social science and medicine*, 42, 843-855.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2003) *Bayesian data analysis (2nd edition)*. Chapman and Hall/ CRC.

Goldie, S.J., Grima, D., Kohli, M., Wright, T.C., Weinstein, M., Franco, E. (2003) A comprehensive natural history model of HPV infection and cervical cancer to estimate the clinical impact of a prophylactic HPV 16/18 vaccine. *International journal of cancer*, 106, 896-904.

Green, P.J. and Richardson, S. (2002) Hidden Markov models and disease mapping. *Journal of the American statistical association*, 97, 1055-1070.

Hammerschmidt, T., Siebert, U., Schwarz, T.F., Schneider, A., Rogoza, R.M., Ferko, N., Welte, R. (2007) A cost-effectiveness analysis of a prophylactic cervical cancer vaccine in Germany: results from a health economic model. *Value in health*, 10, A441-A441.

Harper, P.R. (2002) *Operational modelling for the planning and management of the healthcare resources*. PhD Thesis, University of Southampton.

Harper, P.R., Sayyad, M.G., de Senna, V., Shahani, A.K., Yajnik, C.S., Shelgikar, K.M. (2003) A systems modelling approach for the prevention and treatment of diabetic retinopathy. *European journal of operational research*, 150, 81-91.

Harper, P.R., Phillips, S., Gallagher, J.E. (2005) Geographical simulation modelling for the regional planning of oral and maxillofacial surgery across London. *Journal of the operational research society*, 56, 134-143.

Harper, P.R., Winslett, D.J. (2006) Classification trees: A possible method for maternity risk grouping. *European journal of operational research*, 169, 146-156.

Health Protection Agency (2006), *A complex picture: HIV & other sexually transmitted infections in the United Kingdom*. London, Centre for infections, Health Protection Agency.

http://www.hpa.org.uk/web/HPAwebFile/HPAweb_C/1194947365435

Health Protection Agency (2007), Genital Chlamydia.

<http://www.hpa.org.uk/webw/HPAweb&Page&HPAwebAutoListName/Page/1191942172070?p=1191942172070>

Herbert, A., Smith, J.H.F. (2007) Cervical screening: Women under 25 should be offered screening. *British medical journal*, 334, 273-273.

Hillier, F.S., Lieberman, G.J. (2001) *Introduction to operations research (7th edition)*. New York, McGraw Hill.

Hurvich, C.M., Tsai, C-L. (1989) Regression and time series model selection in small samples. *Biometrika*, 76, 297-307.

Jackson, C., Best, N., Richardson, S. (2006) Improving ecological inference using individual-level data. *Statistics in medicine*, 25, 2136-2159.

Jackson, C., Best, N., Richardson, S. (2008) Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. *Journal of the royal statistical society series A*, 171, 159-178.

- Jarup, L., Best, N., Toledano, M. B., Wakefield, J., Elliott, P. (2002) Geographical epidemiology of prostate cancer in Great Britain. *International journal of cancer*, 97, 695-699.
- Jenkins, D., Sherlaw-Johnson, C., Gallivan, S. (1996) Can papilloma virus testing be used to improve cervical cancer screening? *International journal of cancer*, 65, 768-773.
- Jonse, S.K. (1997) *Mathematical modelling for early detection and treatment of cancer*. PhD Thesis, University of Southampton.
- Kohli, M., Ferko, N., Martin, A., Franco, E.L., Jenkins, D., Gallivan, S., Sherlaw-Johnson, C., Drummond, M. (2007) Estimating the long-term impact of a prophylactic human papillomavirus 16/ 18 vaccine on the burden of cervical cancer in the UK. *British journal of cancer*, 96, 143-150.
- Lai, T.L., Ying, Z.L. (1994) A missing information principle and M-estimators in regression analysis with censored and truncated data. *Annals of statistics*, 22, 1222-1255.
- Lavori, P.W., Dawson, R., Shera, D. (1995) A multiple imputations strategy for clinical-trials with truncation of patient data. *Statistics in medicine*, 14, 1913-1925.
- Lawson, A.B., Clark, A. (2002). Spatial mixture relative risk models applied to disease mapping. *Statistics in medicine*, 21, 359-370.
- Lawson, A.B., Browne, W.J., Rodeiro, C.L. (2003) *Disease mapping with WinBUGS and MLwiN*. Chichester, Wiley.
- Low, N., Bender, N., Nartey, L., Redmond, S., Shang, A., Stephenson, J. (2006) *Revised rapid review of evidence for the effectiveness of screening for genital Chlamydial infection in sexually active young women and men*. National Health Service. www.nice.org.uk

Lunn, D.J., Whittaker, J.C., Best, N. (2006). A Bayesian toolkit for genetic association studies. *Genetic epidemiology*, 30, 231-247.

Martin, E.A. (2000) *Oxford concise medical dictionary*. New York, Oxford.

McCullagh, P., Nelder, J.A. (1992) *Generalized Linear Models*. London, Chapman & Hall.

Meyers. D., Wolff. T., Gregory. K. (2008) USPSTF recommendations for STI screening. *American academy of family physicians*, 77, 819-824.

Minh, D.L. (2000) *Applied probability models*. Pacific Grove, California, Duxbury.

Moore-Higgs, G.J., Almadrones, L.A., Gossfeld, L.M., Eriksson, J.H., Huff, B.C. (2000) *Women and cancer: a gynaecologic oncology nursing perspective (2nd edition)*. Boston, Jones and Bartlett.

Nakaya, T., and Dorling, D. (2005) Geographical inequalities of mortality by income in two developed island countries: a cross-national comparison of Britain and Japan. *Social science and medicine*, 60, 2865-2875.

Nakaya, T., Fotheringham, A.S., Brunsdon, C., Charlton, M. (2005). Geographically weighted Poisson regression for disease association mapping. *Statistics in medicine*, 24, 2695-2717.

National Statistics (2006), Sexual health.

<http://www.statistics.gov.uk/CCI/nugget.asp?ID=1330&Pos=2&ColRank=2&Rank=576>

National Health Service (NHS) (2007), Pregnant.

http://www.chlamydia-screening.nhs.uk/ys/what_if.html

- National Health Service (NHS) (2008 a), Chlamydia introduction.
<http://www.nhsdirect.nhs.uk/articles/article.aspx?articleId=99>
- National Health Service (NHS) (2008 b), Chlamydia treatment.
<http://www.nhsdirect.nhs.uk/articles/article.aspx?articleId=99§ionID=11>
- National Health Service (NHS) (2008 c), Pelvic inflammatory disease.
<http://www.nhsdirect.nhs.uk/articles/article.aspx?articleId=279§ionID=11>
- Pascutto, C., Wakefield, J.C., Best, N., Richardson, S., Bernardinelli, L., Staines, A., Elliott, P. (2000) Statistical issues in the analysis of disease mapping data. *Statistics in medicine*, 19, 2493-2519.
- Patnick, J. (2004) *NHS Cervical Screening Programme: A pocket Guide*. Sheffield, National Health Service Cervical Screening Programme, Department of Health publications.
- Patnick, J. (2005) *Annual Review 2005, Shaping futures, NHS Cervical Screening Programme*. Sheffield, National Health Service Cervical Screening Programme, Department of Health publications.
- Patnick, J. (2008) *Breast and cervical screening: the first 20 years*. Sheffield, National Health Service Cervical Screening Programme, Department of Health publications.
- Pimenta, J.M., Catchpole, M., Rogers, P.A., Perkins, E., Jackson, N., Carlisle, C., Randall, S., Hopwood, J., Hewitt, G., Underhill, G., Mallinson, H., McLean, L., Gleave, T., Tobin, J., Harindra, V., Ghosh, A. (2003 a). Opportunistic screening for genital chlamydial infection. I: acceptability of urine testing in primary and secondary healthcare settings. *Sexually transmitted infections*, 79, 16-21.
- Pimenta, J.M., Catchpole, M., Rogers, P.A., Hopwood, J., Randall, S., Mallinson, H., Perkins, E., Jackson, N., Carlisle, C., Hewitt, G., Underhill, G., Gleave, T., McLean, L., Ghosh, A., Tobin, J., Harindra, V. (2003 b) Opportunistic screening for genital Chlamydia infection. II: Prevalence among healthcare attenders,

outcome, and evaluation of positive cases. *Sexually transmitted infections*, 79, 22-27.

Powell, N.H., Harper, P.R. (2004) Workforce modelling in healthcare. Paper presented at the Seventh National conference of the *UK simulation society*, 66-69.

Primarolo, D. (2006) *Maintaining momentum: Annual report of the National Chlamydia screening programme in England 2006/2007*. London, Health Protection Agency.

http://www.hpa.org.uk/web/HPAwebFile/HPAweb_C/1204013012687

Raffle, A.E. (2004) Cervical screening: recent changes in policy regarding age and frequency are a poor use of resources. *British medical journal*, 328, 1272-1273.

Rasbash, J., Steele, F., Browne, W., Prosser, B. (2005) A user's Guide to MLwiN. Centre for multilevel modelling university of Bristol.

Rees, P., Martin, D., Williamson, P. (2002) *The census data system*. Chichester, Wiley.

Richardson, S., Best, N. (2003) Bayesian hierarchical models in ecological studies of health environment effects. *Environmetrics*, 14, 129-147.

Richardson, S., Thomson, A., Best, N., Elliott, P. (2004) Interpreting posterior relative risk estimates in disease-mapping studies. *Environmental health perspectives*, 112, 1016-1025.

Richart, R.M. (1968) Natural history of cervical intraepithelial neoplasia. *Clinical obstetrics gynecology*, 10, 748-84.

Rogerson, P., Sinha, G., Han, D. (2006) Recent changes in the spatial pattern of prostate cancer in the U.S.. *American journal of preventive medicine*, 30, s50-59.

- Sculpher., M., Drummond., M., Buxton. M. (1997) The iterative use of economic evaluation as part of the process of health technology assessment. *Journal of health services research & policy*, 2, 26-30.
- Sherlaw-Johnson, C., Gallivan, S., Jenkins, D. (1997) Evaluating cervical cancer screening programmes for developing countries. *International journal of cancer*, 72, 210-216.
- Sherlaw-Johnson, C., Gallivan, S., Jenkins, D. (1999) Withdrawing low risk women from cervical screening programmes: mathematical modelling study. *British medical journal*, 318, 356-361.
- Sherlaw-Johnson, C., Gallivan, Steve. (2000) The planning of cervical cancer screening programmes in Eastern Europe: is viral testing a suitable alternative to smear testing? *Health care management science*, 3, 323-329.
- Sherlaw-Johnson. C., Philips. Z. (2004) An evaluation of liquid-based cytology and human papillomavirus testing within the UK cervical cancer screening programme. *British journal of cancer*, 91, 84-91.
- Singer, A., Monaghan, J. M. (2000) *Lower genital tract precancer: colposcopy, pathology and treatment (2nd edition)*. Oxford, Blackwell science.
- Spiegelhalter, D.J., Best, N., Carlin, B.P., Van der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the royal statistical society series B*, 64, 583-616.
- Spiegelhalter, D.J., Abrams, K.R., Myles, J.P. (2004) *Bayesian approaches to clinical trials and health-care evaluation*. Chichester, Wiley.
- Sugiura, N. (1978) Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics, theory and methods*, A7, 13-26.

Taha, H.A. (1997) *Operations research: an introduction (6th edition)*. Upper saddle river, Prentice-hall.

Townsend, P., Davidson, N., Black, D., Morris, J.N., Smith, C. (1982) *Inequalities in health, the black report (1st edition)*. Penguins, Pelican.

Townsend, P., Phillimore, P., Beattie, A. (1988) *Health and Deprivation: Inequalities and the North*. London, Croom helm.

U.S. preventive services task force (2007). Screening for Chlamydia infection: U.S. Preventive services task force recommendation statement. *Annals internal medicine*, 147, 128-34.

Waller, L. A. and Gotway, C. A. (2004) *Applied Spatial Statistics for Public Health Data*. Hoboken, Wiley-interscience.

World Health Organization (WHO) (2007), Vaccinating against cervical cancer.
<http://www.who.int/bulletin/volumes/85/2/07-020207/en/>

Winston, W.L., (1994) *Operations research Applications and Algorithms*, third edition, Duxbury press

Wikimedia (2007) http://en.wikipedia.org/wiki/Cervix_uteri

World Health Organization (WHO) (2008 a), Cancer.
www.who.int/mediacentre/factsheets/fs297/en/

World Health Organization (WHO) (2008 b), Cancer.
<http://www.who.int/cancer/en/>