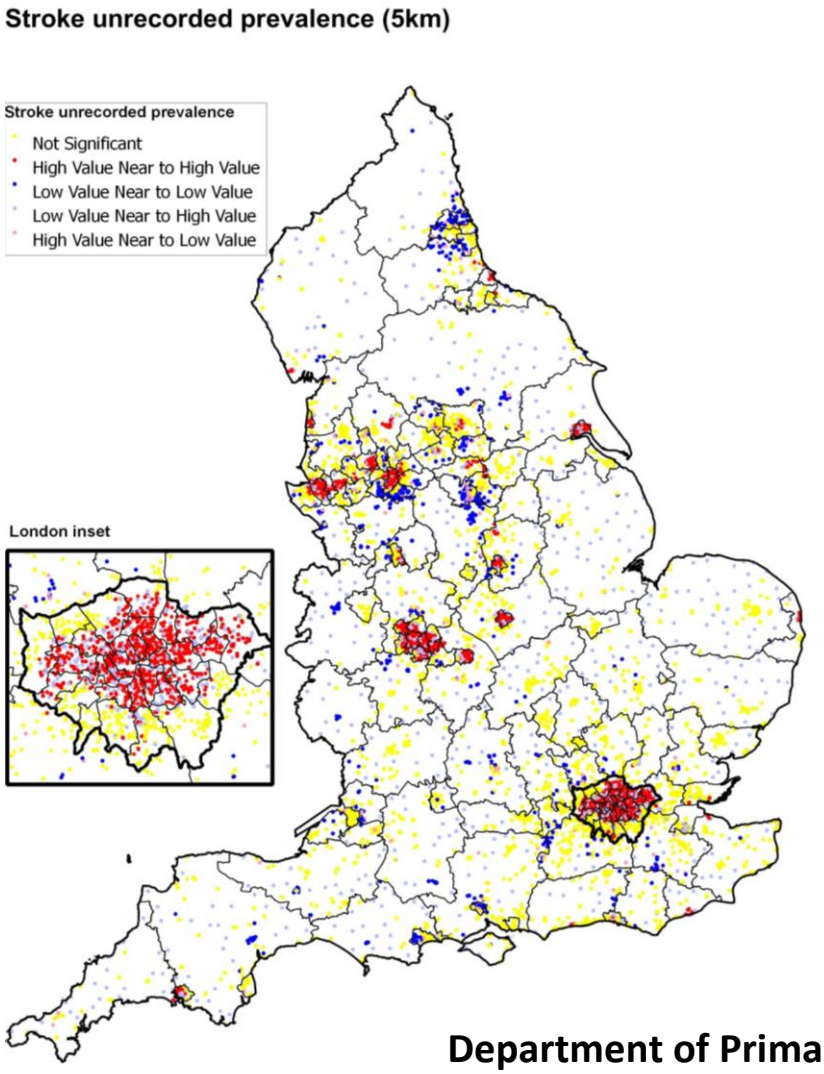
Development and use of methods to estimate chronic disease prevalence in small populations

Thesis submitted to the Faculty of Medicine of Imperial College London for the degree of Doctor of Philosophy



Michael Soljak
Department of Primary Care & Public Health
School of Public Health
Imperial College London 2011

1 Abstract

Introduction
National data on the prevalence of chronic diseases on general practice registers is now available. The aim of this PhD was to develop and validate epidemiological models for the expected prevalence of chronic obstructive pulmonary disease (COPD), coronary heart disease (CHD), stroke, hypertension, overall cardiovascular disease (CVD) and high CVD risk at general practice and small area level, and to explore the extent of undiagnosed disease, factors associated with it, and its impact on population health.

Methods
Multinomial logistic regression models were fitted to pooled Health Survey for England data to derive odds ratios for disease risk factors. These were applied to general practice and small area level population data, split by age, sex, ethnicity, deprivation, rurality and smoking status, to estimate expected disease prevalence at these levels. Validation was carried out using external data, including population-based epidemiological research and case-finding initiatives. Practice-level undiagnosed disease prevalence i.e. expected minus registered disease prevalence, and hospital admission rates for these conditions, were evaluated as outcome indicators of the quality and supply of primary health care services, using ordinary least squares (OLS) regression, geographically-weighted regression (GWR), and other spatial analytic methods.

Results
Risk factors, odds of disease and expected prevalence were consistent with external data sources. Spatial analysis showed strong evidence of spatial non-stationarity of undiagnosed disease prevalence, with high levels of undiagnosed disease in London and other conurbations, and associations with low supply of primary health care services. Higher hospital admission rates were associated with population deprivation, poorer quality and supply of primary health care services and poorer access to them, and for COPD, with higher levels of undiagnosed disease.

Conclusion
The epidemiologic prevalence models have been implemented in national data sources such as NHS Comparators, the Association of Public Health Observatories website, and a number of national reports. Early experience suggests that they are useful for guiding case-finding at practice level and improving and regulating the quality of primary health care. Comparisons with external data, in particular prevalence of disease detected by general practices, suggest that model predictions are valid.

Practice-level spatial analyses of undiagnosed disease prevalence and hospital admission rates failed to demonstrate superiority of GWR over OLS methods. Disease modellers should be encouraged to collaborate more effectively, and to validate and compare modelling methods using an agreed framework. National leadership is needed to further develop and implement disease models. It is likely that prevalence models will prove to be most useful for identifying undiagnosed diseases with a slow and insidious onset, such as COPD, diabetes and hypertension.

2 Declarations

The work in this thesis is my own and where any material could be construed as the work of others, it is fully cited and referenced and/or full acknowledgement is given.

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

3 Acknowledgements

In carrying out this research I have had a great deal of help and support from a number of colleagues.

Firstly, I acknowledge the support of Economic & Social Research Council for funding the PhD studentship which supported me during the PhD programme. The *Informing Healthier Choices* Programme (in particular Professor John Newton), which was funded by the Department of Health, supported my proposal to develop prevalence models, commissioned their development, and funded the Association of Public Health Observatories to make them widely available on their websites. In particular I acknowledge the help of Julian Flowers and Hannah Walford at Eastern Region Public Health Observatory, who made useful practical comments and were closely involved in providing the local data which enabled GP practice level prevalence estimates to be produced, and in making the data available on the APHO website. I also acknowledge the financial support provided by the Care Quality Commission (CQC), which commissioned from Imperial College London (ICL) a report on using geographical information and spatial statistics to support healthcare regulation. The funding was used to process the local data and construct the practice level maps which are in the thesis. I also acknowledge the useful input of Gerrard Abi-Aad (now Organisation for Economic Co-operation & Development), Valerie Olsen and Lisa Annaly from CQC. CQC also provided the practice level population and demographics estimates which have been used in later research.

I thank the ICL Statistical Advisory Service, in particular Dr Fabiana Gordon, whose statistical advice helped me to get started on the analysis of Health Survey data for the models. In addition Dr Alex Bottle has provided helpful ongoing statistical advice at several points in the research. I thank Edgar Samarasundera for introducing me to spatial statistics, in particular to geographically-weighted regression (GWR) techniques, for explaining GWR diagnostics, and for the GWR3 software used in the practice level analyses. In addition Professor Peter Congdon at the Geography Department at Queen Mary's University of London has been very helpful in increasing my knowledge of Bayesian methods, although I have only managed to start using them at the end of the research reported here.

I acknowledge the input of Luis Nacul in the development of the chronic obstructive pulmonary disease (COPD) prevalence model. I was originally Luis's trainer during his period as a Specialist Registrar in public health, when we started on the model, and we subsequently worked together on its development and publications. Andrew Dalton worked with me in a similar way on the cardiovascular disease (CVD) risk model, and we both spent considerable time on the local spreadsheet version, which was funded by the

North East London Cardiac & Stroke Network and its associated primary care trusts (PCTs). The Public Health Action & Support Team (PHAST) provided logistic support and funded access to the QRISK2 risk engine we used for the latter model. Both these models are therefore joint products.

I also thank the Information Centre for health & social care, the NHS Comparators team there, Nick Hopkinson, Sarah Elkin and colleagues at the National Heart & Lung Institute, and others who have been involved with model implementation and case-finding pilots such as that for COPD.

Finally I am particularly grateful to my Supervisor, Azeem Majeed, for his wise and unfailingly prompt advice over the last three years despite his extremely busy schedule.

4 Contents

4.1 Table of contents

1 Abstract 1
2 Declarations 2
3 Acknowledgements 2
4 Contents 4
4.1 Table of contents 4
4.2 Table of tables 7
4.3 Table of figures 10
4.4 Table of abbreviations 12
5 Background/policy context 14
6 Aim, Objectives and Management 16
6.1 Aims 16
6.2 Research Questions 16
6.3 Objectives 16
6.4 Ethics 17
6.5 Outputs 17
6.6 Benefits expected 17
6.7 Risks 17
7 Disease prevalence models: background 19
7.1 Potential uses of disease prevalence estimation 20
7.2 Previous UK prevalence models 21
7.3 Problems with current UK prevalence models 22
7.4 Small area prevalence estimation methodologies 22
7.4.1 Simple multiple logistic regression modelling 23
7.4.2 “Synthetic estimation” 23
7.4.3 Markov models 24
7.4.4 Bayesian approaches 25
7.4.5 Support vector machine modelling 26
7.4.6 Simulation models 27
8 Disease prevalence model development 28
8.1 Chronic Obstructive Pulmonary Disease Model 28
8.1.1 Introduction 28
8.1.2 Methods 29
8.1.3 Results 30
8.1.4 Applications & discussion 33
8.2 Coronary Heart Disease Model 37
8.2.1 Introduction 37
8.2.2 Methods 41
8.2.3 Results 44
8.2.4 Application: PCT, LA and practice level models 52
8.3 Stroke Model 56
8.3.1 Introduction 56
8.3.2 Methods 59
8.3.3 Results 61
8.3.4 Application: PCT, LA and practice level models 65

8.4 Hypertension Model 68
8.4.1 Introduction 68
8.4.2 Methods 71
8.4.3 Results 76
8.4.4 Application: PCT, LA and practice level models 83
8.5 Cardiovascular Disease Model 85
8.5.1 Introduction 85
8.5.2 Methods 86
8.5.3 Results 89
8.5.4 Application: PCT, LA and practice level models 91
8.6 Cardiovascular Disease Risk Model 92
8.6.1 Introduction 92
8.6.2 Methods 93
8.6.3 Results 96
8.6.4 Discussion 99
9 Model Validation 103
9.1 COPD in England: a comparison of expected, model-based prevalence and observed prevalence from general practice data 104
9.1.1 Introduction 104
9.1.2 Methods 104
9.1.3 Results 105
9.1.4 Discussion 112
9.2 Variations in cardiovascular disease under-diagnosis in England: national cross-sectional spatial analysis 116
9.2.1 Introduction 116
9.2.2 Methods 116
9.2.3 Results 118
9.2.4 Discussion 125
9.3 Case-finding strategies for COPD in primary care 127
9.3.1 Background 127
9.3.2 Methods 127
9.3.3 Discussion 130
10 Model Implementation 132
10.1 Evaluation of the uptake of the models by York University Health Economics Consortium 132
10.2 Vascular Programme briefing packs 134
10.3 Case-finding pilots and case studies 137
10.3.1 Case Study 1: Northamptonshire PCT 137
10.3.2 Case Study 2: South West Essex PCT 137
10.3.3 Case Study 3: Leicestershire County and Rutland PCT 138
10.3.4 Case Study 4: South Staffordshire PCT 138
10.3.5 Case Study 5: Nottinghamshire County PCT 139
10.4 Modelling risk factor scenarios and impact on disease prevalence 139
10.4.1 Predicting the impact of risk factor reductions on stroke prevalence 140
10.4.2 Comment on time lags 142
10.5 CVD prevalence modelling by the Care Quality Commission 143
10.6 Does undiagnosed COPD increase hospital admission rates? 145

10.6.1 Introduction....145
10.6.2 Methods....146
10.6.3 Results....149
10.6.4 Discussion....155
10.7 Exploring the utility of mapping and spatial statistics in the context of health & social care regulation....158
10.7.1 Introduction....158
10.7.2 Background/Introduction....161
10.7.3 Project Description/Methods....163
10.7.4 Results....173
10.7.5 Discussion & Conclusions....219
11 Discussion/Conclusions....223
11.1 Discussion....223
11.1.1 How should the prevalence of chronic diseases be estimated? How do different chronic disease prevalence estimation methods compare in terms of their validity?....223
11.1.2 What is the best available methodology given the requirements for prevalence models?....227
11.1.3 How valid are the chronic disease model prevalence estimates?....227
11.1.4 How useful are chronic disease prevalence estimates for case-finding at GP practice (and LSOA) level? Can chronic disease prevalence estimates be used with practice disease register prevalence to estimate the prevalence of undiagnosed disease?....228
11.2 The future....228
11.2.1 linkages with risk factor prevalence data from GP systems....228
11.2.2 A simulation model of the UK population?....231
12 List of related publications....232
13 Technical Appendix....233
14 Pragmatic methods for estimating prevalence in small populations....237
14.1 Prevalence estimates from studies applied to local populations....237
14.2 Regression models using large surveys....237
14.3 Validation, confidence intervals and robustness....238
14.4 Sensitivity testing....238
14.5 Internal validation....239
14.6 Receiver-Operating Characteristic (ROC) Analysis....239
14.7 External validation....239
14.8 Strengths and limitations of prevalence models....239

4.2 Table of tables

Table 1: Summary of small area synthetic estimation methods 24
Table 2: Risk factors for COPD included in the final prevalence model 31
Table 3: Number and proportion of people estimated to have COPD by age group and gender in England (estimates for 2005) 32
Table 4: Estimated prevalence P (percent) of COPD in England and according to area of residence 32
Table 5: Prevalence of coronary heart disease and ECG findings in 25-74 year old men and women 37
Table 6: Percentage prevalence of IHD, by survey year, age and sex [1] 39
Table 7: Ethnic Group Breakdown of HSfE 2004 Dataset & Merged 2003-2004 Dataset 42
Table 8 : Index of Multiple Deprivation Banding 43
Table 9: Variables Included in Merged Dataset 44
Table 10: Respondents Reporting Doctor Diagnosed CHD by Age Band 45
Table 11: Odds Ratios for LOCAL, Model With Only Locally Available Variables 46
Table 12: CHD prevalence by BMI category 47
Table 13: Odds ratios for complete model excluding treatment for hypertension 47
Table 14: Odds ratios for complete model including treatment for hypertension 49
Table 15: Comparison of predictions by model 51
Table 16: UK stroke prevalence rates, estimates from the World Health Organization, men and women per 1000 population 56
Table 17: Prevalence of stroke, by minority ethnic group and sex, aged 16 and over 2004 57
Table 18: Prevalence of stroke, by age within minority ethnic group and sex, aged 16 and over 2004 58
Table 19: Ethnic Group Breakdown of HSfE 2003 & 2004 Datasets & Merged 2003-2004 Dataset 59
Table 20: Index of Multiple Deprivation Banding 60
Table 21: crude stroke prevalence in 2003-4 dataset 61
Table 22: Comparison of area under ROC curves for differing stroke models 61
Table 23: Odds Ratios, Local Model 62
Table 24: Odds Ratios, Complete Model 63
Table 25: complete model variable selection 64
Table 26: Comparison of predictions by model 64
Table 27: Number of practices included in modelling 66
Table 28: Definition of urban-suburban-rural categories 67
Table 29: Distribution of urban-suburban-rural categories 67
Table 30: blood pressure levels, by minority ethnic group and sex, Aged 16 and over with valid blood pressure readings, 2004 69
Table 31: Prevalence of Hypertensive Categories in HSfE 2003 (Adults aged 16 and over with a valid blood pressure reading and data on medication) 71
Table 32: Ethnic Group Breakdown of HSfE 2003 & 2004 Datasets & Merged 2003-2004 Dataset 73
Table 33: Hypertensive Categories of HSfE 2003 & 2004 Datasets & Merged 2003-2004 Dataset 73
Table 34: Variables Included in Merged Dataset 74

Table 35: Index of Multiple Deprivation Banding 75
Table 36: Hypertension Category By Age Band 77
Table 37: Odds ratios for variables included in "local" model 77
Table 38: Odds ratios for Ntiles of IMD score (stepwise regression) 79
Table 39: Odds ratios for variables included in "complete" model 80
Table 40: comparison of observed: predicted hypertension- local model 83
Table 41: comparison of observed: predicted hypertension- complete model 83
Table 42: trends in HSfE prevalence of overall CVD (ever), by survey year, age and sex (all adults) 85
Table 43: Ethnic Group Breakdown of merged dataset 86
Table 44: Comparison of HSfE CVDEF1 & CVDIS variables in model dataset 87
Table 45: frequency of patient-reported doctor-diagnosed diabetes and undiagnosed diabetes in respondents with an FPG 88
Table 46: Crude CVD, diabetes and CVD+diabetes prevalence rates in 2003-5 dataset 89
Table 47: odds ratios, local CVD model 90
Table 48: actual vs predicted CVD 90
Table 49: variables included in QRISK2 and JBS2 algorithms 94
Table 50: major DH model assumptions for the risk factor interventions[166]; † if aged 70 or over 95
Table 51: PCT and practice level estimates of the numbers and proportion of high risk patients 96
Table 52: the prevalence of CVD risk factors in the high risk population, † due to small numbers no age/sex stratification was used in the IGT estimates 98
Table 53: comparison of classical regression and GWR results 111
Table 54: observed (GP-registered) and expected prevalence summary statistics 120
Table 55: comparison of classical regression and GWR results 124
Table 56: summary of primary care COPD screening studies 127
Table 57: objectives and outcomes of pilot study 128
Table 58: definitions of GOLD COPD stages 128
Table 59: smoking status of screened population 129
Table 60: spirometry results and GOLD category of screened patients 130
Table 61: Smoking and COPD prevalence in pilot practice 130
Table 62: Quality and Outcomes Framework indicators for COPD management and Patient Experience in 2008-2009 which were analysed 147
Table 63: characteristics of PCT and practice populations 150
Table 64: Summary of estimated, observed and undiagnosed COPD prevalence and admission rates 153
Table 65: multiple regression analyses 154
Table 66: Analytic scenarios and data sources 159
Table 67: 2008-9 Quality & Outcomes Framework indicators for CHD & stroke/TIA 166
Table 68: analytic scenarios and data sources 169
Table 69: Rationale behind analysis blocks used in the split-variable spatial regression models 171
Table 70: Descriptive analysis 173
Table 71: OLS univariate analysis. Dependent variable: unregistered CHD count 2008-9 174

Table 72: OLS multivariate stepwise analysis. Dependent variable: unregistered CHD count 2008-9 Independent variables: population + institutional factors.....175
Table 73: Collinearity Diagnostics 175
Table 74: descriptive analysis 181
Table 75: OLS univariate analysis. Dependent variable: unregistered stroke count 2008-9 181
Table 76: OLS multivariate stepwise analysis. Dependent variable: unregistered stroke count 2008-9 Independent variables: population + institutional factors 182
Table 77: regression diagnostics for stroke prevalence 190
Table 78: descriptive analysis 192
Table 79: OLS univariate analysis. Dependent variable: observed admission count. 192
Table 80: OLS multivariate stepwise analysis. Dependent variable: observed admission count. Independent variables: population + institutional factors.....193
Table 81: descriptive analysis 195
Table 82: univariate analysis. Dependent variable: observed admission count 195
Table 83: OLS multivariate stepwise analysis. Dependent variable: observed admission count. Independent variables: population + institutional factors.....196
Table 84: regression diagnostics for stroke admissions- all variables (PCT Scores not included).....199
Table 85: Stroke admissions: split variable Akaike Information Criterion scores 200
Table 86: Comparison of registered and unregistered prevalence for COPD, CHD and stroke.....227

4.3 Table of figures

Figure 1: Comparison of AUROC for different CHD prevalence models 51
Figure 2: Prevalence of smoking behaviours in England by age and sex, 2003-5.....54
Figure 3: Prevalence of smoking behaviours by Government Office Region, 2006...54
Figure 4: Comparison of area under ROC curves for differing models 62
Figure 5: modelled proportion hypertensive by age group and sex.....78
Figure 6: ROC Curve for local hypertension model.....82
Figure 7: ROC Curve for complete hypertension model.....82
Figure 8: the prevalence of high risk status, using the QRISK2 score, in the 40-74 year old population in England.....97
Figure 9: percentage of the high risk population with CVD risk factors using the QRISK2 and JBS2 risk scores 99
Figure 10: expected prevalence of COPD 107
Figure 11: diagnosed/QOF registered prevalence of COPD 108
Figure 12: diagnosed over expected prevalence of COPD.....109
Figure 13: local Moran's I cluster significance scores for the diagnosed: expected ratio.....110
Figure 14: local regression coefficients for the association between diagnosed and expected prevalence.....112
Figure 15: clusters and outliers in observed to expected ratios for CHD (red indicates clusters and blue indicates outliers) 121
Figure 16: Clusters and outliers in observed to expected ratios for hypertension (red indicates clusters and blue indicates outliers) 122
Figure 17: Clusters and outliers in observed to expected ratios for stroke (red indicates clusters and blue indicates outliers) 123
Figure 18: study design & flow chart.....129
Figure 19: use of modelled CHD prevalence in Vascular Programme Briefing Packs 135
Figure 20: use of modelled stroke prevalence in Vascular Programme Briefing Packs 136
Figure 21: Stroke Prevalence scenarios, Camden PCT 141
Figure 22: estimated unrecorded CVD prevalence by deprivation at GP practice level 143
Figure 23: average daily quantity of statins per standardised population.....144
Figure 24: expected CHD prevalence rate by practice, England 2008 177
Figure 25: QOF registered CHD prevalence rate by practice, England 2008.....178
Figure 26: unregistered CHD prevalence rate by practice, England 2008 179
Figure 27: observed: expected CHD prevalence rate by practice, England 2008....180
Figure 28: expected stroke prevalence rate by practice, England 2008 184
Figure 29: QOF registered stroke prevalence rate by practice, England 2008.....185
Figure 30: unregistered CHD prevalence rate by practice, England 2008 186
Figure 31: observed: expected CHD prevalence rate by practice, England 2008....187
Figure 32: Local Moran's I (2km) of stroke unrecorded prevalence 188
Figure 33: Local Moran's I (5 km) of stroke unrecorded prevalence 189
Figure 34: Local Moran's I (2km) of stroke admission rates 197
Figure 35: Local Moran's I (5km) of stroke admission rates 198

Figure 36: local R-squared for stroke admission rates, Greater Manchester, population factors only 202
Figure 37: local R-squared for stroke admission rates, Greater Manchester, service delivery variables only 203
Figure 38: local R-squared for stroke admission rates, Merseyside, population factors only 204
Figure 39: local R-squared for stroke admission rates, Merseyside, service delivery variables only 205
Figure 40: local R-squared for stroke admission rates, NW England, population factors only 206
Figure 41: local R-squared for stroke admission rates, NW England, service delivery variables only 207
Figure 42: local R-squared for stroke admission rates, NE England, population factors only 208
Figure 43: local R-squared for stroke admission rates, NE England, service delivery variables only 209
Figure 44: local R-squared for stroke admission rates, West Midlands, population factors only 210
Figure 45: local R-squared for stroke admission rates, West Midlands, service delivery variables only 211
Figure 46: scatterplot of demographic vs primary healthcare R squareds for Tyne & Wear practices 212
Figure 47: scatterplot of demographic vs primary healthcare R squareds for Greater Manchester practices 212
Figure 48: local R-squared for stroke admission rates, NW London, population factors only 213
Figure 49: local R-squared for stroke admission rates, NW London, service delivery variables only 214
Figure 50: local R-squared for stroke admission rates, NE London, population factors only 215
Figure 51: local R-squared for stroke admission rates, NE London, service delivery variables only 216
Figure 52: local R-squared for stroke admission rates, South Coast, population factors only 217
Figure 53: local R-squared for stroke admission rates, South Coast, service delivery variables only 218
Figure 54: Percent of patients with smoking status recorded in last 12 months 230
Figure 55: proportion of smokers by age and sex, comparison of HSfE and QRESEARCH data 231

4.4 Table of abbreviations

ADQ/STAR PU	average daily quantity of statins per standardised population
AF	atrial fibrillation
AIC	Akaike Information Criterion
APHO	Association of Public Health Observatories
AUROC/AUC	area under Receiver Operating Characteristics curve
BMI	body mass index
BRFSS	Behavioural Risk Factor Surveillance Survey
BRHS	British Regional Heart Study
BTS	British Thoracic Society
CCC	concordance correlation coefficient
CfH	Connecting for Health
CHD	coronary heart disease
CI	confidence interval
CKD	chronic kidney disease
COPD	chronic obstructive pulmonary disease
CQC	Care Quality Commission
CVD	cardiovascular disease
DH	Department of Health
DsPH	Directors of Public Health
ECG	electrocardiogram, electrocardiographic
FCE	Finished Consultant Episode
FEV_1	forced expiratory volume in 1 second
FPG	fasting plasma glucose
FVC	forced vital capacity
GHQ	General Health Questionnaire
GP	general practitioner
GPES	GP Extraction Service
GPRD	General Practice Research Database
HDL	high density lipoprotein
HES	Hospital Episode Statistics
HRG	Healthcare Resource Group
HSfE	Health Survey for England
HUM	High impact User Model
IC	Information Centre for health and social care
ICD	International Classification of Disease
ICER	incremental cost effectiveness ratio
IGT	impaired glucose tolerance
IHME	Institute for Health Metrics and Evaluation
IMD	Index of Multiple Deprivation
IRR	incidence rate ratio
JBS	Joint British Societies
JSNA	Joint Strategic Needs Assessment
LA	local authority
LR	likelihood ratio

LSOA	Lower Super Output Area
MI	Mental illness
MI	myocardial infarction
MLM	Multiple logistic regression modelling
MONICA	Multinational MONItoring of trends and determinants in CArdiovascular disease
MSOA	Medium Super Output Area
NatCen	National Centre for Social Research
NHANES	US National Health and Nutrition Examination Survey
NHS	National Health Service
NICE	National Institute for Clinical Excellence
NS-SeC	National Statistics Socioeconomic Class
NSTS	National Strategic Tracing Service
OLS	ordinary least squares
ONS	Office for National Statistics
OR	odds ratio
PAR	population attributable risk
PARR	Patients At Risk of Readmission
PCPH	Department of Primary Care & Public Health
PCSCs	primary care sensitive conditions
PCT	primary care trust
QOF	Quality and Outcomes Framework
RMSE	root mean squared error
ROC	receiver operating characteristics
SBP/DBP	systolic/diastolic blood pressure
SE	Synthetic estimate(s), estimation
SVM	Support vector machine
UK	United Kingdom
UNAIDS	Joint United Nations Programme on HIV/AIDS
VIF	variance inflation factor
WHO	World Health Organization

5 Background/policy context

"A public health department really ought to be looking at doing our QOF analysis and saying actually 'Do you know, you are two per cent lower than the practice next door on thyroid problems. Let's see why that is.' They don't do that. They don't look at my practice… I want to see a public health department that is actively engaged with the practices, that is linked to the practices to say 'OK, we have got a named link for you. You have got a public health-type issue, you come to this guy and he will be the conduit through which the rest of the department will work with you.'

(General practitioner, London)



A major problem with present public health information is the lack of population-based data on key metrics such as risk factor and disease prevalence at local i.e. primary care trust (PCT) and local authority (LA) level- the Health Survey for England (HSfE) produces reliable data only to Regional level.[1] This issue is highlighted in *Securing Good Health for the Whole Population* (the Wanless Report)[2], which states:

> *"Health data are essential for monitoring the health of the population and for evaluating the effects of health interventions. Yet the information collected nationally is often poor and there is no regular mechanism by which a PCT or LA can gather reliable information on its own population… To improve understanding of prevalence of disease and to enable proactive management of personal risk factors, much greater use needs to be made of primary care data systems".*

This challenge was taken up by *Informing Healthier Choices*, the Department of Health (DH) public health information and intelligence strategy,[3] which states that:

- *"reliable data on key contemporary health challenges will be collected and made available centrally as a by-product of care provision (for example information from general practice systems on hypertension or smoking prevalence)*

- *prevalence models will be generated for the common health problems which commissioners need to address. These will allow the current situation in an area or group within the population to be evaluated against an expected level of need".*

My thesis covers the research needed to support these objectives, i.e. to improve the estimation of risk factor and disease prevalence in small areas, and to determine the extent to which primary care data can be used to support estimation.

6 Aim, Objectives and Management

6.1 Aims

My PhD proposal initially contained the two broad aims of developing methods for prevalence estimation, of both disease risk factors and common chronic diseases. The first research aim therefore proposed to assess the fitness of general practice (GP) data for the purpose of risk factor prevalence estimation, and to develop new methodologies to adjust primary care data for sources of bias. In Phase 1 validation was to be undertaken against other estimates, and appropriate adjustment designed. In Phase 2, new uses of this small area/population data was to have been explored, e.g. linkage with datasets including broader health determinants, and modelling future disease prevalence.

However this aim relied heavily upon the availability of data from the GP Extraction Service (GPES), which was to have begun to provide data by late 2008, and then by Spring 2009.[4] This later deadline was missed- and in fact, there was still no data available at the time of writing this thesis. When it became apparent that no data would be available during the course of the degree, I agreed with my Supervisors that the thesis would focus on the second aim, disease prevalence estimation. The remainder of the thesis therefore excludes this aim. However, I retain a close interest in GPES development and presented on this topic at a joint workshop organised by the Faculty of Public Health, NHS Connecting for Health (NHS CfH) and the Information Centre for health and social care (IC) entitled *"The Future of Public Health Information and Intelligence: needs and applications."*[5]

This second aim has proven to be ambitious in scope in itself. For chronic disease prevalence estimation, Phase 1 of the research aimed to further develop the methodology for new and existing prevalence models, and Phase 2 to validate one or more models and develop case-finding strategies, and to explore the impact of undiagnosed disease as estimated by the models.

6.2 Research Questions

- How should the prevalence of chronic diseases be estimated?
- How valid are chronic disease prevalence estimates?
- How do different chronic disease prevalence estimation methods compare in terms of their validity?
- What is the best available methodology given the requirements for prevalence models?
- How useful are chronic disease prevalence estimates for case-finding at GP practice level?
- Can chronic disease prevalence estimates be used with practice disease register prevalence to estimate the prevalence of undiagnosed disease?

6.3 Objectives

For chronic disease prevalence estimation:

- to develop a methodology for chronic disease prevalence modelling, initially by extending logistic regression modelling to examine predictors of chronic obstructive pulmonary disease (COPD), coronary heart disease (CHD), stroke, hypertension, overall cardiovascular disease (CVD) and high CVD risk using the HSfE datasets
- to work with the Association of Public Health Observatories to publish new prevalence estimates and future projections at PCT and GP level and support their use by PCTs, practices and other agencies
- to explore the use of risk factor prevalence to model future disease prevalence
- to investigate the potential for applying Bayesian procedures as developed for HIV/AIDS to other chronic diseases.
- to undertake a validation of one or more disease prevalence models, possibly through testing case-finding strategies at practice level. This could involve data from Brent, Wandsworth, or Hammersmith & Fulham PCTs
- to explore the links between registered and estimated disease prevalence and primary and secondary healthcare utilisation, with the aim of developing age-sex specific utilisation ratios which can be used to project future healthcare capacity requirements.

6.4 Ethics

As the PhD involved secondary analysis of existing data sources (apart from the pilot of COPD screening), ethical issues were dealt with in gaining access to the data for analytic purposes. For example for Hospital Episode Statistics (HES) data, we complied with small cell number requirements for sensitive data items.[6]

6.5 Outputs

For Phase 1, a methodology paper was to be published for each disease prevalence model, and a paper providing results for English PCTs (including maps). Later publications would include validation studies and case-finding strategy results.

For Phase 2, publications were to explore the links between disease prevalence, health determinants and healthcare factors, ideally through fuzzy matching of anonymised person-based data, but if not possible through practice-level analyses. Current person-based ONS survey data is available via the UK Data Archive. It was not clear at the outset whether the same level of access by researchers to person-based data will be permitted for other sources e.g. HES, but this would be an ideal eventual data linkage.

6.6 Benefits expected

The potential benefits of valid disease prevalence models have been outlined above.

6.7 Risks

Risks and management strategies identified at the outset are listed below.

Risk	Risk management
Inability to access person-based data (if needed) from QRESEARCH risk factor extract because of cost	Consider application to DH or to a research funder e.g. Research for Patient Benefit
Lack of suitable local survey data for validation	Use other data sources e.g. Census
Delays in establishment of GPES	This is a major risk, and some implementation delays are possible. In that case it may be possible to obtain some data from current PCT IT service providers e.g. Brent, Wandsworth
Delays in researcher access to GPES data for ethical or administrative reasons	Aim to obtain aggregate data in the interim, possibly from routine early reporting
Low uptake/coverage of GPES	This will affect the level of geography for which the estimates can be produced, but not the integrity of the project so long as robust PCT/LA level estimates can be produced.
Inability to access person-based data from various data sources	Use aggregate data provided it has reasonable age-sex breakdowns
Currently unbudgetable costs for access to GPES, which could be significant	Work with Information Centre to minimise delays; ensure that researcher access to GPES is considered early
Complexity of statistical issues involved e.g. in imputation, and lack of resource	Work with BIAS Project in EPHPC; contract in some expertise; collaborate with national expert e.g. Patrick Royston at Cambridge
Interest/competition from other academic institutions or agencies which do not wish to collaborate	Ensure that there is sufficient expertise within ICFM; ensure that other researchers' needs are addressed
Information Centre refuses to collaborate	Unlikely given past experience.
No 2008-9 DH funding for disease prevalence modelling	Provide input to 2008-9 planning process in DH. Ensure that current models are used frequently
No funding for prevalence model validation or case-finding strategies	Apply to multiple funding sources e.g. NHS Research for Patient Benefit stream, relevant voluntary organisations.

7 Disease prevalence models: background

Parts of this section have been published as a paper in the Journal of Ambulatory Care Management.[7]

Like healthcare systems in other countries, the United Kingdom's NHS is moving from reactive to proactive care, leading to the establishment of programs for the management of chronic diseases. Better management of chronic diseases has consequently been an NHS priority, and there were three chronic disease-related national 2007–2008 Local Delivery Plan targets for English PCTs, the bodies responsible for administering the contracts of GPs in England and for commissioning other healthcare services. These were to reduce emergency bed days, to increase the number of staff in the community providing case management, and to increase the number of patients with chronic diseases and those who are very high intensity users, whose case is being managed by a specialist community-based nurse.

Until now, most attention has been focused on the relatively small number of patients with a high disease burden who use about 30% of healthcare resources. Various tools have been provided to identify these patients, such as the Patients At Risk of Readmission (PARR) case-finding algorithm that uses Admitted Patient Care data to identify patients at high risk of hospitalization,[8] and the High-impact User Model (HUM) developed by Dr Foster Intelligence with support from Imperial College London.[9] These tools use data collected by the NHS to monitor hospital activity, principally HES data.

Other policies that cover the larger group of patients with a single high-impact chronic disease and management of this disease once it has been identified include the National Service Frameworks and the Quality and Outcomes Framework (QOF) of the new General GP Contract, which rewards GPs for improving chronic disease management using data coded into electronic patient records. A great deal remains to be done by PCTs to ensure that chronic disease management programs are functioning optimally, and to prevent patients with chronic diseases moving inexorably into the high disease burden group.

Less attention has been directed toward detecting patients with undiagnosed chronic diseases. Since diagnosis and treatment usually slows disease progression, such patients may be more likely than those already known to general practices to increase the burden of disease and healthcare costs. Moreover, although some admissions for chronic diseases are of high disease burden individuals or those with comorbidities, the majority are of patients in the middle of the "chronic disease pyramid." Recent BMJ and Lancet news items[10] have featured a study comparing disease prevalence and mortality in the US and UK.[11] This found higher disease prevalence among older Americans than their British counterparts. In contrast, age-specific mortality rates are similar in the two countries, with an even higher risk among the English after age 65. The study offers two possible explanations for the higher death rate among the English after the age of 65: that the illnesses looked at in the study resulted in higher mortality in England; or that in England the diseases are diagnosed at a later stage. Both of these explanations, says the report, "imply higher-quality medical

treatment (broadly defined) in the United States than in England, at least in the sense that mortality implications of disease are less severe in the United States."

I first became interested in modelling disease prevalence around 2001 while I was a Director of Public Health for Ealing Hammersmith & Hounslow Health Authority in West London. With its large South Asian population, the prevalence of diabetes and CVD was relatively high. For example, according to the HSfE, the prevalence of diabetes in the English population is 4.9 per cent, but local data I obtained from practice registers showed a prevalence of three per cent. It would clearly be useful to have better local estimates of disease prevalence.

I discussed development of a diabetes prevalence model with Dr Nita Forouhi, a Specialist Registrar in Public Health, and she developed a simple model which included age, sex, ethnicity and obesity prevalence. We discussed the model with Yorkshire and Humber Public Health Observatory (Y&HPHO) and with Sue Roberts, then National Clinical Director for diabetes at the Department of Health (DH). At that stage, data from a diabetes register was being considered as a source of prevalence data. However the advantages of an epidemiologic approach to prevalence estimation were accepted, and Y&HPHO were asked to host the model on their website. This model, with some modifications, has been maintained there since 2002.[12]

Disease prevalence information is also important for both public health monitoring and clinical management. Since diagnosis and treatment usually slows disease progression, patients with undiagnosed long term conditions (chronic diseases) may be more likely than those already known to general practices to increase the burden of disease and healthcare costs.

The QOF has provided for the first time UK population-based data on the detected (defined as having an appropriate diagnosis entered on GP electronic medical record systems according to QOF coding rules) prevalence of disease. It also rewards practices for finding new cases through payments for quality of care that are weighted for the prevalence of disease in a practice i.e. practices with higher prevalence of a chronic disease are paid more than practices with a low prevalence. Recent QOF results show generally small increases in detected prevalence since 2005-6.[13]

7.1 Potential uses of disease prevalence estimation

But how complete are QOF disease registers? Prevalence models, which provide survey-based estimates for small populations, based on local data on known risk factors, may be useful in this regard. They can be used to:

- assess the completeness of disease registers in primary care or assessing the completeness of case finding- the data could be used to guide case-finding at practice level;
- compare complication rates or admission rates after adjustment for variation in expected prevalence;
- compare service provision with population need- for example, if age-sex specific primary and secondary care utilisation ratios are known for specific chronic diseases, these can be applied to population projections to estimate future healthcare

capacity requirements. We used this modelling methodology in Healthcare for London[14];

- undertake health equity audits;
- allocate health service resources;
- link to risk factor prevalence estimates.[7]

7.2 Previous UK prevalence models

The first national UK disease prevalence model for which local data was made available was the "PBS" (Public Health Observatory/Brent/Sheffield) Diabetes Prevalence Model, now hosted by Yorkshire and the Humber Public Health Observatory. The PBS model applies age, sex, and ethnic group-specific estimates of both type 1 and 2 diabetes prevalence rates, derived from epidemiological population studies, to 2001 Census resident populations.[15] Forecasts of 2010 diabetes prevalence are also presented for sub-national areas based on projected population change and trends in obesity. The PBS model estimates that diabetes prevalence in England was 4.37% overall in 2001, and that it will increase to 5.05% in 2010, assuming the prevalence of obesity continues to increase in line with historic trends.

However, 2006–2007 QOF results show an England crude prevalence rate (using the entire population as a denominator as does PBS) of only 3.7%, when the estimated prevalence is about 4.8%.[16] The National Clinical Audit Support Programme's National Diabetes Audit Report for the period 2005–2006, which covered 43% of PCTs, shows that on average for the participating areas, 81.5% of the people predicted to have diabetes by the PBS model are actually recorded as having the disease, which has increased from the 2003–2004 figure of 77%.[17] There is considerable local variation, with the percentage of predicted registered varying across the then 28 English Strategic Health Authorities from 72% to 95%. There is even more variation between general practices, with some recording only of the 15% expected prevalence of diabetes.

Prevalence estimates can be produced for populations registered or enrolled for healthcare, or for resident, geographically based populations (for the former, geographically based risk factor data may need to be applied to the registered/enrolled population). Models are being developed for other chronic diseases, and they can also be produced for other risk factors as well as for hypertension: modelled local estimates have also been developed recently for smoking, excessive alcohol intake, obesity, and fruit and vegetable consumption by the Information Centre for Health and Social Care and the Neighbourhood Statistics Service of the UK Office for National Statistics.[18] These can be used to target area-based public health interventions. The Health Protection Agency for England has developed a sophisticated model for HIV/AIDS prevalence, which uses Bayesian methods to allow multiple evidence sources to be applied to any one parameter,[19]but this has not been applied to sub-national data. The Joint United Nations Programme on HIV/AIDS (UNAIDS) has also developed an HIV model for developing countries as part of its Estimation and Projection Package.[20] Further methodological development is considered below.

Case-finding to close disease prevalence gaps may be cost-effective—the 2003 National Institute for Clinical Excellence (NICE) COPD guidance states that the incremental cost-effectiveness ratio of opportunistic case finding for this purpose is a cost per quality-adjusted life year of £814.[21] But although case finding may be justifiable, population

screening may not. Gillies and co-workers have recently modelled the cost-effectiveness of several diabetes sub-population screening strategies.[22] The overall cost-effectiveness ratio was not sensitive to decisions about which groups to screen, nor the costs of screening or treatment, but was most strongly affected by assumptions about how treatments combine to reduce risk.

This is also likely to be the case for other chronic diseases. Given the large populations involved, more systematic case finding or screening will also not be cheap. There is evidence that the rise in treated disease prevalence, rather than the rise in spending per treated case, was the most important determinant of the recent growth in US healthcare spending.[23] Similarly, reductions in diagnostic or treatment thresholds—the recent Joint British Societies' recommendation to use lower cholesterol treatment targets is but one example—need to be properly justified in terms of their cost effectiveness before they are adopted.[24]

Further work is therefore needed to develop and test systematic case-finding strategies that produce a high yield but do not overburden general practice staff. Moreover, none of the models have yet been validated at general practice level against, for example, numbers of cases detected by case-finding initiatives. There must also be high-quality primary care services in place to manage new cases as they are identified. However, in many deprived areas in the United Kingdom, there are fewer than expected primary care services: a 2004 review shows that the broadly upward trend in GP misdistribution in the United Kingdom from the mid- 1980s continued in the period 1995–2003.[25] Valid prevalence models could be used to allocate primary care resources so that these better match actual need and further support case finding, and work continues to try to achieve this objective.

7.3 Problems with current UK prevalence models

The current models require validation, and cost-effective case-finding strategies are also needed. The relative disease burdens and healthcare impacts of diagnosed and undiagnosed cases of chronic diseases are unknown. Further methodological development is also required. The diabetes, CHD and hypertension models use crude weighting of population data- for example, the HSfE shows that hypertension prevalence is correlated with age, sex and ethnic-group, so calculations were stratified to reflect variations in these factors at PCT level. For the COPD model, the strength of association between each explanatory variable and COPD caseness was used to calculate the relative odds, which were applied to the baseline odds to derive the prevalence estimates. HIV/AIDS models developed by the Health Protection Agency and the Joint UN Programme on HIV/AIDS use Bayesian procedures for prevalence estimation,[19] which could be applied to other chronic diseases. Finally, there is a need for additional prevalence models to cover other QOF (and non-QOF) diseases where there is evidence that early diagnosis improves outcomes.

7.4 Small area prevalence estimation methodologies

The main procedures include direct domain estimation, indirect domain estimation, and small area modelling. Direct domain estimation uses available sample units in the domain to estimate the quantity of interest, leading to unacceptably large standard are related in time and/or space to increase the effective sample size for small domains. Given these problems I will not discuss it further.

Indirect domain estimation includes: synthetic estimation, i.e. using a reliable estimator for a large domain to derive an estimator for the small domain contained within the large domain under the assumption that the small domain has the same characteristics as the large domain; composite estimators, i.e. weighted averages of sample estimates for the same domain but from different surveys); and James-Stein estimators (also called shrinkage estimators because they shrink the mean squared error, sometimes also used in conjunction with the direct estimator in so-called "limited translation estimators").

In contrast to indirect domain estimation, small area modelling is explicit about the assumptions of relatedness in space and/or time and has variably used three strategies to deal with the limited availability of survey and administrative data: pooling data over several years; borrowing strength in space by exploiting spatial correlations; and using structured relationships with covariates to predict the quantity of interest. I used the latter method in this research, but they can be combined.

7.4.1 Simple multiple logistic regression modelling

Multiple logistic regression modelling (MLM) using "classical" regression techniques is an obvious method for local prevalence estimation. However, apart from the modelling described in this PhD, there have been surprisingly few published reports or applications based on MLM. There are some reports from the USA using National Health and Nutrition Survey (NHANES) data.[26;26;27] Other researchers have used more sophisticated methods which are described below.

7.4.2 "Synthetic estimation"

Synthetic estimates (SE) e.g. on the Neighbourhood Statistics website attempt to overcome some of the problems associated with local models by using auxiliary information i.e. more than one data source to produce small area estimates, without or usually with MLM.[18;28;29] As Chambers and Saie[30] explain, auxiliary information may be attributes of surveyed individuals (e.g. age/sex/NS-SeC and ethnicity) and/or of the area in which they reside (e.g. claimant counts or area-based deprivation scores). Estimation and inference approaches based on using the relationship between the variable of interest and auxiliary information are called "model-based SE".

The form of the statistical model used for model-based SE depends on both the survey sampling and availability of auxiliary/covariate information e.g. Census data. As a first stage, regression analysis is performed by modelling the survey data against available predictors. This analysis is conducted for the subset of areas covered by the survey. The output from the first stage is a set of parameter estimates.

At the second stage, the coefficients attached to the predictor variables obtained from the first stage model are attached to the identical set of variables available for all areas to produce an estimate for the area. An alternative technique of multilevel modelling applies a statistical model to survey data that simultaneously accounts for both individual and area-level influences. Prof Nicky Best's team in the Imperial College School of Public Health have refined these methods,[31;32] and they are now in widespread use[33] (see also the BIAS Project website). Composite estimators consist of weighted combinations of direct and synthetic estimators. However it is still necessary to have a sample or direct estimate in the area

concerned at the desired level of geographic concentration i.e. a geographically unclustered survey. Table 1 summarises the data requirements/constraints which enable or prevent the use of alternative small area estimation techniques.

Table 1: Summary of small area synthetic estimation methods

Estimation approach	Data system required to support approach		Level at which model is fitted	Smallest estimation area when applied in the UK
	Sample design	*Covariate availability*		
Direct estimator	Unclustered	-	-	NUTS 4/5
GREG estimator/GREG type EBLUP	Unclustered (1)	Both sample members and all areas	Individual	ditto
Area-level synthetic estimator	Clustered	All areas	Area	ditto
Composite (EBLUP) estimator (1)	Unclustered	All areas	Area	LA
Regression synthetic estimation using area-level covariates (2)	Clustered	All areas	Individual	Ward/LSOA
Regression using individual and area-level covariates (3)	Clustered	Both sample members and all areas	Individual	Ward/LSOA

1 E.g. ONS estimates of ILO unemployment rates.
2 E.g. ONS estimates of average income, NatCen SE I healthy lifestyles.
3 E.g. University of Portsmouth for Health Development Agency.

It is clear from the above that there are a number of methodological problems with SE relating to survey sampling techniques and covariate availability. In addition, with NatCen's current project to produce SEs of risk factors at LA and PCO levels, there are now problems with the detailed cross tabulations of 2001 Census data required becoming outdated. In areas with rapidly growing or changing populations, this could lead to significant errors. Detailed cross-tabulations of Census population projections are thought to be unreliable.

The very wide confidence intervals at small area level are alluded to above. SEs are also inexact and cannot be used for performance monitoring, as they take no account of the impact of local inter-Censal interventions such as NHS stop smoking or primary care services, or regeneration or Neighbourhood Renewal schemes. They are difficult to explain to non-statisticians and require statistical health warnings to prevent over-interpretation and misuse. Finally, they will underestimate local socio-cultural factors such as different health behaviours amongst ethnic minority sub-groups, where data has had to be pooled.

7.4.3 Markov models

Markov modelling is a technique which has been used widely in pharmaco-economics, but also to develop incidence and prevalence models as estimate transitions between various

disease states.[34] For example, Honeycutt et al developed a Markov model to forecast diabetes prevalence in the USA.[35] This framework was used to generate forecasts by age, race and ethnicity, and sex. The model forecasts the number of individuals in each of three states (diagnosed with diabetes, not diagnosed with diabetes, and death) in each year using inputs of estimated diagnosed diabetes prevalence and incidence; the relative risk of mortality from diabetes compared with no diabetes; and estimates of current population, live births, net migration, and the mortality rate of the general population. Other Markov models have been produced for CVD,[36-38] depression,[39] and hypertension.[40] A very useful analysis has compared a Markov model with an agent-based discrete event simulation model.[41]

However the complexity of combining multiple disease states with local population data has limited current models to whole country or large population applications.

7.4.4 Bayesian approaches

Professor Peter Congdon has developed a number of sophisticated prevalence models using Bayesian and multi-level modelling techniques.[42,43] Some of these also use auxiliary information as for SE techniques to refine local estimates. He has used a regression model of prevalence rates by age, sex, region and ethnicity from the 1999 and 2003 HSfE to estimate CHD prevalence for 354 English LA areas.[44] To allow for the impact of social factors on prevalence, survey information on the deprivation quintile in the respondents' micro-area of residence was also used. Allowance was made for area CHD mortality rates (obtained from aggregated vital statistics data) which are positively correlated with, and hence a proxy for, CHD prevalence rates. An application involved assessment of surgical intervention rates in relation to prevalence at the level of 28 Strategic Health Authorities as they then were.

He has used two national mental illness (MI) studies that provide estimates of relative morbidity risk by demographic, socio-economic and ethnic group for major psychiatric conditions to produce an MI model;[45] household/marital and area status also figure in the regression. Relative risk estimates were used, along with suitably disaggregated census populations, to make mental illness prevalence estimates for LAs. Two applications were considered: the first involved analysis of variations in schizophrenia referrals and suicide mortality over English LAs that takes account of prevalence differences, and the second involved assessing hospital referral and bed use in relation to prevalence (for ages 16–74) for a case study area.

He has also estimated diabetes prevalence[46] using a regression analysis of prevalence rates according to age, sex and ethnicity from HSfE (together with Census) data for LAs and 8000 smaller areas (electoral wards).[45] An adjustment for social factors was based on a prevalence gradient over area-deprivation quintiles. A Bayesian estimation approach was used allowing simple inclusion of evidence on prevalence from other or historical sources. He used the prevalence estimates to assess variations between LAs in adverse hospitalization indicators for diabetics and to assess the relationship between diabetes-related mortality and prevalence. Rates of diabetic ketoacidosis and coma were positively correlated with prevalence, while diabetic amputation rates are not.

There are two main problems with this otherwise excellent work. Firstly, the estimates are all based on resident populations. However prevalence estimates are most useful for GP registered populations where they can be compared with QOF registered prevalence. Secondly, to enhance prediction, these models use variables such as QOF prevalence, hospital admissions or mortality to combine with survey data. This obviously compromises their use for primary care performance assessment. As a result they have remained in the socio-geographical literature as interesting mathematical models with no practical implementation.

7.4.5 Support vector machine modelling

Support Vector Machine (SVM) is a supervised machine learning technique that is widely used in pattern recognition and classification problems. Since an SVM is a classifier, then given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that predicts whether a new example falls into one category or the other. Intuitively, an SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. The SVM algorithm performs a classification by constructing a multidimensional hyperplane that optimally discriminates between two classes by maximizing the margin between two data clusters. This algorithm achieves high discriminative power by using special nonlinear functions called kernels to transform the input space into a multidimensional space.[47]

The basic idea behind the SVM technique is to construct an n-1 dimensional separating hyperplane to discriminate two classes in an n-dimensional space. A data point is viewed as an n-dimensional vector. For example, two variables in a dataset will create a two-dimensional space; the separating hyperplane would be a straight line (one dimensional) dividing the space in half. When more dimensions are involved, SVM searches for an optimal separating hyperplane called the maximum-margin separating hyperplane. The distance between the hyperplane and the nearest data point on each side (called support vectors) is maximized. The best scenario is that two classes are separated by a linear hyperplane.

SVM, which has been previously used in bioinformatics, has now been applied to disease prevalence estimation, to classify persons with and without common diseases.[48] The method was used to detect persons with diabetes and pre-diabetes in the 1999-2004 National Health and Nutrition Examination Survey (NHANES) to develop and validate two SVM models: for diagnosed or undiagnosed diabetes vs. pre-diabetes or no diabetes; and undiagnosed diabetes or pre-diabetes vs. no diabetes. The SVM models were used to select sets of variables that would yield the best classification of individuals into these diabetes categories. The discriminative abilities of the SVM models for Classification Schemes I and II, according to the area under the receiver operating characteristic (ROC) curve, were 83.5% and 73.2%, respectively. Validation indicated that the discriminative powers of the two SVM models were comparable to those of the multivariable logistic regression method. However, while SVM appears promising for prevalence estimation, it has not been widely validated, nor has it been shown to be superior to multiple logistic regression methods.

7.4.6 Simulation models

Agent-based or "discrete-event" simulation modelling has developed rapidly in recent years, partly because of the availability of generic software which automates the process.[49-52] A system dynamics model consists of an interlocking set of differential and algebraic equations developed from a broad spectrum of relevant measured and experiential data. A completed model may contain scores or hundreds of such equations along with the appropriate numerical inputs. Modelling is an iterative process of scope selection, hypothesis generation, causal diagramming, quantification, reliability testing, and policy analysis.

For example, in the USA, the National Centers for Disease Control and Prevention used system dynamics simulation modelling to gain a better understanding of diabetes population dynamics and to explore implications for public health strategy.[53] A model was developed to explain the growth of diabetes since 1980 and portray possible futures through to 2050. The model simulations suggest characteristic dynamics of the diabetes population, including unintended increases in diabetes prevalence due to diabetes control, the inability of diabetes control efforts alone to reduce diabetes-related deaths in the long term, and significant delays between primary prevention efforts and downstream improvements in diabetes outcomes. A recent paper has projected diabetes prevalence to 2050.[54] The authors project that annual diagnosed diabetes incidence will increase from about 8 cases per 1,000 in 2008 to about 15 in 2050. Assuming low incidence and relatively high diabetes mortality, total diabetes prevalence (diagnosed and undiagnosed cases) is projected to increase from 14% in 2010 to 21% of the US adult population by 2050. However, if recent increases in diabetes incidence continue and diabetes mortality is relatively low, prevalence will increase to 33% by 2050. The agent-based approach is being taken to its logical conclusion in the USA by Archimedes, a vast simulation model of several diseases in several organ systems modulated by thousands of structural equations.[26;55-57].

However while these agent-based models are valuable tools for high-level policy analysis or population wide intervention, they do not provide support at the local level, where population factors may completely change incidence and prevalence. Moreover in the UK, perhaps because of British caution and suspicion of "big ideas", it appears that the development and application of public health modelling will occur in a classically incremental fashion. For example, the UK Department of Health funded a discrete event simulation model of CHD prevention and treatment, but this was never validated or used for policymaking and soon became outdated.[58] With the disease prevalence models described here I aimed to provide a useful local resource which would raise awareness of modelling methodologies and, through their transparency, avoid raising suspicions of academic black boxes.

8 Disease prevalence model development

8.1 Chronic Obstructive Pulmonary Disease Model

This model was developed with Dr Luis Nacul, while he was a Specialist Registrar in public health while I was at NW London Strategic Health Authority. We later went on to document the methodology in a publication in *Population Health Metrics* about the time I started my PhD.[59]

8.1.1 Introduction

COPD is a chronic condition characterised by progressive airflow obstruction, which is not completely reversible. COPD contributes to nearly 30,000 deaths each year in the United Kingdom (UK), corresponding to 5.7 percent of adult male and 4 percent of adult female deaths, including a significant number of premature deaths. In addition, 1.4% of the population consult their GPs for COPD each year. It accounts for 2% of hospital admission spells and over 3 percent of bed-days in adults,[60] costing the NHS £800 million, and leading to 24 million working days lost each year. [61]

Respiratory function indices have been shown to be predictive of mortality from respiratory disease, cardiovascular disease and all causes.[62] Airflow limitation may precede the development of significant symptoms of COPD by many years and its progression is directly linked to the continuing exposure to risk factors, particularly tobacco smoking. As COPD is difficult to diagnose clinically (without spirometry) in its milder forms, it is often diagnosed late - the average age at diagnosis of COPD in the UK is 67 years.[63] Widespread use of spirometry allowing early detection of airflow obstruction has been increasingly advocated as it enables early management of COPD.[64]

As expected, the prevalence of COPD is higher in smokers and in men, and it increases with age. Other risk factors of public health importance include air pollution, socio-economic deprivation,[65] occupational exposures[66] and possibly ethnicity.[67,68] Stopping smoking prevents the development of COPD, or slows its progress and reduces the risk of hospital admissions.[69] Smoking cessation programmes are highly cost-effective, and crucially, have been specifically shown to be cost-effective when directed to individuals with asymptomatic airway obstruction.[70] This is because smokers may be motivated to attempt to quit when given a diagnosis of airflow limitation.[71]

The incremental cost effectiveness ratio (ICER) of opportunistic COPD case-finding for this purpose is a cost per life year gained of £713.16 and a cost per QALY of £814.56.[60] The magnitude of undiagnosed cases can be ascertained by comparing the model estimates with the recorded prevalence of COPD, to indicate the extent of unmet needs in COPD. In the UK this is facilitated by GP performance payments for COPD management through the QOF of the GP Contract based on an electronic register of all patients with diagnosed COPD. If this is linked to case finding and intervention, there is a potential for reducing the population burden and progression of the disease.

There is considerable variation in the reported prevalence of COPD. Models using smoking rates to estimate COPD prevalence have been previously proposed, but none has direct

relevance to the UK.[72-75] With colleagues I developed a model to estimate the prevalence of COPD based on existing data from the HSfE 2001,[76] which had not been used for this purpose before. It uses the main risk factors for COPD reported in the literature, particularly those that are easily measured and for which information is readily available. This report explains how the model was developed, and uses the population of England to illustrate its application.

8.1.2 Methods

Data sources

The distribution of COPD in the population of England was based on the HSfE 2001 findings for lung function parameters and their association with relevant risk factors. The methods used in the HSfE 2001 are described in detail elsewhere.[76] In brief, the survey included a representative sample of the population who had their lung function evaluated using a portable spirometer with a calibration device (Vitalograph 'Escort Spirometer'). Comprehensive data on risk factors were also collected as part of the survey. The data refer to 5,269 men and 6,133 women over 15 years old with valid lung function measures. This corresponds to 98% of men and 95% of women visited by a research nurse as part of the survey. For 99.3% of these, data were available for all of age-group, smoking status, ethnicity and degree of urbanisation. Data were available for deprivation score and all of the above risk factors in 94.3% of the final sample, which was used for the multivariate analysis. This included 4,970 men and 5779 women. COPD was defined according to British Thoracic Society (BTS) criteria,[77] based on the values of forced expiratory volume in 1 second (FEV1) and the forced vital capacity (FVC) i.e. FEV1/ FVC < 0.70 and FEV1<80% predicted using British reference values derived from the HSfE.[78] The second criterion is not part of the international GOLD definition. The GOLD definition has also been used for prevalence estimates by NICE and the COPD National Strategy. As a result, estimates from these sources are larger.

Model construction

The choice of variables for inclusion in the model was based on logistic regression analysis that examined predictors of COPD using the HSfE 2001 dataset. Explanatory variables obtained from the HSfE dataset and included in the final model were (categories of each variable shown in brackets): gender, age group (15-34 year olds, 10 year age groups from 35 up to 74 years of age and 75 year olds and over), smoking status (smoker, former smoker, never smoker), ethnicity (White, Black or Black British, and Asian or Asian British), area of residence (rural, urban and suburban) and area based index of deprivation (quintile of deprivation score based on Index of Multiple Deprivation).[79] The *baseline odds* of COPD (in non-smokers under 35 years old) were obtained directly from the data. Separate baseline odds were estimated for each gender, and also according to ethnicity, area of residence and area-based deprivation score. The logistic model was used to derive the *odds ratios and prevalence ratios* for COPD for subjects with various combinations of risk factors in relation to baseline. The prevalence in each age group, gender, ethnic group, area of residence and level of deprivation, and smoking status category were derived from the odds, using the formula:

$$prevalence = odds/(1+odds).$$

Model application

The *input variables*, which can be defined by the relevant user, e.g. at a PCT, include age-group, gender, smoking prevalence by gender and age, area of residence, area based deprivation score and ethnic distribution of the population. The main *model outputs* are the prevalence of COPD by gender for the relevant geographic area, at national, regional or local level, as defined by the user. To illustrate the use of the model, we have used here population inputs for England, based on the mid-2005 estimated population distribution[80] and the national smoking prevalence by age-group and gender for 2004-2005.[81]

Model assumptions

The model assumptions include: i) The real prevalence of COPD in non-smokers under 35 years of age (baseline prevalence) is the same as the prevalence in non- smokers of the same age group and gender in the 2001 HSfE population; ii) The ratio of odds and prevalence of COPD in the various age groups compared to the baseline group is the same as in the HSfE for each gender, smoking status and other risk factors in the model; iii) the risks in those falling within each of the risk categories are uniform. We also obtained the prevalence of COPD considering alternative scenarios, which assume that: a) the prevalence of COPD in under 35s or under 40s is uniform across smoking status in each gender, and is equal to the average baseline prevalence found at the HSfE (therefore it does not consider any increase in risk due to smoking in this age group); or b) the prevalence of COPD is zero in under 35s or in under 40s; or c) ethnicity does not have an effect on the risk of COPD, and the risks in white populations apply to all ethnic groups.

8.1.3 Results

Risk factors for COPD and selection of variables for COPD model

Table 2 shows the results of the univariate and final regression logistic models assessing risk factors for COPD. Risk of COPD is significantly lower in women than in men (odds ratio (OR) =0.64; 95% CI= 0.54 - 0.76; $p<0.001$). As gender was shown to modify the effect of other variables on the outcome, the analyses were carried out separately for men and women. No other significant interactions were found in the data, including between age-group and smoking status.

The final model shows age group and smoking history are the strongest predictors of COPD in both genders. Residence in urban areas and possibly black ethnicity are also associated with increased risk, particularly in women. Living in more deprived areas is associated with increased risk in men, but not in women. Being Asian appears to be protective in women, although this association did not reach statistical significance.

Model application

Table 3 shows the prevalence of COPD by age and gender in England. The overall prevalence in the population over 15 years of age was 3.1% (3.9% in men and 2.4% in women). For those over 45 years old, the estimated prevalence was 5.3% (6.8% and 3.9% in men and women respectively). This corresponds to over 1.3 million people in England with COPD, of whom nearly 800,000 or 60% are men. The assumption that ethnicity is not associated with being a case of COPD, i.e. that all population has the same risk of whites, did not change the

total national prevalence estimates considerably (1,297 thousand in 15 year-olds and over and 1,065 thousand over 45s under this assumption). When we considered the risk of COPD in under 35s as equal to the average baseline risk in this age group (in non- smokers), the total number of cases estimated was reduced by 60,800, resulting in an overall prevalence of 1.25 million or 3% (3.8% in men and 2.3% in women). Considering the risk in all over 40s as equal to the average baseline risk, the total number of cases is reduced further to 1,223,200 (3.0% overall prevalence; 3.7% in men and 2.3% in women). If we consider the risk in under 35s as negligible, the total number of estimated cases decreases to 1,185,700 (2.9% overall prevalence; 3.4% in men and 2.2% in women). Assuming the more extreme situation of zero risk in all under 40s, the prevalence comes down to 1,128,550 (2.7% overall; or 3.4% in men and 2.11% in women), representing a decrease of 174,500 compared to the original estimates.

Table 2: Risk factors for COPD included in the final prevalence model

Variable	Odds Ratio (95% CI) (univariate model)	Odds Ratio (95% CI) (final logistic model)
MEN (n=4970)		
Smoking Status	(*P*<0.001)	(*P*<0.001)
- Never smoker	1	1
- Former smoker	3.63 (2.54 - 5.21)	2.18 (1.48 – 3.23)
- Current Smoker	3.81 (2.64 - 5.52)	4.50 (3.01 – 6.74)
Age-group (in years of age)	(*P*<0.001)	(*P*< 0.001)
- <35	1	1
- 35-44	1.65 (0.86 - 3.17)	1.94 (0.99 – 3.78)
- 45-54	2.45 (1.33 - 4.50)	2.66 (1.41 – 4.99)
- 55-64	6.91 (4.02 - 11.89)	7.92 (4.46 – 14.07)
- 65-74	10.40 (6.08 - 17.80)	12.69 (7.12 – 22.60)
- 75+	12.15 (6.78 - 21.76)	16.02 (8.57 – 29.94)
Area of residence	(*P*=0.02)	(*P*=0.25)
- Urban	1	1
- Suburban	0.70 (0.50 – 0.97)	0.74 (0.51 – 1.06)
- Rural	0.58 (0.39 – 0.86)	0.72 (0.46 – 1.15)
Ethnicity	(*P*=0.64)	(*P*=0.95)
- White	1	1
- Black/ Black British	1.20 (0.48 – 2.99)	1.17 (0.44 - 3.10)
- Asian/Asian British	0.69 (0.28 – 1.70)	0.97 (0.37 – 2.51)
Quintile of Multiple Deprivation Score	1.22 (1.11 – 1.34) (*P*<0.001)	1.17 (1.05 – 1.31) (*P*=0.005)
WOMEN (n=5779)		
Smoking Status	(*P*<0.001)	(*P*<0.001)
- Never smoker	1	1
- Former smoker	1.70 (1.07 – 2.64)	1.26 (0.79 – 2.01)
- Current Smoker	3.53 (2.43 – 5.14)	4.11 (2.74 – 6.15)
Age-group (in years of age)	(*P*<0.001)	(*P*<0.001)
- <35	1	1
- 35-44	1.08 (0.53 - 2.19)	1.35 (0.65 – 2.79)

Variable	Odds Ratio (95% CI) (univariate model)	Odds Ratio (95% CI) (final logistic model)
- 45-54	2.34 (1.28 - 4.30)	2.69 (1.42 – 5.09)
- 55-64	4.18 (2.34 - 7.47)	6.03 (3.26 – 11.15)
- 65-74	5.36 (2.99 - 9.61)	8.04 (4.33 – 14.91)
- 75+	5.34 (2.79 - 10.22)	10.40 (5.18 – 20.87)
Area of residence	(*P*=0.01)	(*P*=0.04)
- Urban	1	1
- Suburban	0.62 (0.42 – 0.92)	0.59 (0.39 – 0.90)
- Rural	0.47 (0.29 – 0.77)	0.55 (0.32 – 0.96)
Ethnicity	(*P*=0.19)	(*P*=0.50)
- White	1	1
- Black/ Black British	1.14 (0.42 – 3.12)	1.81 (0.64 – 5.14)
- Asian/Asian British	0.24 (0.03 – 1.76)	0.60 (0.08 – 4.47)
Quintile of Multiple Deprivation Score	1.18 (1.05 – 1.34) (*P*=0.005)	1.11 (0.98 – 1.27) (*P*=0.11)

Table 3: Number and proportion of people estimated to have COPD by age group and gender in England (estimates for 2005)

Age-group (Years)	Men Number (%)*	Women Number (%)	Both sexes Number (%)
15-44	137,530 (1.30)	93,450 (0.89)	230,980 (1.10)
45-54	75,720 (2.38)	64,840 (2.00)	140,560(2.19)
55-64	198,400(6.90)	122,440 (4.11)	320,840 (5.48)
65-74	199,840(10.03)	105,740 (4.81)	305,580 (7.29)
75+	172,700(11.65)	132,400 (5.55)	305,100 (7.89)
Total 15+	784,190 (3.89)	518,870 (2.41)	1,303,060(3.15)
Total 45+	646,660 (6.76)	425,420 (3.92)	1,072,080(5.27)

** Values in brackets indicate age-gender specific prevalence rates (%) of COPD*

Table 4: Estimated prevalence P (percent) of COPD in England and according to area of residence

Area	P men age 15+	P men age 45+	P women 15+	P women 45+
Urban	4.86 (3.65-6.47)	8.41 (6.35-11.12)	3.68 (2.91-4.39)	5.95 (4.71-7.05)
Suburban	3.74 (2.75-4.94)	6.50 (4.79-8.55)	2.23 (1.79-2.69)	3.62 (2.91-4.36)
Rural	3.65 (2.75-4.85)	6.35 (4.80-8.40)	2.05 (1.67-2.53)	3.33 (2.72-4.10)
England	***3.89 (2.89-5.16)***	***6.76 (5.04-8.92)***	***2.41 (1.94-2.92)***	***3.92 (3.15-4.72)***

** Values in brackets correspond to mean values in extreme quintiles of deprivation score, or approximately the 10th and 90th percentiles of the prevalence distribution.*

The latter estimates assume that all cases of airflow obstruction in the younger age groups are due to other diagnoses than COPD, such as asthma. Table 4 shows the estimated prevalence of COPD in urban, suburban and rural England, based on the national population distribution and smoking prevalence. The values in brackets show the estimated average prevalence for areas in the lower and highest quintiles of deprivation. The average prevalence in over 45s varies 4-fold, with the highest values in men in deprived urban areas, and the lowest in women in wealthy rural areas. When the effect of ethnicity is also considered, the variation in prevalence reaches 7-fold, from 1.7% in Asian women from rural areas in the lower quintile of deprivation to 12.5% in black men from urban areas in the upper quintile of deprivation.

8.1.4 Applications & discussion

We developed a mathematical model that enables the prevalence of COPD to be estimated based on information that is easily available to PCTs. The model takes into account the increasing prevalence of COPD with age and smoking, and the modifying effect of gender. It also considers a higher risk of COPD among those living in urban environments and in areas of higher deprivation, and in black ethnicities.

As expected, the risk associated with smoking and former-smoking is higher in men compared to women. This may be related to their longer history and intensity of smoking, as compared to women. The effects of ethnicity and area of residence are more evident in women, among whom deprivation score is not apparently relevant, after other variables are considered. Urban environment increases the risk of COPD, possibly through higher air pollution levels. Social deprivation may increase the risk of COPD through complex mechanisms in addition to the higher prevalence of smoking. This may include different smoking habits (the model does not take into account duration and intensity of smoking as such information is not readily available), and a higher likelihood of exposure to other risk factors, which are not easily measured, such as passive exposure to tobacco smoking, history of respiratory infections, and less access to health services and information. Ethnic differences in susceptibility are less clear and less well understood, but might involve a combination of behavioural, environmental and possibly genetic factors.

We estimated the overall prevalence of COPD in England as 3.1% in people over 15 years old and 5.3% in those over 45 years old. The model illustrates the huge inequalities in the prevalence of COPD across England, with extreme risks in black men in urban deprived areas in one end of the risk spectrum, and Asian women in the lowest deprived rural areas, in the opposite end, between whom the risk of COPD varies 7-fold on average. Thus simpler models that do not take into account such variations in prevalence across population groups, or the extrapolation of national COPD prevalence figures for smaller areas, would be inappropriate for local use.

A systematic review of good quality (the criteria included nationally-representative samples) COPD prevalence studies using spirometry yielded estimates for developed countries of between 4% and 10%.[82] All were produced in Europe or North America, and all except one

were estimates for the adult population (aged ≥ 16 years), The 2001 HSfE data was not included. The Health Needs Assessment report suggests a prevalence of 5% for men and 3% for women of middle age and upwards.[83] Our results are similar to the latter, but not as high as suggested by some of the studies used for the former review. The decreasing trend in smoking prevalence in England is likely to lead to slow reductions in the real prevalence of COPD. However, major causes of variations in estimates include differences in the populations and in the diagnostic criteria used. This is illustrated by the finding of a prevalence of 13.3% in over 35s in the HSfE survey, when a different definition of COPD was used.[84] That study also calculated the prevalence directly from the survey data, differently from our study, where estimates were extrapolated for the population of England.

The figures estimated by the model for England are in general slightly lower than, but comparable with other studies on COPD using the same BTS definition, i.e**.** 4.5% in Norway,[85] 6.8% in the US[86] and 6.8% in white males 40-60 years old in Spain.[87] They are also similar to the overall prevalence of 6.1% found in the NICECOPD study for Belfast white population aged 40 to 69 years.[88] The slightly lower estimated prevalence in our study may be largely explained by the lower smoking prevalence in England, but also by differences in the study populations, and the larger study size of the HSfE.

Comparisons of our results with studies that used other definitions of COPD are difficult to interpret. Estimates based on self-reported symptoms tend to overestimate the prevalence. This is because diagnostic specificity is reduced as other respiratory diseases may be misdiagnosed as COPD, although asymptomatic cases of airflow obstruction will be missed. On the other hand, medically diagnosed COPD tends to under-estimate the true prevalence of the disease, as diagnostic sensitivity is reduced. Compared to other commonly used COPD spirometric-based case definitions of, the BTS is based on quite conservative cut-off points, yielding relatively low estimate values.[75;82]

Prevalence estimates from other sources are often poorly-defined and vary widely depending on the age ranges and exact definitions used. The updated NICE guidance[89] quotes firstly a Healthcare Commission report *("the rate of COPD in the population is estimated at between 2% and 4%, representing between 982,000 and 1.96 million people in England"*) and then Shahab et al *("...it is currently estimated that over 3 million people have the disease and that an estimated 2 million have undiagnosed COPD"*), although the latter report does not estimate overall numbers of cases in England- it estimates national prevalence as 13% of over 35 year olds.[84] The Consultation on a Strategy for Services for COPD states: "Quality and Outcomes Framework (QOF) records (suggest) that approximately 5.9% of the English population, around 3 million people, were affected in 2008".[90] However this is incorrect. The Information Centre QOF prevalence tables for 2008-9 show there were 834,312 people on English COPD registers, and quote this as a raw prevalence of 1.5% of a GP-registered population of around 54 million. Appendix 1 of the Strategy Impact Assessment gives an estimate of 8.2% to 8.3% of the over 16 population. However both the Strategy and Shahab et al use the GOLD definition of FEV1/FVC ratio of <0.7 alone.[91] This has been criticised as failing to take into account normal decreases in the ratio with age. The new NICE guidance states: "consider alternative diagnoses or investigations in: older people without typical symptoms of COPD where the FEV1/FVC ratio is < 0.7; younger people with symptoms of COPD where the FEV1/FVC ratio is ≥ 0.7".[89]

The COPD prevalence model we used is more conservative in that it uses the BTS definition i.e. an additional criterion of FEV1<80% predicted for age using British reference values. It estimates prevalence as 3.1% in the population over 15 years of age (similar age range to Shahab et al). Without this second criterion the prevalence is 11.1%.

A key advantage of the COPD model is that it is based on high quality data from a large representative sample of the population, and uses standard and specific diagnostic criteria for COPD, which is based on lung function rather than symptoms. Response rates were high in the survey with the achieved samples matching the target populations closely.[92] Prevalence estimates are based on the strength of association between key risk factors for COPD, including the effects of ethnicity, area of residence and deprivation, which were shown to be independent risk factors for COPD in the HSfE survey. This represents a significant advantage in relation to previous COPD prevalence models,[73;74] which were based only on smoking status, age and gender (also used in the COPD model) of mostly white populations outside the United Kingdom.

The input data are usually readily available at local level. The model uses current smoking status (never, former, current) as surrogates for total exposure to cigarette smoke, and is therefore not ideal to predict short term effects of changes in smoking prevalence (e.g. following intervention), due to long latent periods, large time lag between intervention and effects, and irreversibility of disease. However, the model is not static and will be updated over time, as parameters' values change, e.g. smoking prevalence (in England there are regular estimates of smoking prevalence at sub-regional level). Noteworthy, since intensity and duration of smoking (and thus smoking associated risk of COPD) tend to be lower in younger populations, the model may slightly overestimate the prevalence in young people.

Another reason for prevalence overestimation in young ages is a possible misclassification of cases of asthma into COPD (note reversibility test was not used in the HSfE). We dealt with these by providing alternative estimates for the prevalence of COPD, applying baseline or nil prevalence rates in all those under 35 and under 40 years old. These brought the overall prevalence estimates down by 0.1% to 0.4%. The model also relies on the quality of smoking and other data, and does not take into account competing causes of morbidity and mortality in the population, e.g. cardiovascular disease and lung cancer, which may affect prevalence of COPD. Moreover, it still needs correction for populations with significant occupational exposures.

There is some degree of imprecision in the estimates, which are larger when the rates are estimated for smaller populations or sub-groups, such as ethnic minorities and specific age-groups. Therefore we recommend that it is used primarily to derive overall population prevalence estimates, rather than estimates within population sub-groups. Further validation with a representative sample of the UK population including large proportions of people from ethnic minorities is still needed, before it may be reliably used in small population groups, such as GP practices and in ethnic minorities. We are currently planning to validate the model at practice level in North West London and investigating the feasibility of two COPD case-finding strategies.

The average QOF-diagnosed prevalence of COPD in England reported in 2004-5 through the Quality and Outcomes Framework of the GP Contract was only 1.4%.[16] This indicates that around 600,000 or nearly half of the 1.3 million COPD cases remain undiagnosed. A relatively large number of these individuals live in London or North of England (data not shown). Many of such cases will continue their risk behaviours and eventually present to the health services at later stages as more severe cases, possibly through emergency hospital admissions. Many of them will become high intensity users of health services, with considerable costs to individuals, the NHS and society. An audit of 80,000 COPD admissions showed that 70% of them are of patients not previously admitted with the condition (Bird M, personal communication).

The COPD model may be freely obtained directly from the authors and is publicly available on the Eastern Region Health Observatory website. It may be used, with the qualifications stated above, by general practices and PCTs, in England, and indeed in other user defined populations in the country and probably in other countries of the UK i.e. Wales, Scotland and Northern Ireland. General practices can use their own data on the population distribution by age-group and gender, smoking prevalence, ethnicity, degree of urbanisation and deprivation score for the area where the practice is located, to obtain indicative figures on the prevalence of COPD in their population. Data from the Primary Care Trust where the practice is located may be used as a proxy (with any relevant adjustments) in cases where practice specific data are not available. In addition it may be valid in Western populations that are not too dissimilar from England.

Conclusions

We believe that compared to previous models and prevalence estimates, the COPD model offers the most reliable estimates for England and the United Kingdom. It recognises deprivation, urban living and ethnicity as independent risk factors for COPD, which are taken into account in the estimates derived, in addition to smoking, age and gender. The model gives prevalence estimates for areas of varying sizes, including large populations at local level, however, the precision of the estimates will be higher for larger areas. The overall prevalence of COPD in England is estimated as 1.3 million, of whom as many as 600,000 people may be unaware of their diagnosis, therefore missing the opportunity of benefiting from early interventions. This emphasises the importance of active case finding and the model can be used to identify areas with a high level of unmet needs, i.e. with a high proportion of undiagnosed disease, where the benefits of case finding would be optimised. This strategy may also have an impact on reducing health inequalities, due to the socio-economic class gradient in COPD prevalence. The model should be validated, and case-finding strategies using the model should be evaluated for their cost-effectiveness.

8.2 Coronary Heart Disease Model

8.2.1 Introduction

CHD is a relatively common condition, for which there are several evidence-based treatments that reduce mortality. The HSfE for 2006 estimates that, based on respondents' self-reports of doctor-diagnosed CHD, the prevalence is about 6.5 per cent in men and 4.0 per cent in women [93], and this increases markedly with age. This prevalence has remained static over the last ten years. However the QOF, covering over 8,000 practices and 53 million patients, shows a GP-registered unadjusted prevalence of only 3.5 per cent [94] (but note that unadjusted prevalence rates show these registers as a percentage of the total practice list size i.e. for all ages).

The disparity between CHD prevalence estimates from large surveys, in particular the HSfE, and the number of patients diagnosed with CHD and registered in QOF led to demand for a CHD prevalence model at PCT and Local Authority level that gives an accurate estimate of true prevalence. The Association of Public Health Observatories (APHO) published a simple prevalence model to support development of 2007-08 Local Delivery Plans (1). However, it was acknowledged that this was a crude model and APHO subsequently asked me to develop a more robust model.

Defining prevalence

There are differences between various methods of estimating CHD prevalence. Questionnaire responses such as those used in HSfE to define CHD, may be less accurate than clinical diagnosis. Conversely, reliance on a medical diagnosis may underestimate prevalence, as patients with unrecognised angina or very mild symptoms may not attend (or be correctly identified by) their GP.

A Belgian analysis of the records of four large Belgian epidemiological studies during the past 30 years compared clinical and electrocardiographic (ECG) findings (see Table 5)[95]. Q wave patterns, ST segment depression or elevation, T wave inversion or flattening, and left bundle branch block are often seen as indications of silent myocardial ischaemia. The occurrence of ischaemia-like findings on the ECG was comparable between men and women (9.0% *v* 9.8%). The results from this and other studies consistently show that ischaemia-like ECG changes are associated with an approximately twofold increased risk of dying of CHD.

Table 5: Prevalence of coronary heart disease and ECG findings in 25-74 year old men and women

	Sex	25-34 years	35-44 years	45-5 years	55-64 years	65-74 years	25-74 years*	Odds ratios (men v women) (95% CI)
Angina pectoris	M	2.50 %	3.10 %	4.90 %	7.90%	13.10 %	5.00%	0.51 (0.46 to 0.56) for age < 55 years,
	F	4.00 %	6.70 %	8.50 %	8.40%	11.90 %	6.00%	1.00 (0.85 to 1.18) for age>55 years
History of	M	0.00	0.60	2.40	6.30%	12.80	3.60%	2.66 (2.20 to 3.22)

	Sex	25-34 years	35-44 years	45-5 years	55-64 years	65-74 years	25-74 years*	Odds ratios (men v women) (95% CI)
acute MI		%	%	%		%		
	F	0.10 %	0.20 %	0.60 %	3.00%	5.60%	1.50%	
Minnesota codes $I_{1,2}$	M	0.40 %	0.90 %	1.70 %	3.00%	5.00%	1.80%	2.02 (1.63 to 2.48)
	F	0.00 %	0.40 %	0.80 %	1.60%	3.30%	0.90%	
Coronary heart disease	M	2.90 %	4.30 %	7.80 %	13.60 %	22.70 %	8.30%	0.70 (0.64 to 0.77) for age < 55 years
	F	4.00 %	7.10 %	9.60 %	11.40 %	17.60 %	7.60%	1.22 (1.06 to 1.39) for age > 55 years
Major ECG findings	M	1.60 %	2.80 %	4.80 %	9.30%	19.00 %	6.00%	1.42 (1.27 to 1.57)
	F	0.80 %	2.00 %	3.40 %	6.70%	14.60 %	4.30%	
Minor ECG findings	M	3.40 %	5.50 %	9.20 %	16.60 %	29.40 %	10.40 %	1.13 (1.05 to 1.22)
	F	3.50 %	3.90 %	8.30 %	16.10 %	29.50 %	9.50%	

Source: De Bacquer D, De Backer G, Kornitzer M. Prevalence of ECG findings in large population based samples of men and women. Heart 2000 December 1;84(6):625-33.

In the British Regional Heart Study (BRHS), there was considerable overlap of questionnaire and ECG evidence of CHD, and high agreement between self-report and medical record for diagnosed CHD: for example, 80 per cent of men with a GP record of angina reported their diagnosis, and 70 per cent of men who reported an angina diagnosis had confirmation of this from the record review [96;97]. The prevalence of diagnosed angina in 1992 in these older men was 10.1 per cent according to self-reported history and 8.9 per cent according to GP record review.

Nevertheless, more than half of those in the BRHS with possible myocardial infarction (MI) combined with angina had no resting electrocardiographic evidence of CHD, and half of those with definite myocardial infarction on electrocardiogram had no history of chest pain at any time [98]. Only half of those with a definite MI on an electrocardiogram could recall a medical diagnosis of CHD [99]. Even in patients with severe (grade 2) angina, 40 per cent could not recall being told that they had heart disease. Overall, only one in five of those regarded as having CHD was able to recall such a diagnosis having been made by a doctor, and these were likely to be those most severely affected.

However there was substantial agreement between self-report and GP record of angina. The BRHS subsequently combined two questionnaire-based definitions to define prevalence: either *current angina symptoms*, which were defined as a positive response to standard World Health Organization (Rose) questionnaires (overall prevalence 11.1 per cent); or *history of diagnosed CHD* was defined as subject recall of ever having had a doctor's diagnosis of either angina or heart attack (overall prevalence also 11.1 per cent) [100].

The HSfE, however, uses only the latter questionnaire definition in its prevalence estimates. Trends in prevalence are shown in Table 6. These are not dissimilar to the BRHS given the latter only included males aged 40-59.

Table 6: Percentage prevalence of IHD, by survey year, age and sex [1]

Age	1994		1998		2003 Unweighted		2006 Unweighted		2003, Weighted		2006 Weighted	
	M	F	M	F	M	F	M	F	M	F	M	F
16-24	-	0.2	0.1	-	-	0.2	0.2	0.1	-	0.3	0.1	0.1
25-34	0.3	0.1	0.4	0.3	-	-	0.2	0.2	-	-	0.2	0.1
35-44	0.5	0.3	0.9	0.6	0.9	0.4	0.7	0.3	1.0	0.5	0.6	0.3
45-54	3.0	2.3	4.3	1.8	3.5	2.0	3.7	1.3	3.4	1.9	3.6	1.3
55-64	10.3	5.9	13.6	6.3	11.1	5.9	10.7	3.4	11.1	5.8	10.6	3.5
65-74	21.0	10.5	20.2	12.5	21.5	9.7	20.6	10.2	21.6	9.7	20.8	10.0
75+	22.7	15.9	23.4	18.4	26.4	18.4	28.5	19.3	26.5	18.1	28.4	19.3
All	6.0	4.1	7.1	4.6	7.4	4.5	6.2	3.0	6.4	4.1	6.5	4.0

Source: HSfE, Information Centre

Other prevalence surveys

While CHD mortality has greatly declined in the last four decades, the use of age-adjusted rates to describe CHD mortality obscures the fact that the decline largely represents the postponement of CHD deaths until older age. In fact, the overall burden of CHD is increasing in parallel with the increase in life expectancy. As the burden of prevalent CHD is increasing, identifying persons with CHD, measuring its incidence and outcome and how these vary over time and across populations is essential to understand the determinants of the trends in CHD. This in turn is crucial to define the relative contributions of risk factor reduction and therapeutic improvements, which is necessary to design effective interventions to reduce CHD.

I undertook a literature search for recent (post-1996) CHD prevalence surveys. I included only articles in English, which covered groups represented in the UK population in significant numbers. We also excluded surveys which covered CHD risk factor prevalence, or which sampled only sub-populations e.g. those with diabetes.

Community surveillance is a comprehensive approach designed to track disease at the community level, and is less costly and more efficient than cohort studies. In the USA,

several community surveillance studies have reported on temporal trends in CHD prevalence e.g. the Atherosclerosis Risk in Communities study, the Minnesota Heart Survey the Olmsted County Study, and the Worcester Heart Attack Study [101]. An analysis of US NHANES data on participants aged ≥ 40 years who attended the medical examination, the age-adjusted prevalence of angina pectoris, self-reported myocardial infarction, and ECG-defined myocardial infarction were 5.8% of 9255, 6.7% of 9250, and 3.0% of 8206 participants, respectively [102]. The age-adjusted prevalence of coronary heart disease defined by the presence of any of these conditions was 13.9% among men and 10.1% among women. These studies suggested that in the US medical care of clinical CHD was the main contributor to the mortality decline [103].

Outside the USA, the World Health Organization (WHO) MONICA (Multinational MONItoring of trends and determinants in CArdiovascular disease) Project was established in the early 1980s to monitor trends in cardiovascular diseases and to relate these to risk factor changes. Its central goal was to explain the trends in cardiovascular disease mortality observed from the 1970s. There were 32 MONICA centres in 21 countries. In these populations, the decline in coronary disease mortality is mostly related to the decline in CHD events, thereby pointing to primary prevention as the main source [104]. However the study populations excluded over 65s in whom most CHD occurs.

In a survey of a rural Indian population, CHD was diagnosed based on past documentation, response to WHO-Rose questionnaire, or changes in ECG. The prevalence of CHD (clinical + ECG criteria) was 3.4% in males and 3.7% in females. According to ECG criteria only, it was 2.8% in males and 3.3% in females and according to Q-waves only, it was 1.6% in males and 0.9% in females [105]. In a Finnish population survey Ahto et al found the prevalence of angina symptoms was 9.1% among men and 4.9% among women aged 64-97 [106]. Ischaemic ECG findings were common: 32.9% of men and 39.3% of women had such changes. An international systematic review and meta-analysis found that angina prevalence varied widely across populations, from 0.73% to 14.4% (population weighted mean 6.7%) in women and from 0.76% to 15.1% (population weighted mean 5.7%) in men [107].

In the UK, Carroll et al used GP records in London and found a prevalence of 8 per cent of men and 5 per cent of women over 44 years of age- although this may be lower than the true national average [108]. There was a history of myocardial infarction in 30 per cent of men and 22 per cent of women with CHD. Lampe and colleagues examined trends in the prevalence of CHD in men participating in the BRHS [100]. The authors demonstrated a decrease in the prevalence of current angina symptoms: the age adjusted annual percentage change in odds was -1.8%. However, there was no evidence of a trend in the prevalence of history of diagnosed CHD.

A study by Davies et al examined trends in CHD incidence, prevalence, and mortality in the UK between 1996 and 2005, using the THIN GP database (a total of 5 million patients). The results indicate that, while CHD mortality declined, CHD incidence decreased less than mortality, resulting in an increase in CHD prevalence [109]. From 1996 to 2005, age-standardized incidence of CHD decreased by 2.2% in men and 2.3% in women per year (average percentage change). Age-standardized all-cause mortality among those with CHD decreased by 4.5% in men and 3.4% in women per year (average percentage change). Age-

standardized prevalence increased by 1.3% in men and 1.7% in women per year (average percentage change). The decline in incidence had some impact on limiting the increase in prevalence, but its effect was offset by the increase in prevalence occurring as a result of improved survival among people with CHD. Although patients with nitrate prescriptions were also included, this study relied mainly on CHD diagnostic codes which may underestimate actual prevalence.

Previous CHD prevalence modelling

An epidemiologic CHD prevalence model was first developed in the UK to assist case-finding in Sheffield PCTs. Subsequently, English PCTs were required to set targets for CHD case-finding in their 2007-8 Local Delivery Plans negotiated with strategic health authorities. To assist them, a simple PCT-based prevalence model was developed rapidly by the APHO. Stage 1 of the modelling predicts the number of people with identified CHD within each population, taking account only of the demographic distribution of the population. The prevalence of patient-reported doctor-diagnosed CHD in each age/sex stratum is based on national data from the HSfE. Stage 2 takes account of deprivation levels in different PCTs in England. In the absence of sufficiently precise published data on the relationship between deprivation and CHD prevalence, the model makes the assumption that areas with higher CHD mortality rates have comparably higher prevalence of CHD. Using data for all local authorities in England, a linear relationship was calculated between 2002-04 SMRs for CHD and a deprivation score (UV67) derived from the 2001 Census Classification of Deprivation:

$$\text{CHD SMR} = (2.604 \times \text{UV67}) + 25.97$$

Using UV67 scores calculated for each PCT, the above formula gives a multiplying factor for each PCT. For example, a PCT with a UV67 score of 40% (very deprived) has a multiplying factor of 1.3.

APHO accepts that this current model is rather crude, and that a prevalence model based on a comprehensive regression model using HSfE data would be more robust. This model was commissioned by APHO from the Department of Primary Care & Public Health (PCPH) Imperial College London.

Congdon has also produced a CHD prevalence model using earlier HSfE data [44]. Data from the 1999 and 2003 HSfEs were used to provide model based rates of CHD prevalence by age, sex, ethnic group, region, and area deprivation category. To take into account the effect of socio-economic factors, the HSfE model gradient over deprivation quintiles was applied to scale area prevalence estimates specific for age, sex and ethnicity. The final stage of the prevalence estimation procedure incorporated proxy information from mortality. To adjust the HSfE-based prevalence rates to take account of interdependence of area mortality and prevalence, prevalence and mortality were taken as joint (i.e. correlated) outcomes in a Bayesian model allowing also for the spatial patterning of both outcomes, using WinBUGs software.

8.2.2 Methods

Data sources

The HSfE was used as the data source, and the outcome of interest was patient-reported doctor-diagnosed CHD (called IHD in HSfE), which is the variable used for CHD prevalence in HSfE. It was also chosen because, based on previous research, it appears to be the best **single** proxy for true prevalence. (An alternative would be to combine patient-reported doctor-diagnosed CHD with a positive Rose angina questionnaire as in BRHS.)

Because ethnicity is a known CHD risk factor, it was necessary to use a sample containing data from a large number of ethnic minority respondents. The HSfE 2004 was the last survey to include an ethnic minority boost, and the boost sample was used for the modelling. However, there were relatively small numbers of whites in the HSfE 2004 sample, and they were not asked to respond to many questions, including those on CHD, presumably to save resources for the boost itself. The HSfE 2004 boost sample was therefore merged with the HSfE 2003, which was the year with the largest number of identical variables.

Table 7 shows the numbers of respondents in the original HSfE 2004 ethnic boost, and the number in the merged 2003-2004 dataset used for the model where a response for the CHD outcome variable was obtained. Note that it was necessary to collapse two of the HSfE 2003 ethnic group variables to use the same classification as HSfE 2004.

Table 7: Ethnic Group Breakdown of HSfE 2004 Dataset & Merged 2003-2004 Dataset

Ethnic Group	HSfE 2004			Merged 2003-4 Dataset		
	Freq.	Percent	Cum.	Freq.	Percent	Cum.
No answer/refused	21	0.21	0.21	16	0.08	0.08
Don't know	3	0.03	0.24	2	0.01	0.08
White	1,597	15.79	16.03	14,575	68.58	68.67
Mixed ethnic group	623	6.16	22.19	308	1.45	70.12
Black or Black British	2,468	24.40	46.59	1,991	9.37	79.48
Asian or Asian British	4,764	47.10	93.69	3,725	17.53	97.01
Any other group	638	6.31	100.00	635	2.99	100.00
Total	10,114	100.00		21,252	100.00	

(Note that there were 34 missing values.)

Model construction: data issues

The choice of variables for original inclusion in the merged dataset included all those known to be CHD risk factors. The variable names and labels are shown in Table 3 below. The HSfE dataset has a nested or hierarchical structure so three variables related to the sampling strata were included: area (sample point), cluster (stratification level), and hserial (serial number of household). These were used in the model to adjust for clustering of respondents. In the analysis variables cholest and hdlchol were combined to give a lipid ratio.

The bandings of Index of Multiple Deprivation 2004 scores were slightly different between the two years, but raw scores were not provided in the dataset so it was necessary to assume that the banding were the same (see Table 8 : Index of Multiple Deprivation Banding below).

The Stata10 software package was used for analysis. All variables were re-coded to drop negative values for estimation purposes (in HSfE various non-response categories are assigned negative values). The methodology applied was multinomial logistic regression with the "cluster" adjustment option for households (see above). For analysis of two categories as here, multinomial logistic regression is reduced to binomial logistic regression. However the reason for not using other logistic regression routines that take into account nested structures is that the other options available in Stata produced an estimation error, probably because of the small percentage of disease-positive respondents in the sample.

Table 8 : Index of Multiple Deprivation Banding

Rank	IMD Band	IMD		Number	Per Cent	Cum Per Cent
		HSfE 2003	HSfE 2004			
least	1	0.59-8.35	0.55-9.02	3,803	17.87	17.87
	2	8.35-13.72	9.03-14.14	3,573	16.79	34.65
	3	13.72-21.16	14.15-21.17	3,788	17.8	52.45
	4	21.16-34.21	21.18-33.52	4,551	21.38	73.83
most	5	34.21-86.36	33.53-85.69	5,571	26.17	100
			Total	21,286	100	

The Stata software package was used for analysis. As for the CHD model the methodology applied was multinomial logistic regression with the "cluster" option (see above). As previously the prevalence in each age group, gender, ethnic group, area of residence and level of deprivation, and smoking status category were derived from the odds, using the formula: prevalence = odds/(1 + odds).

Model construction: interactions between variables

Effect modification or interaction occurs if the effect of one exposure or risk factor on the outcome varies according to the level of another risk factor. This can be tested using a χ^2 test of heterogeneity e.g. Mantel-Haenszel odds ratios, Wald or likelihood tests, or by introducing interaction terms or parameters into the regression model. These allow the effect of one variable to be different in different categories of other variables. In Stata, the xi command expands terms containing categorical variables into indicator (also called dummy) variable sets by creating new variables and estimates interactions and main effects.

An initial examination for interactions was carried out using Mantel-Haenszel odds ratios, Wald and likelihood ratio tests. Where these reached significance interaction terms were created as indicator variables in a regression model. None of the interaction terms tested in this way showed consistent statistical significance. The inclusion of interaction terms in the model was therefore rejected.

Model construction: internal validation

Ideally the best prediction should result from utilising the most risk factor information in the regression model. However, only a limited range of HSfE variable data is either available or can be estimated at the PCT or LA level, so there is no purpose in including other variables

(see Table 9). We decided to validate the local model by comparing it, in terms of prediction, to a model including all available and significant HSfE variables. In addition, however, the amount of missing data affects the prediction of a model. In the complete HSfE variables the largest proportion of missing data occurred in those variables related to drug treatment for high blood pressure. We therefore excluded these variables in a second version of the model. Hence the model COMP 1 included the "complete" list of variables, including BP drugs; model COMP 2 included the "complete" list of variables, but excluded BP drugs.

8.2.3 Results

LOCAL only used locally available data. I included smoking and BMI variables on the basis that local synthetic estimates are now available. Local GHQ-12 or Limiting Longstanding Illness score data will be available locally from 2009 and could be included later.

Table 9: Variables Included in Merged Dataset

COMP 1 = "Complete" variables with BP drugs
COMP 2 = "Complete" variables without BP drugs
LOCAL= only using locally available data

Name	Label	COMP 1	COMP 2	LOCAL
ihdis	had IHD (angina or heart attack)	X	X	X
aceinh	ace inhibitors (blood pressure)	X		
addnum	address number	N/A	N/A	N/A
adtot30	adults: total days/4 weeks active 30 mins + moderate +	X	X	
age	age last birthday	X	X	X
area	sample point	N/A	N/A	N/A
beta	beta blockers (blood pressure)	X		
bmival	valid body mass index	X	X	X
calciumb	calcium blockers (blood pressure)	X		
cholest	total cholesterol result (blood data)	X	X	
cigst1	cigarette smoking status - never/ex-reg/ex-occ/current	X	X	X
cluster	stratification level	N/A	N/A	N/A
diabete2	doctor diagnosed diabetes (excluding pregnant)	X	X	
diur	diuretics (blood pressure)	X		
ethnici	ethnic group	X	X	
famcvd	family history of cvd	X	X	
fatvala	fat score	X	X	
fldlchol	LDL cholesterol result (fasting blood data)	X	X	
ftrigl	triglycerides result (fasting)	X	X	
ghq12scr	General Health Questionnaire score- 12 point scale	X	X	
glucval	glucose result (fasting)	X	X	
hdlchol	HDL cholesterol result (blood data)	X	X	
hserial	serial number of household	N/A	N/A	N/A
imd2004	index of multiple deprivation (SOA level)	X	X	X
limitill	limiting longstanding illness	X	X	
nssec8	National Statistics Socioeconomic Class (8 variable)	X	X	
obpdrug	other drugs affecting BP	X		

Name	Label	COMP 1	COMP 2	LOCAL
omdiaval	omron valid mean diastolic BP	X	X	
omsysval	omron valid mean systolic BP	X	X	
porftvg	grouped portions of fruit (inc.orange juice) & veg yesterday	X	X	
roseanmi	angina or MI (Rose angina questionnaire)	N/A	N/A	N/A
sex	sex	X	X	X
topqual3	highest educational level	X	X	

The prediction of the three models was assessed in two ways:

- by generating a receiver operating characteristics (ROC) curve using the predicted probabilities of the CHD outcome compared to the observed cases
- by deriving predicted probabilities of the CHD outcome in Stata from the three models, and comparing these to the observed cases

Model construction: external validation

An early external validation will be carried out by examining the association between PCT/LA level CHD prevalence estimates and QOF registered prevalence. The regression-based model will also be validated against a prevalence model obtained by Bayesian strategies using WinBUGS. Finally, funding will be sought to undertake a validation of practice-based prevalence estimates against registered prevalence supplemented by active case finding. Table 10 shows the frequency of the CHD outcome by age group in the merged dataset.

Table 10: Respondents Reporting Doctor Diagnosed CHD by Age Band

	Age Band							
	16-24	25-34	35-44	45-54	55-64	65-74	75+	Total
Don't know	0	0	0	0	0	1	2	3
Yes	2	6	43	97	262	350	344	1,104
No	2,500	3,855	4,409	3,379	2,792	1,928	1,316	20,179
Total	2,502	3,861	4,452	3,476	3,054	2,279	1,662	21,286

Local model

The regression model for risk factors for CHD in the "local" prevalence model is shown in Table 11. As expected ORs increase strikingly with increasing age in all models. In the prevalence predictions using coefficients (not shown in these tables) this results in age-related increases in prevalence which closely match the crude overall prevalence in Table 6. Surprisingly the only significant comparison for smoking is for category 3 "used to smoke regularly" i.e. this group is more likely to report CHD compared to the group "never smoked cigarettes at all". There is a significant comparison for male sex. ORs, p values and confidence intervals are generally similar to the "Complete" models. Unfortunately, however, local synthetic estimates of smoking prevalence do not include categories for occasional/regular smokers.

Table 11: Odds Ratios for LOCAL, Model With Only Locally Available Variables

Risk factor	RRR	Std Error	z	P>z	[95% Conf.	Interval]
Age 25-34	1.000					
Age 35-44	6.867	2.998	4.41	0	2.919	16.157
Age 45-54	19.514	8.217	7.06	0	8.549	44.541
Age 55-64	65.698	27.217	10.1	0	29.169	147.972
Age 65-74	122.864	50.719	11.65	0	54.707	275.936
Age 75+	191.252	79.128	12.7	0	85.003	430.307
Female sex	1.000					
Male sex	1.849	0.139	8.16	0	1.595	2.143
Never smoker	1.000					
Used to smoke occasionally	0.757	0.131	-1.61	0.107	0.539	1.062
Used to smoke regularly	1.484	0.119	4.91	0	1.267	1.737
Current smoker	1.072	0.109	0.68	0.495	0.878	1.308
Index of multiple deprivation 0.59-8.35	1.000					
Index of multiple deprivation 8.35-13.72	1.226	0.143	1.75	0.08	0.976	1.541
Index of multiple deprivation 13.73-21.16	1.350	0.154	2.63	0.009	1.079	1.689
Index of multiple deprivation 21.17-34.21	1.645	0.183	4.49	0	1.323	2.044
Index of multiple deprivation 34.22-86.36	2.420	0.256	8.36	0	1.967	2.978
White	1.000					
Mixed	1.264	0.851	0.35	0.727	0.338	4.726
Black/BB	0.763	0.168	-1.23	0.218	0.496	1.173
Asian/AO	1.511	0.243	2.56	0.01	1.102	2.071
Other	0.168	0.170	-1.76	0.079	0.023	1.227

I examined the data for interactions between variables e.g. if there is an interaction between sex and ethnicity, there will be two separate effects of ethnicity on CHD: one for males and another for females. We tested for interactions between local variables initially using Mantel-Haenszel odds ratios, and likelihood ratio and Wald tests. These indicated a possible interaction between sex and ethnicity. However a regression analysis showed significance for only one sex-ethnicity indicator level. We therefore decided not to include an interaction variable in the model.

Complete model excluding treatment for hypertension

The regression model for risk factors for CHD in the "Complete" model excluding hypertension either systolic or diastolic BPs or treatment for hypertension) is shown in Table 12. This shows well the impact of additional variables on ORs for "local" variables. BMI is now recognised as an independent risk factor for CVD (although its effect is mediated largely through changes in "physiological" risk factors such as cholesterol: high density lipoprotein (HDL) ratio [110-112]). Although an expected higher OR is found in underweight

patients, this is not the case for overweight/obese categories. This may be because this is cross-sectional rather than longitudinal data and that the prevalence of obesity in older age groups, where CHD is more common, is very low. The regression model for risk factors for CHD in the "complete" model without BP drugs is shown in Table 13.

Table 12: CHD prevalence by BMI category

	BMI <18.51	BMI >18.50 & BMI<25	BMI >25 & BMI <30	BMI >30 & BMI <40	BMI >40	Total
CHD	12	187	386	268	251	1,104
No CHD	306	6,693	6,502	3,483	3,194	20,178
Total	318	6,880	6,888	3,751	3,445	21,282

Table 13: Odds ratios for complete model excluding treatment for hypertension

Risk factor	RRR	Std Error	z	P>z	[95% Conf.	Interval]
Age 25-34	1.000					
Age 35-44	7.131	6.891	2.03	0.042	1.073	47.396
Age 45-54	18.447	17.436	3.08	0.002	2.893	117.624
Age 55-64	50.099	46.710	4.2	0	8.058	311.490
Age 65-74	121.219	112.515	5.17	0	19.655	747.578
Age 75+	216.341	200.711	5.8	0	35.111	1333.034
Female sex	1.000					
Male sex	2.258	0.320	5.74	0	1.710	2.981
Never smoker	1.000					
Used to smoke occasionally	1.039	0.320	0.13	0.9	0.569	1.900
Used to smoke regularly	1.512	0.238	2.63	0.009	1.111	2.057
Current smoker	1.114	0.232	0.52	0.604	0.740	1.676
Index of multiple deprivation 0.59->8.35	1.000					
Index of multiple deprivation 8.35-13.72	1.215	0.241	0.98	0.325	0.824	1.791
Index of multiple deprivation 13.73-21.16	1.025	0.217	0.12	0.907	0.677	1.551
Index of multiple deprivation 21.17-34.21	1.101	0.230	0.46	0.646	0.730	1.659
Index of multiple deprivation 34.22-86.36	1.424	0.323	1.56	0.119	0.913	2.219
White ethnic group	1.000					
Mixed ethnic group	0.208	0.144	-2.26	0.024	0.054	0.810
Black or Black British ethnic group	0.831	0.355	-0.43	0.666	0.360	1.922
Asian or Asian British ethnic group	1.553	0.436	1.57	0.117	0.896	2.693
Any other ethnic group	1.049	0.665	0.08	0.939	0.303	3.631
BMI <18.51	1.000					
BMI >18.50 & <25	0.538	0.568	-0.59	0.557	0.068	4.268

Risk factor	RRR	Std Error	z	P>z	[95% Conf.	Interval]
BMI >25 & BMI <30	0.890	0.936	-0.11	0.912	0.113	6.984
BMI >30 & BMI <40	1.026	1.085	0.02	0.98	0.129	8.143
BMI >40	0.693	0.734	-0.35	0.729	0.087	5.515
Total cholesterol:HDL ratio	0.748	0.049	-4.44	0	0.658	0.850
Diabetes; no	1.000					
Diabetes; yes	0.686	0.144	-1.8	0.072	0.455	1.035
Family History of CVD; yes	1.000					
Family History of CVD; no	0.622	0.103	-2.87	0.004	0.450	0.860
Limiting longstanding illness	1.000					
Non limiting longstanding illness	0.625	0.102	-2.89	0.004	0.454	0.859
No limiting longstanding illness	0.318	0.058	-6.24	0	0.222	0.456
Rose questionnaire: angina and MI	1.000					
Rose questionnaire: neither angina MI	0.007	0.004	-9.06	0	0.002	0.021
Rose questionnaire: angina, but not MI	0.074	0.044	-4.36	0	0.023	0.239
Rose questionnaire: MI, but not angina	0.174	0.096	-3.16	0.002	0.059	0.514
Top qualification nvq4/nvq5/degree	1.000					
Top qualification higher ed below degree	1.479	0.428	1.35	0.176	0.839	2.607
Top qualification nvq3/gce a level equiv	1.122	0.407	0.32	0.751	0.551	2.283
Top qualification nvq2/gce o level equiv	1.227	0.334	0.75	0.453	0.719	2.092
Top qualification nvq1/cse other grade equiv	1.443	0.504	1.05	0.294	0.728	2.861
Top qualification foreign/other	0.875	0.305	-0.38	0.701	0.442	1.732
Top qualification no qualification	1.458	0.355	1.55	0.121	0.905	2.349

Complete model including treatment for hypertension

The regression model for risk factors for CHD in the "complete" model with BP drugs is shown in Table 14. Hypertension is a well-established risk factor, so it is desirable to include variables related to it in a "complete" model. Treatment with ACE inhibitors, beta blockers and calcium blockers are highly significant. However these may also be used to treat established CHD, so the association may be unrelated to hypertension. In addition, the HSfE relies upon patient recall for drug treatment. Not unexpectedly, much of the data for these variables is missing. I added systolic and diastolic BP as ordinal variables to the model, but this resulted in major changes to ORs for other variables. This may simply be related to model instability because of the large numbers of variables included. The addition of a single variable for hypertension, either treated or untreated, will be explored as a further later step in model development.

Table 14: Odds ratios for complete model including treatment for hypertension

Risk factor	RRR	Std Error	z	P>z	[95% Conf.	Interval]
Age 25-34	1.000					
Age 35-44	3.975	3.522	1.56	0.119	0.700	22.566
Age 45-54	8.606	7.259	2.55	0.011	1.647	44.959
Age 55-64	18.559	15.529	3.49	0	3.600	95.676
Age 65-74	36.219	30.104	4.32	0	7.103	184.684
Age 75+	61.082	50.587	4.97	0	12.050	309.643
Female sex	1.000					
Male sex	2.273	0.360	5.18	0	1.666	3.101
Never smoker	1.000					
Used to smoke occasionally	0.989	0.342	-0.03	0.975	0.502	1.948
Used to smoke regularly	1.450	0.246	2.18	0.029	1.039	2.023
Current smoker	1.201	0.278	0.79	0.43	0.762	1.892
Index of multiple deprivation 0.59->8.35	1.000					
Index of multiple deprivation 8.35-13.72	1.214	0.273	0.86	0.39	0.781	1.887
Index of multiple deprivation 13.73-21.16	1.033	0.248	0.13	0.893	0.645	1.653
Index of multiple deprivation 21.17-34.21	1.174	0.270	0.7	0.485	0.748	1.843
Index of multiple deprivation 34.22-86.36	1.394	0.366	1.26	0.206	0.833	2.331
White ethnic group						
Mixed ethnic group	0.353	0.240	-1.53	0.125	0.093	1.336
Black or Black British ethnic group	0.304	0.169	-2.14	0.032	0.102	0.902
Asian or Asian British ethnic group	1.316	0.391	0.92	0.356	0.734	2.357
Any other ethnic group	0.894	0.816	-0.12	0.902	0.149	5.345
BMI <18.51	1.000					
BMI >18.50 & <25	0.435	0.441	-0.82	0.411	0.060	3.171
BMI >25 & BMI <30	0.588	0.591	-0.53	0.597	0.082	4.218
BMI >30 & BMI <40	0.680	0.689	-0.38	0.704	0.093	4.954
BMI >40	0.519	0.525	-0.65	0.517	0.072	3.763
Total cholesterol:HDL ratio	0.742	0.053	-4.14	0	0.644	0.854
Diabetes; no	1.000					
Diabetes; yes	1.034	0.250	0.14	0.888	0.645	1.660
Family History of CVD; yes	1.000					
Family History of CVD; no	0.630	0.115	-2.54	0.011	0.441	0.900
Limiting longstanding illness	1.000					
Non limiting longstanding illness	0.592	0.102	-3.05	0.002	0.422	0.829
No limiting longstanding illness	0.578	0.128	-2.48	0.013	0.375	0.892
Rose questionnaire: angina and MI	1.000					

Risk factor	RRR	Std Error	z	P>z	[95% Conf.	Interval]
Rose questionnaire: neither angina MI	0.007	0.004	-8.49	0	0.002	0.022
Rose questionnaire: angina, but not MI	0.063	0.039	-4.4	0	0.018	0.215
Rose questionnaire: MI, but not angina	0.182	0.107	-2.91	0.004	0.058	0.574
Top qualification nvq4/nvq5/degree	1.000					
Top qualification higher ed below degree	1.655	0.549	1.52	0.129	0.864	3.173
Top qualification nvq3/gce a level equiv	1.396	0.593	0.79	0.432	0.607	3.211
Top qualification nvq2/gce o level equiv	1.510	0.457	1.36	0.173	0.834	2.733
Top qualification nvq1/cse other grade equiv	1.768	0.665	1.52	0.129	0.847	3.694
Top qualification foreign/other	0.922	0.371	-0.2	0.84	0.419	2.031
Top qualification no qualification	1.496	0.408	1.48	0.14	0.876	2.553
ACE inhibitor: no	1.000					
ACE inhibitor: yes	1.875	0.335	3.52	0	1.321	2.661
Beta blocker: no	1.000					
Beta blocker: yes	3.467	0.570	7.57	0	2.512	4.785
Calcium blocker: no	1.000					
Calcium blocker: yes	2.735	0.468	5.88	0	1.956	3.825
Diuretic: no	1.000					
Diuretic: yes	0.832	0.140	-1.09	0.276	0.598	1.158
Other BP drug: no	1.000					
Other BP drug: yes	0.869	0.279	-0.44	0.662	0.463	1.631

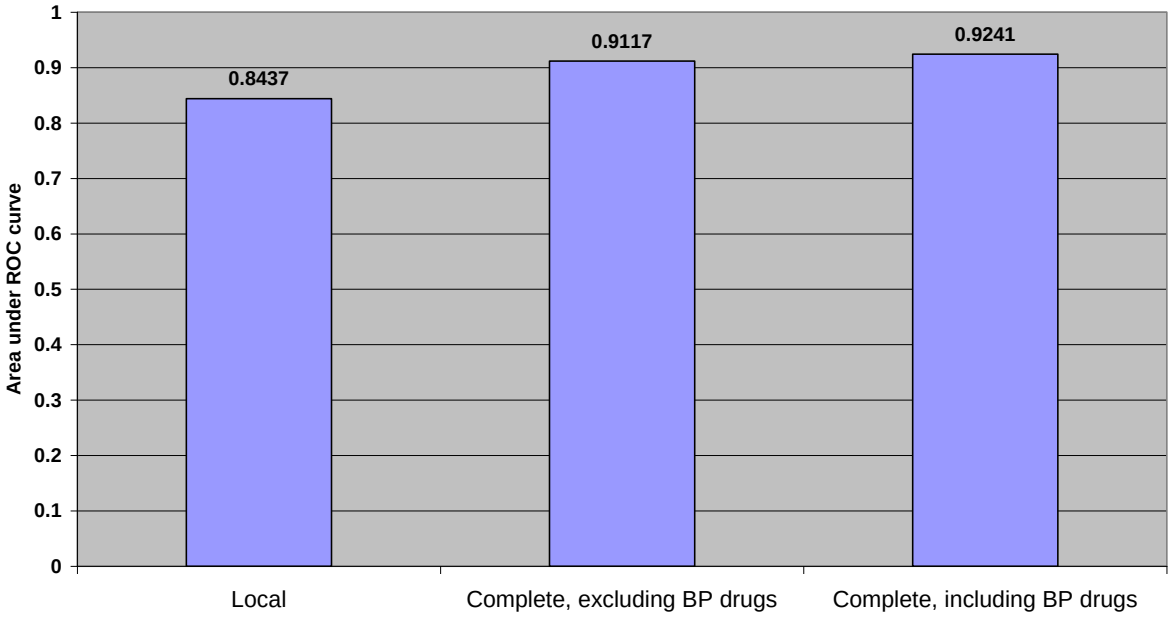
Internal validation: area under Receiver Operating Characteristics curve (AUROC)

ROC analysis was originally developed during World War II to analyse classification accuracy in differentiating signal from noise in radar detection. Recently, the methodology has been adapted to several clinical areas heavily dependent on screening and diagnostic tests, in particular, laboratory testing, epidemiology, radiology, and bioinformatics [113;114]. ROC analysis is a useful tool for evaluating the performance of diagnostic tests and more generally for evaluating the accuracy of a statistical model (e.g. logistic regression, linear discriminant analysis) that classifies subjects into one of two categories, diseased or non-diseased, as in this model [115]. Its function as a simple graphical tool for displaying the accuracy of a medical diagnostic test is one of the most well-known applications of ROC curve analysis.

An ROC curve is a plot of sensitivity on the *y* axis against (1-specificity) on the *x* axis for varying values of the threshold, t. The 45° diagonal line connecting (0,0) to (1,1) is the ROC curve corresponding to random chance. The ROC curve for the gold standard is the line connecting (0,0) to (0,1) and (0,1) to (1,1). Generally, ROC curves lie between these two extremes. The area under the ROC curve is a summary measure that essentially averages diagnostic accuracy across the spectrum of test values. The area under the curve is an overall summary of diagnostic accuracy. AUC equals 0.5 when the ROC curve corresponds to random chance and 1.0 for perfect accuracy. On rare occasions, the estimated AUC is <0.5, indicating that the test does worse than chance.

AUROCs for the three models tested above were estimated using Stata10. These are shown in Figure 1. If both sensitivity and specificity are of importance in a CHD model, the optimal threshold of t would be 0.75, where sensitivity and specificity equal 0.77. The local model exceeds this level, although the two more complete models have even better performance.

Figure 1: Comparison of AUROC for different CHD prevalence models



Validation: model prediction

Another method of assessing performance is to use the regression model to predict the response for each subject. These predictions are called fitted values. The difference between the fitted and the observed values are called residuals. These can then be tabulated against the observed presence of CHD to assess "misclassification" by each model. The results for the three models are shown in Table 15: Comparison of predictions by model.

Table 15: Comparison of predictions by model

	Observed	Predicted		
Model		No CHD	CHD	Total
Local	No CHD	17,614	3	17,617
	CHD	1,098	1	1,099
	Total	18,712	4	18,716
Complete, no BP drugs	No CHD	3,879	64	3,943
	CHD	1	230	219
	Total	4,109	283	4,392
Complete	No CHD	3,883	60	3,943
	CHD	203	246	449
	Total	4,086	306	4,392

8.2.4 Application: PCT, LA and practice level models

The local model (which includes only those variables that are available at population level i.e. age, sex, ethnicity, smoking status and deprivation score) has been applied to Local Authorities to create prevalence estimates of hypertension in those aged 16+ for 2005–2020. Models for PCTs have been created for 2006-2020, along with practice level estimates for 2007.[116]

Assumptions of the modelled estimates

It is assumed that:

- the proportion of smokers, ex-smokers and never-smokers is the same across ethnic groups.
- the proportion of ex-smokers in each age-sex group is the same in all areas.
- the smoking prevalence rates from the model-based estimates of lifestyle behaviours are reliable.
- the prevalence of CHD in those aged 16-24 is negligible.

Due to lack of data, it was not possible to treat ex-regular-smokers and ex-occasional-smokers separately. Ex-occasional-smokers are treated as non-smokers.

Populations

The CHD prevalence model uses ONS 2005 mid-year LA population estimates by ethnic group, age and sex. The Office for National Statistics (ONS) publishes the data by broad age band, but supplied full quinary age-band data to APHO for the prevalence modelling project. Five ethnic groups were used: white, black, Asian, mixed and other. The model uses ONS 2006 mid-year PCT population estimates by ethnic group, age and sex. ONS publishes the data by broad age band, but supplied full quinary age-band data to APHO for the prevalence modelling project. Five ethnic groups were used: white, black, Asian, mixed and other.

In order to calculate estimate prevalence of CHD in the future, population projections were incorporated into the model. ONS has not published population projections by ethnic group, so the 2005 (LA) or 2006 (PCT) distribution of ethnic groups was used to generate population estimates to 2020. For 2006 and 2007, LA population projections the distribution of ethnic groups in 2005 was applied to ONS mid-year population estimates by quinary age band. For 2007 PCT population projections, the distribution of ethnic groups in 2006 was applied to ONS mid-year population estimates by quinary age band. For 2008 and 2009, the distribution of ethnic groups in 2005 (LA) or 2006 (PCT) was applied to ONS 2006-based population projections by quinary age band for LAs and PCTs. For 2010, 2015 and 2020 the distribution of ethnic groups in 2005 (LA) or 2006 (PCT) was lagged by 5, 10 and 15 years respectively and combined with ONS 2006-based population projections by quinary age band for LAs and PCTs. For example, the ethnic proportions for the age band 40-45 in 2005 were applied to the population aged 50-55 in 2015. With the exception of this 'ageing' of the ethnic population, no other changes to the distribution of ethnic groups were considered.

Practice populations by age and sex were derived from a National Strategic Tracing Service (NSTS) extract taken on 31/03/2007. Patients with date of death before 31/03/07 were

excluded. to improve the data quality GP codes were re-mapped to practices using NSTS files of February 2007 (available from the NHS Organisation Data Service). Some practices have missing PCT and SHA codes in the Feb 07 NSTS files, so these were populated using the May 07 NSTS files. Practices with a population of <1000 patients are excluded from modelling due to the uncertainty associated with such small numbers.

Ethnicity

The proportion of practice population in ethnic groups was supplied by the Care Quality Commission (CQC). The proportions were derived from HES admissions in 2005/06-2006/07, excluding mental health and maternity admissions and records with unknown ethnicity. HES is the only routine dataset which includes GP practice and patient ethnicity, and has high levels of completion and data quality. It is assumed that the hospital admissions (excluding maternity and mental health) reflect the true ethnic population of the practice and there is no systematic bias. The proportions by ethnic group for each practice were calculated by dividing admissions within each ethnic category by the total admissions for the practice. The same ethnic distribution is applied across all age bands as there are insufficient hospital admissions to robustly calculate the distribution of ethnic groups by age and sex for practices.

Smoking status

National (England) proportions of smokers, ex-smokers and current smokers by age and sex are taken from HSfE (2003-2005 pooled, see Figure 2). These proportions were then adjusted for each LA/PCT using the synthetic estimates of smoking prevalence for 2003-2005, using the following algorithm.

Local proportion of smokers in age-sex category = national prevalence of smoking in age-sex category * local overall smoking prevalence / national overall smoking prevalence

$$S_{asl} = S_{asn} \times \frac{S_l}{S_n}$$

Local proportion of ex-smokers in age-sex category is not adjusted

$$E_{asl} = E_{asn}$$

Local proportion of never-smokers in age-sex category = 1 – proportion of ex-smokers in age-sex category – local proportion of smokers in age-sex category

$$N_{asl} = 1 - E_{asl} - S_{asl}$$

Where:

S = proportion of population who are smokers
E = proportion of population who are ex-regular-smokers
N = proportion of population who have never smoked
l = local
n = national
as = by age and sex

This approach assumes that the proportion of ex-smokers in each age-sex category is fixed and the number of never-smokers increases as the number of smokers decreases. Regional analysis of the relationship between prevalence of smokers and ex-smokers in the HSfE shows no systematic relationship and therefore it was decided that the ex-smoking rate should not be locally adjusted (see Figure 3).

Figure 2: Prevalence of smoking behaviours in England by age and sex, 2003-5

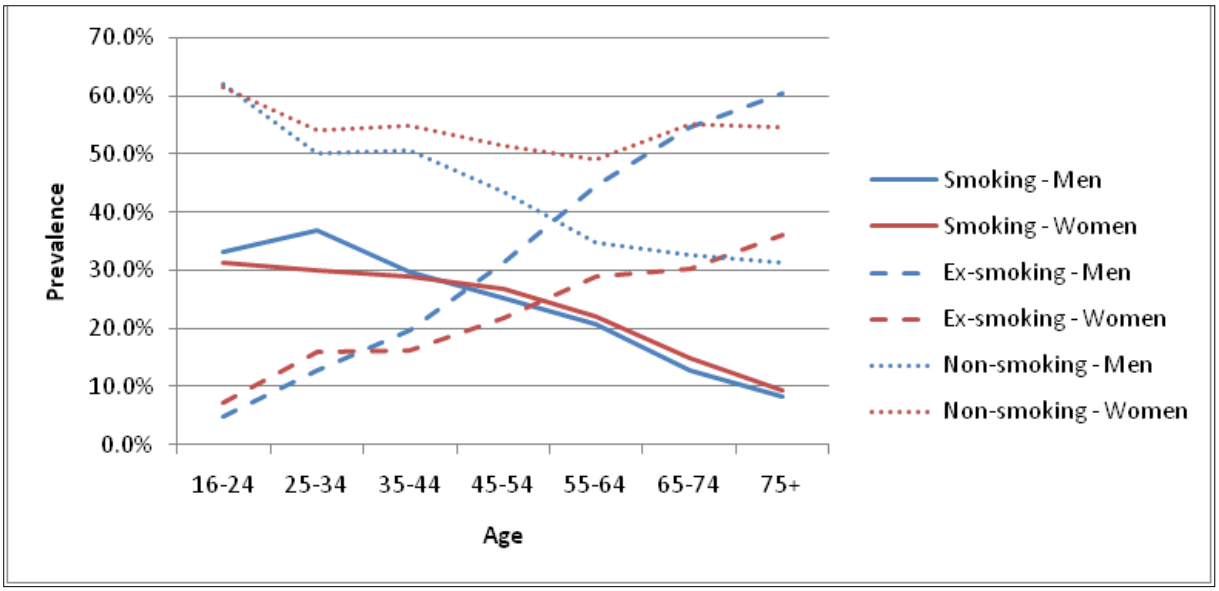
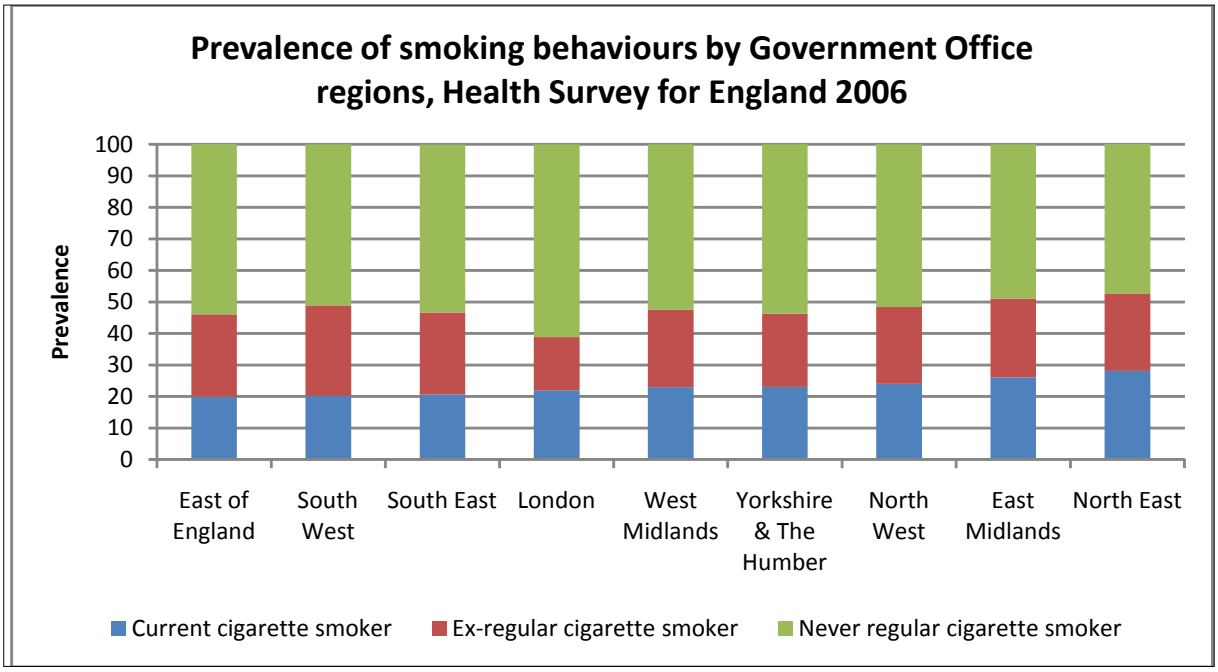


Figure 3: Prevalence of smoking behaviours by Government Office Region, 2006



The same smoking prevalence rates are applied across all ethnic categories. Future changes in smoking prevalence are not taken into account in the CHD prevalence projections. This is because of the uncertainty associated with predictions of smoking prevalence, and the lag time between smoking cessation and improved health. Even if there was a rapid drop in the number of smokers over the next few years, any associated decrease in CHD would not be seen for many years.

Deprivation

Deprivation scores are taken from IMD 2004. Deprivation scores for PCTs were calculated by taking a population weighted average of the scores for each Medium Super Output Area

(MSOA) (which in turn were calculated by taking a weighted average of the IMD2004 scores of each Lower Super Output Area (LSOA) within the PCT. Five deprivation categories are used in the model. Note that these categories are based on quintiles of IMD score at LSOA level. When the cut-offs are applied to larger geographies (LA or PCT) there is not an even distribution across the categories.

8.3 Stroke Model

8.3.1 Introduction

Prevalence of stroke

A report from WHO estimated stroke incidence and prevalence for each European country from routine mortality statistics[117]. Rates from studies that met the 'ideal' criteria were compared with WHO's estimates. Forty-four incidence studies and 12 prevalence studies were identified. WHO stroke estimates were in good agreement with results from 'ideal' stroke population studies. According to the WHO estimates, the number of stroke events in these selected countries is likely to increase from 1.1 million per year in 2000 to more than 1.5 million per year in 2025 solely because of the demographic changes.

In men, the lowest stroke prevalence rates are estimated for Cyprus, Lithuania, Poland, and Slovakia, whilst the highest rates are estimated for Czech Republic, Greece, Portugal, and Slovenia. In women, low prevalence rates are estimated for Cyprus, France, Lithuania, Poland, and Slovakia, whilst high prevalence rates are estimated for Czech Republic, Greece, Hungary, and Portugal. Table 16: UK stroke prevalence rates, estimates from the World Health Organization, men and women per 1000 population below shows the WHO prevalence estimates for the UK.

Table 16: UK stroke prevalence rates, estimates from the World Health Organization, men and women per 1000 population

	Men	Women
25–34	0.33	0.93
35–44	0.63	1.77
45–54	4.55	9.52
55–64	8.47	20.21
65–74	20.62	50.16
75–84	39.11	79.18
85+	56.39	93.15

Source: Truelsen T, Piechowski J, Bonita R, Mathers C, Bogousslavsky J, Boysen G. Stroke incidence and prevalence in Europe: a review of available data. European Journal of Neurology 2006; 13(6):581-598

Another international review of stroke epidemiology found nine high quality prevalence studies[118]. Overall, 8,788 strokes were reported, with age-specific prevalence increasing with age. The age-standardised prevalence for people aged 65 years or more ranged from 46·1 to 73·3 per 1000 population, but ranged from 58·8 to 92·6 per 1000 population for men, and from 32·2 to 61·2 per 1000 population for women. The small variation in age-specific and age-standardised prevalence of stroke across the populations (five to ten per 1000) was consistent with the geographical similarity in stroke incidence and case-fatality.

A systematic review of stroke epidemiology in South America found stroke prevalence rates ranging from 1.74 to 6.51 per 1000, and annual incidence rates from 0.35 to 1.83 per 1000[119]. Community-based studies showed crude stroke prevalence rates ranging from 1.74

to 6.51 per 1000 and annual incidence rates from 0.35 to 1.83 per 1000. In a population survey in Berlin of a total of 75,720 households (28,090 persons responded), a total of 4.5% reported a physician-diagnosed stroke (women 4.3%; men 4.9%)[120]. Combining reported stroke history with reported impaired vision and/or articulation problems, the prevalence of stroke increased to 7.6% (men 8.4%; women 7.2%).

There have been three recent published UK prevalence studies. A point prevalence study using postal questionnaires (n=18,000) in northern England found that prevalence increased with age and, apart from the very elderly, males had a higher prevalence than females[121]. Overall prevalence was found to be 46.8/1,000 (95% CI 42.5, 51.6). Full recovery from stroke was reported by 23% of respondents. Cognitive impairments (33%), problems with lower limbs (33% for right leg; 27% for left leg) and speech difficulties (27%) were the most common residual impairments.

In another UK prospective study, all incident cases of neurological disorders were ascertained in an unselected urban population based in 13 general practices in the London area[122]. A population of 100,230 patients registered with the practices was followed prospectively for the onset of neurological disorders, using multiple methods of case finding. Lifetime prevalence rates, expressed as rate per 1,000 persons with 95% CI), were: completed stroke, 9/1,000 (CI: 8, 11); and for transient ischaemic attacks, 5/1000 (CI: 4, 6).

A two-stage point prevalence study in Newcastle used a valid screening questionnaire to identify stroke survivors from a stratified sample[123]. This was followed by assessment of stroke patients with scales of disability and handicap. The overall prevalence of stroke was found to be 17.5/1,000 (95% CI 17.0-18.0). The prevalence of stroke-associated dependence was 11.7/1,000 (95% CI 11.3-12.1).

Table 17 and Table 18 show percentage prevalence of stroke in ethnic minority groups and the general population, and in various age groups, for the HSfE 2004[124]. Informants are classified as having IHD or stroke if they reported having angina, or a heart attack or a stroke, confirmed by a doctor. In summary, prevalence is estimated in these studies to be between one and five per cent, with the HSfE estimate falling midway in this range.

Table 17: Prevalence of stroke, by minority ethnic group and sex, aged 16 and over 2004

	Black Caribbean	Black African	Indian	Pakistani	Bangladeshi	Chinese	Irish	General population
Men								
Observed %	3.4	-	1.1	1.8	1.8	0.7	4.5	2.4
Standardised risk ratios	1.26	0.00	0.59	1.06	2.05	0.71	1.98	1
Standard error of the ratio	0.55	0.00	0.25	0.46	1.02	0.43	0.70	
Women								
Observed %	1.8	0.5	1.2	1.7	1.8	0.4	2.7	2.2
Standardised risk ratios	1.31	0.69	0.72	2.25	2.73	0.22	1.20	1
Standard error of the ratio	0.42	0.43	0.27	0.86	1.02	0.17	0.44	

Table 18: Prevalence of stroke, by age within minority ethnic group and sex, aged 16 and over 2004

	Men				Women			
	16-34	35-54	55+	All	16-34	35-54	55+	All
	%	%	%	%	%	%	%	
Black Caribbean	-	-	11.5	3.4	-	1.1	5.6	1.8
Black African	-	-	[-]	-	0.4	0.4	[1.5]	0.5
Indian	-	-	5.2	1.1	-	1.0	4.2	1.2
Pakistani	-	1.1	9.6	1.8	0.2	0.9	10.1	1.7
Bangladeshi	-	1.9	[9.2]	1.8	0.3	1.6	11.9	1.8
Chinese	-	0.8	2.2	0.7	-	0.5	0.8	0.4
Irish	-	2.2	9.4	4.5	0.9	0.6	6.3	2.7
General population (2003)	0.3	0.7	6.4	2.4	0.3	0.7	5.2	2.2

Source: HSfE 2004

The GP practice ascertainment study in South London may therefore have underestimated prevalence because practices were unaware of a proportion of stroke victims. This is consistent with 2006-7 QOF data, which shows an observed/registered crude prevalence (whole population denominator) of only 1.61 per cent.

Incidence and outcomes of stroke

In the review quoted above, eight population-based studies assessing secular trends in stroke incidence in a given population were identified[118]. Although the studies covered different periods, several common themes were evident. Most of the studies showed a decline in stroke incidence, through to the late 1970s or early 1980s. In several studies, however, this decline seemed to have reached a plateau or even reversed in the late 1980s and early 1990s. Of the few population-based studies that reported time-trend data for stroke mortality, the consistent finding was of a decrease in rates from the 1970s through to the 1990s. About half to three-quarters of cases had stroke-related disability.

In South London, stroke incidence decreased over a 10-year time period[125]. The greatest decline in incidence was observed in black women, but ethnic group disparities still exist, indicating a higher stroke risk in black people compared to white people. Total stroke incidence was higher in blacks compared to whites (IRR 1.27, 95% CI 1.10-1.46 in men; IRR 1.29, 95% CI 1.11 to 1.50 in women), but the black-white gap reduced during the 10-year time period (IRR 1.43, 95% CI 1.13-1.82 in 1995 to 1996 to 1.18, 95% CI 0.93-1.49 in 2003 to 2004). Age- and sex-adjusted IRRs for haemorrhagic stroke were higher in Black Africans (IRR, 2.80; 95% CI, 2.00 to 3.91) than in Black Caribbeans (IRR, 1.46; 95% CI, 1.07 to 1.99) which could be explained by prestroke hypertension being more common among young blacks[126].

The South London study also used capture-recapture models including covariates to estimate incidence of stroke[127]. This suggested that the stroke register was 88% complete. Adjusting for under-ascertainment increased the estimated incidence from 1.31 (95% CI : 1.21-1.42) to 1.49 (95% CI : 0.38-2.60) per 1000.

Previous stroke prevalence modelling

There has apparently been very little previous published modelling of stroke prevalence. A simple prevalence model was included in the ASSET stroke package released by the Department of Health, but it is no longer available. It used only South London study data.

8.3.2 Methods

Data Sources

The stroke model described here uses data from the 2003 and 2004 Surveys. The 2003 HSfE contains data on a total of 18,553 individuals. The 2004 data consists of two individual level files of which one (10,114 records) contains data for all individuals in the Ethnic Boost Sample and informants in the General Population Sample who were of the specified ethnic groups in co-operating households who gave a full interview. It contains information from the household questionnaire, main individual schedule, self-completions and the nurse visit (where one occurred). Data on under 16s was dropped from both these raw data files for the regression modelling.

Because the HSfE showed that stroke prevalence varies with ethnicity[1], it was necessary to use a sample containing data from a large number of ethnic minority respondents. The HSfE 2004 was the last Survey to include an ethnic minority boost, and the boost sample was used for the modelling. However there were relatively small numbers of Whites in the HSfE 2004 sample, and BP was not measured for most respondents in the 2004 general population sample, presumably to save resources for the boost itself. The HSfE 2004 boost sample was therefore merged with the HSfE 2003 data, which measured BP and which was the year with the largest number of identical variables. Table 19 shows the ethnic group breakdown of the two samples. Note that it was necessary to collapse two of the HSfE 2003 ethnic group variables to use the same classification as HSfE 2004.

Table 19: Ethnic Group Breakdown of HSfE 2003 & 2004 Datasets & Merged 2003-2004 Dataset

	HSfE 2003		HSfE 2004		HSfE 2003+2004	
	Freq.	**Percent**	**Freq.**	**Percent**	**Freq.**	**Percent**
white	13,445	92.11	1,130	16.98	14,575	68.64
mixed ethnic group	86	0.59	222	3.34	308	1.45
black or black british	308	2.11	1,683	25.29	1,991	9.38
asian or asian british	594	4.07	3,131	47.04	3,725	17.54
any other group	163	1.12	472	7.09	635	2.99
Total	**14,596**	**100**	**6,638**	**100**	**21,234**	**100**

NB excludes no answer/refused/don't know

Model construction: data issues

The choice of variables for original inclusion in the merged dataset included all those known to be stroke risk factors. The variable names and labels are shown in below. The HSfE dataset has a nested or hierarchical structure so three variables related to the sampling strata were included: area (sample point), cluster (stratification level), and hserial (serial number of household). These were used in the model to adjust for clustering of

respondents. In the analysis variables cholest and hdlchol were combined to give a lipid ratio.

There was a problem with merging the variable for deprivation (IMD 2004). The bandings of IMD scores were slightly different between the two years, but raw scores were not provided in the dataset so that it was necessary to assume identity (see Table 20 below).

Table 20: Index of Multiple Deprivation Banding

Rank	IMD Band	IMD		Number	Per Cent	Cum Per Cent
		HSfE 2003	HSfE 2004			
least	1	0.59-8.35	0.55-9.02	3,803	17.87	17.87
	2	8.35-13.72	9.03-14.14	3,573	16.79	34.65
	3	13.72-21.16	14.15-21.17	3,788	17.8	52.45
	4	21.16-34.21	21.18-33.52	4,551	21.38	73.83
most	5	34.21-86.36	33.53-85.69	5,571	26.17	100
			Total	**21,286**	**100**	

The Stata software package was used for analysis. The Stata software package was used for analysis. As for the CHD model the methodology applied was multinomial logistic regression with the "cluster" option (see above). As previously, the prevalence in each age group, gender, ethnic group, area of residence and level of deprivation, and smoking status category were derived from the odds, using the formula: prevalence = odds/(1 + odds).

Model construction: interactions between variables

Effect modification or interaction occurs if the effect of one exposure or risk factor on the outcome varies according to the level of another risk factor. This can be tested using a χ^2 test of heterogeneity, or by introducing interaction terms or parameters into the regression model. These allow the effect of one variable to be different in different categories of other variables. In Stata the xi command expands terms containing categorical variables into indicator (also called dummy) variable sets by creating new variables and estimates interactions and main effects.

Model construction: validation

Ideally, the best prediction should result from utilising the most information in the regression model. However only a limited range of HSfE variable data is either available or can be estimated at the PCT level. We decided to validate the local model by comparing it, in terms of prediction, to a model including all available HSfE variables. In addition, however, the amount of missing data affects the prediction of a model. For example, in the complete HSfE variables the largest proportion of missing data occurred in those variables related to particular drug treatment for high blood pressure. Some clinical tests also have a high proportion of missing values. Hence the "complete" model included the complete list of variables, including differing definitions of high blood pressure (history of high blood pressure, on drugs for high blood pressure, or clinically measured hypertension >140/90 at the time of the Survey.

The local model only used locally available data. I included smoking because local synthetic estimates are now available. Local General Health Questionnaire 12 (GHQ-12) or Limiting Longstanding Illness score data will be available locally from 2009 and could be included later.

8.3.3 Results

Table 21 shows crude stroke prevalence in the HSfE 2003-4 merged dataset.

Table 21: crude stroke prevalence in 2003-4 dataset

	Age 16-24	Age 25-34	Age 35-44	Age 45-54	Age 55-64	Age 65-74	Age 75+	Total
No stroke	2,498	3,850	4,429	3,435	2,972	2,131	1,476	20,791
%	99.84	99.72	99.48	98.82	97.31	93.51	88.86	97.68
Stroke	4	11	23	41	82	148	185	494
%	0.16	0.28	0.52	1.18	2.69	6.49	11.14	2.32
Total	2,502	3,861	4,452	3,476	3,054	2,279	1,661	21,285

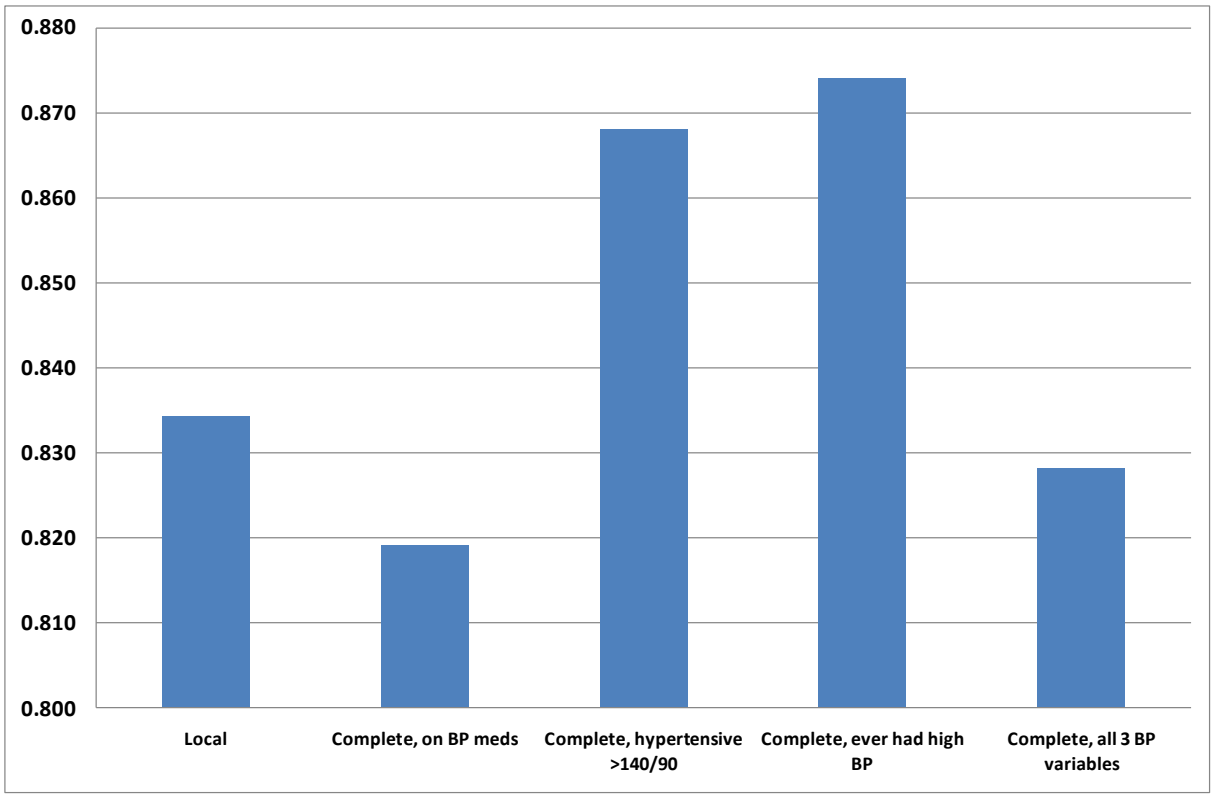
Sensitivity/Specificity: ROC Curve

Sensitivity and specificity were calculated using the area under the ROC curve. Area under ROC curves for the local and complete models tested were estimated using Stata10. These are shown in Table 22 and Figure 4 below. The local model, with an AUROC of 0.834 exceeds this level, although the complete model using a history of high blood pressure has even better performance, with an AUROC of 0.874. This data would also be easiest to estimate at a local level (possibly using the hypertension model). The table illustrates well the impact of missing data on area under ROC curve. Exact binomial and Hanley confidence intervals were calculated. Hanley specifies that the standard error for the area under the ROC curve be calculated using the method suggested by Hanley and McNeil. Otherwise, standard errors are obtained as suggested by DeLong and Clarke-Pearson.

Table 22: Comparison of area under ROC curves for differing stroke models

Model/BP Variable(s)	ROC Test	Obs (n)	Area under ROC	Std. Err.	Lower 95% C. I.	Upper 95% C. I.
Local	Binomial Exact	21,136	0.834	0.009	0.829	0.839
	Hanley-Asymp Normal	21,136	0.834	0.009	0.818	0.851
Complete, 3 BP variables	Binomial Exact	4,915	0.828	0.015	0.817	0.839
	Hanley-Asymp Normal	4,915	0.828	0.015	0.799	0.858
Complete, on BP meds	Binomial Exact	6,003	0.819	0.014	0.809	0.829
	Hanley-Asymp Normal	6,003	0.819	0.014	0.793	0.846
Complete, ever had high BP	Binomial Exact	16,669	0.874	0.010	0.869	0.879
	Hanley-Asymp Normal	16,669	0.874	0.010	0.854	0.894
Complete, hypertensive >140/90 Omron	Binomial Exact	10,227	0.868	0.013	0.861	0.875
	Hanley-Asymp Normal	10,227	0.868	0.013	0.843	0.894

Figure 4: Comparison of area under ROC curves for differing models



Local Model

ORs, p values and confidence intervals are generally similar to the more complete models (see Table 23: Odds Ratios, Local Model). Variables which should be included as inputs to the local model are highlighted. Unfortunately, however, local synthetic estimates do not have separate categories for occasional/regular smokers.

Table 23: Odds Ratios, Local Model

	RRR	Std. Err.	z	P>z	Lower 95% CI	Upper 95% CI
Female	1.000					
Male	1.144	0.111	1.39	0.166	0.946	1.383
Age 16-24	1.000					
Age 25-34	1.712	0.999	0.92	0.357	0.545	5.375
Age 35-44	3.254	1.764	2.18	0.03	1.124	9.418
Age 45-54	7.329	3.852	3.79	0	2.616	20.532
Age 55-64	17.652	9.092	5.57	0	6.432	48.443
Age 65-74	43.780	22.307	7.42	0	16.128	118.846
Age 75+	82.058	41.833	8.65	0	30.212	222.878
Never smoker	1.000					
Used to smoke occasionally	0.947	0.212	-0.24	0.809	0.611	1.469
Used to smoke regularly	1.353	0.150	2.72	0.006	1.088	1.682
Current smoker	1.389	0.189	2.41	0.016	1.063	1.814
IMD 2004 0.59->8.35 [least deprived]	1.000					

	RRR	Std. Err.	z	P>z	Lower 95% CI	Upper 95% CI
IMD 2004 8.35->13.72	1.233	0.199	1.3	0.193	0.899	1.692
IMD 2004 13.72->21.16	1.236	0.199	1.32	0.188	0.902	1.694
IMD 2004 21.16->34.21	1.397	0.217	2.15	0.032	1.030	1.895
IMD 2004 34.21->86.36 [most deprived]	2.009	0.291	4.83	0	1.513	2.668

Table 24 shows odds ratios and other statistics for the complete model, using history of high blood pressure as the variable with the highest area under ROC curve. Included variables were selected by reverse stepwise selection using repeated likelihood ratios and Wald tests as recommended by Kirkwood and Sterne. A threshold probability value of 0.2 was used (see Table 25).

Table 24: Odds Ratios, Complete Model

	RRR	Std. Err.	z	P>z	Lower 95% CI	Upper 95% CI
Age 16-24	1.000					
Age 25-34	1.434	0.853	0.61	0.544	0.447	4.599
Age 35-44	1.785	1.003	1.03	0.303	0.593	5.369
Age 45-54	2.907	1.591	1.95	0.051	0.995	8.499
Age 55-64	5.724	3.040	3.29	0.001	2.022	16.207
Age 65-74	15.196	8.026	5.15	0	5.397	42.787
Age 75+	21.734	11.547	5.8	0	7.672	61.571
Limiting long-term illness	1.000					
Non limiting long-term illness	0.387	0.065	-5.6	0	0.278	0.539
No long-term illness	0.291	0.051	-7.1	0	0.207	0.410
IMD 2004 0.59->8.35 [least deprived]						
IMD 2004 8.35->13.72	1.305	0.250	1.39	0.165	0.896	1.900
IMD 2004 13.72->21.16	1.008	0.202	0.04	0.968	0.680	1.494
IMD 2004 21.16->34.21	1.012	0.202	0.06	0.951	0.685	1.497
IMD 2004 34.21->86.36 [most deprived]	1.322	0.259	1.42	0.154	0.900	1.942
Never had high BP						
Ever had high BP	2.767	0.385	7.31	0	2.106	3.635
Never smoker						
Used to smoke occasionally	1.012	0.305	0.04	0.967	0.561	1.828
Used to smoke regularly	1.392	0.209	2.2	0.028	1.037	1.868
Current smoker	1.638	0.296	2.74	0.006	1.150	2.333
BMI under 20						
BMI 20-25	0.822	0.278	-0.6	0.561	0.424	1.594
BMI 25-30	0.494	0.169	-2.1	0.039	0.252	0.966
BMI Over 30	0.661	0.225	-1.2	0.224	0.339	1.289
Female sex						
Male sex	1.187	0.150	1.36	0.175	0.926	1.522

	RRR	Std. Err.	z	P>z	Lower 95% CI	Upper 95% CI
GHQ12 Score=0 [best psychosocial health]						
GHQ12 Score=1	1.369	0.238	1.8	0.071	0.973	1.926
GHQ12 Score=2	1.149	0.294	0.54	0.587	0.696	1.898
GHQ12 Score=3	1.332	0.361	1.06	0.291	0.783	2.267
GHQ12 Score=4	2.192	0.613	2.81	0.005	1.267	3.791
GHQ12 Score=5	2.710	0.746	3.62	0	1.580	4.647
GHQ12 Score=6	1.614	0.623	1.24	0.215	0.758	3.439
GHQ12 Score=7	1.686	0.658	1.34	0.181	0.785	3.621
GHQ12 Score=8	1.307	0.642	0.54	0.586	0.499	3.424
GHQ12 Score=9	3.814	1.267	4.03	0	1.989	7.312
GHQ12 Score=10	2.586	1.164	2.11	0.035	1.070	6.250
GHQ12 Score=11	1.700	0.843	1.07	0.285	0.643	4.495
GHQ12 Score=12 [worst psychosocial health]	0.969	0.547	-0.1	0.956	0.321	2.930

Table 25: complete model variable selection

Variable Name	LR Test		Wald Test	
	χ^2	Pr	χ^2	Pr
Age Group	741.25	0	522.16	0
IMD 2004	27.3	0	27.99	0
Limiting illness	64.46	0	58.13	0
Hypertensive	14.55	0.0001	13.92	0.0002
Ever had high blood pressure	10.99	0.0009	10.92	0.001
On drugs for BP	10.41	0.0013	9.85	0.0017
Smoking	12.33	0.0063	12.24	0.0066
BMI	9.78	0.0205	10.01	0.0185
Sex	2.69	0.1012	2.69	0.1008
GHQ 12 Screen	18.3	0.1068	20.75	0.0542

Internal validation: prediction

Prediction by the models is shown in Table 26 below.

Table 26: Comparison of predictions by model

	Observed	Predicted no stroke	Predicted stroke	Total
Local	No stroke	20,740	0	20,740
	Stroke	493	0	493
Complete, 3 BP variables	No stroke	4,728	1	4,729
	Stroke	183	3	186
Complete, hypertensive	No stroke	10,029	0	10,029

	Observed	Predicted no stroke	Predicted stroke	Total
	Stroke	198	0	198
Complete, ever had high BP	No stroke	16,366	1	16,367
	Stroke	302	0	302
Taking drugs prescribed for BP	No stroke	5,772	0	5,772
	Stroke	230	1	231

8.3.4 Application: PCT, LA and practice level models

The same practice, LA and PCT population data used for other models was used here. In order to calculated estimate prevalence of hypertension in the future, population projections were incorporated into the model. Modelled estimates of disease prevalence at practice level were provided for the adult population (aged 16+ years) using a similar methodology to previous models. By assuming that the prevalence of disease is 0% in the under 16s, prevalence estimates for all ages were calculated. These all-age prevalence are more closely equivalent to QOF prevalence. For each practice the model required the following input data:

- Population by age and sex
- Distribution of ethnic groups by age and sex (not required for stroke model)
- Prevalence of smoking and ex-smoking by age and sex (not required for hypertension model)
- IMD deprivation score
- Urban-suburban-rural category (only required for COPD model)

The following assumptions are made in the modelling process. For some practices, particularly those with a large BME population, these assumptions may not hold and the model may not be an accurate reflection of true prevalence.

- The distribution of ethnic groups within each age-sex group is the same as the overall distribution of ethnicity for the practice.
- Smoking prevalence is the same across ethnic groups
- Ex-smoking prevalence by age group and sex is the same in all practices.
- Disease prevalence in population aged <16 is 0%.

The urban-suburban-rural categorisation of the MSOA where the practice is located is representative of the catchment population of the practice. Only practices with a complete input dataset could be included in modelling. Table 27 summarises the number of practices included in models for 2007.

Table 27: Number of practices included in modelling

	Number of practices with valid data
CHD model	8178
Stroke model	8180
Hypertension model	8178
COPD model	8170

Data sources

As before practice populations by age and sex were derived from a National Strategic Tracing Service (NSTS) extract taken on 31/03/2007. Patients with date of death before 31/03/07 were excluded. In order to improve the data quality, GP codes were re-mapped to practices using NSTS files of February 2007 (available from the NHS Organisation Data Service). Some practices have missing PCT and SHA codes in the Feb 07 NACS files, so these were populated using the May 07 NSTS files. Practices with a population of <1000 patients are excluded from modelling due to the uncertainty associated with such small numbers.

Ethnicity

The proportion of practice population in ethnic groups was supplied by the CQC, as for the other models. The proportions were derived from HES admissions in 2005/06-2006/07, excluding mental health and maternity admissions and records with unknown ethnicity. The proportions by ethnic group for each practice were calculated by dividing admissions within each ethnic category by the total admissions for the practice.

Deprivation

This model used IMD 2007 scores for practices, derived by South East PHO in August 2008. These scores are weighted MSOA IMD scores based on the March 2008 Attribution Data Set. There are 40339 registrations (0.07 per cent of the total of 57,374,247 for England and Wales) for which the postcode was either missing, obviously invalid, or failed to generate all the required area codes when matched to recent postcode directories. These registrations were re-assigned to the postcode geography of the lists of their practice, retaining the total number of registrations in the file. The resulting file of raw registrations was aggregated to give the numbers of registrations per practice per LSOA. This was matched to a file of IMD domain scores for LSOAs to compute weighted average scores for each practice.

Smoking status

Practice smoking prevalence is based on National Centre for Social Research (NatCen) synthetic estimates of smoking at MSOA level 2003-2005 (available from Neighbourhood Statistics). Practice smoking prevalence was calculated as the weighted average of prevalence in MSOAs in the practice catchment area, determined from postcodes in cleaned 2007 NSTS data (as above) for patients aged 16+. MSOA catchment weighting was limited to patients in England only (e.g. practices with some catchment in Wales have smoking prevalence derived from English part of catchment only). The prevalence models require smoking status including prevalence of ex-smoking and non-smoking. However, this information is not available at practice level so national (England) proportions of smokers, ex-smokers and current smokers by age and sex are taken from HSfE 2003-2005 pooled as for previous models and adjusted for each practice.

Rurality

The Urban-suburban-rural categorisation is based on the Rural and Urban area classification of Super Output Areas, 2004 (Table 28, available from Neighbourhood Statistics).

Table 28: Definition of urban-suburban-rural categories

Urban	DEFRA MSOA category "Urban >10k" AND DEFRA LA category LU or MU
Suburban	(DEFRA MSOA category "Urban >10k" AND DEFRA LA category not LU or MU) OR DEFRA MSOA category "Town and Fringe"
Rural	DEFRA MSOA category "Village Hamlet and isolated dwelling"

This definition gives a distribution of categories that differs from the 2001 HSfE distribution of LAs used in the model development (Table 29), but there is no other logical definition available.

Table 29: Distribution of urban-suburban-rural categories

Category	MSOA distribution	LA distribution used in COPD model
Rural	10.09%	23%
Suburban	42.04%	49%
Urban	47.87%	27%

The urban-suburban-rural category of the MSOA in which the practice is physically located is used in the COPD model.

8.4 Hypertension Model

8.4.1 Introduction

Hypertension is defined as a persistently raised blood pressure (BP) above a designated threshold (which has reduced over time). The current NICE guidance[128] defines this further as follows:

- measured at the past two GP visits i.e. an average and
- systolic or diastolic pressure (SBP/DBP) or both are above 140/90 mmHg.

Prevalence Studies

Hypertension is an important public health challenge in both economically developing, and developed, countries. An international review estimated that 26·4% of the adult population in 2000 had hypertension[129]. The estimated total number of adults with hypertension in 2000 was 972 million, of which 333 million are in economically developed countries. The number of adults with hypertension in 2025 was predicted to increase by about 60% to a total of 1·56 billion. Worldwide, hypertension caused 7.6 million premature deaths in 2001[130]. Hypertension is one of the main risk factors for heart attack and stroke[131]. Throughout middle and old age, usual blood pressure is strongly and directly related to vascular (and overall) mortality, without any evidence of a threshold down to at least 115/75 mm Hg[132].

Numerous national and regional studies have been conducted worldwide to estimate the prevalence of hypertension: hypertension rates have varied from 3.4% in rural Indian men to as high as 72.5% in Polish women[133;134]. Within developed countries, average BP is higher in Europe than North America, even after adjusting for treatment differences[135]. A meta-analysis of national studies showed a mean BP of 136/83 mm Hg in six European countries and 127/77 mm Hg in Canada and the United States among men and women combined who were 35 to 74 years of age. This difference already existed among younger persons (35-39 years) in whom treatment was uncommon (i.e., 124/78 mm Hg and 115/75 mm Hg, respectively), and the slope with age was steeper in the European countries. For all age groups, BP measurements were lowest in the United States and highest in Germany. The age- and sex-adjusted prevalence of hypertension was 28% in the North American countries and 44% in the European countries at the 140/90 mm Hg threshold.

Hypertension is more common among African Caribbeans and South Asians in Britain. The table below from HSfE 2004 report shows that Black Africans and Caribbeans have the highest standardised risk ratios for hypertension (see Table 30). Many studies in the UK and elsewhere have also shown that mortality rates for stroke, CHD, and end-stage renal failure are higher in South Asians than white Europeans[133;136]. The NICE guidelines recommend a reduction of BP to <140/85, but give lower target BP levels (<130/80) for patients with diabetes or chronic renal failure[128]. The majority of UK-based surveys have shown that people of Black African or Black Caribbean origin have higher average BP levels and rates of hypertension than their white counterparts. Socioeconomic disadvantage was associated with the greatest risk of developing hypertension in whites and in blacks[137-139]. Aging-related increases in systolic blood pressure were inversely associated with socioeconomic position in whites.

There is strong evidence, including a systematic review and a meta-analysis, from diverse populations showing that BP "tracks" from childhood into adulthood. Childhood BP is associated with BP in later life, and early intervention is important[137;140].

Table 30: blood pressure levels, by minority ethnic group and sex, Aged 16 and over with valid blood pressure readings, 2004

	Minority ethnic group					Gen population
	Black Caribbean	Black African	Pakistani Bangladeshi	Chinese	Irish	
Me						
Observed						
Normotensive untreated[b]	62	75	80	8	6	68
Hypertensive controlled[b]	10	4	6	4	7	5
Hypertensive uncontrolled[b]	10	5	2	5	7	6
Hypertensive untreated[b]	19	16	11	1	2	20
All with hypertension	38	25	20	2	3	32
Hypertensive untreated[c]	2	4	2	2	5	5
Standardised risk						
All with hypertension	1.37	1.21	0.98	0.7	1.1	1
Standard error of the ratio	0.19	0.22	0.19	0.	0.	
Wome						
Observed						
Normotensive untreated[b]	68	81	85	8	7	71
Hypertensive controlled[b]	9	5	4	6	4	6
Hypertensive uncontrolled[b]	10	2	5	3	9	8
Hypertensive untreated[b]	13	12	5	7	1	16
All with hypertension	32	19	15	1	2	29
Hypertensive untreated[c]	1	4	3	2	3	5
Standardised risk						
All with hypertension	1.58	1.71	1.01	1.1	0.9	1
Standard error of the ratio	0.19	0.37	0.18	0.	0.	
Bases (weighted)						
Men	169	136	159	6	66	4420
Women	249	183	207	6	92	4702
Bases						
Men	155	123	162	15	24	4108
Women	243	154	207	16	32	5075

Source: HSfE 2004

Effects of reducing BP

Epidemiological, clinical and animal experimental evidence show a direct relationship between dietary electrolyte consumption and blood pressure. Clinical trials have shown that reduction of salt (sodium chloride, NaCl) intake reduces BP levels in normotensive and hypertensive populations and prevents the development of hypertension. Data from randomized controlled trials, in which mean age at event was approximately 70 years, indicate that a 10 mm Hg reduction in systolic BP is associated with a reduction in risk of stroke of approximately one third[131]. The results of an international study provide evidence that antihypertensive treatment in persons 80 years of age or older is beneficial[141].

Drug treatment is recommended in all people with sustained levels of blood pressure ≥160/100, and at lower levels in those with additional risk factors, such as diabetes, or with end-organ damage[132]. The threshold for offering drug treatment is defined by NICE[128] as:

- blood pressure of more than 160/100 mmHg or
- isolated systolic hypertension (systolic blood pressure of more than 160 mmHg) or
- blood pressure of more than 140/90 mmHg and:
    - 10-year CVD risk of at least 20%, or
    - existing CVD or target organ damage.

Prevalence of treated and controlled hypertension

The experience of high-income countries indicates that sustained interventions can achieve at least the required 4% annual average decline in stroke mortality for people age 60-69 years[142]. In Canada, 16% of participants were treated and controlled; 23% were treated and not controlled; 19% were not treated and not controlled; and 42% were unaware of their hypertension[143]. In a French population, more than two-thirds of those with hypertension were aware of their diagnosis and 81% were treated with antihypertensive drugs[144]. In the US, among adults with hypertension, 83.0% were diagnosed, 72.7% were treated, and 47.1% had hypertension controlled. Of those treated, 64.8% had hypertension controlled[145]. These results seem superior to UK figures in the table above. Lower treatment thresholds and more intensive treatment contribute to better hypertension control in the United States compared with the western European countries (including the UK) studied[146].

Defining modelled prevalence

The concept of "pre-hypertension" is gaining credence in the USA. The Seventh Report of the US Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure provides a new classification of blood pressure levels[147]. According to the new report, normal BP is defined as systolic BP (SBP) less than 120 mm Hg and diastolic BP (DBP) less than 80 mm Hg; a SBP of 120 to 139 mm Hg or a DBP of 80 to 89 mm Hg is defined as pre-hypertension. The low-end threshold for pre-hypertension is lower than the previous designation of high-normal BP (i.e., SBP/DBP: 130/85. Approximately 60% of American adults (67% of men and 50% of women) have pre-hypertension or hypertension and 27% had hypertension.

However, there is little separation between the distributions of the risk factors in people who over a specified period do or do not have a disease event. With such closely overlapping distributions there are no cut-off levels that include most people who will have disease events but few of those who will not have them, but rather there is a spectrum of risk as BP increases[148]. In addition, management of overall CVD risk rather than managing individual risk factors is now recommended. The definitions used by NICE and the HSfE have therefore been used for this model. The prevalence of various categories of hypertension is shown in Table 31 below, taken from the HSfE 2003.

Table 31: Prevalence of Hypertensive Categories in HSfE 2003 (Adults aged 16 and over with a valid blood pressure reading and data on medication)

2003 (unweighted)[c]	16-24	25-34	35-44	45-54	55-64	65-74	75+	Total
Normotensive untreated	94.3	91.4	84.9	69.8	50.2	33.2	26.8	65.6
Normotensive treated	0.1	0.2	1.5	5.3	10.2	14.3	10.7	5.9
Hypertensive treated	-	0.3	1.0	3.5	11.5	19.4	26.1	8.0
Hypertensive untreated	5.5	8.0	12.6	21.4	28.1	33.0	36.4	20.5
All with high blood pressure	5.7	8.6	15.1	30.2	49.8	66.8	73.2	34.4
2003 (weighted)[b]								
Normotensive untreated	93.9	90.7	84.1	69.6	49.3	33.1	27.1	67.9
Normotensive treated	0.1	0.2	1.6	5.3	10.5	14.9	10.4	5.4
Hypertensive treated	-	0.4	1.0	3.4	11.9	19.5	26.9	7.3
Hypertensive untreated	6.0	8.8	13.3	21.7	28.3	32.6	35.5	19.4
All with high blood pressure	6.1	9.3	15.9	30.4	50.7	66.9	72.9	32.1
Bases								
2003 (unweighted)	849	1272	1800	1536	1625	1194	907	9183
2003 (weighted)	1201	1460	1765	1457	1346	1032	861	9122

Previous hypertension prevalence modelling

In the US, self-reported data on hypertension diagnosis from the Behavioral Risk Factor Surveillance System was used to obtain unbiased state-level estimates of blood pressure and uncontrolled hypertension as benchmarks for priority setting and for designing and evaluating intervention programmes[149]. A crude epidemiologic hypertension model based on age and sex-specific prevalence was first developed in the UK by the Faculty of Public Health. Subsequently, English PCTs were required to set targets for hypertension case-finding in their 2007-8 Local Delivery Plans negotiated with strategic health authorities. To assist them a simple PCT-based prevalence model was developed rapidly by APHO.

In the APHO model, numbers of persons predicted to be hypertensive were derived by multiplying April 2006 PCT registered populations by hypertension prevalence rates identified in the 2003 and 2004 HSfE, modified by ethnic-group age-standardised risk ratios from the 2004 HSfE. Calculations were stratified to reflect variations in these factors at PCT level. The PCT-registered populations were derived by aggregating April 2006 GP practice populations from the Exeter System using a practice to new PCT lookup table. In the absence of age by sex by ethnic-group PCT populations, the age by sex registered populations of the current PCTs were attributed the ethnic-group distributions of their constituent former PCT/s resident populations from the 2001 Census.

APHO accepts that this current model is rather crude, and that a prevalence model based on a comprehensive regression model using HSfE data would be more robust. This model was commissioned from the Department of PCPH Imperial College London.

8.4.2 Methods

Data Sources

The hypertension model described here uses data from the 2003 and 2004 HSfEs. The 2003 data in file hse03ah.dta contains data on 18,553 individuals. The 2004 data consists of two

individual level files and one household level file: HSfE04gpa.sav (8,354 records) contains data for all individuals in the General Population Sample in co-operating households who gave a full interview. This was not used for the reasons outlined above. HSfE04etha.sav (10,114 records) contains data for all individuals in the Ethnic Boost Sample and informants in the General Population Sample who were of the specified ethnic groups in co-operating households who gave a full interview. It contains information from the household questionnaire, main individual schedule, self-completions and the nurse visit (where one occurred). Data on under 16s was dropped from both these raw data files for the regression modelling.

In HSfE 2003, a new oscillometric automated device, the Omron HEM 907, was introduced to measure blood pressure, as a replacement for the Dinamap 8100, which had become obsolete. HSfE 2004 measures blood pressure as a one-off measurement in a cross sectional survey and does not provide information on whether high BP is sustained over time (the clinical definition in the NICE guidance), so this survey may slightly overestimate the prevalence of hypertension. BP measurement by GPs or practice nurses may, however, increase BP slightly (so-called "white coat hypertension"). The HSfE 2004 report introduced a change to a clearer terminology from previous years' reports. The specific definitions of the four levels used in the HSfE 2004 report are:

> **Normotensive untreated** SBP <140mmHg and DBP <90mmHg and not taking medicine prescribed for high blood pressure
>
> **Hypertensive controlled** SBP <140mmHg and DBP <90mmHg and taking medicine prescribed for high blood pressure (in 2003 this was called "normotensive-treated")
>
> **Hypertensive uncontrolled** SBP ≥140mmHg and/or DBP ≥90mmHg and taking medicine prescribed for high blood pressure (in 2003 this was called "hypertensive-treated")
>
> **Hypertensive untreated** SBP ≥140mmHg and/or DBP ≥90mmHg and not taking medicine prescribed for high blood pressure (in 2003 this was called "hypertensive-untreated")

However the actual definition did not change.

Because hypertension prevalence is known to vary with ethnicity, it was necessary to use a sample containing data from a large number of ethnic minority respondents. The HSfE 2004 was the last Survey to include an ethnic minority boost, and the boost sample was used for the modelling. However there were relatively small numbers of Whites in the HSfE 2004 sample, and BP was not measured for most respondents in the 2004 general population sample, presumably to save resources for the boost itself. The HSfE 2004 boost sample was therefore merged with the HSfE 2003 data, which measured BP and which was the year with the largest number of identical variables. Table 32 shows the ethnic group breakdown of the two samples. Note that it was necessary to collapse two of the HSfE 2003 ethnic group variables to use the same classification as HSfE 2004.

Table 32: Ethnic Group Breakdown of HSfE 2003 & 2004 Datasets & Merged 2003-2004 Dataset

	HSfE 2003		HSfE 2004		HSfE 2003+2004	
	Freq.	**Percent**	**Freq.**	**Percent**	**Freq.**	**Percent**
white	13,445	92.11	1,130	16.98	14,575	68.64
mixed ethnic group	86	0.59	222	3.34	308	1.45
black or black british	308	2.11	1,683	25.29	1,991	9.38
asian or asian british	594	4.07	3,131	47.04	3,725	17.54
any other group	163	1.12	472	7.09	635	2.99
Total	**14,596**	**100**	**6,638**	**100**	**21,234**	**100**

NB excludes no answer/refused/don't know

In HSfE 2003, valid blood pressure readings were obtained from 80.8% of men and 80.3% of women aged 16 and over who were visited by a nurse. Response rates increased with age for both sexes. The main reason for exclusion was that the informant ate, drank or smoked in the half hour before the measurement (17.0% among men and 15.2% among women). The remainder were excluded because they were pregnant, (4.9% of women aged 16-44 were excluded due to pregnancy) or because three valid readings had not been obtained or the measurement had been refused or not attempted.

In 2004, nurse visits were offered to all individuals in minority ethnic groups in the boost sample (NB the general population data presented in the HSfE 2004 report are actually from HSfE 2003, as the general population did not receive a nurse visit in 2004). Response rates for BP measurements and urine samples are based on those receiving a nurse visit, though women who were pregnant were not asked for a urine sample. Section 6.3 in Volume 2 of the 2004 report gives more details response rates to the different stages of the survey. Among those who participated in a nurse visit, the proportion with three valid blood pressure readings ranged from 69% of Bangladeshi men, to 84% of Chinese and Indian men. Among women, the proportion with three valid blood pressure readings ranged from 71% of Black African women to 84% of Indian women. As in 2003, informants were excluded if they were pregnant or had eaten, drunk, exercised vigorously or smoked in the half hour before the measurement; a proportion of the remainder refused to have their blood pressure taken or failed to complete the measurements or measurements were not possible for technical reasons. Table 33shows the numbers in each survey with valid BPs and their hypertensive categories.

Table 33: Hypertensive Categories of HSfE 2003 & 2004 Datasets & Merged 2003-2004 Dataset

	HSfE 2003		HSfE 2004		HSfE 2003+2004	
	Freq.	**Percent**	**Freq.**	**Percent**	**Freq.**	**Percent**
Refused/not applicable	5,569	38.1	3966	59.6	9,535	44.8
normotensive untreated	6,037	41.3	1,983	29.8	8,020	37.7
normotensive treated	579	4.0	180	2.7	759	3.57

	HSfE 2003		HSfE 2004		HSfE 2003+2004	
	Freq.	Percent	Freq.	Percent	Freq.	Percent
hypertensive treated	699	4.8	171	2.6	870	4.09
hypertensive untreated	1,746	11.9	356	5.4	2,102	9.88
Total with valid BP	**9,061**	**61.9**	**2,690**	**40**	**11,751**	**55.2**
Total	**14,630**	**100**	**6,656**	**100**	**21,286**	**100**

Model construction: data issues

The choice of variables for original inclusion in the merged dataset included all those known to be high blood pressure risk factors. The variable names and labels are shown in Table 34 below. The HSfE dataset has a nested or hierarchical structure so three variables related to the sampling strata were included: area (sample point), cluster (stratification level), and hserial (serial number of household). These were used in the model to adjust for clustering of respondents. In the analysis variables cholest and hdlchol were combined to give a lipid ratio.

Table 34: Variables Included in Merged Dataset

Variable	Description	Source
ADDRESS5	Address number	Sample
ADTOT30	Adults: Total days/ 4week active 30min + moderate	Derived
ADTOT30C	Adults: Total days per week active 30min + moderate	Derived
AG16G10	(D) Age 16+ in ten year bands	Derived
AGE	Age last birthday	Individual
BIRTWT	Birth weight (kg)	Individual
BMI	BMI - inc unreliable measurements	Derived
BMIVAL	Valid BMI	Derived
BMIVG4	Valid BMI (grouped:<20,20-25,25-30,30+)	Derived
BP1	Doctor diagnosed high blood pressure (excluding pregnant)	Derived
BPMEDC	Whether taking drugs affecting blood pressure	Derived
BPMEDD	Whether taking drugs prescribed for blood pressure	Derived
CIGST1	Cigarette Smoking Status - Never/Ex-reg/Ex-occ/Current	Derived
COOKSALT	Salt added at when cooking? SC 16+	Individual
CVDDEF	Had cardiovascular condition	Derived
CVDDEF1	Had cardiovascular condition (excluding diabetes/high BP)	Derived
EQMEAN	EQ-5D social preference weight (mean)	Derived
ETHCIND	Ethnic group	Individual
EVERBP	Ever had high BP	Individual
GHQ12SCR	GHQ Score - 12 point scale	Derived
HBP140OM	Whether hypertensive: 140/90: all prescribed drugs for BP (Omron reading)	Derived
HHOLD	Household	Sample
HIBP1OM	Whether hypertensive: all prescribed drugs for BP (Omron reading)	Derived
HY140OM	Hypertensive categories: 140/90: all prescribed drugs for BP (Omron reading)	Derived

Variable	Description	Source
HYPER1OM	Hypertensive categories: all prescribed drugs for BP (Omron reading)	Derived
IMD2004	Index of multiple deprivation (SOA level)	Derived
LIMITILL	Limiting longstanding illness	Derived
MEDCINBP	Take medicines for high BP	Individual
NOBPCVD	Had CVD excludes those with high BP	Derived
NSSEC5	NS-SEC 5 variable classification (individual)	Derived
NSSEC8	NS-SEC 8 variable classification (individual)	Derived
OMDIAST	Omron Diastolic BP (mean 2nd/3rd) inc. invalid	Derived
OMDIAVAL	Omron Valid mean diastolic BP	Derived
OMMAP	Omron Mean arterial pressure (mean 2nd/3rd) inc. invalid	Derived
OMMAPVAL	Omron Valid mean arterial pressure	Derived
OMPULS	Omron Pulse pressure, systolic-diastolic inc. invalid	Derived
OMPULVAL	Omron Valid pulse pressure	Derived
OMSYST	Omron Systolic BP (mean 2nd/3rd) inc. invalid	Derived
OMSYSVAL	Omron Valid mean systolic BP	Derived
POINT	Sample point	Sample
SERIALH	Serial number of household	Hhold
SERIALI	Serial number of individual	Individual
SEX	Sex	Individual
STRATA	Stratification level	Individual
TABSALT	Salt added at table? SC 16+	Individual
TOPQUAL2	(D) Highest Educational Qualification - Students separate	Derived
TOPQUAL3	(D) Highest Educational Qualification	Derived

There was a problem with merging the variable for deprivation (Index of Multiple Deprivation 2004). The bandings of IMD scores were slightly different between the two years, but raw scores were not provided in the dataset so that it was necessary to assume identity (see Table 35: Index of Multiple Deprivation Banding below).

Table 35: Index of Multiple Deprivation Banding

Rank	IMD Band	IMD		Number	Per Cent	Cum Per Cent
		HSfE 2003	HSfE 2004			
least	1	0.59-8.35	0.55-9.02	3,803	17.87	17.87
	2	8.35-13.72	9.03-14.14	3,573	16.79	34.65
	3	13.72-21.16	14.15-21.17	3,788	17.8	52.45
	4	21.16-34.21	21.18-33.52	4,551	21.38	73.83
most	5	34.21-86.36	33.53-85.69	5,571	26.17	100
			Total	**21,286**	**100**	

The Stata software package was used for analysis. As for the CHD model the methodology applied was multinomial logistic regression with the "cluster" option (see above). As previously the prevalence in each age group, gender, ethnic group, area of residence and

level of deprivation, and smoking status category were derived from the odds, using the formula: prevalence = odds/(1 + odds).

Model construction: interactions between variables

Effect modification or interaction occurs if the effect of one exposure or risk factor on the outcome varies according to the level of another risk factor. We tested for interactions using a χ^2 test of heterogeneity (Mantel-Haenszel Odds Ratio) and Wald test. Where significant interaction was found, we performed a regression analysis of a generated interaction term. Where evidence of interaction was found across a range of strata, we introduced interaction terms or parameters into the final regression model. These allow the effect of one variable to be different in different categories of other variables. In Stata the xi command expands terms containing categorical variables into indicator (also called dummy) variable sets by creating new variables and estimates interactions and main effects.

Model construction: internal validation

Ideally the best prediction should result from utilising the most information in the regression model. However only a limited range of HSfE variable data is either available or can be estimated at the PCT level. We decided to validate the local model by comparing it, in terms of prediction, to a model including all available and significant HSfE variables. In addition, however, the amount of missing data affects the prediction of a model.

The local model used only used locally available data. In future body mass index (BMI) could be included on the basis that local synthetic estimates are now available and good local data from primary care may soon be. Local Limiting Longstanding Illness score data will be available locally from 2009 and could also be included later.

Model construction: external validation

An early external validation will be carried out by examining the association between PCT/LA level high blood pressure prevalence estimates and QOF registered prevalence. The regression-based model will also be validated against a prevalence model obtained by Bayesian strategies using WinBUGS. Finally, funding will be sought to undertake a validation of practice-based prevalence estimates against registered prevalence supplemented by active case finding.

Model application

The model input variables, which could be defined by the relevant user, e.g. at PCT or practice level, or pre-populated, include age-group, gender, smoking prevalence by gender and age, area of residence, area based deprivation score and ethnic distribution of the population.

8.4.3 Results

In developed countries, increasing age is by far the most important risk factor for high blood pressure, although this is not the case in non-developed agrarian societies. The age-related increase in hypertension prevalence appears after migration to urban environments. Table 36 shows the increasing numbers of treated and hypertensive untreated respondents in the dataset.

Table 36: Hypertension Category By Age Band

	Age Band							
	16-24	**25-34**	**35-44**	**45-54**	**55-64**	**65-74**	**75+**	**Total**
normotensive untreated	1,047	1,663	2,132	1,408	979	498	293	8,020
normotensive treated	1	7	45	123	239	224	120	759
hypertensive treated	0	6	31	93	225	265	250	870
hypertensive untreated	49	125	283	400	487	417	341	2,102
Total	1,097	1,801	2,491	2,024	1,930	1,404	1,004	11,751

"Local" model

The regression model for hypertension risk factors for in the "local" prevalence model is shown in Table 37. As expected ORs increase strikingly with increasing age in all models. In the Excel-based prevalence predictions using coefficients (not shown in these tables), this results in age-related increases in prevalence which closely match those crude overall prevalence in Table 37. An age-sex interaction was found (see Figure 5: modelled proportion hypertensive by age group and sex) and so interaction terms for age and male sex were created. As expected a significant association was found between Black or Black British ethnicity and the outcome, and also with increasing deprivation.

Table 37: Odds ratios for variables included in "local" model

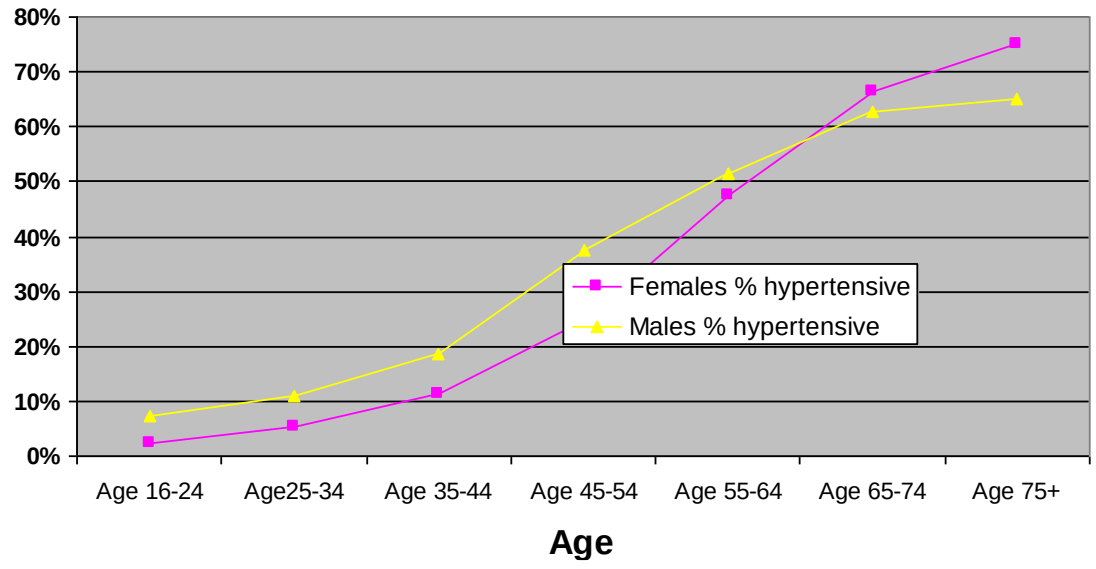
	RRR	**Std. Err. of RRR**	**z**	**P>z**	**[95% Conf.**	**Interval]**
Female Sex	1.000					
Male sex	3.197	1.010	3.680	0.000	1.721	5.937
Females Age 16-24	1.000					
Females Age 25-34	2.271	0.675	2.760	0.006	1.267	4.068
Females Age 35-44	5.201	1.431	5.990	0.000	3.034	8.917
Females Age 45-54	14.069	3.814	9.750	0.000	8.270	23.933
Females Age 55-64	39.730	10.720	13.650	0.000	23.413	67.420
Females Age 65-74	87.067	23.899	16.270	0.000	50.840	149.107
Females Age 75+	134.375	37.612	17.510	0.000	77.636	232.580
Interaction Male Sex & Age 16-24	1.000					
Interaction Male Sex & Age 25-34	0.684	0.251	-1.030	0.301	0.333	1.404
Interaction Male Sex & Age 35-44	0.572	0.191	-1.670	0.095	0.297	1.102
Interaction Male Sex & Age 45-54	0.574	0.190	-1.680	0.093	0.300	1.098
Interaction Male Sex & Age 55-64	0.370	0.121	-3.030	0.002	0.195	0.703
Interaction Male Sex & Age 65-74	0.265	0.089	-3.970	0.000	0.137	0.510
Interaction Male Sex & Age 75+	0.191	0.066	-4.800	0.000	0.097	0.376
White ethnic group	1.000					
Mixed ethnic group	0.920	0.256	-0.300	0.764	0.534	1.586

	RRR	Std. Err. of RRR	z	P>z	[95% Conf.	Interval]
Black or Black British ethnic group	1.706	0.174	5.240	0.000	1.397	2.083
Asian or Asian British ethnic group	1.054	0.083	0.660	0.509	0.902	1.230
Any other ethnic group	0.972	0.155	-0.180	0.858	0.711	1.328
Index of multiple deprivation 0.59-8.35	1.000					
Index of multiple deprivation 8.35-13.72	0.996	0.073	-0.060	0.953	0.862	1.150
Index of multiple deprivation 13.73-21.16	1.159	0.085	2.010	0.044	1.004	1.339
Index of multiple deprivation 21.17-34.21	1.210	0.092	2.520	0.012	1.043	1.404
Index of multiple deprivation 34.22-86.36	1.346	0.105	3.810	0.000	1.156	1.568

Interactions between variables

We tested for interactions between variables. Significant interactions were found on Mantel-Haenszel Odds Ratio and Wald tests for sex and ethnicity, and sex and age. We therefore generated interaction terms for these variables in a regression model. Only age and sex showed a number of significant associations across strata, for the three oldest age-sex categories ($p \leq 0.001$-0.003). We therefore included an interaction term in the final local model. This means that there are two separate effects of age on the outcome: one for males and another for females. The interaction terms shown represent the interaction between sex and age, and show the combined effect of being male and of a particular age group. The interaction is clearly demonstrated in the chart below of the prevalence of hypertension by age and sex. The prevalence of hypertension in females crosses over that in males in the older age groups.

Figure 5: modelled proportion hypertensive by age group and sex



"Complete" model

The regression model for risk factors for high blood pressure in the "complete" model with BP drugs is shown in Table 38: Odds ratios for Ntiles of IMD score (stepwise regression). Stepwise addition of variables in addition to age and sex was undertaken and only independent variables which were significant in this case were included in the final complete model. Variables which were eliminated through this process were:

- Cigarette Smoking Status - Never/Ex-reg/Ex-occ/Current
- Salt added at when cooking? SC 16+
- EQ-5D social preference weight (mean)
- GHQ Score - 12 point scale
- NS-SEC 5 variable classification (individual)

Although EQ-5D social preference weight (mean) and GHQ Score - 12 point scale were marginally significant when tested stepwise, they ceased to be so in the complete model and were excluded. On the other hand, although ORs increase with increasing deprivation. Index of multiple deprivation (SOA level) also ceased to be significant in the complete model, but has been retained as it is included in the local model. ORs for IMD score from stepwise regression are shown in Table 38 below for comparison, and are similar to the local model. This could be explained by collinearity with highest educational qualification in the complete model.

Table 38: Odds ratios for Ntiles of IMD score (stepwise regression)

	RRR	Std. Err.	z	P>z	[95% Conf.	Interval]
0.59->8.35 [least deprived]	1.000					
8.35->13.72	0.999	0.073	-0.02	0.987	0.865	1.153
13.72->21.16	1.178	0.086	2.24	0.025	1.021	1.360
21.16->34.21	1.267	0.095	3.15	0.002	1.093	1.467
34.21->86.36 [most deprived]	1.458	0.109	5.05	0	1.260	1.688

As expected, ORs increase markedly with increasing age in all models, with confidence intervals above 1 from age 35. Note that age group 16-24 has been defined as the baseline odds. There is a significantly lower overall OR for female sex and for most categories of increasing physical activity. This latter relationship has been described in numerous other studies and is one of the most important pathways through which physical activity reduces CVD risk. There is also a strong significant increase in ORs with increasing BMI. The result for Salt added at table is unexpected and counter-intuitive, with a significant increase in ORs for "rarely, or never, add salt at the table". There was a good response rate for this question (14,478/21,286).

Table 39 shows odds ratios for the "complete" model. ORs are significantly higher in the Black or Black British ethnic groups. There are some associations between measures of deprivation and educational level and the high blood pressure outcome.

Table 39: Odds ratios for variables included in "complete" model

	RRR	Std. Err.	z	P>z	[95% Conf.	Interval]
Females	1.000					
Males	3.173	1.057	3.47	0.001	1.651	6.096
Females Age 16-24	1.000					
Females Age 25-34	1.834	0.584	1.9	0.057	0.982	3.423
Females Age 35-44	4.101	1.202	4.82	0	2.309	7.284
Females Age 45-54	9.937	2.881	7.92	0	5.630	17.540
Females Age 55-64	26.757	7.746	11.35	0	15.171	47.191
Females Age 65-74	54.697	16.224	13.49	0	30.583	97.824
Females Age 75+	99.816	30.945	14.85	0	54.363	183.270
Interaction Male Sex & Age 16-24	1.000					
Interaction Male Sex & Age 25-34	0.787	0.306	-0.62	0.537	0.367	1.686
Interaction Male Sex & Age 35-44	0.589	0.208	-1.5	0.134	0.295	1.177
Interaction Male Sex & Age 45-54	0.600	0.210	-1.46	0.144	0.302	1.190
Interaction Male Sex & Age 55-64	0.375	0.130	-2.82	0.005	0.190	0.741
Interaction Male Sex & Age 65-74	0.286	0.102	-3.52	0	0.143	0.574
Interaction Male Sex & Age 75+	0.187	0.070	-4.49	0	0.090	0.389
Ethnic Group						
white	1.000					
mixed ethnic group	0.943	0.297	-0.19	0.853	0.510	1.747
black or black british	1.668	0.190	4.49	0	1.334	2.086
asian or asian british	1.092	0.095	1	0.315	0.920	1.296
any other group	1.095	0.197	0.51	0.612	0.770	1.558
Total days/4 weeks active 30 mins + moderate +						
none	1.000					
less than 1	0.762	0.066	-3.16	0.002	0.644	0.902
1 or 2 a week	0.832	0.063	-2.44	0.015	0.717	0.965
3 or 4 a week	0.750	0.069	-3.13	0.002	0.626	0.898
5 or more a week	0.694	0.052	-4.86	0	0.599	0.804
Body mass index (BMI)						
under 20	1.000					
20-25	1.860	0.318	3.64	0	1.331	2.600
25-30	2.996	0.506	6.5	0	2.153	4.171
Over 30	5.413	0.931	9.82	0	3.864	7.582
Salt added at table						
generally add salt to food without tasting it first	1.000					
taste the food, but then generally add salt	1.121	0.103	1.24	0.217	0.935	1.343
taste the food, but only occasionally add salt	1.078	0.089	0.91	0.365	0.917	1.268
rarely, or never, add salt at the table	1.225	0.093	2.66	0.008	1.055	1.423
Ntiles of IMD score						
0.59->8.35 [least deprived]	1.000					

	RRR	Std. Err.	z	P>z	[95% Conf.	Interval]
8.35->13.72	0.959	0.075	-0.54	0.591	0.822	1.118
13.72->21.16	1.105	0.087	1.26	0.206	0.946	1.291
21.16->34.21	1.118	0.094	1.32	0.186	0.948	1.318
34.21->86.36 [most deprived]	1.163	0.102	1.72	0.086	0.979	1.381
Highest educational qualification separate						
NVQ4/NVQ5/degree or equiv	1.000					
higher ed below degree	1.014	0.100	0.14	0.891	0.835	1.231
NVQ3/GCSE A level equiv	1.159	0.123	1.39	0.164	0.942	1.426
NVQ2/GCSE O level equiv	1.157	0.100	1.69	0.09	0.977	1.370
NVQ1/CSE other grade equiv	1.212	0.163	1.43	0.152	0.932	1.577
foreign/other	1.440	0.195	2.7	0.007	1.105	1.878
no qualification	1.262	0.108	2.71	0.007	1.066	1.494
Full time student	1.363	0.184	2.29	0.022	1.046	1.776
Limiting long-term illness						
limiting long-term illness	1.000					
non limiting long-term illness	1.560	0.109	6.35	0	1.360	1.789
no long-term illness	0.882	0.055	-2.01	0.045	0.781	0.997

Internal validation: area under Receiver Operating Characteristics curve (AUROC)

AUROCs for the local and complete models tested above were estimated using Stata10. These are shown in Figure 6and Figure 7 below. If both sensitivity and specificity are of importance in a high blood pressure model, the optimal threshold of t would be 0.75, where these 2 accuracy measures equal sensitivity and specificity equal 0.77. The local model, with an AUROC of 0.8071 exceeds this level, although the complete model has even better performance, with an AUROC of 0.8304.

Figure 6: ROC Curve for local hypertension model

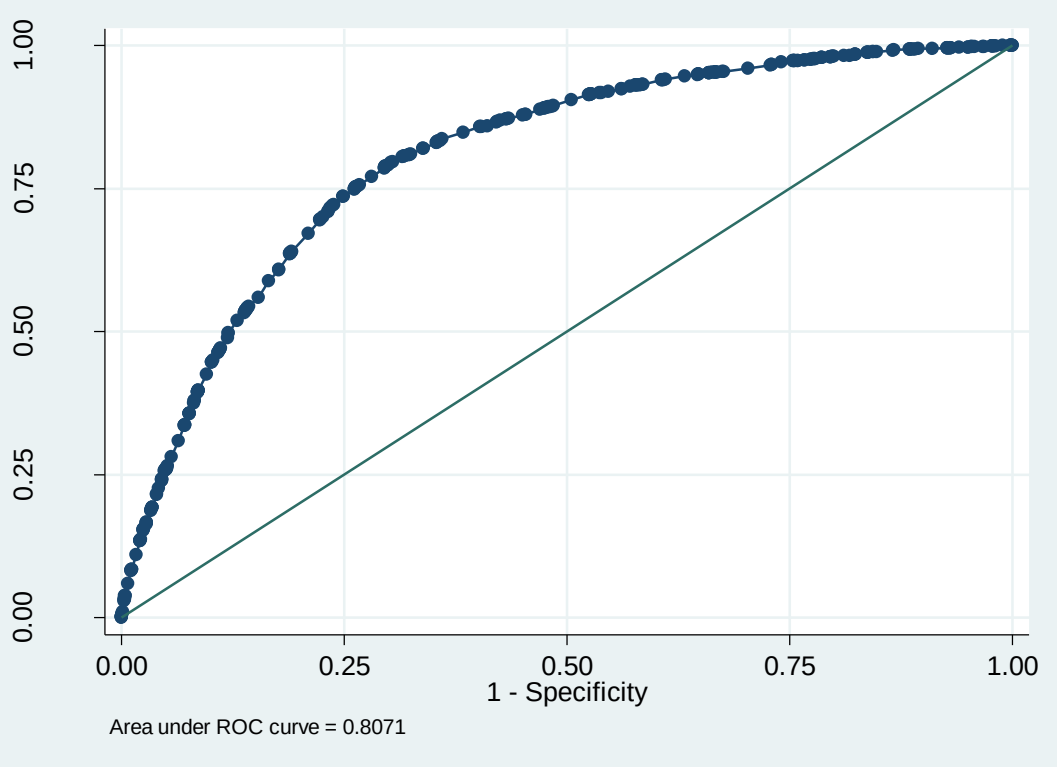
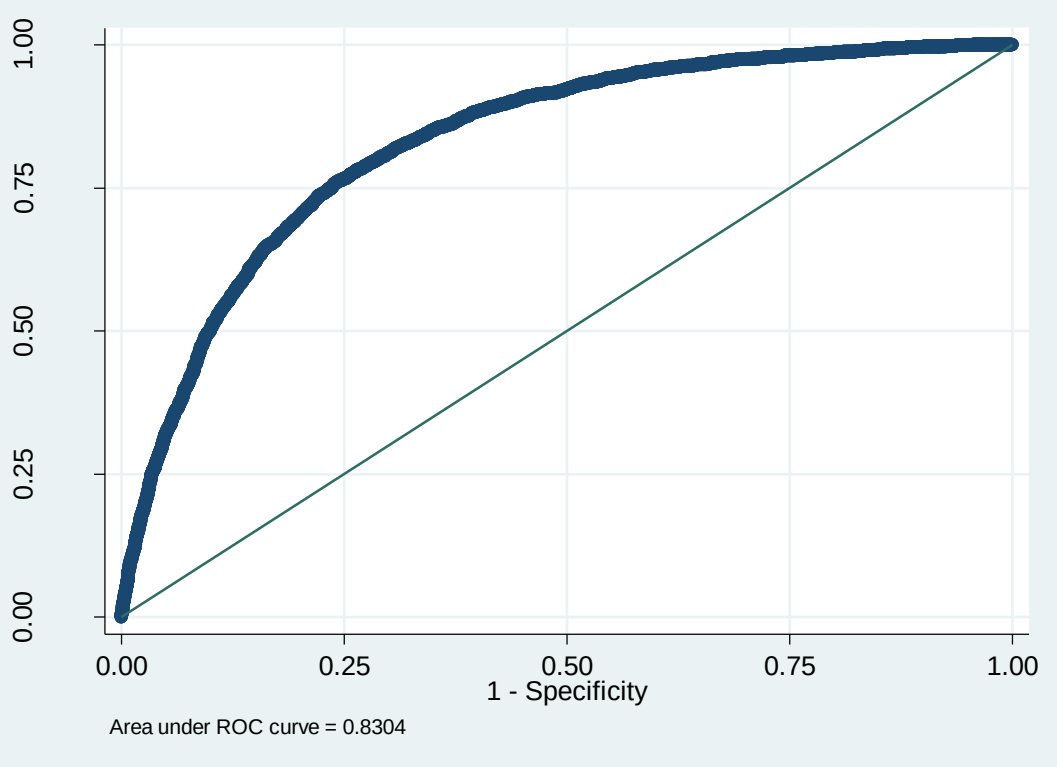


Figure 7: ROC Curve for complete hypertension model



Validation: model prediction

Residuals were tabulated against the observed presence of high blood pressure to assess "misclassification" by each model (in shaded cells). The results for the local and complete models are shown in Table 40 and Table 41 below. The complete model shows slightly lower misclassification but this is minimal in view of the lower overall numbers in this model.

Table 40: comparison of observed: predicted hypertension- local model

		Predicted		
		Not_hypertensive	**Hypertensive**	**Total**
Observed	**Not_hypertensive**	6,886	1,129	8,015
	Hypertensive	1,717	2,012	3,729
	Total	8,603	3,141	11,744

Table 41: comparison of observed: predicted hypertension- complete model

		Predicted		
		Not_hypertensive	**Hypertensive**	**Total**
Observed	**Not_hypertensive**	6,446	938	7,384
	Hypertensive	1,424	1,878	3,302
	Total	7,870	2,816	10,686

8.4.4 Application: PCT, LA and practice level models

The same practice, LA and PCT population data used for the other models was used here. In order to estimate prevalence of hypertension in the future, population projections were incorporated into the model. Modelled estimates of disease prevalence at practice level were provided for 2007 for the adult population (aged 16+ years) using a similar methodology to previous models. By assuming that the prevalence of disease is 0% in the under 16s, prevalence estimates for all ages were calculated. These all-age prevalence are more closely equivalent to QOF prevalence. For each practice the model required the following input data:

- Population by age and sex
- Distribution of ethnic groups by age and sex (not required for stroke model)
- Prevalence of smoking and ex-smoking by age and sex (not required for hypertension model)
- IMD deprivation score
- Urban-suburban-rural category (only required for COPD model)

The same assumptions have been made as previously in the modelling process. For some practices, particularly those with a large minority ethnic population, these assumptions may not hold and the model may not be an accurate reflection of true prevalence.

- The distribution of ethnic groups within each age-sex group is the same as the overall distribution of ethnicity for the practice.
- Smoking prevalence is the same across ethnic groups
- Ex-smoking prevalence by age group and sex is the same in all practices.

- Disease prevalence in population aged <16 is 0%.

The urban-suburban-rural categorisation of the MSOA where the practice is located is representative of the catchment population of the practice. Only practices with a complete input dataset could be included in modelling.

Data sources

As before practice populations by age and sex were derived from a NSTS extract taken on 31/03/2007. The proportion of practice population in ethnic groups was supplied by the CQC, as for the other models. The proportions were derived from HES admissions in 2005/06-2006/07, excluding mental health and maternity admissions and records with unknown ethnicity. The proportions by ethnic group for each practice were calculated by dividing admissions within each ethnic category by the total admissions for the practice.

This model also used IMD2007 scores for practices, derived by SEPHO in August 2008. These scores are weighted MSOA IMD scores based on the March 2008 Attribution Data Set. Practice smoking prevalence is based on NatCen synthetic estimates of smoking at MSOA level 2003-2005 (available from Neighbourhood Statistics). Practice smoking prevalence was calculated as the weighted average of prevalence in MSOAs in the practice catchment area, determined from postcodes in cleaned 2007 NSTS data (as above) for patients aged 16+. The Urban-suburban-rural categorisation is based on the Rural and Urban area classification of Super Output Areas, 2004 (available from Neighbourhood Statistics) shown previously. The urban-suburban-rural category of the MSOA in which the practice is physically located is used in the COPD model.

8.5 Cardiovascular Disease Model

8.5.1 Introduction

The CQC commissioned Eastern Region PHO to develop prevalence models for cardiovascular disease and cardiovascular disease including diabetes, to complement models already developed for stroke, CHD and hypertension. They in turn commissioned the CVD model development from ICL. These models were also developed from person-specific data from HSfE. For the CVD and CVD+diabetes models, data for the 2003, 2004 and 2005 Surveys were merged, as all three cover (to a variable extent) established CVD and known CVD risk factors. The 2003 HSfE covered CVD in detail, the 2004 Survey included a boost for ethnic minorities (some of which are known to have higher prevalence of CVD), and the 2005 HSfE covered older people- age is the most important risk factor for CVD. The outcome variables CVDDIS, a derived variable, defined as "has had CVD (Angina, Heart Attack or Stroke only)", and excluding heart murmur, irregular heartbeat and other heart disease was selected from the HSfE datasets for modelling. Review of CVD, stroke & CHD Prevalence

Overall CVD Prevalence

Unadjusted trends in CVD prevalence from the HSfE are shown in Table 42. This table uses the same definition of CVD used for modelling.

Table 42: trends in HSfE prevalence of overall CVD (ever), by survey year, age and sex (all adults)

	1994	1998	2003	2006	2003 weighted1	2006 weighted
16-24	0.1	0.3	0.3	0.2	0.3	0.2
25-34	0.3	0.5	0.3	0.3	0.3	0.2
35-44	0.6	1.3	1.0	0.8	1.1	0.8
45-54	3.0	3.8	3.5	3.3	3.4	3.3
55-64	9.8	11.5	10.0	8.6	10.1	8.7
65-74	18.5	19.6	19.3	18.7	19.5	18.6
75+	22.8	26.7	28.4	31.6	28.4	31.3
All adults	**6.1**	**7.2**	**7.5**	**5.8**	**6.8**	**6.8**

Source: The Information Centre for Health & Social Care [150]

I performed a literature search of the MEDLINE database using "cardiovascular disease" and prevalence as keyword and title search terms. The HSfE is unusual in that most cross-sectional prevalence studies have focused on IHD/CHD or stroke separately. In the US equivalent of HSfE, the 1999-2002 NHANES,[151] a history of myocardial infarction and stroke was reported by 3.29% and 2.41%, respectively, of adults, which is very similar. Other published reports of overall CVD prevalence are rare. The Framingham Study is the best known long-term cohort study of CVD incidence and associated risk factors [152]. Although

1 Data up to 2002 are unweighted; from 2003 onwards data have been weighted for non-response. From 2003 onwards data are shown both unweighted (for consistency with previous years) and weighted (shaded columns).

most initial work was on CHD events,[153] and stroke was classified as cerebrovascular disease rather than CVD, data collection also included stroke[152]. However, only stroke/CVA incidence data, gathered from hospitals and death certificates, was published.

The SCORE project assembled a pool of datasets from 12 European cohort studies, mainly carried out in general population settings [154]. There were 205,178 persons representing 2.7 million person years of follow-up. There were 7934 CVD deaths of which 5652 were deaths from CHD. However the published SCORE report also does not report the prevalence of CVD, and it may be that respondents with pre-existing CVD were excluded from the analysis. The WHO MONICA Project was established in the early 1980s to monitor trends in CHD and to relate these to risk factor changes, but it did not cover stroke [155]. Studies of overall CVD prevalence have been carried out in New Zealand.[156] Other papers which studied CHD or stroke separately are reviewed in the sections describing these models.

Previous CVD prevalence modelling

There has apparently been very little previously published modelling of CVD prevalence. A report from New Zealand estimated the prevalence of CVD and distribution of CVD risk in New Zealanders for healthcare planners, funders, and providers [157]. Survey data provided estimates of CVD prevalence and absolute CVD risk distributions, and these were applied to population data using the Framingham CVD risk prediction equation. About 7% of the population were estimated to have suffered a non-fatal heart attack or stroke or have angina. A further 13% were estimated to have a 5-year CVD risk greater than 15% based on New Zealand CVD risk charts.

8.5.2 Methods

Data Sources

The CVD model described here uses data from the 2003, 2004 and 2005 HSfE. Because the HSfE has shown that CVD prevalence varies with ethnicity [1], I used a sample containing data from a large number of ethnic minority respondents. The HSfE 2004 was the last Survey to include an ethnic minority boost, and the boost sample was used for the modelling. However, there were relatively small numbers of Whites in the HSfE 2004 sample, and BP was not measured for most respondents in the 2004 general population sample, presumably to save resources for the boost itself. The HSfE 2004 boost sample was therefore merged with the HSfE 2003 and 2005 data, both of which measured BP. Table 43 shows the ethnic group breakdown of the merged dataset. Note that it was necessary to collapse two of the HSfE 2003 ethnic group variables to use the same classification as HSfE 2004 and 2005.

Table 43: Ethnic Group Breakdown of merged dataset

Ethnic Group	N	Percent
White	24,025	76.62
Mixed	384	1.22
Black or Black_British	2,146	6.84
Asian or Asian British	4,107	13.1
Chinese or Other	696	2.22
Total	**31,358**	**100**

NB excludes no answer/refused/don't know

Definitions

There were 31,458 respondents in the merged dataset, of whom around 31,430 answered questions on CVD. The definitions used in the relevant HSfE variable and for the models are as follows (variable name in capitals):

- CVDIS: a derived variable, defined as "has had CVD" including angina, heart attack or stroke. **This variable was used for the overall CVD prevalence model**
- CVDDEF: a derived variable, defined as "has had a cardiovascular condition", including heart murmur, diabetes, high blood pressure, angina, heart disease, irregular heartbeat, other heart disease, or stroke
- CVDDEF1: a derived variable, defined as "has had a cardiovascular condition excluding diabetes/high BP", but including heart murmur, angina, heart disease, irregular heartbeat, other heart disease, or stroke.
- NOBPCVD: a derived variable, defined as "has had a cardiovascular condition excluding those with high blood pressure", but including heart murmur, diabetes, angina, heart disease, irregular heartbeat, other heart disease, or stroke. **This variable was used for the overall CVD+diabetes prevalence model**
- IHDIS: a derived variable, defined as "has had IHD/CHD" including only angina or heart attack. **This variable was used for the CHD prevalence model**
- STRODEF: a derived variable, defined as "doctor-diagnosed stroke". **This variable was used for the stroke prevalence model**

All these variables are based on a patient-reported but doctor-diagnosed history of the disease. There are differences between various methods of estimating CVD prevalence. The positive predictive value of questionnaire responses such as those used in HSfE may be sub-optimal in comparison with clinical diagnosis. Conversely, reliance on a medical diagnosis may underestimate prevalence, as patients with unrecognised disease or very mild symptoms may not attend (or be correctly identified by) their GP.

Table 44: Comparison of HSfE CVDEF1 & CVDIS variables in model dataset below compares the HSfE definitions of CVDIS and CVDDEF1 in the merged dataset. The definition of CVDDEF1 is broader than CVDIS, incorporating heart murmur, irregular heartbeat and other heart disease (this presumably including heart failure). The greater proportion of the latter two categories will have been originally caused by IHD/CHD. However CVDIS has therefore been used in the model fitting because this corresponds most closely to the definitions used in the development of the Framingham/Joint British Societies and QRISK2 risk scoring algorithms.

Table 44: Comparison of HSfE CVDEF1 & CVDIS variables in model dataset

		CVDIS: has had CVD (angina, heart attack or stroke)		**Total**
		yes	**no**	
CVDDEF1: Has had CVD condition (excluding diabetes/high BP)	no	0	27,325	**27,325**
	yes	2,452	1,654	**4,106**
	Total	**2,452**	**28,979**	**31,431**

In the HSfE, 'diabetes' means self-reported doctor-diagnosed diabetes, excluding pregnancy (DIABETE2). The HSfE interview makes no distinction between type 1 and 2 diabetes. For HSfE classification purposes, type 1 diabetes is defined as being diagnosed before the age of 35 and being on insulin therapy at the time of the survey. For those respondents who had a fasting blood sample taken, a fasting blood plasma glucose level ≥ 7mMol/l in the absence of a reported diagnosis of diabetes is considered indicative of 'undiagnosed diabetes'. The prevalence of undiagnosed diabetes is reported separately in the HSfE for the subgroup with a valid fasting blood sample and aged over 35. The prevalence of 'controlled diabetes' is assessed by measuring glycated haemoglobin, but the HSfE bases were too small to examine levels in subgroups.

In the merged 2003-5 dataset, the crude prevalence of self-reported doctor-diagnosed diabetes is 4.83% (1,518 of 31,457 respondents). Blood samples were tested for glucose only in the 2003 and 2004 Surveys. In the merged dataset, only 2,148 respondents had a fasting plasma glucose (FPG), of whom 90 had a result ≥ 7mMol/l (Table 45). Of these, only 35 had undiagnosed diabetes. I considered imputing FPG values for other patients in the dataset. This could be done if requested but will require more methodological development. We would expect about 1.74% of patients without patient-reported doctor-diagnosed diabetes to have undiagnosed diabetes.

Table 45: frequency of patient-reported doctor-diagnosed diabetes and undiagnosed diabetes in respondents with an FPG

	FPG < 7.0mMol/L	FPG > 7.0mMol/L	Total
No patient-reported diabetes	2,007	35	2,042
%	97.52%	38.89%	95.07%
Patient-reported diabetes	51	55	106
%	2.48%	61.11%	4.93%
Total	2,058	90	2,148

Model construction: data issues

The choice of variables for original inclusion in the merged dataset included all those known to be CVD risk factors. The HSfE dataset has a nested or hierarchical structure so three variables related to the sampling strata were included: area (sample point), cluster (stratification level), and hserial (serial number of household). These were used in the model to adjust for clustering of respondents. In the analysis variables CHOLEST and HDLCHOL were combined to give a lipid ratio. IMD 2004 score bands were slightly different between the two years, but raw scores were not provided in the dataset so it was necessary to assume identity.

The Stata software package was used for analysis. The methodology applied for model fitting was multinomial logistic regression with the "cluster" option as for previous models. The prevalence in each age group, gender, ethnic group, area of residence and level of deprivation, and smoking status category were derived from the odds, using the formula:

prevalence = odds/(1 + odds). Variables for both "local" and "complete" models were selected by reverse stepwise selection using likelihood ratio (LR) and Wald tests i.e. the model was fitted using all available exposure variables, omitting each in turn and recording p values [158]. The variable with the highest p value was omitted and the tests were repeated. Since the models were to be used for prediction rather than hypothesis testing, a p value of 0.2 was used. No significant interactions were found using χ^2 tests.

8.5.3 Results

The local model only used locally available data. I included smoking as practice synthetic estimates are now available. Local GHQ-12 or Limiting Longstanding Illness score data will be available locally from 2009 and could be included later. Ideally, the best prediction should result from utilising the most information in the regression model. However only a limited range of HSfE variable data is either available or can be estimated at the PCT level. We decided to validate the local model by comparing it, in terms of prediction, to a model including all available HSfE variables. In addition, however, the amount of missing data affects the prediction of a model. For example, in the complete HSfE variables the largest proportion of missing data occurred in those variables related to particular drug treatment for high blood pressure. Some clinical tests also have a high proportion of missing values. Hence the "complete" model included the complete list of variables, including differing definitions of high blood pressure (history of high blood pressure, or clinically measured hypertension >140/90 at the time of the Survey). Table 46 shows crude CVD, diabetes and CVD+diabetes prevalence in the merged dataset, using the stated variables.

Table 46: Crude CVD, diabetes and CVD+diabetes prevalence rates in 2003-5 dataset

	Age in 10 year bands							
	16-24	**25-34**	**35-44**	**45-54**	**55-64**	**65-74**	**75+**	**Total**
No CVD	99.82	99.66	98.95	97.34	92.64	80.57	71.14	86.94
CVD	0.18	0.34	1.05	2.66	7.36	19.43	28.86	13.06
No diabetes	99.6	99.32	98.16	96.53	93.91	87.69	89.84	95.17
Diabetes	0.4	0.68	1.84	3.47	6.09	12.31	10.16	4.83
No CVD or diabetes	97.04	95.9	93.21	90.29	83.58	64.96	55.73	83.86
CVD+diabetes	2.96	4.1	6.79	9.71	16.42	35.04	44.27	16.14
Raw numbers	**3,273**	**5,023**	**5,817**	**4,808**	**4,313**	**4,629**	**3,567**	**31,430**

Local CVD Model

Table 47 shows ORs, p values and confidence intervals for the local CVD model. Values are generally similar to the more complete model. Unfortunately, however, local synthetic estimates do not have separate categories for occasional/regular smokers.

Table 47: odds ratios, local CVD model

	RRR	Std. Err.	z	P>z	Lower 95% CI	Upper 95% CI
Female	1.000					
Male	1.550	0.073	9.32	0	1.414	1.700
Age 16-24	1.000					
Age 25-34	1.790	0.852	1.22	0.221	0.704	4.549
Age 35-44	6.119	2.628	4.22	0	2.637	14.198
Age 45-54	15.591	6.539	6.55	0	6.853	35.470
Age 55-64	47.533	19.727	9.3	0	21.073	107.214
Age 65-74	144.217	59.426	12.06	0	64.309	323.416
Age 75+	255.112	105.356	13.42	0	113.554	573.138
Never smoker	1.000					
Used to smoke occasionally	0.858	0.094	-1.4	0.161	0.692	1.063
Used to smoke regularly	1.384	0.077	5.87	0	1.241	1.542
Current smoker	1.140	0.083	1.81	0.07	0.989	1.314
IMD 2004 0.59->8.35 [least deprived]	1.000					
IMD 2004 8.35->13.72	1.184	0.089	2.24	0.025	1.021	1.373
IMD 2004 13.72->21.16	1.295	0.100	3.35	0.001	1.113	1.506
IMD 2004 21.16->34.21	1.497	0.114	5.31	0	1.290	1.737
IMD 2004 34.21->86.36 [most deprived]	2.176	0.163	10.36	0	1.879	2.522
White	1.000					
Mixed	0.809	0.320	-0.54	0.591	0.372	1.756
Black/BB	0.781	0.104	-1.85	0.064	0.602	1.014
Asian/AO	1.613	0.150	5.15	0	1.345	1.935
Other	0.692	0.190	-1.34	0.181	0.404	1.186

Sensitivity/Specificity: ROC Curve

Sensitivity and specificity were calculated using the area under the receiver operating characteristics (ROC) curve. The AUC for the local CVD model is 0.8601

Internal validation: prediction

Another method of assessing performance is to use the regression model to predict the response for each subject. These predictions are called fitted values. The difference between the fitted and the observed values are called residuals. These can then be tabulated against the observed presence of CHD to assess "misclassification" by each model. Prediction by the models is shown in Table 48: actual vs predicted CVD below.

Table 48: actual vs predicted CVD

		Predicted		
Model		**No CVD**	**CVD**	**Total**
Actual	**No CVD**	27,095	11	27,106
	CVD	4,073	7	4,080

		Predicted		
Model		**No CVD**	**CVD**	**Total**
	Total	**31,168**	**18**	**31,186**

8.5.4 Application: PCT, LA and practice level models

The local model (which includes only those variables that are available at population level i.e. age, sex, ethnicity, smoking status and deprivation score) has been applied to Local Authorities to create prevalence estimates of coronary heart disease in those aged 16+ for 2005–2020. Models for PCTs have been created for 2006-2020.

Assumptions of the modelled estimates
It is assumed that:

- the proportion of smokers, ex-smokers and never-smokers is the same across ethnic groups.
- the proportion of ex-smokers in each age-sex group is the same in all areas.
- the smoking prevalence rates from the model-based estimates of lifestyle behaviours (43) are reliable.
- the prevalence of CVD in those aged under 16 is negligible.

Due to lack of data, it was not possible to treat ex-regular-smokers and ex-occasional-smokers separately. Ex-occasional-smokers are treated as non-smokers.

Populations

The CVD prevalence model uses ONS 2005 mid-year population estimates by ethnic group, age and sex as for other models. Population projections were incorporated into the model. For 2008 and 2009, the distribution of ethnic groups in 2005 (LA) or 2006 (PCT) was applied to ONS 2006-based population projections by quinary age band for LAs and PCTs. National (England) proportions of smokers, ex-smokers and current smokers by age and sex were taken from HSfE (2003-2005 pooled). Practice smoking prevalence is based on NatCen synthetic estimates of smoking at MSOA level 2003-2005 (available from Neighbourhood Statistics). These proportions were then adjusted for each practice/LA/PCT as above. The same smoking prevalence rates are applied across all ethnic categories. This model also used IMD2007 scores for practices, derived by SEPHO in August 2008. These scores are weighted MSOA IMD scores based on the March 2008 Attribution Data Set.

8.6 Cardiovascular Disease Risk Model

The development of this model was commissioned by NE London Cardiac & Stroke Network on behalf of NE London PCTs, through the Public Health Action & Support Team, a community interest company. I worked on model development with Andrew Dalton, and we have also submitted it for publication.

8.6.1 Introduction

The NHS Health Checks Programme[159] is a high profile national policy in England designed to CVD mortality and health inequalities. It aims to lower CVD risk factors in those without pre-existing CVD or diabetes[160] through risk assessment and treatment of the entire 40-74 year old English population- about 22.4 million people- over a five year period. The local Health Check process is loosely specified by national guidance[159;161], although PCTs have scope for innovation and originality to suit their population's need. A minimum battery of procedures has been prescribed[159], with further details following the release of a ratified minimum dataset in early 2011.

The Health Checks, along with guidance on the use of statins for the primary prevention of CVD[162], rely heavily on the notion of stratifying patients by absolute risk of developing CVD. Those designated at high risk will be offered a number of risk lowering interventions; including statins and weight management; and will be recalled for annual follow up. High risk is defined as an absolute or "global" risk of a CVD event within ten years of over 20%, designated by an appropriate risk calculator. The concept of global risk has changed CVD prevention in recent years and been shown to be a more cost effective strategy than managing individual risk factors[163;164]. There are many CVD risk scores available, although in the UK NICE has specified the use of either the Joint British Societies 2[24] (JBS2) or QRISK2[111] algorithms. The JBS2 algorithm is a modification of the original Framingham risk score, which was developed from a North American cohort study,[153] whereas QRISK2 was derived from a large sample of the English population using the QRESEARCH database (derived from the electronic medical records of general practices that use the EMIS system). The unmodified Framingham risk score is known to over-estimate CVD risk in European populations.[154]

High CVD risk status is central to the Health Checks programme, yet there are few estimates of the numbers at high risk in England, merely estimates from small populations[111;165]. Although the programme commenced in April 2009 and general practices will be capturing data electronically, there are as yet no national data. Before the programme roll out, DH issued economic modelling [166] presenting the cost and cost effectiveness of different methods of CVD screening. They demonstrated the preference for a total population-based approach to screening, with risk reduction interventions for everyone found to have a risk factor whether high risk or not. Recent work, however, has suggested a more targeted approach may be of value[167;168]. DH modelling did not quote an estimate of the numbers expected to be at high risk, or estimates of the prevalence of CVD risk factors within the high risk population. Further, when estimating CVD risk factors in the high risk population, prevalence estimates from the general population were used. The population with high CVD risk will have a higher prevalence of risk factors than the general population; therefore prevalence in this group will be underestimated.

We were commissioned to develop this model by North East London Cardiac & Stroke Network, whose funding is gratefully acknowledged. Previous work has demonstrated the ability to use HSfE data in the modelling of CVD risk[169]. We used HSfE data to model the number of patients with a QRISK2 score of ≥20% in each general practice in England, and the number of patients within this group with CVD risk factors. We estimated the resultant costs of the programme, the reductions made in both hospital admissions and mortality concerning CVD, and finally compared the findings to those had the JBS2 algorithm[24] been applied.

8.6.2 Methods

Data sources

We merged person-specific data from the HSfE from 2003 to 2006 inclusive. Of 71,717 informants, 28,209 were aged 40-74. We removed patients with diagnosed CVD, diabetes and hypertension i.e. those not eligible for the Health Checks. All further work was carried out using this population. In addition to the standard HSfE data we obtained the 2001 Townsend deprivation score from the postcode of residence of each participant.

Imputation of Missing Data

One method of dealing with missing data in surveys or GP systems is imputation. Multiple imputation aims to allow for the uncertainty about missing data by creating multiple copies of the dataset under analysis, in which the missing values are replaced by imputed values drawn using a statistical prediction based on the observed data. Standard statistical analyses are performed on each of the imputed datasets and the results are combined in an appropriate way to obtain final conclusions.[170;171]

For example, this method was used in the development of QRISK2.[110] Their model was initially fitted using patients without any missing data (complete case analysis), and the fractional polynomial terms were obtained from the complete case analysis using a multivariable approach. However, because patients with complete data might have a different health status and CVD risk compared with those with missing data, they fitted the principal models on the basis of multiple imputed datasets using Rubin's rules to combine effect estimates and estimate standard errors.[158;172;173] They subsequently improved their implementation of multiple imputation for missing data by including more variables.

The validity of imputation depends on the modelling being done carefully and appropriately, and it is not possible to automate the procedure. A number of pitfalls can arise. In particular, when an analysis explores the association between one or more predictors and an outcome, but some of the predictors have missing values, the outcome carries information about the missing values of the predictors, and this information must be used in the imputation procedure. Given the relatively large sample size, it is uncertain to what extent imputation will improve the estimation of population prevalence from GP data. We used multiple imputation in the development of our CVD risk model.

We applied the QRISK2[111] and JBS2[24] CVD risk algorithms to each informant's data. The variables in each algorithm are shown in Table 49 below. Age, sex, ethnicity, smoking status and deprivation values were complete; 18,426 (65.3%) had blood pressure recorded, 17,373 (61.6%) BMI and 12,621 (44.7%) had lipid ratios (total cholesterol and high density

lipoprotein). For the remainder we used multiple imputation based on complete data to estimate missing values[173]. QRISK2 contains additional data concerning chronic conditions which increase CVD risk. In the algorithm these can be assumed to be absent if data are missing[174]. Family history of CVD was near complete, with the 973 missing assumed null; there are no data in the HSfE for chronic kidney disease (CKD) or rheumatoid arthritis therefore these were excluded. Finally there were no data on atrial fibrillation (AF), however using estimates from a population-based survey of ECG abnormalities[95] we estimated the number of cardiac arrhythmias (found in HSfE) caused by AF, and applied this prevalence at random to the population with arrhythmias.

Table 49: variables included in QRISK2 and JBS2 algorithms

Variable	QRISK2	JBS2
age	X	X
sex	X	X
body mass index	X	
deprivation score (Townsend)	X	
systolic blood pressure	X	X
cholesterol ratio	X	X
family history of coronary heart disease	X	X
smoking status	X	X
treated hypertension	X	
type 2 diabetes	X	X
atrial _fibrillation	X	
rheumatoid arthritis	X	
chronic renal disease	X	
history of cardiovascular disease	X	
ethnicity	X	X*

* NICE guidance recommends that, if JBS2 is used, the risk score should be increased by a factor of 1.4 for men with a South Asian background, by a factor of 1.5 if one first-degree relative has a history of premature CHD, and by a factor of up to 2.0 if more than one first-degree relative has a history of premature CHD

High risk models

We partitioned the population based on age (40-49, 50-54, 55-64, 65-74); gender; ethnicity (white, south Asian, black, other) and deprivation (national thirds of Townsend score), giving 96 data cells. We applied the risk algorithms to the data then calculated proportions at high risk (≥20%) in each cell. Any cell with fewer than 10 entrants was combined with the adjacent cell (the next age group).

We obtained age and sex breakdowns of every general practice in England[175] and subtracted the number of patients on the CHD, Stroke/TIA, diabetes and hypertension registers[13]. Based on the HSfE data, we assumed 40% of diabetes patients appear on a CVD register and 37% of the hypertension register have CVD or diabetes. Using HSfE data we estimated the age distribution of each condition, and hence the proportion in the 40-74 age group. We

obtained estimates of the ethnic breakdown of each practice from the CQC, which were derived by assuming the proportions of ethnic minority populations admitted to hospital reflected the ethnic composition of LSOA resident populations, and applying these to practice populations. We applied the ethnicity proportions to our population data assuming the ethnic breakdown was equal over age/ sex groups. We used CQC Townsend Deprivation scores for each practice, based on the residency of their registered population, and assigned each practice to a national third. We assumed that the whole practice population lay in the same third. Each practice was broken down into the same 96 cells as the risk proportions and we applied the risk proportions to each, giving the number of patients with a ≥20% score for each risk score; these were summed to give PCT level estimates. We then mapped the estimated PCT prevalence of high risk status in England.

CVD risk factors in high risk group

From within the HSfE population found to be high risk we modelled the number of patients with individual CVD risk factors. These were BMI greater than 25; undiagnosed hypertension (assuming 50% of patients with a raised blood pressure proceed to be diagnosed with hypertension[166]); physical inactivity; undiagnosed impaired glucose tolerance (IGT) (fasting plasma glucose between 6 and 7) and smoking. Additionally, we assumed all patients ≥20% risk were prescribed statins [162] and estimated the prevalence of undiagnosed diabetes (fasting plasma glucose greater than 7). Taking the high risk population, we found the proportion in the HSfE data of patients in each age/ sex category with risk factors, and transposed the proportions onto the PCT high risk estimates.

Table 50: major DH model assumptions for the risk factor interventions[166]; † if aged 70 or over

Intervention	Uptake	Compliance	Attribution	RRR of CVD	Cost
Smoking	19%	15%	51%	0.36	£ 176.80
Anti-Hypertensive	40%	87%	24%	0.24	£ 29.64 (£70†)
Exercise	77%	23%	63%	0.14	£ 33.50
IGT lifestyle	85%	90%	90%	0.09	£ 462.00
Statin prescription	85%	70%	50%	0.31	£ 60.52
Weight management	85%	68%	47%	0.36	£ 51.00

Costs and benefits

We used the DH costs of the basic NHS Health Check and initial interventions and referrals (related to the major risk factors in Table 50 above). Model assumptions were derived from the DH impact assessment[166]. We assumed 70% attendance at the initial health check based on DH assumptions; additional model assumptions are shown in the appendix (*box A1*). We estimated the costs of the basic health check; including time, diagnostic tests and resultant referrals but not additional time for the diagnoses.

By applying formula 1, where RRR_i = relative risk reduction for intervention i; N_i = the number of patients completing intervention i; R_x = the mortality or admission rate, we estimated the reduction in mortality or admissions as a result of each intervention.

$$(1)\ RRR_i \times N_i \times R_x$$

Using the numbers taking up and adhering to the interventions we estimated savings in hospital admissions and mortality. We obtained PCT level hospital mortality and admission data, using International Classification of Disease (ICD) codes I20-I25, I63 and I64 for mortality and the associated Healthcare Resource Group (HRG) codes for admissions, the case mix groupings used to fund NHS acute hospitals according to a standard national tariff. We assumed an immediate impact of the interventions on CVD risk and estimated the reductions in mortality and hospitals costs after the whole population has been screened. All data analyses were carried out using Stata 11.1SE.

8.6.3 Results

Prevalence of high CVD risk

We estimate a total of 3,020,786 people in England, aged 40-74, have a QRISK2 score ≥20% out of a population of 22,397,278 (13.5% [95% CI= 13.47-13.50]) (Table 51). The mean number per PCT was 19,874 [95% CI= 18,019- 21,729], and the median 16,392 (range = 6,007- 71,239; IQR= 12,350). As a percentage of the 40-74 population the PCT mean was 13.6% [95% CI = 13.39-13.81]; median =13.7% (range= 10.3%- 17.3%; IQR= 1.95%). Table 51 further presents the practice level numbers and percentages of the 40-74 population at high risk. Figure 8 shows the geographical distribution of high risk status in England, with the highest prevalence is centred on the major urban areas in northern England.

Table 51: PCT and practice level estimates of the numbers and proportion of high risk patients

	Total	PCT-level				
		Mean	(CI)	Min - Max	Median	IQR
Number High Risk	3,020,786	19,874	18,019-21,729	6,007-71,239	16,392	12,350
Prevalence High Risk	13.5%	13.6%	13.4% -13.8%	10.3%-17.3%	13.7%	1.95%
		Practice-level				
		Mean	**(CI)**	**Min - Max**	**Median**	**IQR**
Number High Risk		367	362-372	0-2,272	316	332
Prevalence High Risk		13.6%	13.5%-13.7%	0%-30.8%	13.6%	3.0%

Prevalence of risk factors in the high risk population

Within the high risk population the number with individual CVD risk factors can be seen in

Table 52. A large proportion of the population were both physically inactive (79.1%) and overweight or obese (81.6%). We estimate Health Checks will find 768,405 undiagnosed hypertensive patients and 438,015 with IGT, and will detect 126,876 patients with undiagnosed diabetes, a mean of 835 [757-913] per PCT.

Figure 8: the prevalence of high risk status, using the QRISK2 score, in the 40-74 year old population in England

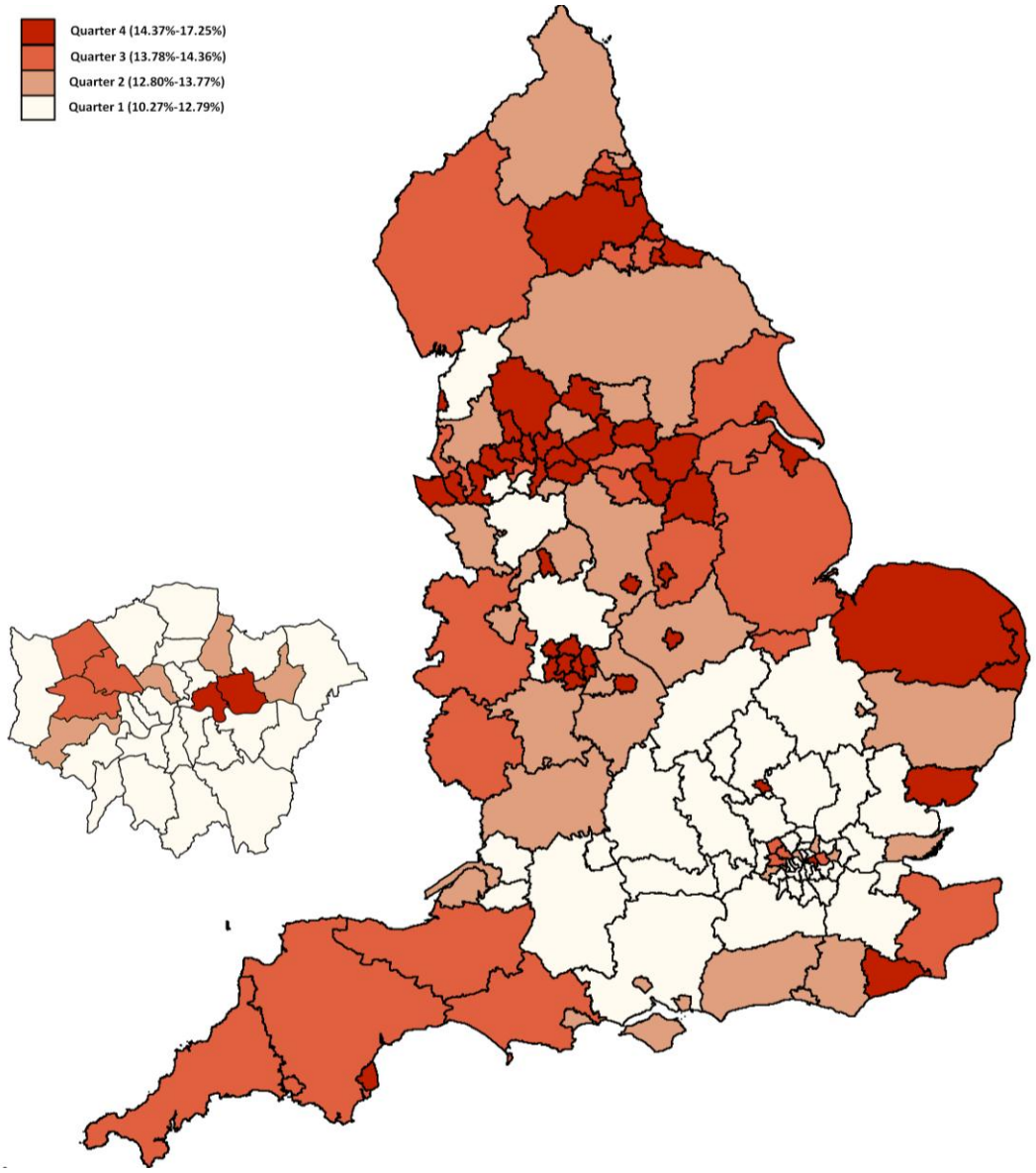


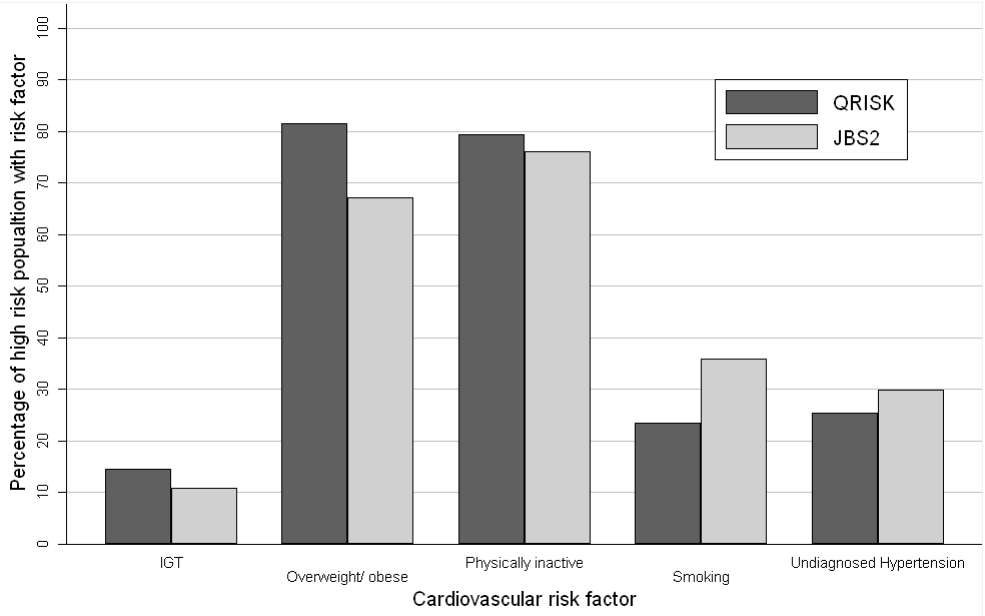
Table 52: the prevalence of CVD risk factors in the high risk population, † due to small numbers no age/sex stratification was used in the IGT estimates

	QRISK2				JBS2
	No. in England	**Mean per PCT [95% CI]**	**Median per PCT [Range; IQR]**	**% of high risk per PCT [95% CI]**	**Mean per PCT [95% CI]**
High risk	3,020,786	19,874 [18,019-21,729]	16,392 [6,007-71,239; 12,350]	-	30,493 [27,698-33,289]
Smoking	706,078	4,645 [4,239-5,052]	3,959 [1,446-15,824; 2,855]	23.8% [23.5%-24.0%]	10,943 [10,012-11,873]
Undiagnosed hypertension	768,405	5,055 [4,587-5,524]	4,174 [1,535-18,067; 3,147]	25.5% [25.5%-25.5%]	9,081 [8,255-9,906]
Physically inactive	2,396,899	15,769 [14280-17259]	12,982 [4,751-57,153; 9,804]	79.1% [78.98%-79.3%]	23,179 [21,029-25,329]
Overweight/ obese	2,462,821	16,203 [14,699-17,707]	13,363 [4,912-57,804; 10,115]	81.6% [81.6%-81.7%]	20,464 [18,591-22,338]
Undiagnosed IGT	438,015	2,882 [2,613-3,151]	2,377 [871-10,330; 1,791]	†	3,302 [3,000-3,605]

Comparison to JBS2

Using the JBS2 algorithm, there were a larger number of people classified as high risk; 4,634,965 aged 40-74 (20.69% [20.67-20.71]), with a mean of 30,493 [27,698-33,289] per PCT, and a median 24,707 [range =9,197-106,010; IQR= 18,147]. There were more patients with CVD risk factors (*table 3*); these did not increase uniformly with the increase in numbers at high risk. Figure 9 shows the proportions of the high risk population with each of the five risk factors using the two risk scores. The JBS2 population had an especially large proportion of smokers (35.9%) and undiagnosed hypertension (29.7%), whilst QRISK2 had a higher prevalence of overweight/ obesity (81.5%).

Figure 9: percentage of the high risk population with CVD risk factors using the QRISK2 and JBS2 risk scores



Costs and risk reductions

In the 3,020,786 high risk patients, with a total of 2,114,539 screened (70% uptake), we estimate the total cost of the basic health check for the high risk population in England to be £ 34,063,143. The average cost per PCT was £224,100 and median £184,847, ranging from £67,757 to £803,236. The total cost, including the six risk factor interventions, was £358,900,728. The mean per PCT was £2,361,189, median= £1,947,958, ranging from £731,440-£8,652,724.

Screening the high risk patients in England, taking 2008-9 admission rates as the baseline, will save 6,824 admissions per year, out of a total of 661,199 (1.0%). Per PCT there will be 44.9 [40.3-49.4, median =36 (range 10-142)] admissions saved per annum. Health Checks in the high risk population will reduce CVD mortality by 1,252 deaths per annum in England (1.1% reduction), with a median of 7 per PCT (range = 2-28). In comparison, the total costs using JBS2 would be £ 469,359,340, ranging from £845,812- £9,701,207, with the cost of screening alone £48,414,242. There would be 9,988 admissions saved per year and 1,832 deaths.

8.6.4 Discussion

Main findings

Our model using the QRISK2 risk score, estimates 3 million patients aged 40-74 at high (≥20%) risk of CVD in England, 13.5% of the population. There was wide geographic variation in the prevalence. Using the JBS2 risk score to risk stratify patients, results in considerably larger numbers at high risk. Within the high risk population we estimate a high prevalence of cardiovascular risk factors; for example 82% of the high risk will be overweight or obese (BMI >25), 79% physically inactive and over 400,000 with undiagnosed impaired glucose tolerance. The levels of cardiovascular risk factors captured by the high risk population were dependent on the risk score used; with high levels of obesity and physical inactivity with QRISK2 and smoking with JBS2. Including the five risk interventions the total cost of

screening the high risk population using the QRISK2 score was £359 million; with the cost per PCT ranging from £67,757 to £803,236, whilst using JBS2 would be £469 million.

Our estimates of the total cost of Health Checks in the high risk population, 16% of the eligible age group, are greater than the DH estimates of £180-243 million per year. The NHS Health Checks once fully implemented will be a five year rolling programme, with 20% of the eligible population screened per year. The high risk cohort identified here is fewer in number than the annual target for screening, although they posses increased levels of risk factors which the costs of management. Previous estimates suggest the majority of the programme cost will come from follow-up intervention and not the basic health check[166]. Both the cost of screening and of the entire programme is larger using the JBS2 score, although the total programme is relatively less costly than using QRISK2. The QRISK2 score identifies a large number of patients requiring IGT, weight management and physical activity interventions within the high risk group. These are expensive interventions (compared to statin and anti-hypertensive prescription) and we assume a higher uptake, inflating the programme costs despite fewer at high risk overall.

Although risk scores were derived as clinical tools for the treatment of single patients, they are also of use when comparing levels of risk between populations[176]. There is a large amount of geographic variation in the distribution of high cardiovascular risk across England, with prevalence higher in the urban centres of the north. There have been lasting health inequalities in England, with high prevalence of both poor health outcomes and the causes of ill health in the northern regions[92]. Our data provide further evidence of this divide. The NHS Health Check will create a large workload for general practice from cardiovascular risk assessment[177]. Further to this, we estimate 316 patients designated as high risk per general practice, with a high prevalence of risk factors; management of these further adds to the workload created.

NHS Health Checks focus on cardiovascular and chronic kidney disease, although the programme may have wider impacts, especially for diabetes[166]. Our data estimate that within the high risk population there will be a large amount of undiagnosed diabetes; early diagnosis can lead to appropriate treatment and a reduction in complications[178]. In the population with undiagnosed impaired glucose tolerance, interventions to reduce the progression from IGT to diabetes are effective[179] and can have prolonged effects.[180] Slowing progression to diabetes is both beneficial to patients in terms of increases in life expectancy, and to the health system in terms of cost savings[181].

The profiles of CVD risk factors differ between the high risk population defined by the two risk scores, an artefact of the populations in which the scores were derived. The most prevalent risk factors in the derivation dataset are the most common in a high risk population defined by that risk score[182]. The JBS2[24], based on a Framingham risk score[153], was produced using data from the USA in the 1970s where smoking prevalence was higher than today[183]. Likewise the QRISK2 dataset was from the UK between 1993 and 2003, a time with higher levels of inactivity and obesity, with for example a threefold increase in the prevalence of obesity from 1980 to 1998[184].

What is already known?

Modelled data presented here are to our knowledge the first estimates of the prevalence of high risk status in England, using a now common definition from national guidance[161;162]. Our estimates of 13.5% of the 40-74 population using QRISK2 and 20.7% using JBS2 are in line with previous estimates from smaller populations. Using Framingham risk scores there have been estimates of 15.8%[185] and 19.1%[111] in the 35-74[185], 15.5% in the 30-74 and 17.8% in the 40-70[186]; with 13.3% of the 35-74 population for QRISK2[111]. Estimates from DH put the cost of the NHS Health Checks at £180-243 million per year. The higher costs presented here are the result of increased interventions required in the high risk population- the DH modelling assumed that prevalence of risk factors was the same as in the general population. The evaluation of a pilot CVD risk assessment service in pharmacies found 70% of the population screened required a referral to general practice because of a raised risk factor[186]. This population was not at high risk, but still had a large need for referral.

Strengths and limitations

NHS Health Checks, although a national programme in England, are being carried out by local service providers. With no national call-recall system or national minimum data set, the programme is likely to remain fragmented into the near future. There has been no collation of data from across the whole of England, especially at a patient level. This makes modelling, such as the data presented here, a vital element in the evaluation of the effectiveness of the programme. Our modelling estimates CVD risk factors using prevalence estimates bespoke to the high risk population, instead of using general population estimates. We also use population data which captured local level differences in the population make-up, with modelling taking place at a general practice level.

Although our data from the HSfE contained a large amount of missing data for risk factors, this is not uncommon when using large population based datasets. In the modelling carried out by DH, [166] there were also considerable missing data. The recording of blood pressure in the HSfE data was poor; we used multiple imputation to replace missing data, though with hypertension, one of the Health Check exclusion criteria more complete data would be preferred. We only estimate the cost of the Health Checks in the high risk population, not the entire eligible cohort, although they will be the priority in the screening programme, and arguably any risk factor interventions should be based on global risk not the presence of a single risk factor. Our modelling was dependent on a number of assumptions. We based these on evidence from either the HSfE dataset or other literature, although a number including an equal ethnic breakdown across all age groups were unavoidable due to the availability of data, and no formal sensitivity analysis was carried out due to the complexity of the model assumptions. Finally we feel a need for more complete analysis to be carried out on the cost effectiveness of the programme.

Implications for practice

DH and NICE have specified the use of either QRISK2 or JBS2 in NHS Health Checks or when using statins for primary prevention of CVD[162]. Our modelling suggests the choice of risk score is important, with the two scores capturing different profiles of high risk patient. Although risk scores have been adapted and used in a wide range of settings, they perform better in a population similar to their derivation cohort[187]. Recent work has suggested the QRISK2 score may outperform Framingham scores in England[188], and a screening programme using it may cost less.

The high risk population in England suffers from a far greater burden of obesity and physical inactivity- two risk factors seeing a sharp rise in the developed world- than smoking and hypertension. Both are risk factors for a variety of other health conditions[189] and constitute health concerns in their own right. Their management must focus on more than a solely clinical approach[128], with wider lifestyle and community based interventions crucial. Local service providers involved in the NHS Health Checks must ensure there is sufficient provision to manage the community referrals generated by the programme. With the extent of both obesity and physical inactivity, it may be necessary to use a wider population approach to manage the societal causes, not simply manage individual cases as they arise.

The screening of high risk patients is estimated to make national reductions in cardiovascular mortality and admissions, greater than DH estimates of 650 deaths saved per year. Targeted screening has recently been suggested as a more cost effective approach than the universal method of the NHS Health Checks[167;168]. Risk prediction using data from patients' electronic records can be efficient in prioritising treatment[190];has been shown to be more accurate with improved data records[185], with some relevant areas of medical records largely complete[177]. The targeted approach may also be effective in reducing geographic health inequalities in England if there is successful risk reduction in the high risk population. Modelling presented includes the offer of a statin to all patients at a greater than 20% risk of CVD; this will generate large prescribing costs, with over 3 million patients eligible for therapy. Recent evidence has questioned the effectiveness of statins in their role in primary prevention[191], and further work is required to ensure the cost effectiveness of this national expenditure.

Conclusions

There is a high prevalence of patients at high risk of CVD in England, and therefore eligible for a series of interventions through the NHS Health Check. Within the high risk population there is a high prevalence of CVD risk factors, especially obesity in physical inactivity. The choice of CVD risk score will affect the nature of the high risk population identified by the programme and the costs of managing the high risk population.

9 Model Validation

Prevalence model validation is problematic for a number of reasons. There is no other major or definitive source of the national HSfE prevalence data we used for the models, apart from population-based prevalence research for specific diseases. The literature search before developing each model revealed what studies there were, and was useful in initially validating the models. Similarities and differences are discussed in the sections dealing with model development. Nor is there any similar local data for validation purposes. QOF data is useful to an extent, but as a major objective of developing the models was to support case-finding in practice populations, it ought not to be used to validate them. In general, we would expect modelled prevalence estimates to exceed QOF registered prevalence for 90 per cent of practices, with previously-described model limitations leading to under-prediction in perhaps ten percent of practices.

So how should the "fitness for purpose" of prevalence models, including validity, be defined? In the case of disease risk scores, the technical criteria are widely agreed.[188;192] The accuracy of a risk prediction score can be judged on two main components—calibration and discrimination. A well-calibrated score is one in which the predicted risk is similar to the observed risk. The more important component of accuracy is discrimination or the ability of a score to differentiate between people who will have an event from those who will not, over a defined period of time. There are two types of discrimination measures—summary measures and threshold measures. ROC curves are summary measures of discrimination, which can also be used for prevalence models as illustrated here. Threshold measures compare discrimination at a previously- agreed risk threshold. Net reclassification improvement at this threshold is commonly used in comparisons.

In the case of prevalence models, there is no such agreement as yet on technical criteria for ***internal validity***- a gold standard. This issue is considered in more depth in the Discussion. However this Section considers a number of other analyses of ***external validity*** using prevalence model data.

One method of assessing external validity is to map observed and expected prevalence and investigate associations with a low ratio of observed to expected cases, at both local authority, PCT, and/or practice levels. We did this for both COPD (Section 9.1) and cardiovascular diseases (Section 9.2, funded by the CQC). The gold standard validation would be comparison of model predictions with a comprehensive population survey or extensive case-finding efforts in a number of differing populations: deprived and advantaged, rural and urban etc. In the absence of local or national disease screening programmes (although this is planned for CVD risk), we obtained funding from the Lung Foundation to carry out a pilot COPD screening programme in Westminster PCT (section 9.3). The initial results are also presented in this Chapter.

9.1 COPD in England: a comparison of expected, model-based prevalence and observed prevalence from general practice data

9.1.1 Introduction

This section reports a national comparison of modeled COPD prevalence, using the model described previously,[59] with QOF-registered COPD prevalence. It was published recently in the Journal of Public Health. We present the estimated population prevalence of COPD for LAs in England, and compare them with the prevalence of GP-diagnosed COPD, as reported in national GP Contract QOF data.[16] We investigated geographical patterns in this ratio using classical and GWR to compare the methods.

9.1.2 Methods

Model development has been reported in *Population Health Metrics*[59] The HSfE is a representative population-based annual survey which in 2001 included the assessment of respiratory function using spirometry, as well as comprehensive data on risk factors. COPD was defined using the British Thoracic Society (BTS) criteria:[193] FEV_1) divided by FVC under 0.70, and FEV_1<80%of predicted using reference values from the HSfE[78]. However in the HSfE, spirometry was not carried out after bronchodilator challenge. The variables included in the model, based on their association with COPD in logistic regression analysis, were age-group, gender, ethnicity, smoking prevalence, area of residence (rural, suburban or urban) and area-based deprivation score.[79]

We previously externally validated the model by comparing COPD expected prevalence results to an alternative model, based on a survey of prevalence studies. The comparative prevalence was similar. We also applied the model to the Belfast, Northern Ireland population, and compared the results with those from a population survey of the same population.[88] Our results were slightly lower, but within the 95% confidence interval of those estimated from the survey (4.9% total prevalence in 40—69 year olds compared with 6.1% (95% CIs= 4.5 – 7.7) in the survey). Our prevalence estimates for the total England population were also similar to those in the Health Needs Assessment Report[194] and to other studies that used the BTS definition of COPD, albeit ours were in general slightly lower. The significant correlations between expected prevalence, and both diagnosed COPD and COPD mortality, gives us further reassurance of validity.

We also present data on the GP-diagnosed and registered prevalence of COPD, obtained from the QOF,[16] and COPD-specific directly standardised mortality rates for comparison.[195] QOF COPD prevalence estimates are based on populations registered with GPs. We derived residence-based registered prevalence estimates for LAs using a lookup table- a pooled extract of England GP registers- from the National Strategic Tracing Service (NSTS),[196] which matched GP practice populations to LA areas as at January 2006. We apportioned counts of COPD patients by practice to LAs, in accordance with the proportion of each practice population in that area, assuming that COPD prevalence was geographically uniform across a practice population. We divided this count of COPD patients by the mid-year LA population to give estimated crude prevalence. Where less than 50 patients fell into an LA, the numbers were excluded from the look-up process. Three LAs could not be mapped due to discrepancies between QOF and NSTS datasets.

Comparison of diagnosed with expected prevalence

The ratio of the QOF COPD observed or diagnosed prevalence to the expected prevalence was used to estimate unmet needs for COPD diagnosis. We compared the ratios across LAs, English Regions and within population sub-types (rurality and IMD quintiles), to identify possible risk factors for under-diagnosis. Simple and multiple linear regression analysis were used to investigate correlations between diagnosed and expected prevalence; and between these and directly standardised mortality rates for COPD.[158] Geographical clustering in the ratio was identified using the Local Moran's I cluster test, which measures associations with neighbouring areas.[197] The data were analysed using Microsoft Excel 2007©, Stata 9© and ArcGIS 9© software.

In classical regression we assume that a modelled relationship holds consistently across the study area. In many situations this is not the case, a phenomenon known as spatial non-stationarity.[198] We used GWR3© software to perform GWR modelling to investigate spatial variation in the association between the diagnosed: expected ratio. GWR is a local version of spatial regression that generates parameters disaggregated by the spatial units of analysis.[199-201] This allows assessment of spatial heterogeneity in estimated relationships between independent and dependent variables. Mapping GWR local correlation coefficients (strength of association), residuals (the difference between observed and predicted data) and Cook's D values (the effect of removing a single data observation from the dataset, a test of model robustness) provides an elegant method for modelling geographical variation in regression relationships. The disaggregation of the full dataset into spatial units for analysis is performed using a probability kernel which moves across the study area.

In order to estimate the optimal kernel bandwidth we used the Akaike Information Criterion.[202] We compared classical regression with GWR for the diagnosed: expected ratio by testing for spatial non-stationarity using a Monte Carlo simulation.[199] The results of the GWR analysis were mapped to visualize any geographic variation in the diagnosed: expected relationship. Two GWR models were run, the first a basic diagnosed: expected regression, and the second a bivariate regression involving diagnosed prevalence as the dependent variable with expected prevalence and primary care supply as the independent variables. The number of full-time GP equivalents per practice, sourced from the Information Centre,[203] was mapped to geographic areas in the same way as QOF prevalence above, using the NSTS lookup table. This was used as a proxy for primary care supply, to explore any potential association between COPD prevalence and physician availability. Unfortunately national data on practice nurse and other staff supply is only available at PCT level, and so could not be mapped as accurately to geographic areas.

9.1.3 Results

Expected prevalence of COPD

The overall estimated/expected prevalence of COPD in people over 15 years old in England was 3.58% i.e. just over 1.4 million people. LA prevalence ranged from 1.88% to 6.02%, with a median of 3·14% and inter-quartile range (IQR) 2·69 – 3.78 (see Figure 10: expected prevalence of COPD). The North West and North East regions had some of the highest prevalence observed. The overall expected prevalence in all age groups was 2.58% (95% CIs= 2.49 - 2.66). The rate for 15- 45 year olds was 1.32%, for 45-64s 4.22%, and for those

65-74 and over 75 7.93% and 8.72%, respectively. The mean expected prevalence by region varied from 2·90% in the South East of England to 4·02% in the North East.

Diagnosed prevalence of COPD

The mean prevalence of QOF-diagnosed COPD in LAs was 1·37% (95% CIs 1·33 – 1·42), and varied between 0.65% and 3.13% (median 1.29%, IQR 1·05 – 1·61). Figure 11: diagnosed/QOF registered prevalence of COPD shows the geographical distribution of diagnosed COPD. There is a north-south gradient; the prevalence of diagnosed COPD was generally low in southern England.

Ratio of diagnosed to expected prevalence

The ratio of diagnosed to expected prevalence in each LA varied from 0·20 to 0·95, with a mean of 0·52 (95% CIs= 0·50 – 0·53). There was pronounced variation ($p<0.001$) between urban (mean ratio 0.38, 95% CIs 0.36-0.40) and rural areas (mean ratio 0.56, 95% CIs 0.54-0.58). The mean ratio of diagnosed to expected cases in London was lowest at 0·31 (95% CIs= 0·29 – 0·33), illustrated in Figure 12.

Figure 10: expected prevalence of COPD

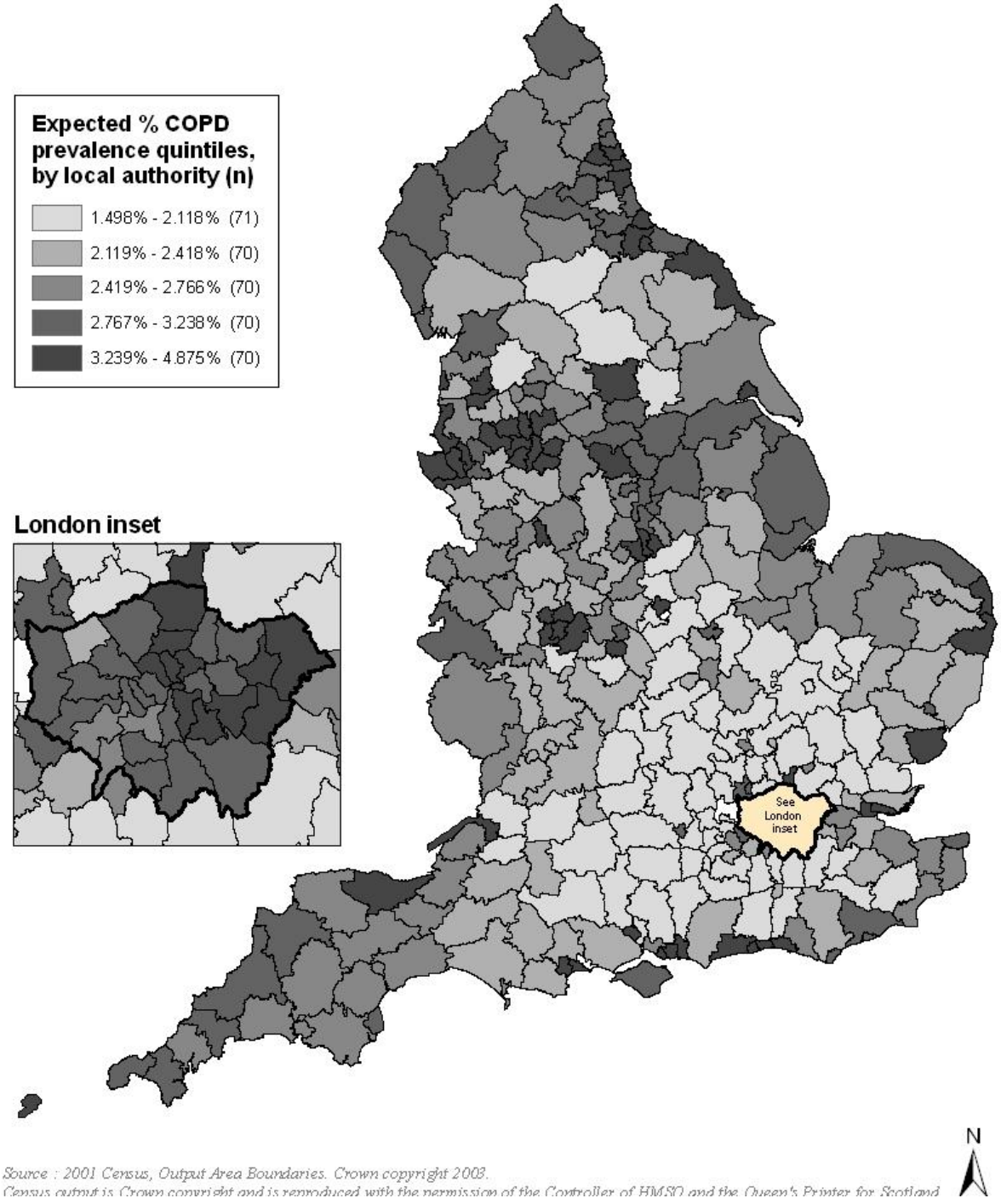
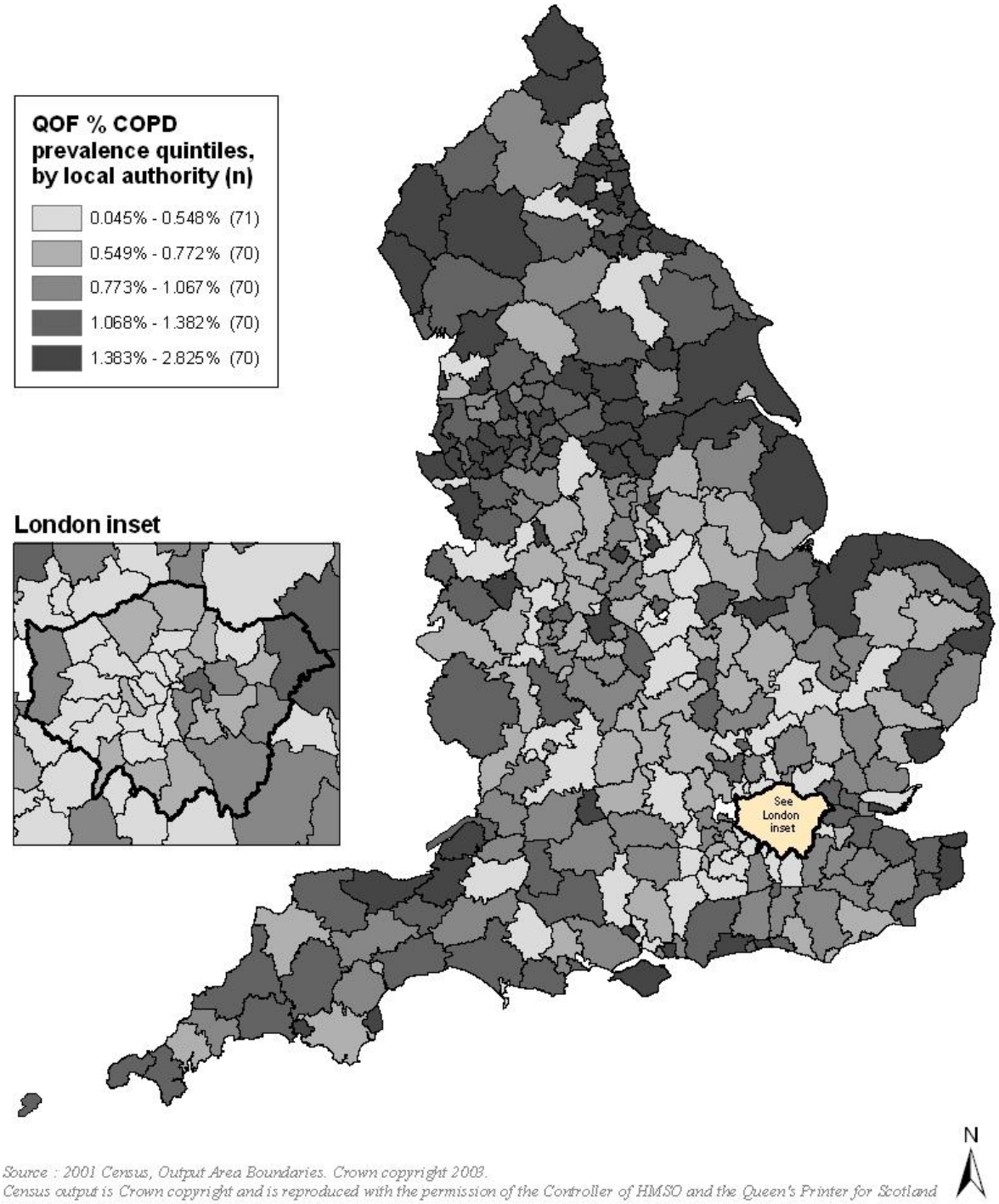
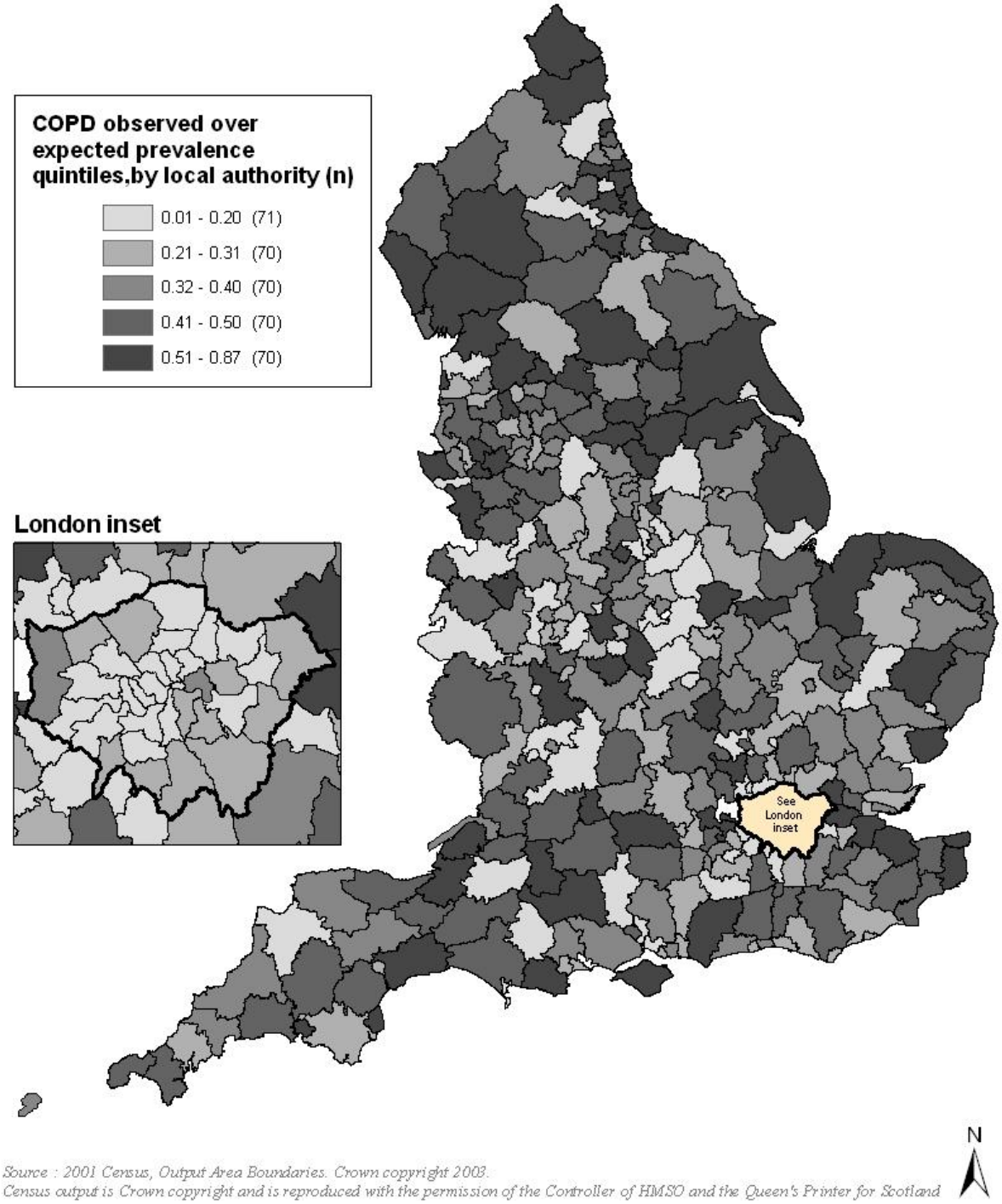


Figure 11: diagnosed/QOF registered prevalence of COPD



Source : 2001 Census, Output Area Boundaries. Crown copyright 2003.
Census output is Crown copyright and is reproduced with the permission of the Controller of HMSO and the Queen's Printer for Scotland

Figure 12: diagnosed over expected prevalence of COPD

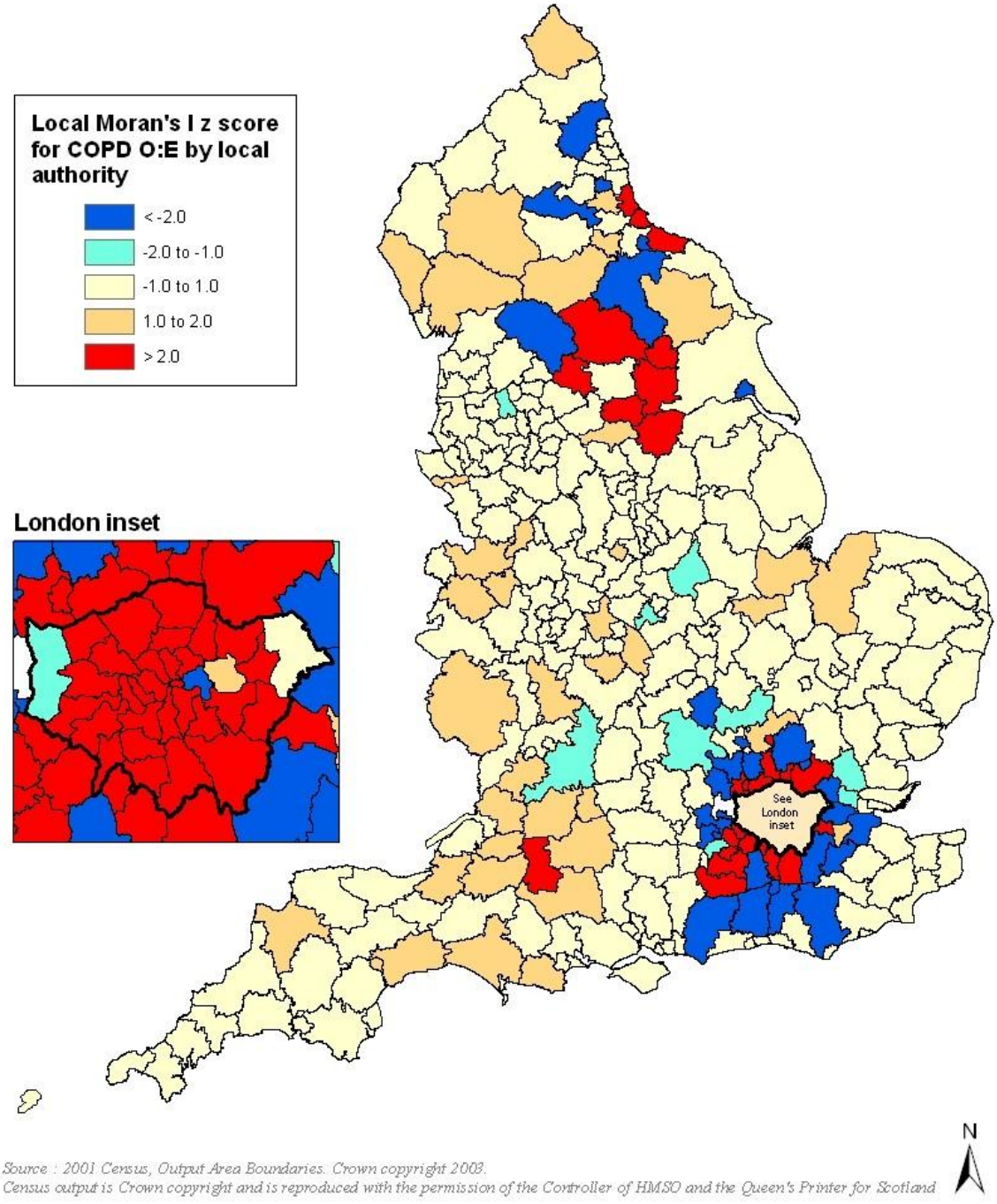


Comparisons between diagnosed and expected prevalence, and mortality

Figure 10, Figure 11, and Figure 12 illustrate the geography of diagnosed and expected prevalence. We found a moderate correlation between diagnosed and expected prevalence of COPD in people over 15 years old ($r=0{\cdot}48$, $p<0{\cdot}001$). The expected prevalence also showed a significant association with mortality risk ($r=0{\cdot}24$, $p<0{\cdot}001$) and diagnosed prevalence similarly ($r=0{\cdot}21$, $p<0{\cdot}001$). Figure 13: local Moran's I cluster significance scores

for the diagnosed: expected ratio. The z scores from the Local Moran's I analysis (Figure 13) indicated that there are statistically significant clusters in the diagnosed: expected ratio for London and its hinterland. Positive spatial autocorrelation is observed for London (coloured red) and negative spatial correlation for much of its hinterland (coloured blue) suggesting that the low diagnosed: expected ratios for London is a genuine cluster with sharp contrast to its hinterland. Some other parts of England, particularly in parts of Yorkshire and the North East, also showed significant positive autocorrelation, though unlike London these clusters were associated with relatively high diagnosed: expected ratios.

Figure 13: local Moran's I cluster significance scores for the diagnosed: expected ratio



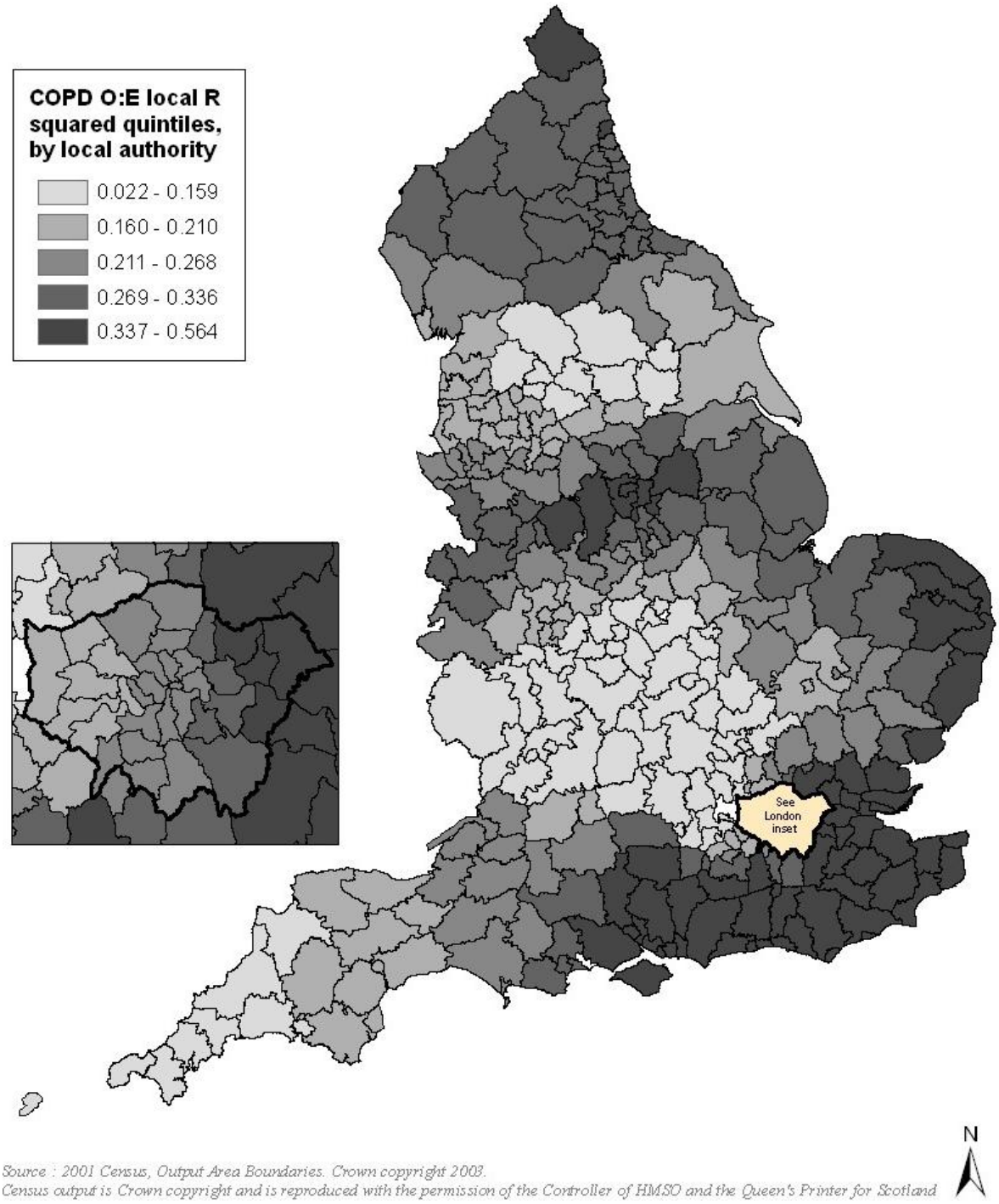
Comparison of correlation coefficients between classical regression results and those of GWR, as well as the use of analysis of variance and Akaike Information Criterion methods, confirmed that GWR modelled the dataset more accurately than classical regression (Table 53: comparison of classical regression and GWR results). The Cook's D and standardised residuals were small, the former indicating robust local regression results and the latter indicating no obvious outliers. Neither test showed evidence of geographic clustering. The Monte Carlo tests confirmed strong spatial non-stationarity in the diagnosed: expected association ($p<0.000001$). GWR analyses revealed clear geographic variation in the association at the LA level (Figure 14: local regression coefficients for the association between diagnosed and expected prevalence) with the weakest associations in the West Midlands, Cornwall and part of Yorkshire, and strongest in the South East and East Anglia. The inclusion of GP supply as an additional independent variable increased the local correlation coefficient for most LAs in both classical regression and GWR, particularly the latter (see Table 53: comparison of classical regression and GWR results) and especially for the East Midlands.

Table 53: comparison of classical regression and GWR results

Diagnostics	Classical regression		Geographically-weighted regression	
	O:E ratio	O:E+GP	O:E ratio	O:E+GP
Residual sum of squares	0.007210	0.006633	0.004940	0.003428
Standard deviation	0.004545	0.004366	0.003878	0.003329
Akaike Information Criterion	-2786.19513	-2813.43807	-2874.37046	-2955.98465
Correlation coefficient	0.17881	0.24456	0.43733	0.60958
Adjusted correlation coefficient	0.174100	0.23803	0.39862	0.55693
Sum of squares	0.0	0.0	0.0	0.0
Degrees of freedom	2.0	3.0	328.47	309.41

O:E: ratio of observed to expected prevalence of COPD
O:E+GP: inclusion of GP supply as an additional independent variable

Figure 14: local regression coefficients for the association between diagnosed and expected prevalence



9.1.4 Discussion

Main findings of this study

We compared diagnosed COPD prevalence in English LAs with those predicted by a model derived from HSfE data. The main findings were that both diagnosed and expected prevalence varied widely, and that the discrepancy between diagnosed and expected prevalence, as a measure of unmet need, also varied considerably, being significantly

greater in urban than rural areas. However, the highly mobile populations of all these urban areas also may be contributing to the low level of diagnosed prevalence. Local Moran's analysis revealed a statistically significant clustering of similar ratios comprising much of London, surrounded by a hinterland of dissimilar ratio values, indicating COPD under-diagnosis in this metropolitan area. Similar but less extensive clusters are evident in metropolitan areas of Yorkshire and the North East.

Differing patterns are evident from the Local Moran's I and GWR analyses, with the Moran's test highlighting clusters and outliers within Regions, whilst the GWR analyses summarise variation between Regions. They suggest that Regional boundaries may not be optimal for epidemiological purposes and that further analyses at finer geographic scales e.g. practice level may be warranted. The inclusion of GP supply as an additional independent variable in the GWR analyses yielded a stronger regression relationship for most parts of the country, supporting the notion of GP numbers as a factor affecting diagnosis levels, with implications for resource allocation.

What is already known on this topic

Our findings are consistent with a number of recent small primary care screening studies showing COPD under-diagnosis. In a survey in the Netherlands, the prevalence of self-reported asthma or COPD (9.7%) was almost twice as high as the prevalence based on GP information (5.2%).[204] Vandevoorde et al screened patients in six Belgian semi-rural practices if they were current smokers between ages 40-70, and had a smoking history of at least 15 pack-years.[205] Excluding those with known COPD, spirometry revealed a 29.5% prevalence, with greater under-diagnosis in younger age categories. Tinkelman et al recruited patients via random mailing from a primary care practice in Aberdeen, Scotland, and Denver, Colorado, who were current and former smokers aged over 40 with no prior diagnosis.[206] Of 818 patients, 18.9% had a post-bronchodilator diagnosis of COPD, which was mild in 57.4%, moderate in 36.8%, and severe in 5.8%.

What are the benefits of early diagnosis? There is some evidence that effective management of COPD may reduce the risk of hospital admission. Gotfredsen et al found a gradient of relative risk of hospital admission with increasing tobacco exposure.[69] Telling smokers their lung age significantly improves the likelihood of them quitting smoking, Practices could be incentivised to deliver smoking cessation advice and prescriptions, or to refer to stop smoking services. The updated NICE COPD guidelines that are currently out for consultation recommend that pulmonary rehabilitation should be made available to all appropriate patients with COPD, including those who have had a recent hospitalisation for an acute exacerbation, and that newer long-acting drugs be prescribed to appropriate patients.[207] In addition, combination treatment with long-acting beta-agonists and inhaled corticosteroids reduces the annual rate of COPD exacerbations and improves health status and spirometric values.[208;209] A nurse-led self-management programme for COPD patients reduced unscheduled practice attendances.[210]

What this study adds

Diagnosed COPD prevalence apparently grossly underestimates true population prevalence, a finding that supports the use of case-finding strategies. Under-diagnosis of COPD varies systematically and is more pronounced in urban areas, where the risks of COPD are also

higher. Under-diagnosis appears particularly severe in London, with up to five unknown cases for each case diagnosed in the worst-performing LA. However, relatively wide confidence intervals are involved, and combined with the results of the GWR analyses this suggests the model's results for London are less certain than those for the West Midlands, Yorkshire and North East conurbations.

GP supply appears to be a key factor affecting diagnosis levels. This suggests greater emphasis on GP-based COPD management would be beneficial, particularly in local authorities such as in the East Midlands, for which the inclusion of GP supply in our model markedly improved local correlation coefficients. Nevertheless there is considerable unexplained variation in the observed to expected relationship even after GP supply and the model variables are taken into consideration, especially in Cornwall and the West Midlands.

While opportunistic case-finding may be cost-effective, screening of even high risk groups may not. Practices may lack the resources to extend spirometry to invited patients. The variability in the ratio of diagnosed to expected prevalence may reflect patient and/or provider factors, and in highlights the need for additional case-finding support for practices in low diagnosis areas; many younger patients may attend infrequently and may need to be invited for screening. A strategy we are exploring in London involves applying the model at practice level and providing additional lung function testing alongside active patient recall. There is a strong case for evaluating the feasibility and yield of several such case-finding strategies and using the data to further validate the predictions of the prevalence model, especially at practice level.

Limitations of this study

The estimates need to be interpreted with caution and are subject to a degree of imprecision, particularly when applied to small populations. We used robust Census-based demographic data for each LA, but QOF data refers to the PCTs of patients' GPs, which does not always coincide with their LA area of residence, although they overlap considerably. We attempted to overcome this by apportioning practice populations to LAs, but this may have been erroneous in some areas. Limitations of the study may result from inaccuracies in the model parameters or assumptions, and from the quality of the input data. For example, we used current smoking status as a proxy for lifetime tobacco exposure. Post-bronchodilator spirometry is not part of the BTS case definition and was not carried out in the HSfE respondents, so the model results will slightly overestimate prevalence. The screening studies above show that in older age groups the proportion with bronchodilator-responsive airflow obstruction is small. We have confirmed this is a case-finding pilot study in London. However, a recent analysis of the BOLD study data found that that the spirometric prevalence of COPD was reduced by 25 per cent after bronchodilator challenge.[211] We carried out a sensitivity analysis by adjusting the expected prevalence accordingly, which changed the minimum value of the observed:expected ratio from 0.015 to 0.020, the maximum from 0.872 to 1.163, and the median to 0.363 to 0.484. This adjustment did not alter the findings of our study.

Misclassification of diagnosis at primary care level may also have had additional minor influence on the estimates of under-diagnosis. Some people not meeting the BTS definition may be misdiagnosed as having COPD by their GPs. This would increase the diagnosed but

not the expected prevalence, leading to an underestimation of the proportion with undiagnosed COPD. On the other hand, the numerator (diagnosed prevalence) will not include patients with known COPD criteria who do not appear in the QOF statistics e.g. non-NHS patients or those for whom diagnostic Read codes have not been entered. Although there are financial rewards for GPs to record diagnosis of COPD, if a GP believes that fulfilling QOF required criteria for payment is not in the interests of an individual patient they may avoid entering READ codes which include the patient in the QOF denominator. For a QOF included diagnosis such as COPD, where clinical judgement is explicitly sanctioned by NICE, deliberate systematic under-recording is inevitable.

Much of practices' screening activity is carried out by practice nurses. However, we were unable to examine any effects of the supply of practice nurses or other staff because national data is not collected in sufficiently disaggregated form. Some PCTs commission services delivered by specialist COPD nurses, but these staff mainly provide care for patients with severe COPD, so they are unlikely to affect initial diagnosis levels

Conclusions

GPs should examine how clinical judgment might account for the substantial differences in observed and expected prevalence we have shown. The distinct geographical variation suggest that these are not simply the results of good clinical judgment. Ultimately, the validity and utility of the model depends on its identification of practices where there is under-diagnosis. The COPD prevalence model now available on the APHO website now contains estimates for GP practices' registered populations, a number of PCTs are already using it as a tool for COPD case-finding, and it may also be useful for health planning, and to direct resources to areas of greater need. In other countries, national health surveys can be used to develop prevalence models in a similar way.

Two online resources based on the model are available at:
http://tools.erpho.org.uk/copdprev.aspx and
http://www.erpho.org.uk/viewResource.aspx?id=18024

The former allows the user to enter relevant populations, smoking prevalence estimates and other variables to generate COPD prevalence estimates for that population. The latter contains pre-calculated COPD estimates for English local authorities and primary care trusts, based on post October 2006 boundaries, with projections to 2020.

9.2 Variations in cardiovascular disease under-diagnosis in England: national cross-sectional spatial analysis

This section results from work on mapping CVD prevalence commissioned by the CQC, with help from Edgar Samarasundera, who provided GIS expertise and invaluable training in GWR including the use of GWR3 software. We were also joined by Tejal Indulkar, who carried out the actual mapping. This section has been submitted as a paper to BioMed Central Cardiovascular Disorders, and is still out for peer review.

9.2.1 Introduction

Geographic variation in the incidence and prevalence of CVD is a well-known phenomenon in population surveys. Such variations in both CHD and stroke incidence and prevalence are largely explained by area- and person-specific factors such as population socio-economic composition, demographic structure and ethnic diversity,[107;133;212;213] mediated through established CVD risk factors.[214] In the case of England, this is exemplified at a regional level by a North-South gradient in prevalence and outcomes, with higher prevalence in the North.[215]

From a health services perspective, it is important to ensure that as high a proportion as possible of actual disease prevalence has been diagnosed and is well-managed when secondary prevention is known to be effective, as is the case for CVD.[216;217] Under-diagnosis and/or under-treatment of CVD has been suspected in some previous UK studies,[218;219] but it was not possible to differentiate these causes. The recent availability of QOF registered prevalence data, has provided accurate counts of diagnosed disease prevalence, and assurance that identified CVD is being increasingly well-managed. For example, in 2008-9, achievement by English general practices of all the 89 QOF points available for CHD management was 99.1 per cent.[220] The QOF also provides a financial reward for registering new cases of disease.

A recent study which compared observed (QOF-registered) and expected prevalence of CHD and hypertension suggested that under-diagnosis may vary geographically.[221] However, the CHD model used by the authors was derived from General Practice Research Database (GPRD) data, resulting in a lack of independence between the observed and predicted datasets. A number of disease prevalence models are now available on the APHO website.[222] These provide independently-derived prevalence estimates for resident populations of English LA, PCT, and practice populations, which can be compared with QOF disease registers.[222] Using these models we investigated the observed and expected prevalence of three cardiovascular conditions (CHD, hypertension and stroke) at LA level to identify unmet population health care needs, their geographical variation, and population and healthcare predictors which might influence diagnosis.

9.2.2 Methods

Data sources: registered prevalence and primary care supply

Counts of practice-registered prevalence of CHD, stroke and hypertension in April 2007 were obtained from English practice disease registers, produced for the purposes of incentivizing practices for achievement of QOF treatment targets. QOF prevalence rates are based on total populations registered with practices, but to compare observed prevalence with resident population-based contextual data from the Census and other sources e.g.

deprivation and proportion ethnic minority population, we derived residence-based QOF prevalence estimates for LAs using a lookup table- a pooled extract of England practice registers- from the NSTS which matched practice populations to LA areas as at January 2006.[223]

We apportioned counts of CVD patients by practice to LAs, in accordance with the proportion of each practice population in that area, assuming that CVD prevalence was geographically uniform across a practice population. We divided this count of CVD patients by total mid-2006 LA population estimates to give estimated crude prevalence. Where less than 50 patients fell into an LA, the numbers were excluded from the look-up process. Three LAs could not be mapped due to discrepancies between QOF and NSTS datasets. In order to investigate the effect of healthcare supply upon diagnosis levels we included a measure of general practitioner availability in the form of the number of GPs per thousand LA population, calculated in the same way.[203]

Data sources: expected/estimated prevalence

Expected prevalence for each LA was obtained from the APHO epidemiologic models, which are based on the socio-demographic and behavioural characteristics of respondents with the respective conditions in the HSfE. To produce the models, HSfE data for the years 2003-6 was pooled to obtain sufficient cases of diseases. Surveys over this period included boosts for ethnic minority and older people, and focused on CVD and its risk factors. The outcome variables were patient-reported doctor-diagnosed CHD and stroke, and for hypertension, normotensive-treated, hypertensive-treated but uncontrolled, and hypertensive-untreated groups i.e. a combination of patient-reported and objectively-measured variables. Patient reports of doctor-diagnosed CHD and stroke have been extensively validated elsewhere.[96;97;224-226] Ordinary least-squares (OLS) logistic regression models were fitted and explanatory variables for each disease outcome identified by reverse stepwise selection.

The baseline odds of each disease were obtained directly from the HSfE dataset. The strength of association between each explanatory variable and disease caseness was then used to calculate the relative odds, which were applied to the baseline odds to derive the prevalence estimates for sub-groups of risk factors. The variables which can be included in each local model are limited by the availability of local data for them. The core model variables are (ten-year age band), gender (male and female), ethnicity (Asian/Asian British, Black/Black British and White) and deprivation (quintiles of Index of Multiple Deprivation 2004 scores).[79] The models use 2006 mid-year quinary and ethnic population estimates from the UK Office for National Statistics (ONS).[80] The latter dataset provides estimates in wider age bands, so the models assume that the distribution of ethnic populations is uniform across quinary age bands. LAs are stratified into deprivation score quintiles and other categories based on the ONS system used in the HSfE.

In the case of CHD smoking status is also included. Local smoking prevalence estimates are not available from the HSfE because of small sample sizes, so the CHD model uses synthetic estimates from the Neighbourhood Statistics website.[227] Model assumptions include that the proportion of smokers, ex-smokers and never-smokers is uniform across ethnic categories and that the proportion of ex-smokers in each age-sex group is constant across areas. Further technical details of the models are available on the APHO website.[222]

Spatial analyses

Observed: expected prevalence ratios for LAs were calculated in Excel 2007 and mapped using the geographic information systems package ArcGIS 9. Two exploratory spatial data analysis methods commonly used in geographical studies were used to investigate patterns in O:E relationships, Local Moran's I (LMI) analysis and geographically weighted regression (GWR). The LMI technique is used to identify geographic clusters and outliers in data by testing for randomness in spatial distribution across a dataset, localities with significance scores (Z scores) greater than two standard deviations being considered to be either clusters or outliers.[197] Strongly positive Z scores indicate statistically significant similar values in close geographic proximity hence the presence of a cluster; a strongly negative Z score demonstrates a locality with a significantly dissimilar value in relation to its neighbouring localities thus indicating an outlier.

GWR is a form of spatial statistics which disaggregates geographic data into spatial blocks using a probability distribution kernel, which moves from location to location across the dataset to test for geographic variation in regression relationships. In situations where there is geographic variation in the strength of a regression relationship, a phenomenon referred to as spatial non-stationarity, GWR will produce higher correlation coefficients, lower residuals and higher degrees of freedom than traditional ordinary least squares (OLS) regression.[198;199]

In this section, we used GWR to assess whether a linear regression relationship between observed and expected prevalence existed and if so whether it varied in strength over space, the purpose of which should be viewed as distinct from that of mapping observed to expected ratios, as the latter aims to measure equality between two variables rather than to assess predictability of an association. Both OLS and GWR models were run in the software package GWR3 to test for spatial non-stationarity. The optimal bandwidth for the kernel was estimated using the Akaike Information Criterion.[202] Two rounds of regression were performed, the first a univariate regression involving expected prevalence as the independent variable and observed prevalence as the dependent variable; the second bivariate including whole-time equivalent GP supply as an additional variable.

9.2.3 Results

Observed/diagnosed prevalence of CVD

Observed (O) and expected (E) prevalence summary statistics are shown in Table 54. Total English population was used as the denominator for consistency with standard reporting of QOF prevalence. The mean prevalence of QOF-diagnosed CHD in LAs was 2.57% (95% CIs 2.45-2.69), and the expected prevalence was 4.52% (95% CIs 4.42-4.61), giving an O:E ratio of 0.57 i.e. about 60 per cent of expected cases are diagnosed. Although the prevalence of stroke is less than half that of CHD, the O:E ratio is very similar. The expected prevalence of hypertension is much higher (about 24% of over 16s), and the O:E ratio is only 0.37 i.e. less than 40% of cases may be diagnosed. There was wide variation in the O:E ratio between LAs, with less than five per cent of expected cases diagnosed in some areas for all three diseases.

Spatial analysis

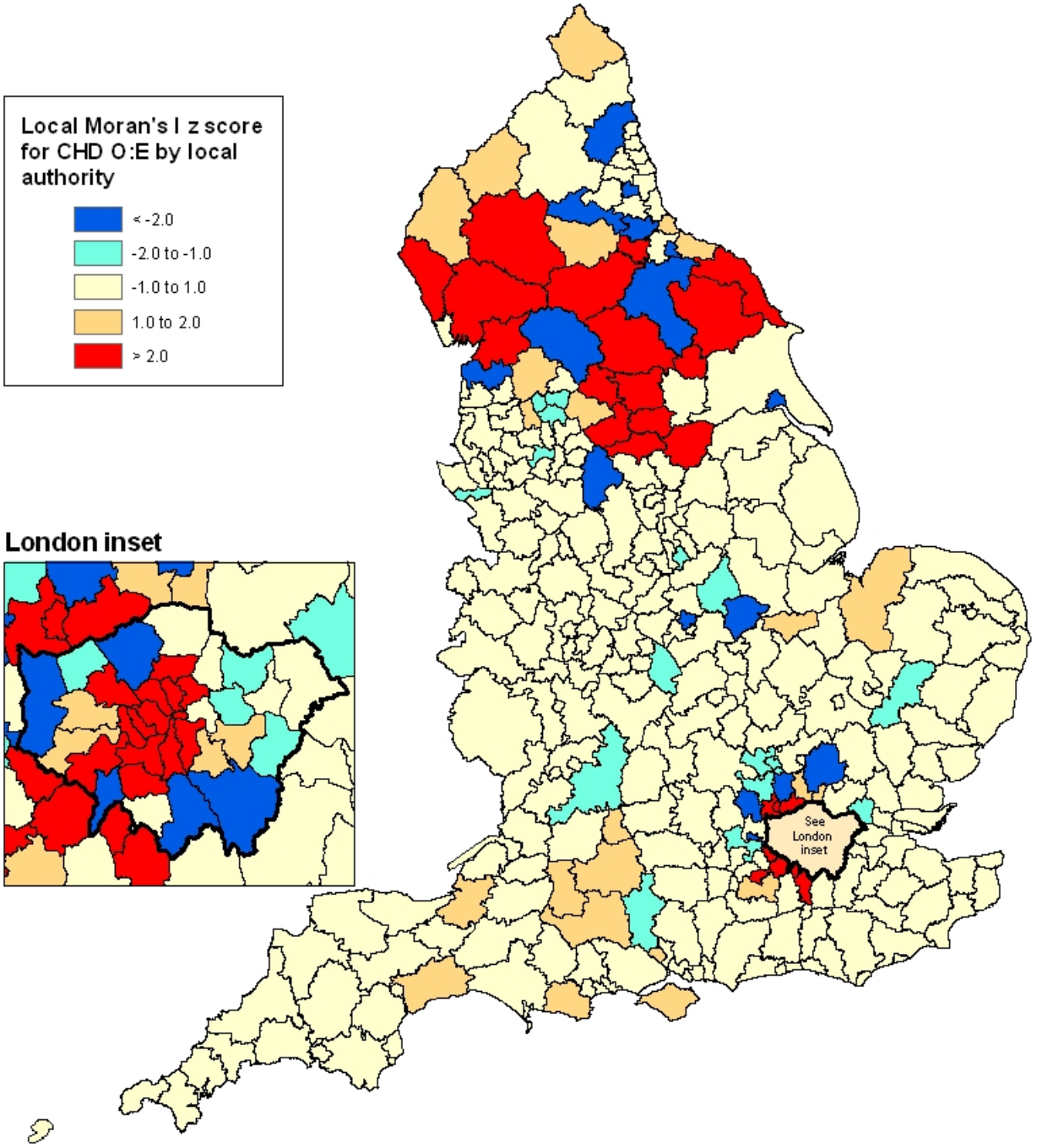
Mapping of observed and expected prevalence demonstrated that observed and predicted prevalence for both diseases showed strong north-south and southwest-southeast gradients, with the southeast generally showing lowest disease levels, especially in the case of expected prevalence. However, within these geographical trends in both QOF and modelled prevalence, mapping of O:E ratios demonstrated spatial variation between neighbouring LAs. LMI Z scores indicated the presence of some statistically significant clusters and outliers in O:E ratios (red indicates clusters and blue indicates outliers), as shown in Figure 15, Figure 16 and Figure 17).

Table 54: observed (GP-registered) and expected prevalence summary statistics

	CHD			Hypertension			Stroke		
	Observed Prevalence	**Expected Prevalence**	**O:E Ratio**	**Observed Prevalence**	**Expected Prevalence**	**O:E Ratio**	**Observed Prevalence**	**Expected Prevalence**	**O:E Ratio**
Number	1,299,601	2,286,597		4,530,369	12,356,995		588,500	1,009,144	
Percent (95% CIs)2	2.57% (2.45%-2.69%)	4.52% (4.42%-4.61%)	0.57 (0.55-0.59)	8.95% (8.59%-9.32%)	24.42% (24.13%-24.72%)	0.37 (0.35-0.38)	1.16% (1.11%-1.22%)	1.99% (1.96-2.03%)	0.58 (0.56-0.61)
Median	2.40%	4.48%	0.55	8.77%	24.77%	0.36	1.11%	1.99%	0.57
Maximum	5.47%	7.78%	1.10	16.62%	35.17%	0.63	2.48%	3.26%	1.08
Minimum	0.14%	2.61%	0.04	0.57%	17.71%	0.02	0.06%	1.24%	0.03
1st quartile	1.56%	3.89%	0.38	5.92%	23.06%	0.25	0.74%	1.77%	0.40
3rd quartile	3.31%	5.10%	0.69	11.10%	26.75%	0.43	1.49%	2.23%	0.73
IQR	1.75%	1.22%	0.32	5.18%	3.69%	0.18	0.75%	0.47%	0.33
Standard deviation	0.01147	0.00912	0.21	0.03486	0.02822	0.13	0.00513	0.00364	0.22

2 The England population denominator used for all three diseases was 50,598,200

Figure 15: clusters and outliers in observed to expected ratios for CHD (red indicates clusters and blue indicates outliers)



Source : 2001 Census, Output Area Boundaries. Crown copyright 2003.
Census output is Crown copyright and is reproduced with the permission of the Controller of HMSO and the Queen's Printer for Scotland

Figure 16: Clusters and outliers in observed to expected ratios for hypertension (red indicates clusters and blue indicates outliers)

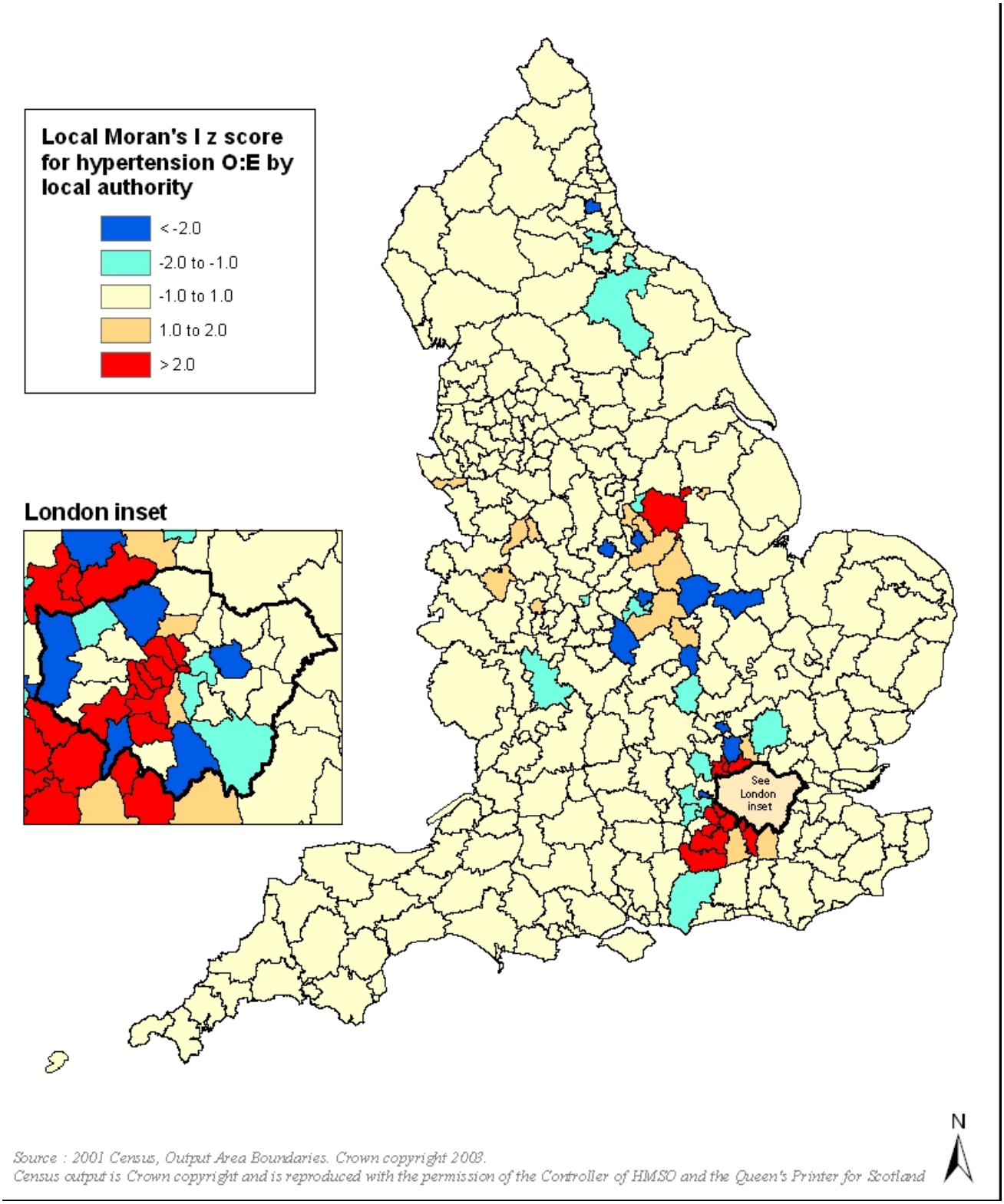
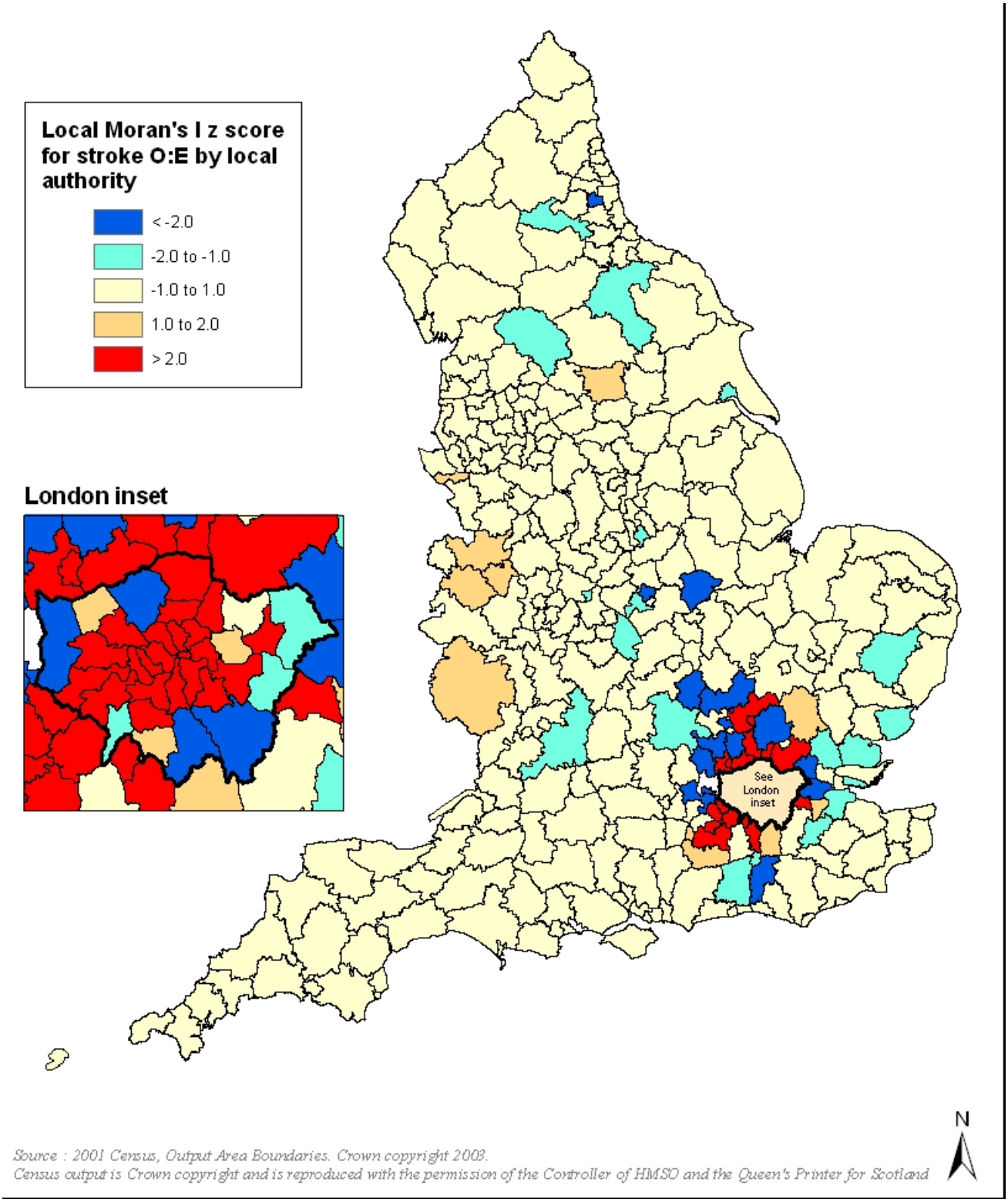


Figure 17: Clusters and outliers in observed to expected ratios for stroke (red indicates clusters and blue indicates outliers)



For all three conditions, particularly CHD and stroke, London and much of its hinterland showed statistically significant discrepancies in O:E ratios, with observed prevalence much lower than the epidemiological models predicted. For CHD there were also significant clusters and outliers in parts of northern England, especially in Cumbria and Yorkshire,

though unlike for the London area the clusters typically related to O:E ratios tending towards unity.

Further geographic analysis with GWR revealed significant spatial variation in the O:E relationship for CHD, hypertension and stroke, with GWR describing the dataset more accurately than traditional OLS regression. The linearity of the relationship between observed and expected prevalence was weakest for hypertension with the lowest overall regression coefficients and highest residual sum of squares for observed against expected prevalence; stronger associations were observed for CHD and stroke.

Even maximum coefficient values for hypertension, found along the south and east coasts, were less than 0.26; coefficients were lowest in the north and a pronounced north-south gradient was observed. The pattern for stroke was somewhat more complex with coefficients typically less than 0.21 for much of the northern parts of the country and the Midlands, yet with parts of Kent and the Sussex Counties having coefficients up to 0.69. CHD followed a similar geographic structure to stroke, with values of less than 0.08 for much of the Midlands and north, yet rising to 0.66 in eastern and southern coastal areas. There were no statistically significant patterns in Cook's D and standardised residual results for any of the three conditions examined. It should be noted that although O:E ratios tended towards 1 in the north and declined further south, the GWR results showed that the predictability of the relationship between observed and expected prevalence decreased with an increasing latitude. The addition of GP supply as a variable yielded stronger regression results for all three conditions (Table 55: comparison of classical regression and GWR results).

Table 55: comparison of classical regression and GWR results

CHD	**Classical regression**		**GWR**	
	O against E	**O = E+GP**	**O against E**	**O = E+GP**
Residual sum of squares	0.032246	0.025831	0.026362	0.016685
Standard deviation	0.009612	0.008615	0.008964	0.007374
Akaike Information Criterion	-2260.431534	-2336.249712	-2285.749801	-2393.809008
Correlation coefficient	0.299110	0.438554	0.427014	0.637331
Adjusted correlation coefficient	0.295081	0.433700	0.386880	0.584988
Sum of squares	0.0	0.0	0.0	0.0
Degrees of freedom	2.00	3.00	328.09	306.86
Hypertension	**Classical regression**		**GWR**	
	O against E	**O = E+GP**	**O against E**	**O = E+GP**
Residual sum of squares	0.373755	0.249962	0.361588	0.241345
Standard deviation	0.032725	0.026801	0.032380	0.026486
Akaike Information Criterion	-1400.410427	-1539.568186	-1403.498576	-1543.667517
Correlation coefficient	0.121033	0.412159	0.149645	0.432425
Adjusted correlation coefficient	0.115981	0.407077	0.134488	0.420914
Sum of squares	0.4	0.2	0.4	0.2
Degrees of freedom	2.00	3.00	344.87	344.04

CHD	Classical regression		GWR	
	O against E	O = E+GP	O against E	O = E+GP
Stroke	**Classical regression**		**GWR**	
	O against E	**O = E+GP**	**O against E**	**O = E+GP**
Residual sum of squares	0.006790	0.005595	0.005318	0.003484
Standard deviation	0.004411	0.004010	0.004049	0.003392
Akaike Information Criterion	-2807.247922	-2873.159418	-2838.923853	-2932.935545
Correlation coefficient	0.261902	0.391824	0.421924	0.621315
Adjusted correlation coefficient	0.257660	0.386566	0.374319	0.560939
Sum of squares	0.0	0.0	0.0	0.0
Degrees of freedom	2.00	3.00	324.37	302.87

O against E: ratio of observed against expected prevalence
O = E+GP: inclusion of GP supply as an additional independent variable

9.2.4 Discussion

Main findings

The findings presented here indicate that despite almost universal access to free primary healthcare services, and a significant financial incentive for these services to find and register new cases on their computer systems, there may be significant under-diagnosis of all three CVD conditions in many areas in England. Under-diagnosis is most marked for hypertension, with strikingly wide variation in diagnosis levels between areas. Hypertension and stroke showed similar geographic variation to CHD, but without the clusters/outliers for northern England; additionally the contrast between the ratios for London and LAs immediately to the north was more pronounced for stroke.

An obvious question is: why was there such a discrepancy between expected and practice-registered disease prevalence, when the model-based CHD and stroke prevalence estimates were based on patient reports of doctor-diagnosed disease? Reasons are likely to include inadequate searching of practice records (which are now in the UK mainly electronic in primary care) for previous diagnoses or CVD-related prescriptions,[228;229] lack of linkage of practice and hospital records (practice still have to enter codes for hospital admissions manually), and high population mobility in urban areas with concomitant changes (or delays) in registration with general practices.[230] In the case of hypertension, for which the modelled estimates used a combination of doctor-diagnosed disease and blood pressure measurements, the gap is greatest in younger and middle age groups, when males in particular are less likely to be seen by their practices.[184] In contrast, diagnosis appears to be more complete in the north of England.

From a spatial analytic perspective, there are significant geographic discrepancies between QOF prevalence and modelled prevalence for all three conditions in much of London and its hinterland. The inclusion of a measure of GP supply in the GWR analyses suggests that, despite needs-based healthcare resource allocation in the UK, persistent differences in availability of primary care services is an important limiting factor in diagnosis. An analysis of Gini coefficients to measure geographical equity in GPs per 100,000 population in England

and Scotland showed that equity in England rose between 1974 and 1994, but then decreased, and in 2006 it was below the 1974 level.[231]

The results also provide some contrast to a study investigating revascularisation rates in males, which found that even after adjusting for the higher CHD burden in the north of the country related to socio-demographic composition, the likelihood of receiving surgery during the 1990s decreased outside of southern England, suggesting a geographic imbalance in the provision of tertiary cardiology care.[232] Considerable efforts have been made subsequently to ensure that access to tertiary care is more equitable. However our analysis suggests that from a primary care perspective, it is London and its hinterland which would benefit most from increased resource allocation, although GP supply itself seems to be only one influential factor affecting diagnosis levels. Further work will need to investigate the effectiveness and yield of strategies for CVD case-finding by practices, and to validate the predictions of the prevalence models at practice level.

Strengths and limitations of the study

Strengths of the study include the use of new data from practice disease registers and recently-developed geographical analytical techniques and software. Limitations include the fact that it was carried out at LA level, which may conceal much wider variation at lower levels. The spatial scale of the analyses includes many LAs which are highly heterogeneous in socio-demographic composition, so the study may have missed significant small area variation; this is especially likely in northern areas where regression associations were weakest. The prevalence models are constrained by the availability of local data on risk factors, some of which rely on 2001 Census data, although this will be improved by, for example, more accurate and timely data on smoking prevalence from the new ONS Integrated Household Survey.

Conclusions

Despite the absence of barriers to primary healthcare, it is likely that there is significant and highly variable under-diagnosis of CVD across England, which can be partially explained by persistent inequity in GP supply. However, it is also plausible that the distinctive composition and dynamics of London's population adds complexities to the identification of CVD in that region. Studies of disease management should consider the impact of this "iceberg" of undiagnosed disease on hospital utilisation and population health outcomes. Spatial analytic techniques can provide additional information about geographical variation compared to classical regression modelling, and can suggest where additional healthcare resources may be most needed.

9.3 Case-finding strategies for COPD in primary care

This section results from work carried out with respiratory physician colleagues at ICL and Imperial College Healthcare Trust- Drs Nick Hopkinson, Sarah Elkin and Irem Patel. The pilot study itself was carried out by Julia Kelly, who is based at Royal Brompton & Harefield NHS Trust.

9.3.1 Background

This pilot study was funded by the British Lung Foundation and had an overall aim of testing the feasibility of targeted COPD case-finding or "targeted screening" in an urban, deprived general practice to assess the challenges involved. It also offered the potential of validating the COPD prevalence model. As noted previously, risk factors for COPD include smoking, age, area of residence, occupation, ethnicity and deprivation. The practice, situated in Westminster PCT, was in a deprived area with a multi-ethnic and highly mobile population, which was thought to have a high prevalence of smoking.

There are numerous obstacles to COPD case-finding. For example, 89 per cent of people in UK have never heard of COPD.[233]. As noted previously (Section 9.1), there is geographical clustering of observed: expected prevalence ratios which is particularly prominent in London, possibly because of high population mobility. There have been a number of primary care COPD screening studies in other European countries, but none in England (see Table 56 below). Various age bands were screened. An important issue is which definition of COPD to use: the BTS definition[77] FEV_1) divided by FVC under 0.70, and FEV_1<80%of predicted using reference values from the HSfE) is much more conservative than the GOLD definition[91], and so gives a lower prevalence.

Table 56: summary of primary care COPD screening studies

Authors	Location	Study type	Sample size	Age groups	Prev-alence
Mohangoo et al[204]	National, Netherlands	Random patient population survey (self reports)	19,685	All	9.7%
Vandevoorde et al[205]	Semi-rural, Belgium	Systematic opportunistic screen of current smokers	146	40-70	29.5%
Tinkelman et al[206]	NS, Scotland & Colorado	Random mailing of smokers & ex-smokers	818	>40	18.9%
Bednarek et al[234]	Semi-rural, Poland	Mailing of whole practice population	1,960	>40	11%
Stratelis et al[235]	Small city, Sweden	Practice posters & advertisements, current smokers	512	40-55	17%

9.3.2 Methods

Objectives

The objectives and outcomes defined for the project are in Table 57 below.

Table 57: objectives and outcomes of pilot study

Objective	Outcome measure
to validate smoking status recorded in electronic patient records	smoking prevalence in GP system and at screening
to estimate the prevalence of COPD in a diverse urban population	frequency of abnormal spirometry
to assess the feasibility of two directed COPD screening strategies compared to current practice	response rate to letter/telephone invitation or invitation on attendance for other reasons
to provide initial validation of COPD prevalence model predictions	post-study observed: expected prevalence
to provide initial data on the relationship between respiratory symptoms and muscle strength	associations between spirometry, respiratory symptoms and quadriceps strength
to act as a pilot for an RCT of the cost-effectiveness of COPD screening	response rates, costs

Data collection

The study aimed to perform spirometry screening of 320 over 35s with a smoking history in the practice. Patients were informed that a "Lung Health Screen" was being offered. This involved a short symptom questionnaire, smoking history and MRC dyspnoea score , a medication discussion, height, weight and BMI measurement, spirometry including 'lung age' calculation and COPD staging (Table 58), bioelectrical impedance analysis, quadriceps strength using a hand-held dynamometer, feedback to patient, including smoking cessation advice, referral if results abnormal including post-bronchodilator spirometry.

Table 58: definitions of GOLD COPD stages

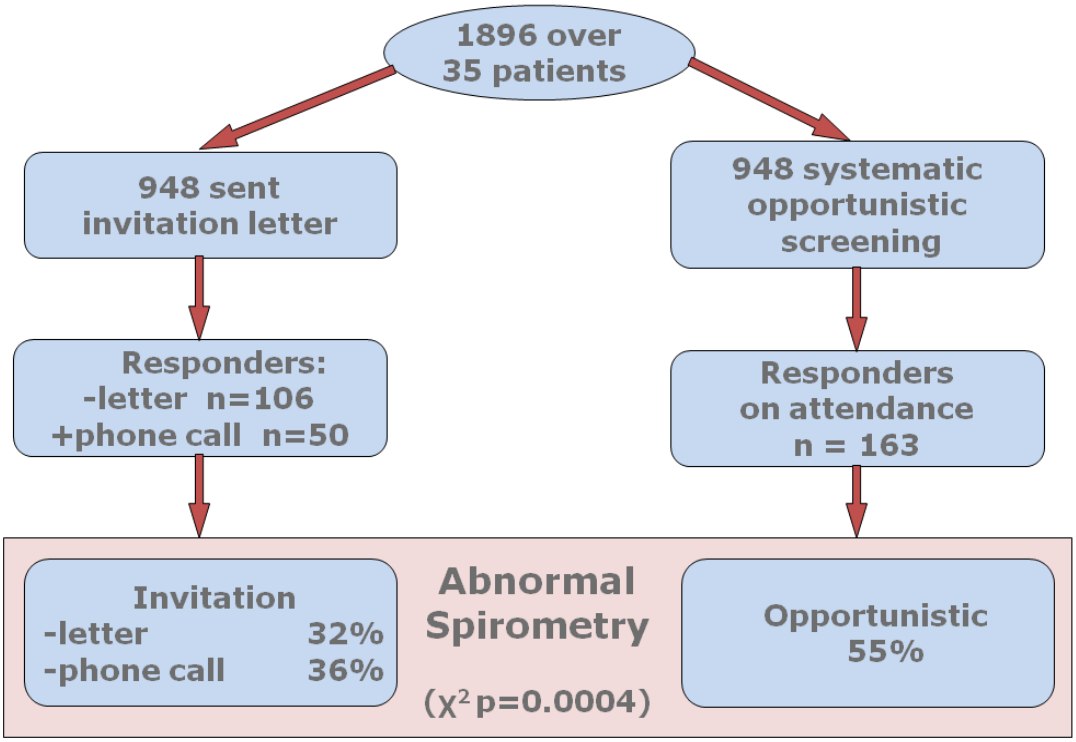
GOLD Stage	Spirometry	Symptoms
0: At Risk	Normal spirometry	Chronic symptoms (cough, sputum production)
I: Mild COPD	FEV1/FVC <70%, FEV1≥80% predicted	With or without chronic symptoms (cough, sputum production)
II: Moderate COPD	FEV1/FVC<70%, 30%<FEV1≤80% predicted	With or without chronic symptoms (cough, sputum production, dyspnoea)
III: Severe COPD	FEV1/FVC<70%, FEV1≤30% predicted	or the presence of respiratory failure, or clinical signs of right heart failure

Results

A total of 1896 patients over 35 were selected from the practice register. Of these, 969 (51%) were recorded as smokers, but because of concern over accuracy of this data, all 1896 randomised to either a letter/phone call or invitation for opportunistic screening when they

attended (see Figure 18: study design & flow chart). Non-smokers were asked to disregard the letter.

Figure 18: study design & flow chart



Smoking status of screened population is shown in Table 59 below. Concordance between the smoking status at time of spirometry and practice computer system data (n=257) is shown in highlighted cells. Only 13 smokers (four per cent of the over 35 sampled population) would be misclassified as non-smokers if the practice data were used.

Table 59: smoking status of screened population

Spirometry Screening	GP system			
	non-smoker	ex-smoker	smoker	Total
Non-smoker	21	1	0	22
Ex-smoker	26	89	20	135
Smoker	4	9	147	160 (50.4%)
Total	**51**	**99**	**167 (52.6%)**	**317**

Initial spirometry results and GOLD category are shown in Table 60 below. As expected patients with more severe disease (GOLD Stage 3 and 4) were more likely to be on the practice COPD register.

Table 60: spirometry results and GOLD category of screened patients

	Spirometry n (%)	On COPD register n (%)
Normal	177 (55.5)	-
Restrictive pattern	45 (14.1)	-
Obstructive pattern	97 (30.4)	24 (24.7)
GOLD stage 1	33 (10.3)	2 (6.1)
GOLD stage 2	46 (14.4)	10 (21.7)
GOLD stage 3	14 (4.4)	9 (64.3)
GOLD stage 4	4 (1.3)	3 (75.0)
Total	**319 (100)**	**24 (7.5)**

Based on this sample, what is the actual prevalence of smoking and COPD in this practice? Study and practice system results are both very similar, giving a smoking prevalence of over 50 per cent. This seems very high, but the CQC practice-level smoking prevalence estimates also predict some practices have prevalence of 50-60 per cent. Before screening, there were 42 patients on the practice COPD register. The COPD prevalence model predicted a prevalence of 5.2 per cent- but according to the more conservative BTS definition. The detected prevalence in the sample was 15.3 per cent using the GOLD definition. At this stage, we do not have data on the age-based predicted FEV1 from this sample, but these figures correspond with national estimates quoted in Section 10.4.3 Results using both definitions.

Table 61: Smoking and COPD prevalence in pilot practice

Population	Smoking prevalence[1]	GP registered/ observed COPD prevalence[1]	Modelled COPD prevalence[1]
England	24.1%	1.8%	3.1%
Westminster PCT	23.1%[2]	0.8%	3.6%
Harrow Road pre-study	23.1%[2]	1.2%	3.8%
Harrow Road post-study	51%?	2.8%	5.18%
Harrow Road possible	51%?	15.3%?[3]	N/A

1 all denominators = 15+ or 16+ population
2 NatCen modelled smoking prevalence
3 assuming study sample is unbiased

9.3.3 Discussion

In summary, of 319 over 35 current/ex smokers 177 (55.5%) had normal spirometry, and 97 (30.4%) had an obstructive pattern. Only 24 of these were previously known to have COPD. Only 106/948 invited patients responded to a letter, and a further 50 to a follow-up phone call- 163 were screened using systematic, opportunistic screening at the time of an attendance for other reasons. There was abnormal spirometry in 32 per cent of letter responders, 36 per cent of call responders and 55 per cent of patients screened opportunistically (χ^2 test p=0.0004). All methods effectively detected cases, but the yield was greatest from systematic, opportunistic screening at the time of an attendance.

Systematic, opportunistic screening may be the most efficient strategy, but this can only be confirmed by a cost-effectiveness analysis based on a much larger study.

This pilot study has a number of limitations, not least its small size. It is based on a single practice, which may be unrepresentative. There may be sampling bias, in that responders to letter/phone invitations and practice attenders may be more likely to have disease. Diagnosis based on spirometry will over-estimate actual prevalence: complete post-bronchodilator diagnosis data is awaited, but the great majority so far have an irreversible/obstructive pattern. Practice data under-estimated the number of ex-smokers, and practice data missed a small proportion of smokers, which may have implications for screening. On the other hand, it demonstrated the feasibility of screening in a deprived urban practice. In rural or suburban areas response rates to invitations are likely to be much higher. A logical third strategy may be to include screening with NHS (Vascular) Health Checks.

In terms of prevalence model validation, it did not over-predict in this practice. This will be confirmed by including the second BTS criterion in the data analysis. A letter has recently been published in Thorax which [236] reports on financial incentives for COPD case-finding in Kensington & Chelsea PCT. Between 2005 and 2008 there was a linear increase in COPD prevalence in K&C (r^2=0.997). If the preceding trend had continued, the predicted COPD prevalence for 2009 would have been 0.87% (95% CI 0.84% to 0.90%), whereas following the introduction of the LES it was 0.98%. The modelled COPD prevalence for the PCT in 2008 is 3.9%.

10 Model Implementation

The four disease prevalence models I developed have had significant impact on public health practice since they were disseminated. Dissemination has occurred through the APHO website as previously noted, and through NHS Comparators, a website established by CfH.[237] This Chapter explores how the models have been implemented and used to support primary care quality improvement nationally.

10.1 Evaluation of the uptake of the models by York University Health Economics Consortium

An evaluation of the uptake of the models was undertaken by York University Health Economics Consortium (Dr Peter West and Sophie Tatlock) in 2009.[238] A national telephone survey of 104 PCTs (out of 145) was carried out to ascertain how many PCTs were using the models and how many were not. For those PCTs who had confirmed that they had used the models, public health analysts in 75 PCTs were interviewed regarding types of use, barriers to use and suggested improvements; 14 Directors of Public Health (DsPH) were interviewed regarding the outcomes of model use and the potential impacts upon the service planning and commissioning within PCTs. A thematic analysis was conducted on all qualitative data. Themes and codes from interviews in Phase 1 were quantified for frequency analyses to be conducted. Frequency analyses were not conducted on Phase 2 interview data due to the small sample size and therefore limited amount of data obtained.

Findings from the interviews in Phase 1 indicated the models were being used by the majority of PCTs in a sample of 104 and that users appreciated the potential benefits that the models offered. The most frequent forms of use reported were:

- comparisons of prevalence estimates with QOF registered prevalence figures
- case finding
- service planning
- performance monitoring
- health needs assessment
- strategic commissioning plans
- health equity audit; and
- annual public health reports.

However, the extent of model use to inform or change World Class Commissioning processes was dependent upon a number of limiting factors and reported barriers to use. These fell into four main categories, which also serve as indicators of requests for improvement:

Flexibility: the PBS Diabetes model was reported to be used most frequently and was valued for its additional flexibility in comparison to the other models. The PBS Diabetes model goes down to ward level data and enables the user to input their own (most commonly practice level) data, whilst at the time of the survey the other models only went down to PCT or LA level. Although the extent of use varied and was often dependent

upon the users' knowledge of model usage outcomes, basic comparisons of prevalence figures with QOF registered prevalence was the most frequently reported form of use. Therefore, the ability to work with practice level data in the Diabetes model led to its further use, and requests for more flexibility in the other models were common in the responses. The inability to extract practice level prevalence estimates from the other 5 models was reported as a limitation upon the extent of use.

The Doncaster PCT prevalence models were reported to be used the most frequently by users, when not using APHO models. The Doncaster PCT prevalence models produce estimates at practice level, which was reported as an attractive option for analysts, with QOF comparisons being the most frequently reported form of use. However, one of the authors of the Doncaster models has confirmed that they were developed rapidly for local use based on rapid scans of literature, and they should not be assumed to be the most accurate models. It would be appropriate for them to be replaced by a set of robustly-researched, nationally agreed models offering similar functionality.

It was also noted that practice profile tools are increasingly being developed by PHOs. These tools offer prevalence estimates at practice level for the associated PCTs, and were not known to most respondents. This is perhaps indicative of the need for APHO to 'advertise' their available tools and resources to PCTs in a more effective manner.

Transparency: a lack of methodological detail and clarity of assumptions within the documents which accompany each model was reported as a limitation by users. This reportedly reduced the "statistical" confidence of users and therefore limited the extent of use. A frequent issue reported was that of approaching GPs for case finding visits to discuss differences between prevalence estimates and QOF figures. With GPs questioning the figures and analysts being unable to provide enough methodological detail to support the figures, confidence was reduced and therefore the productivity of GP meetings, in stimulating case finding initiatives, was also reduced. This issue in particular regards the Diabetes model only, as it is the only model to offer the function of inputting local practice level data.

A lack of transparency of the assumptions within the models, as well as the associated caveats and confidence intervals, were reported as a limitation of use by some analysts. A commonly reported challenge was other members of staff with non-statistical backgrounds misusing the estimates and treating them as "true" figures within documents, due to a conceptual lack of understanding of confidence intervals. A clearer presentation of the assumptions and caveats, as requested by users, would reduce the chances of this misuse, as well as the inclusion of confidence intervals within the actual models themselves.

A lack of a clear help desk function was also reported as a limitation to use, with many users being unsure as to whom to contact for what were often simple queries. A more transparent help desk function would reduce the amount of time taken to answer simple queries. A 'Frequently Asked Questions' page on the APHO website might also have been helpful.

Consistency: limitations to use regarding consistency were reported in three main areas:

Accessibility: The models were reportedly difficult to locate, both within the APHO website itself and through internet search engines. The links for all models can be found on the APHO website but the models themselves are hosted on different sites (mainly Eastern Region PHO and Yorkshire and Humber PHO). To compound these issues, the models have undergone various Phases of development, and various methods of approaching the models led to users reporting uncertainty regarding which model Phase they were using. There were frequent requests for the models to be centrally hosted on the APHO site, to improve accessibility and certainty of using the most up to date tools and resources.

Control and updates: further to the issues of accessibility, users reported the lack of control over the models and updates as a barrier to use, in particular being notified of new and upcoming work. Requests for user registration upon download of the models, as well as email notifications of any problems with the models or upcoming work were frequent. This would assure users they were using the most up-to-date models, and enable them to maintain contact with APHO through version control and a release policy.

Model format: the various formats of the model (with the Diabetes model seen as the most user-friendly and intuitive) were reported as barriers to use, in particular when other members of staff from non-statistical backgrounds attempted to use the models or their outputs. A single interface was proposed to be used for all models. This would ensure all users were comfortable with the interface and would provide a familiar APHO format for public health analysts within PCTs.

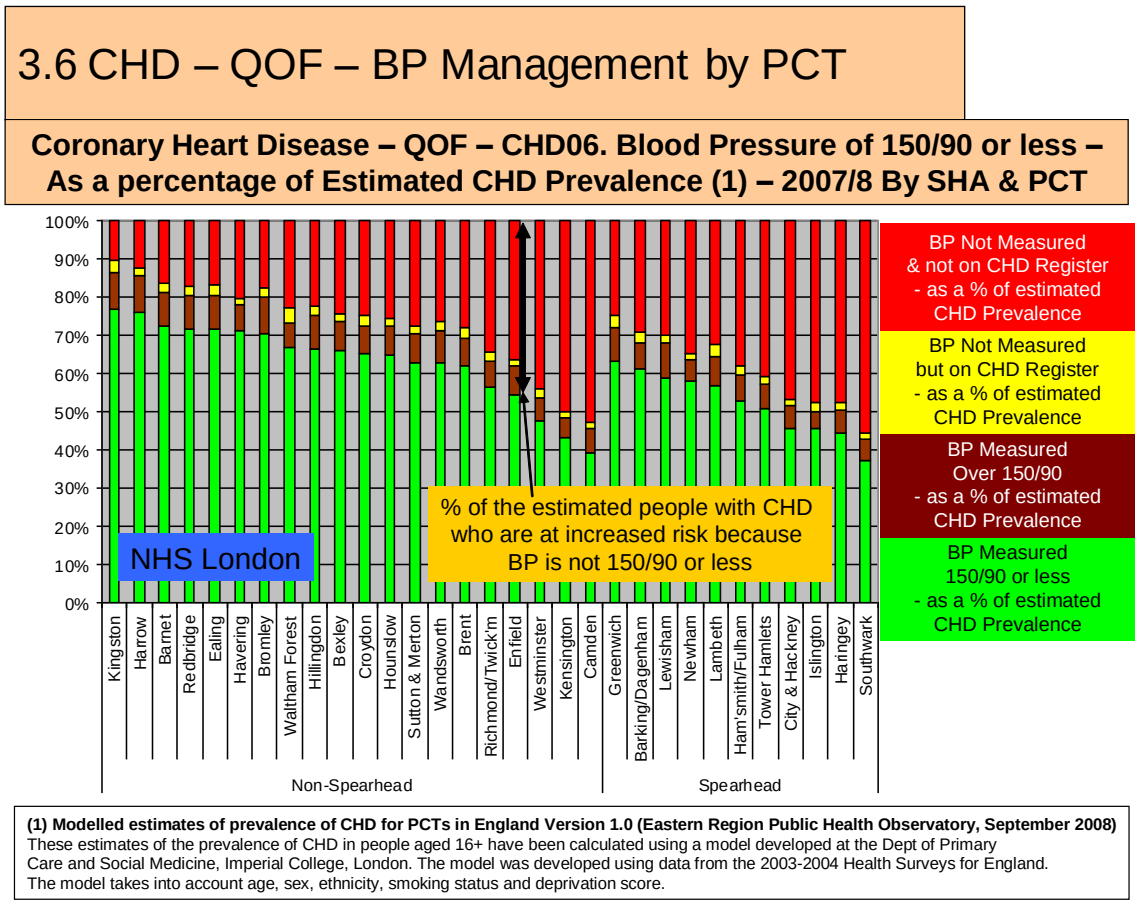
Training and Support: the qualifications, experience and skill base of public health analysts varied greatly across the sample. The various levels of skills were potentially linked to the fact that some users encountered practical issues with the models whilst others did not. A lack of full communication in PCTs was also noted, with analysts often not aware of the use of their outputs. This led to two recruitment Phases to find out more about the use of outputs from DsPH. Some DsPH interviewed also raised issues about the skills of staff and about being generally understaffed within their public health departments.

In response, a proportion of analysts requested further training to be offered to them, to streamline their career pathway within the PCT (currently not viewed by most as apparent) and to also improve their skill base. In particular, there were requests for training to develop their own models, which can be tailored towards the requests of clinical leads and commissioning. Not only would this enable PCTs to produce more localised data but it would also potentially save costs of commissioning third parties (i.e. external consultants) to develop models on their behalf. Basic statistical training was also suggested as an improvement for all staff members who encounter prevalence estimates and models within PCTs, including members of commissioning teams, where "misuse" was reported. This would improve the understanding of estimates and their role within the world class commissioning process.

10.2 Vascular Programme briefing packs

In 2009, the DH Vascular Policy Programme incorporated prevalence estimates from the CHD and stroke models into its Vascular Programme Briefing Packs.[239] PCT-level modelled estimates were used as denominators for QOF indicators (rather than disease register totals) so that the extent of undiagnosed disease was highlighted (see Figure 19).

Figure 19: use of modelled CHD prevalence in Vascular Programme Briefing Packs

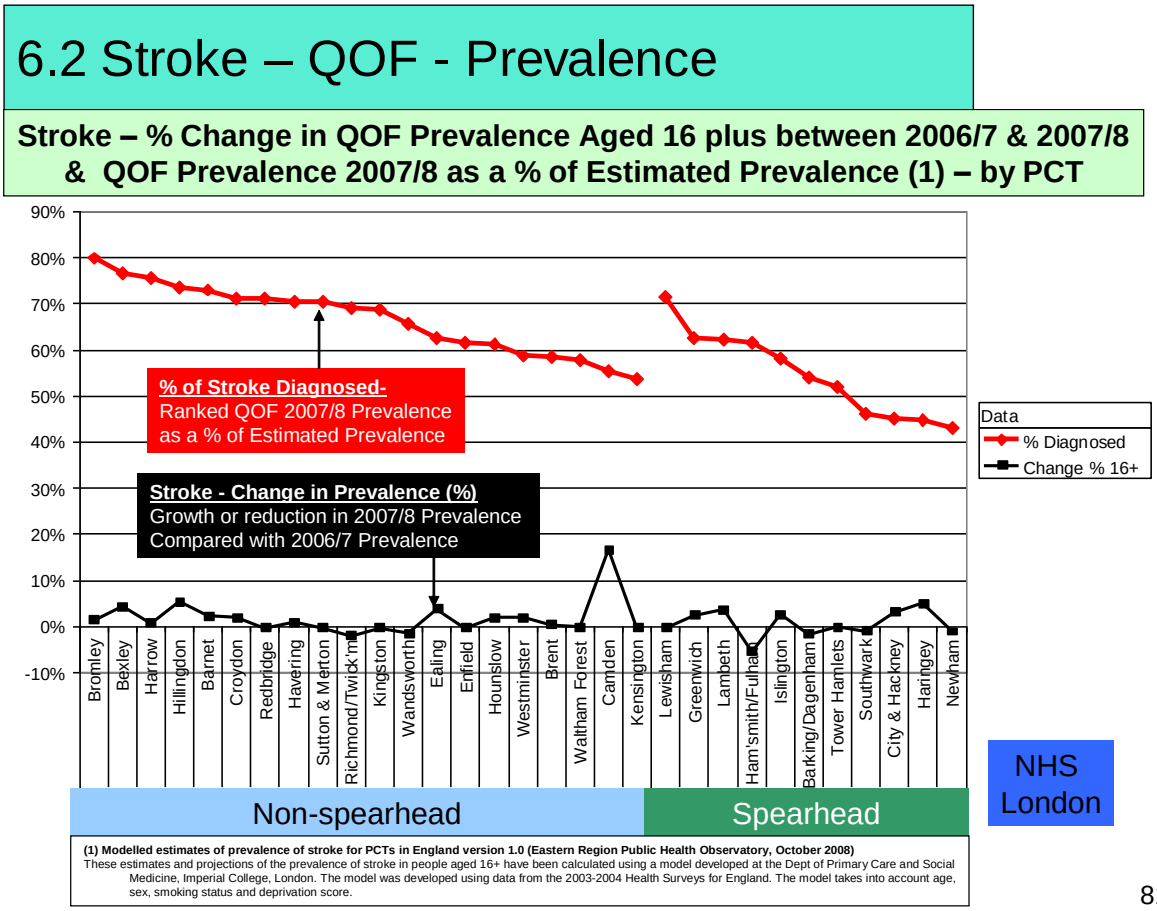


(1) Modelled estimates of prevalence of CHD for PCTs in England Version 1.0 (Eastern Region Public Health Observatory, September 2008)
These estimates of the prevalence of CHD in people aged 16+ have been calculated using a model developed at the Dept of Primary Care and Social Medicine, Imperial College, London. The model was developed using data from the 2003-2004 Health Surveys for England. The model takes into account age, sex, ethnicity, smoking status and deprivation score.

16

Similarly, expected prevalence was used to show the extent of unrecorded disease by Spearhead and non-Spearhead PCT (see Figure 20).

Figure 20: use of modelled stroke prevalence in Vascular Programme Briefing Packs



82

10.3 Case-finding pilots and case studies

Five case studies explain how the disease prevalence models are being used by PCTs, for Joint Strategic Needs Assessment (JSNA), case-finding, preparing for NHS Health Checks, and targeting investment to improve the health of communities**.**

10.3.1 Case Study 1: Northamptonshire PCT

Northamptonshire PCT used the COPD model to inform a programme of work to identify and address significant health inequalities. They were able to identify possible under-diagnosis of COPD in a number of GP practices. As a result, the PCT has commissioned additional screening for certain groups and in certain geographical areas of the trust, and in particular in those areas where the model suggests COPD would be more prevalent, because of levels of deprivation, smoking and even some industrial pollution issues.

This should lead to significant health benefits from earlier diagnosis. The strength of the model was that it enabled the trust to target resources at a specific problem with confidence that they had correctly identified what was going on and where. If the tool had not been available, the Trust's analysts would have had to try to calculate expected prevalence from a range of disparate published sources.

The work would have been more labour intensive, and the analysts would perhaps only have been able to look at COPD as part of the current JSNA. Instead they were able to look in detail at prevalence of not only COPD, but of all the diseases for which models have been made available.

David Pearce, Senior Public Health Analyst at Northamptonshire, explains: *"Having centrally produced numbers from a respected source, using robust methodology, makes a real difference and is less open to challenge and criticism. It gives a better understanding of what is going on in the Trust area and provides a reality check. By comparing the figures from the model to the figures emerging in the QOF, it became clear that there were significant gaps in diagnosis in some areas. The updated tools allowing comparisons at GP Practice level will enhance their use even further."*

10.3.2 Case Study 2: South West Essex PCT

South West Essex PCT has been using the COPD model with data on QOF prevalence and emergency admissions, to identify GP practices who might have incomplete disease registers, or those who have complete disease registers but are not managing their COPD patients as well as others. Quadrant charts are produced that categorise GP practices into 4 groups:

Quadrant 1: Lower than expected recorded prevalence and high emergency admissions
Quadrant 2: Higher than expected recorded prevalence and high emergency admissions
Quadrant 3: Lower than expected recorded prevalence and low emergency admissions
Quadrant 4: Higher than expected recorded prevalence and low emergency admissions

This information is to be used to support GP practices depending on the needs of the patients. The PCT also identified high intensity users of emergency admissions for COPD (and other disease areas) in the GP practices in Quadrants 1 and 2, so that the PCT could

report the information to the GP who can in turn review whether they have the person on their disease register and whether they have had a recent medication review or other clinical intervention.

Emma Sanford, Senior Epidemiologist at South West Essex, added: *"The next step is to pass the lists to specialist COPD nurses to make sure the patients are registered on their case load. It was also possible, using the costing templates, to attach costs (tariffs) to these emergency admissions, which we can then ensure are offset against any costs of improving the COPD care pathway."*

10.3.3 Case Study 3: Leicestershire County and Rutland PCT

Leicestershire County and Rutland PCT has used all of the models and has them embedded in the JSNA process. CVD mortality is a key World Class Commissioning issue for the PCT, and they are using the CHD, hypertension and PBS diabetes models to try to help plan for the possible outcomes of extended NHS health checks.

They were able to identify a gap between those with CVD who have been picked up already by GPs, and the numbers that the new health checks are likely to. The costing templates then enabled them to estimate the costs of picking up these additional cases. They were able to calculate the likely impact on the medicines budget of ensuring that people are put on to the right medical regime.

The model tools have created an evidence base unavailable before, and Janine Dellar, Associate Director of public health intelligence at Leicestershire and Rutland says that without them, the PCT would have had to implement NHS Health Checks with less understanding of the impacts.

"Before the models were produced, we would have used a range of other, but less satisfactory models, none of them as well worked through as the models produced by APHO. Over time we will need to be clear about how good they are by reviewing the outcomes. It will also be important to be sure that there are user guides so that people are using them properly. The fact that some of the tools are now available to operate at GP practice level is a major step forward. It is important for APHO to be very proactive in letting users know about such useful developments."

10.3.4 Case Study 4: South Staffordshire PCT

South Staffordshire PCT has used the models for CHD, stroke, hypertension, COPD and CKD to compare estimates of expected prevalence with QOF data on recorded prevalence on GP disease registers and also to predict the impact of an ageing population on the numbers with these long term conditions. The data was used to produce a report which will help commissioners planning health services across different local authorities in Staffordshire. They are also included as part of the JSNA.

Without the model, the analysts would have estimated prevalence from HSfE data and other epidemiological reports, but with serious health warnings about their not accounting for differences in smoking, deprivation and ethnicity levels. The PCT particularly appreciated the fact that estimates of smoking, deprivation and other factors affecting

diseases are used within the models and also that the methods and assumptions are comprehensively documented.

Divya Patel, Public Health Epidemiologist at South Staffordshire, added: *"The recent enhancement of practice level estimates will enable us to drill down further and will be useful in terms of monitoring health inequalities and identifying unmet need. It is also useful to have comparative local authorities and PCTs against which to benchmark."*

10.3.5 Case Study 5: Nottinghamshire County PCT

Nottinghamshire County PCT took the 2006/07 QOF data and compared it with the model estimates across four conditions, by practice and by geography, to help map unmet need. They found the surprising result that the amount of undiagnosed COPD in 2007 was higher in parts of the southern, less deprived area of the County. This borders an area of high deprivation where there is less undiagnosed COPD, suggesting that the GPs there have done better in targeting and preventing the disease. The next step was to reapply the model to the current QOF data, looking at unmet need by GP Practice, including looking at GP Practices in other bordering PCTs, who have patients who live in Nottinghamshire PCTs area – quite a significant number.

This created an important opportunity to work in partnership with bordering PCTs, contacting all GP Practices with patients in the Nottingham PCT wherever they are located. This was only made possible because the model is available for all GP practices in England. David Gilding, Health Improvement Practitioner at Nottinghamshire County PCT, explained: *"We could only do this because we had the models for the whole of England and for small geographical areas, such as wards. The models are helping us to build a really detailed local picture. We've used the tools to look at what is happening in QOF and because the data comes from a validated source, people are more likely to accept and have confidence in it."*

10.4 Modelling risk factor scenarios and impact on disease prevalence

This section comes from an unpublished report on stroke needs commissioned by London PCTs from PHAST. The lead author of the report was Dr John Hayward.

The concept of population attributable risk (PAR) is fundamentally important for evidence-based public health decision-making, because it allows for the quantification of the impact of a risk factor on a population, as opposed to an individual level.[240] PAR measures potential impact of control measures in a population. Thus, the considerable increase of studies addressing the PAR in the course of time is encouraging.[241] It is usually defined as:

PAR= (prevalence exposure (RR-1))/(1+ prevalence exposure(RR-1))

where RR = relative risk. Because stroke is a comparatively rare outcome, we can assume RR= odds ratio[240]

Estimates of the attributable number of deaths (AD) from all causes can also be obtained by first estimating PAR adjusted for confounding covariates, and then multiplying the PAR by the number of deaths that occurred in the population for a specific time period. Proportional hazard regression estimates of adjusted relative hazards obtained from

mortality follow-up data from a cohort is combined with a joint distribution of risk factor and confounders to compute an adjusted PAR.[242] Current and future disease prevalence can be attributed to individual risk factors by calculating PAR. This has been done by WHO to attribute overall disease burden to individual risk factors.[243]

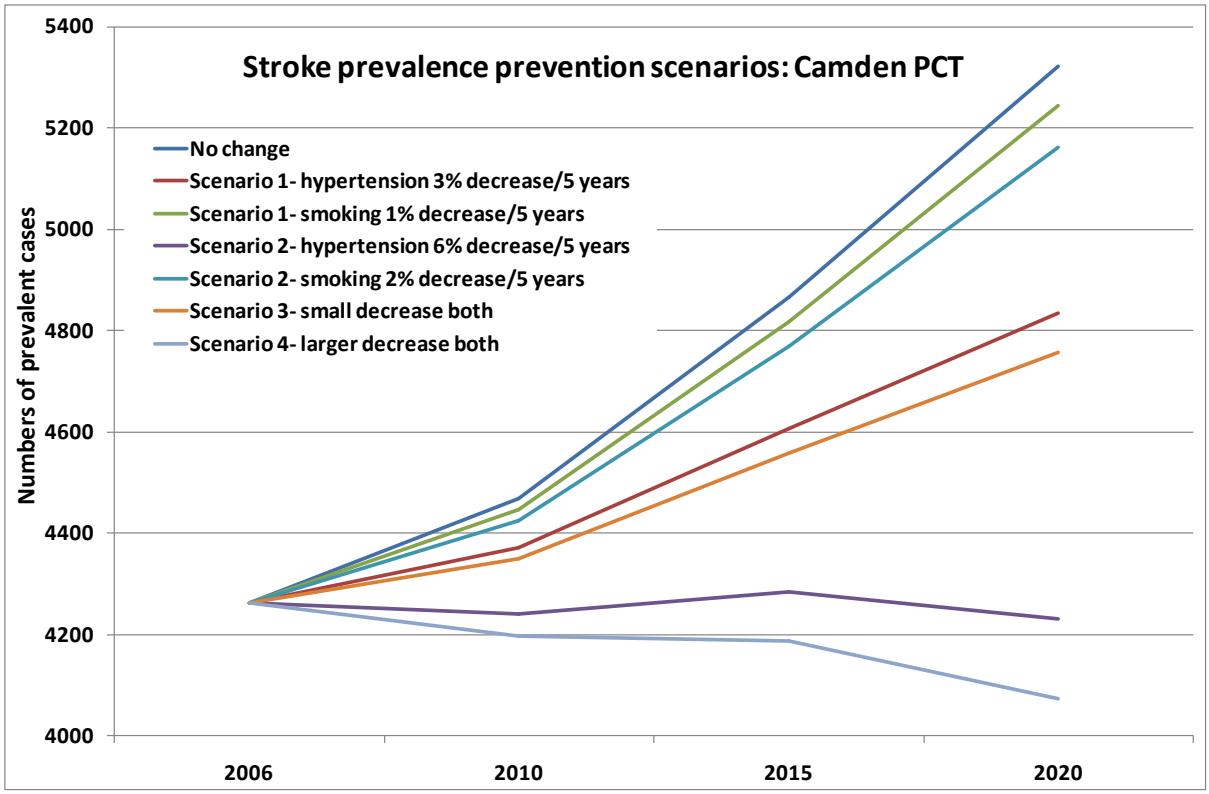
10.4.1 Predicting the impact of risk factor reductions on stroke prevalence

WHO has also looked specifically at the effect of changes in hypertension prevalence on stroke prevalence and disease burden.[244] We were commissioned by London Cardiac and Stroke Networks to predict the likely impact of reducing two important stroke risk factors by differing extents – hypertension and smoking. To do this, I used the stroke prevalence model and PAR estimates. The PAR of smoking for stroke in England is 13.3% and of hypertension 34.8%. Thus about half the risk for stroke is preventable from controlling these two risk factors together, without considering the additional contribution of atrial fibrillation or obesity. Using this information, I modelled the impact of four feasible primary prevention scenarios on stroke prevalence:

- **Scenario 1a** assumed that smoking prevalence falls by 1% from the current level by 2010, a further 1% by 2015 and a further 1% by 2020, but the prevalence of hypertension is unchanged.
- **Scenario 1b** assumed that the prevalence of hypertension is reduced by 3% over each five-year period, but the prevalence of smoking is unchanged.
- **Scenario 2a** assumed that smoking prevalence falls by 2% from the current level by 2010, a further 2% by 2015 and a further 2% by 2020, but the prevalence of hypertension is unchanged.
- **Scenario 2b** assumed that the prevalence of hypertension is reduced by 6% over each five-year period but the prevalence of smoking is unchanged.
- **Scenario 3** assumed that smoking prevalence falls by 1% from the current level by 2010, a further 1% by 2015 etc, **and** that the prevalence of hypertension is reduced by 3% over each five year period
- **Scenario 4** assumed that smoking prevalence falls by 2% from the current level by 2010, a further 2% by 2015 etc, **and** that the prevalence of hypertension is reduced by 2% over each five-year period.

An example of the impacts is shown in Figure 21: Stroke Prevalence scenarios, Camden PCT; other PCT charts (and charts for the London sectors as a whole) were also produced. National trends in smoking and hypertension prevalence were included, along with an Excel-based modelling package which allowed PCTs to enter their own scenarios.

Figure 21: Stroke Prevalence scenarios, Camden PCT



The figures shown on the vertical axis are the predicted prevalence numbers for the PCT concerned – (i.e. number of individuals in that PCT who will be living having had a stroke by the year concerned).

In the absence of a change in risk factors, a substantial increase in stroke numbers is expected as the population ages. The reason for this is the dominant impact of demographic change between now and 2020 across the whole of London. The important element of that change is the increase in population age during the period. Age is a very powerful determinant of overall stroke risk, and the prevalence model shows that it will have a dominant impact on future stroke prevalence unless there is a significant reducing effect from smoking cessation or better blood pressure control. Reductions in smoking prevalence would reduce the rate of stroke increase, but changes in hypertension prevalence would have greater impact. Best impact can be achieved in Scenario 4 (serial reduction of 2% in smoking, plus 2% reduction of hypertension every five years). In this scenario, in most PCTs' stroke prevalence would have fallen slightly between 2006 and 2020.

This methodology can also be used to apply multifactorial exposure situations and embraces various types of attributable fractions including adjusted, sequential and average attributable fractions. Scaled Venn diagrams can be used to illustrate total disease risk and excess disease risk attributable to the exposures as areas in a unit square. This allows display of simple pie charts of attributable fractions summing to 1 (or 100%).[245;246] However it is necessary to employ adjustment techniques to account for the effect of other –

confounding or effect-modifying – factors to obtain valid estimates for the effect of the factor under consideration.

10.4.2 Comment on time lags

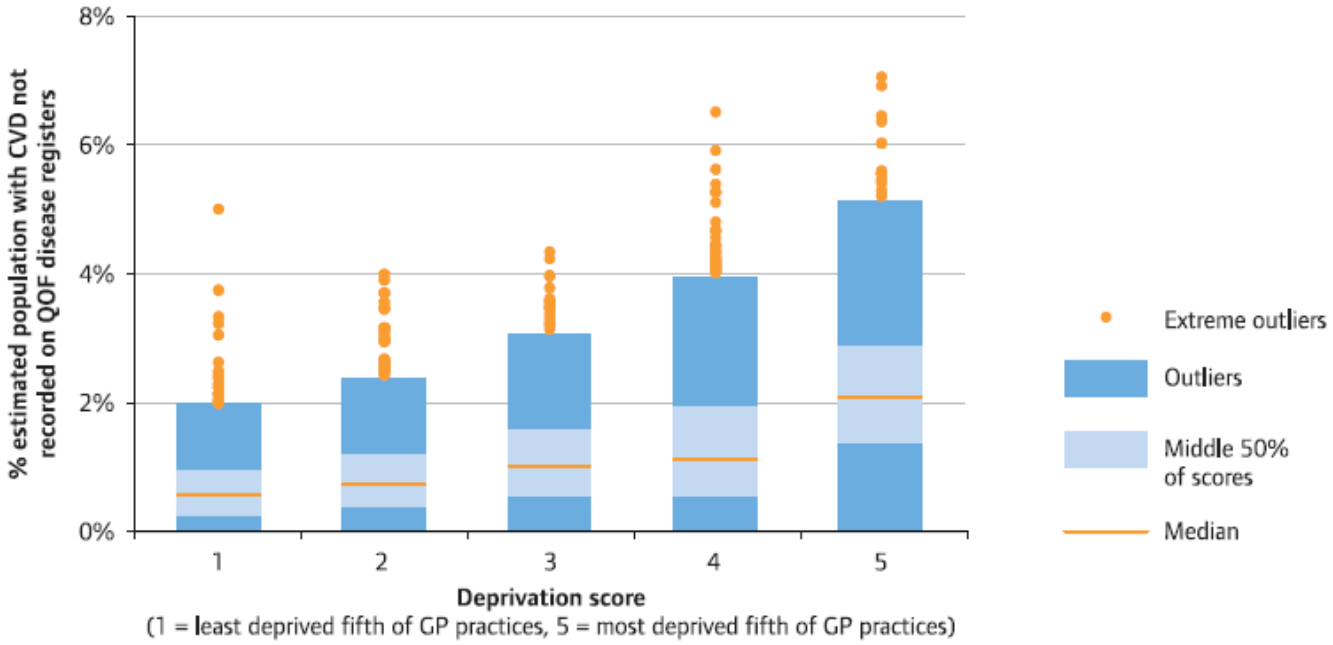
I did not incorporate any time lags for the effects of prevention greater than five years in the model, although this is possible with a Markov model,[35;40] because there is evidence that reduction in CVD risk occurs rapidly after smoking cessation or control of hypertension. The US Nurses Study showed that the risk decline after smoking cessation occurred for CHD and total CVD within two years, and for stroke after 2-4 years.[247] Similarly, the Japanese JACC study showed that for mortality alone, much of the reduction in the excess risk were realized within the first five years for CHD disease and stroke.[248] Sixty-one percent of the full potential benefit of quitting in regard to CHD mortality and 42% of the full potential benefit of quitting in regard to stroke mortality was realised within the first 5 years of quitting smoking. There is less evidence about risk reduction rates after control of hypertension, but it is biologically plausible that this happens quite quickly.

The hypertension model uses as an outcome overall hypertension (treated controlled, treated uncontrolled and untreated). A rapid reduction in overall prevalence would only result from improving management of treated hypertension and finding new cases of established hypertension, and not from "primordial" prevention e.g. by increased physical activity. In other words, a rapid reduction in overall stroke prevalence requires clinical action *in addition* to broad population health improvement initiatives.

10.5 CVD prevalence modelling by the Care Quality Commission

The CVD prevalence model was commissioned by the CQC to assist them in assessing the performance of PCTs in supporting practices in the task of identifying and managing CVD. The CQC published a report on their investigations, entitled *Closing the gap: Tackling cardiovascular disease and health inequalities by prescribing statins and stop smoking services*.[249] The model was used to help answer the question: does estimated unrecorded CVD prevalence increase with deprivation? Figure 3.2 in the report (shown here as Figure 22) shows that the level of estimated unrecorded prevalence in each practice increased with greater deprivation.

Figure 22: estimated unrecorded CVD prevalence by deprivation at GP practice level



Source: Care Quality Commission. *Closing the gap: Tackling cardiovascular disease and health inequalities by prescribing statins and stop smoking services*

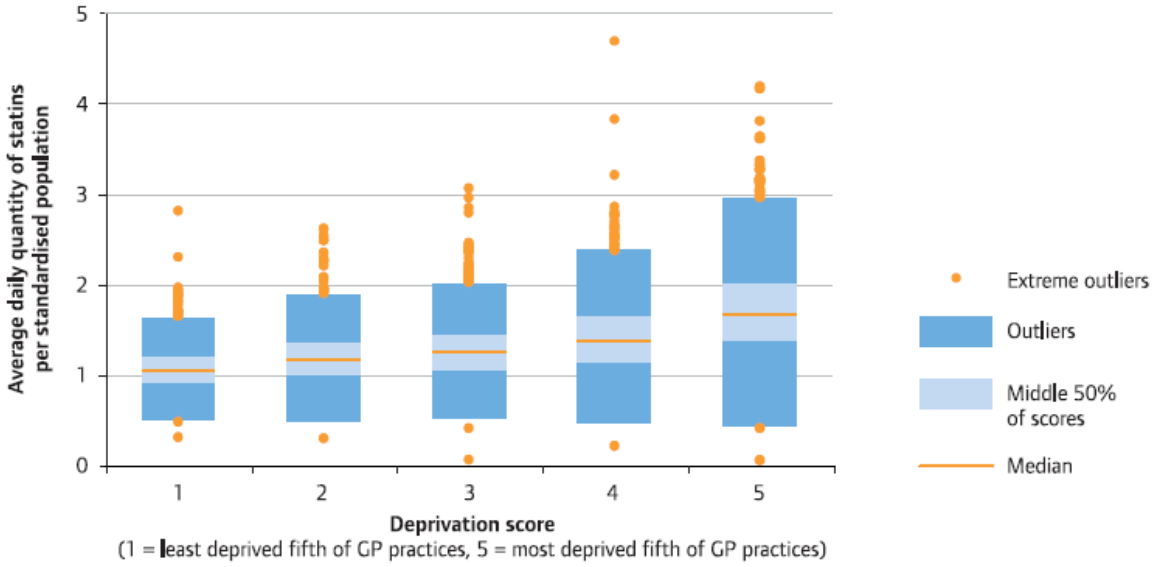
The variation in unrecorded prevalence also increased markedly with deprivation. Overall, the median level of estimated unrecorded prevalence was just over 1% and a large number of GP practices had median unrecorded prevalence rates of over 2%. The report concluded that QOF was not designed to be an epidemiological disease register, and is therefore likely to miss a proportion of those with established CVD. Some reasons given why CVD prevalence may be unrecorded include patients who are registered with a GP but whose CVD has not yet been diagnosed, people in the community who are not registered with a GP but who have CVD, or because some GP practices in deprived inner city areas may not have systems in place to capture the data in electronic format for the QOF.

The median level of unrecorded prevalence appeared to be low at just over 1%, but this is significant in public health terms, and it is worrying that the level of unrecorded prevalence increases with increasing deprivation. People who have CVD and who also live in relatively poor social and economic circumstances are more likely to be in greater need of

preventative treatment because of the added burden of deprivation. Variations in the performance of GP practices are inevitable. However, high levels of variation can indicate that they are either not consistently observing national guidelines, they do not have a focused approach or that there are challenges associated with deprivation in the local area.

The report also addressed the question: does statin prescribing increase with recorded CVD prevalence and deprivation? Volumes of statin prescribing as average daily quantity of statins per standardised population (ADQ/STAR PU) increased only marginally as the prevalence of CVD increases. However, the increase in prescribing was more noticeable with increasing deprivation (see Figure 23: average daily quantity of statins per standardised population). There appeared to be a wider variation in statin prescribing in GP practices with a lower estimated CVD prevalence. Conversely, in the context of deprivation, variation in statin prescribing is greater in the more deprived GP practices.

Figure 23: average daily quantity of statins per standardised population



The report concluded that statin prescribing increases more noticeably with increasing deprivation than with increasing CVD prevalence. This may be because deprivation is a better predictor of the prevalence of CVD than QOF (as stated above, QOF was not designed to be an accurate epidemiological disease register). Prescribing statins to people who are not yet recorded as having established CVD might explain why it appears that there is more statin prescribing in deprived areas than in areas of higher CVD prevalence. It suggested that GPs are undertaking *some* preventive statin prescribing to patients at high risk of developing CVD. If this is the case, it is good news for people's health and fits in with both the Department of Health's policy and guidance from NICE on reducing risks of CVD, although not consistently so. Further research may clarify how much of the above prescribing is to those whose CVD is unrecorded and to those who are at high risk.

This application of the CVD prevalence model demonstrates its utility in combination with other indicators of population needs and healthcare use.

10.6 Does undiagnosed COPD increase hospital admission rates?

This sub-section is part of a paper (Association of population and primary healthcare factors with hospital admission rates for chronic obstructive pulmonary disease in England: national cross-sectional study) published by Thorax.[250] It explores the associations between registered, estimated and undiagnosed disease prevalence (using prevalence models), primary healthcare quality, and hospital utilisation. There are a number of potential benefits in evaluating these associations, for example:

- to establish a firmer evidence base for the identification of primary care sensitive conditions (PCSCs)
- to model future healthcare capacity requirements, and the impact of quality improvements on them
- to achieve better value in healthcare resourcing, and improve healthcare outcomes.

10.6.1 Introduction

The total annual cost of COPD to the English National Health Service (NHS) is over £800 million ($1,215 million), and it has become the second largest cause of emergency hospital admissions in the UK.[251] DH for England has released a COPD national strategy[252] and it also has a range of policies to improve long term condition management in primary care, with the aim of improving quality and length of life and reducing hospital admissions.

Reducing unnecessary hospital admissions will play a critical role in ensuring the NHS and other health systems meet the challenge of real terms reductions in resources - the 2010 UK Budget report factors in potential savings of £2.4bn from this source.[253] COPD is one of a number of PCSCs, for which it is hypothesised that risk of hospital admission will be reduced by effective primary care management.[254-256] However PCSCs have mainly been identified by expert opinion rather than empirically.[257;258] Previous research has also been hampered by lack of detailed data about practice populations and the quality of primary care services, such as is now available through QOF, or has been carried out only at PCT level .[259-261]

Some population factors- especially deprivation - are strongly associated with PCSC admission rates, and there are weaker associations with primary care access and supply factors.[262] Effective management of COPD may reduce the risk of hospital admission. Gotfredsen et al found a gradient of relative risk of hospital admission with increasing smoking prevalence.[69] In addition, combination treatment with long-acting beta-agonists and inhaled corticosteroids reduces the annual rate of COPD exacerbations and improves health status and spirometric values.[208;209] A systematic review demonstrated that patients with COPD who received interventions with two or more components of the Chronic Care Model had lower rates of hospitalizations.[263]

The aim of this study was to examine the associations between population characteristics, including both diagnosed and undiagnosed prevalence, primary healthcare factors, and COPD admission rates at PCT and practice levels in England. Both levels of analysis are desirable, as aggregation to PCT level could mask important associations present at practice

level. Specific objectives were firstly to investigate the associations between rates of hospital admission for COPD and population factors (deprivation, smoking prevalence, QOF registered/diagnosed prevalence rates, and undiagnosed i.e. expected minus QOF registered prevalence rates). The second objective was to investigate associations between admission rates and a range of healthcare factors, after adjustment for population factors. These included access factors (general practitioner and practice nurse supply and QOF access indicators), COPD-specific QOF quality indicators, and smoking cessation advice or referral, to explore their relative impacts. As all UK residents are entitled to primary care consultations which are free at the point of use, and as the vast majority of residents are registered with a practice, there are no apparent barriers to access. However, despite NHS resources being allocated on a needs-weighted population basis, there are quite wide variations in access as measured by patients' experience of the availability of unscheduled care and appointments - hence the inclusion of these indicators in QOF.

10.6.2 Methods

Hospital Episodes Statistics Data

NHS-funded hospitals in England are responsible for providing data to the HES database. A hospital episode forms all or part of an overall hospital spell (one complete admission). We used hospital spells for this analysis to avoid multi-counting, Admissions were selected where the primary diagnosis i.e. main reason for admission met ICD version 10 J44.0-44.9 (Other chronic obstructive pulmonary disease, excluding asthma and other non-COPD) for all patients 15 years of age and over admitted during financial years 2006-7, 2007-8, 2008-9 to maximise the number of admissions per practice. It was not possible to validate the HES diagnosis against data from practice COPD registers as only aggregate QOF data is available.

Quality & Outcomes Framework Data

We used QOF data from the IC website for April 2008 - March 2009.[220] QOF indicators are split between four quality domains (clinical, organisational, patient experience and additional services). The QOF clinical domain included 129 evidence-based indicators of the clinical quality of care including five indicators for COPD (Table 62: Quality and Outcomes Framework indicators for COPD management and Patient Experience in 2008-2009 which were analysed). We used the number of COPD patients recorded by general practices (registered prevalence), the total list sizes of patients per practice as denominators (the definition used by the Information Centre), the total QOF points achieved for the COPD domain and individual indicators for specific COPD care (i.e. spirometric confirmation of diagnosis, recent record of forced expiratory volume in one second (FEV1), inhaler technique and influenza immunisation offered). Since stopping smoking affects the prognosis of COPD, we also included the percentage of patients with a chronic disease whose notes record that smoking cessation advice or referral to a specialist service has been offered within the previous 15 months (Smoking 4). To measure broad access to care from the patient perspective we used two new patient experience (PE) indicators introduced in the 2008-9 QOF, which were derived from the results of the national GP Patient Survey, and which measure ability to access a GP consultation within two days (PE7), and ability to book an appointment with a GP more than two days ahead (PE8).

Table 62: Quality and Outcomes Framework indicators for COPD management and Patient Experience in 2008-2009 which were analysed

Indicator	Points	Payment thresholds*
COPD quality indicators		
Records		
COPD 1 The practice can produce a register of patients with COPD	3	
Initial diagnosis		
COPD 12 The percentage of all patients with COPD diagnosed after 1st April 2008 in whom the diagnosis has been confirmed by post bronchodilator spirometry	5	40-80%
Ongoing management		
COPD 10 The percentage of patients with COPD with a record of FeV1 in the previous 15 months	7	40-70%
COPD 11 The percentage of patients with COPD receiving inhaled treatment in whom there is a record that inhaler technique has been checked in the previous 15 months	7	40-90%
COPD 8 The percentage of patients with COPD who have had influenza immunisation in the preceding 1 September to 31 March	6	40-85%
Smoking quality indicator		
Smoking 4 The percentage of patients with coronary heart disease, stroke or TIA, hypertension, diabetes, COPD, CKD, asthma, schizophrenia, bipolar affective disorder or other psychoses who smoke whose notes contain a record that smoking cessation advice or referral to a specialist service, where available, has been offered within the previous 15 months	35	40-90%
Patient Experience indicators		
PE 7 Patient experience of access (1) The percentage of patients who, in the appropriate national survey, indicate that they were able to obtain a consultation with a GP (in England)	23.5	70-90%
PE 8 Patient experience of access (2) The percentage of patients who, in the appropriate national survey, indicate that they were able to book an appointment with a GP more than 2 days ahead	35	60-90%

** The lower and upper achievement levels required to receive the minimum and maximum payment.*

Practice staffing

The IC provided data on GP and nurse supply (GPs and practice nurses per 100,000 population), at practice and PCT level respectively, as of September 2007.[203] Practice nurse numbers do not include other nurses working in the community, such as "community matrons" or district nurses, who may have contact with COPD patients.

Undiagnosed prevalence data

We used practice-level predictions of COPD prevalence from the APHO website, which uses a model developed by Nacul et al.[59] This uses spirometry data from the 2001 HSfE to produce prevalence estimates for population subgroups. It uses the British Thoracic Society definition of COPD i.e. FEV1 FVC under 0.70, and FEV_1 less than 80% of predicted, a more stringent definition than the National Strategy. The model includes practice-level values of age, sex (population data as of March 2007), ethnicity, deprivation, smoking and urbanisation to estimate overall COPD prevalence, and has been validated against estimates from population surveys. Practice-registered COPD prevalence was then subtracted to estimate undiagnosed prevalence. Further information about the COPD model development is provided on the APHO website.[116]

Population Data

The CQC (CQC) provided estimates by sex and five-year age band of registered patients by practice and PCT in England, as of April 2008. The Department of Communities & Local Government Index of Multiple Deprivation (IMD) 2004 data for England was used to measure deprivation[79]- high scores indicate a high level of deprivation. IMD data were available for 152 PCTs. A deprivation weighting for each practice was obtained from the CQC. This was produced by aggregating IMD scores from postcodes of individual patients in practice populations.

Statistical Analysis

The basic unit of analysis was the general practice population. Indirect standardisation is generally accepted as more robust when counts are relatively small, so indirectly standardised admission rates were calculated at the practice level, to produce overall expected admission rates for each practice based on the England age/sex specific rates. These were compared to the observed practice-level admission numbers. Baseline PCT characteristics were then calculated including registered population (the sum of registers of general practices located within PCTs' geographic boundaries), age distribution, IMD, practices per PCT and practice list size. Practice level characteristics (QOF points and indicator breakdown) were aggregated to PCT level, as required. As admission counts were larger at PCT level, directly standardised (age and sex) admission rates were calculated per 100,000 GP registered population, using the total England population as the standard with which to weight the PCT rates. Three year averages of registered and estimated COPD prevalence were summarised and undiagnosed prevalence calculated.

At practice and PCT levels, bivariate associations between admission rates and the above continuous independent variables were first assessed using Spearman's correlation coefficient. For the PCT level analysis least squares linear multivariate regression models were then fitted to examine the effects of all independent variables. At the practice level, because of low admission numbers, we applied Poisson regression, where standard errors have been scaled using the square root of the deviance-based dispersion. The dependent

variable was observed count of hospital admissions and the offset was the expected number of admissions. Poisson regression returns incidence rate ratios (IRRs), which in this context can be considered as analogous to odds ratios. Microsoft Excel and Access were used to manage the data and the software package Stata10 was used for all statistical analysis.

10.6.3 Results

The 152 English PCTs had a total GP-registered population of 53,547,488 patients (Table 63: characteristics of PCT and practice populations). The GP-registered population in each PCT varied from 94,635 to 1,294,596 (mean 352,286). We were able to match all the data required for 8,064 (90.3%) GP practices from a total of 8,932. List size varied from 984 to 38,343 (mean 6,603). Estimated smoking prevalence at practice level varied from 13.2 to 51.7%. QOF scores achieved were generally high, for influenza immunisation, inhaler technique and spirometry, although there was a wider variance in PCTs recording spirometry (76.0% - 96.4%) and recording FEV1 (75.5% - 89.3%). The range of QOF Patient Experience scores also tended to be lower.

Table 63: characteristics of PCT and practice populations

	Mean	PCT-level				Practice-level			
		(SD)	Min - Max	Median	IQR	(SD)	Min - Max	Median	IQR
PCT population	352,286	193,433	94,635 - 1,294,596	292,118	225,238 - 422,895	-	-	-	-
Practice population	6,603	-	-	-	-	4,011	984 - 38,343	5,774	3,394 - 9,011
Percent of people 15-34y	27.5	25.7	19.6 - 43.1	43.1	25.7	7.8	10.5 - 98.5	25.7	22.4 - 30.0
Percent people 35-74y	48.1	3.8	34.2 - 53.7	49.3	45.5 - 50.9	6.6	1.3 - 64.4	49.5	45.3 - 52.5
Percent people 75+y	7.2	1.8	2.9 - 12.3	7.2	6.3 - 8.2	2.9	0.0 - 26.4	7.2	5.3 - 8.9
IMD (deprivation) score	23.7	9.1	8.1 - 48.3	23.6	16.1 - 29.8	12.6	2.2 - 71.9	21.3	13.4 - 32.0
Smoking prevalence	24.7	4.3	14.5 - 40.9	24.3	21.7 - 27.1	6.4	13.2 - 51.7	23.7	19.7 - 28.9
GPs/100,000 population	60	7	44 - 81	59	55 - 64	17	8 - 334	55	10 - 180
Practice nurses/100,000 population	28	7	0 - 60	28	24 - 32	-	-	-	-
QOF Clinical Indicators*									
Total COPD score	96.2%	2.6%	83.7% - 99.9%	96.8%	94.8% - 98.2%	9.9%	0.0% - 100%	100%	99.4% - 100%
Influenza immunisation	91.6%	1.2%	86.9% - 95.5%	91.8%	90.9% - 92.4%	6.3%	0.0% - 100%	92.7%	89.5% - 96.1%
Record of FEV1	82.4%	2.6%	75.5% - 89.3%	82.4%	81.0% - 84.2%	12.1%	0.0% - 100%	83.8%	77.3% - 89.7%
Record of inhaler technique	91.6%	1.6%	87.4% - 95.1%	91.8%	90.6% - 92.7%	8.9%	0.0% - 100%	93.4%	90.9% - 96.2%
Record of post-bronchodilator spirometry	90.1%	3.4%	76.0% - 96.4%	90.9%	88.8% - 92.2%	25.0%	0.0% - 100%	100%	86.2% - 100%
Smoking cessation advice or referral to a specialist service	92.7%	1.1%	89.2% - 95.3%	92.9%	92.1% - 93.5%	5.3%	26.6% - 100%	93.1%	91.1% - 95.8%

QOF Patient Care Experience Indicators*									
Ability to obtain a consultation within 2 working days	84.8%	8.4%	0.0% - 92.8%	85.6%	82.9% - 88.3%	14.1%	0.0% - 100%	87.3%	79.3% - 93.0%
Ability to book an appointment with a GP more than 2 days ahead	77.3%	8.4%	0.0% - 89.0%	77.8%	75.2% - 81.0%	17.6%	0.0% - 100%	80.8%	67.3% - 90.3%

** Total Points / Available %*

Table 64 summarizes the observed, estimated and calculated undiagnosed prevalence levels for COPD at the PCT and practice level. The England mean prevalence of registered COPD in the whole population is 1.5% (SD 0.0%), and undiagnosed COPD is 2.2% (SD 0.1%). Mean (over three years) annual hospital admission rates for COPD per 100, 000 population varied from 124.7 to 646.5 (median 241.5, inter-quartile range 178.1 to 331.4) at PCT level and 0.0 to 2175.2 (median 179.0, inter-quartile range 111.1 to 288.2) at general practice level, illustrating the greater variance observable in a practice level analysis. Over 98% of admissions were in the 45 years and over age group.

Regression modelling

In the bivariate regression of exposure variables against hospital admission rates at PCT level, deprivation and smoking prevalence were the variables with highest explanatory power, accounting for 59.3% and 51.4% of the total variance respectively ($p<0.001$). Estimated prevalence accounted for 45.1% of the variance ($p<0.001$) compared to the observed (registered) COPD prevalence (21.9%, $p<0.001$). About 14.4% of the variance in hospital admissions was explained by GP supply ($p<0.001$). The QOF COPD overall clinical quality score was shown to account for only 2.1% of the variance ($p=0.034$). Among all specific clinical QOF indicators, the score for influenza immunisation showed the highest explanatory capacity (6.9%, $p<0.001$). In the multivariate regression model using reverse stepwise variable selection ($R^2=75.0\%$), the undiagnosed disease prevalence as well as all QOF clinical indicators were not retained in the model using a probability criterion of 0.1 (Table 64).

Table 64: Summary of estimated, observed and undiagnosed COPD prevalence and admission rates

			PCT-level			Practice-level			
	Mean	**(SD)**	**Min - Max**	**Median**	**IQR**	**(SD)**	**Min - Max**	**Median**	**IQR**
Populations	352,286	193,433	94,635 - 1,294,596	292,118	225,238 - 422,895	4,011	984 - 38,343	5,774	3,394 - 9,011
Estimated prevalence	3.8%	0.9%	2.1% - 6.0%	3.6%	3.1% - 4.4%	1.1%	0.7% - 7.5%	3.6%	2.9% - 4.6%
Observed prevalence	1.5%	0.4%	0.7% - 2.5%	1.5%	1.3% - 1.8%	0.8%	0.0% - 7.5%	1.4%	1.0% - 2.0%
Estimated-observed (undiagnosed) prevalence	2.2%	0.9%	0.6% - 4.3%	2.0%	1.5% - 3.0%	1.0%	-2.6% - 6.0%	2.1%	1.4% - 3.0%
Estimated/observed prevalence ratio	2.6	0.7	1.4 - 5.0	2.4	2.0 - 3.0	8.4	0.5 - 577.4	2.5	1.9 - 3.5
Undiagnosed/observed prevalence ratio	1.6	0.7	0.4 - 4.0	1.4	1.0 – 2.0	8.4	-0.5 - 576.4	1.5	0.9 - 2.5
Admissions/100,000 population*	265.3	104.2	124.7 - 646.5	241.5	178.1 - 331.4	170.8	0.0 – 2175.2	179.0	111.1 - 288.2

** Admission rates are directly standardized at PCT-level and adjusted for age and sex (indirectly standardized) at practice-level.*

Table 65 also shows the practice-level analysis using Poisson regression with reverse stepwise variable selection . The highest incidence rate ratio (IRR) was estimated for observed prevalence (IRR=1.224, p<0.001), and undiagnosed prevalence, deprivation, smoking prevalence, GP supply and access to a GP consultation, and influenza immunisation status were also retained in the final model using reverse stepwise variable selection and a probability criterion of 0.1. Overall COPD score, and clinical quality indicators COPD 10, COPD 11, COPD 12 and Smoking 4 all dropped out of this model.

Table 65: multiple regression analyses

PCT level[a]	Regression coefficient	95% CIs	p-value
Observed prevalence	1.325	1.204-1.459	<0.001
IMD (deprivation) score	1.023	1.018-1.028	<0.001
GPs/100,000 population	0.995	0.991-1.000	0.052
Practice nurses/100,000 population	0.992	0.987-0.996	0.001
Smoking prevalence	1.023	1.012-1.033	<0.001
QOF patient experience 7	0.678	0.466-0.985	0.042
Practice level[b]	**Incidence rate ratio**	**95% CIs**	**p-value**
Observed prevalence	1.224	1.206-1.243	<0.001
Undiagnosed prevalence	1.045	1.032-1.059	<0.001
IMD (deprivation) score	1.009	1.006-1.012	<0.001
Smoking prevalence	1.036	1.031-1.042	<0.001
GPs/100,000 population	0.998	0.998-0.999	<0.001
QOF patient experience 7	0.790	0.730-0.855	<0.001
QOF patient experience 8	0.902	0.850-0.957	0.001
COPD clinical quality 8 (influenza immunisation)	0.825	0.690-0.987	0.036

[a] *Stepwise linear regression. Adjusted R-squared= 0.75. Dependent variable: admission rates for COPD.*

[b] *Stepwise Poisson regression, Dependent variable: observed admission count.*

PE 7: Score for ability to obtain a consultation within 2 working days.

PE 8: Score for ability to book an appointment with a GP more than 2 days ahead.

COPD 8: The percentage of patients with COPD who have had influenza immunisation in the preceding 1 September to 31 March.

10.6.4 Discussion

In this cross-sectional study of hospital admissions for COPD, registered COPD prevalence, undiagnosed prevalence, smoking prevalence and deprivation were all risk factors for hospital admission ($p<0.001$). Primary healthcare factors such as being offered influenza immunisation, patient-reported access to primary care consultations within two days, and primary care supply (GPs and practice nurses per 100,000 population) were protective factors ($p<0.05$). There was an overall consistency in the findings at organisation and practice levels, but with additional independent variables being retained in the practice level model, and with stronger associations apparent. At practice level, after adjustment for deprivation, undiagnosed COPD prevalence as well as diagnosed disease was strongly associated with the rate of hospital admission. Primary healthcare factors such as being offered influenza immunisation, patient-reported good access to GP consultation, and GP and practice nurse supply were also significantly associated with reduced hospital admission rates.

Hospital admissions for COPD were strongly associated with deprivation scores and smoking prevalence at both PCT and practice level. This relationship has been documented previously,[264;265] but associations with primary care supply less often. Saxena et al studied the association at London PCT level between population and practice factors with hospital admission rates for PCS conditions (including COPD).[262] They found some associations between primary healthcare factors and admission rates, but not with practice resourcing. In contrast, a practice level analysis of hospital admissions for asthma in East London, found that practice characteristics were the strongest predictor of the admission rate, but the most likely explanation for this result could be that the geographically small region under investigation was uniformly deprived.[261] In our study, GP supply was inversely associated with hospital admission rates at the practice (but not PCT) level. Conversely, practice nurse supply was associated with lower admission rates at PCT level, but unfortunately these data are not readily available at practice level. These and similar findings, together with the impacts demonstrated in our QOF indicator analysis, highlight the need for adequate primary care provision in deprived areas. Geographical inequalities in GP supply have been persistent despite needs-based funding of PCTs.[231]

Our study was strengthened by using data from a national population, giving a high level of statistical power, and by combining practice-level data and estimates from a number of sources which were previously unavailable, or only available locally. This revealed associations which may have been masked in previous studies which used the organisation as the level of analysis. This study is also the first to examine estimates of undiagnosed disease and its association with hospital admissions.

The BTS definition of COPD used for the prevalence estimates is stringent, which may lead to underestimation of the effect of undiagnosed disease prevalence. Conversely, the COPD model does not include lack of bronchodilator responsiveness in the definition of COPD, so this will slightly overestimate prevalence. Moreover, practice registers can include patients with no spirometry. We were able to include several new QOF clinical quality and patient experience indicators. Our analysis also illustrates the greater benefits of using practice level rather than organisation level data. Our findings would be strengthened by similar

analyses of other long term conditions, which would also validate the selection of PCS conditions, and assist in regulation of primary care.

There are several limitations. Hospital or general practice diagnostic coding may be incorrect. However, the advent of diagnosis-based payment of hospitals has accelerated improvement in data quality. We only considered primary or principal diagnosis, and 87.2% of primary diagnoses for hospital admissions in England are coded correctly.[266] Counts of annual hospital admissions per practice were relatively low, even using means of three years' data (mean per year=12.9, median per year=10.0). We used appropriate methods (indirect standardisation and Poisson regression) to deal with this, and were reassured by the overall consistency between the results at organisation and practice levels. Furthermore, a number of UK agencies use practice-level admission data to support commissioning and quality improvement, so its use should prompt similar scrutiny to that which followed use of similar data at hospital level. In the case of COPD, general practices may add patients to COPD registers without necessary confirmatory spirometry- only 82.4 per cent have a record of this test. This will overestimate the diagnosed prevalence.

Reported practice populations may overestimate actual numbers, especially in urban areas with high mobility. Conversely, a small number of residents may not be registered with a practice. There is also a high level of achievement of many QOF indicators which results in a "ceiling effect", which may reduce the ability to detect real differences in clinical quality. Many QOF indicators measure care processes which may not be strongly related to disease outcomes (as our analysis suggests). The impact of recent changes in quality of care on outcomes such as hospital admission may be delayed, but changes in QOF scores between 2005 and 2007 were small. Other data inevitably covered slightly different periods and dates. HES data does not include privately-funded hospital admissions, but the vast majority of unplanned admissions for COPD are to NHS hospitals. We did not analyse readmissions separately as the focus was on factors affecting exacerbations in the community. Practice-level COPD and smoking prevalence estimates are based on the limited range of risk factor data available locally.

There are now several disease prevalence models for England which enable comparisons of QOF-registered and expected prevalence counts. We showed an association between admission rates and undiagnosed prevalence. The National COPD Audit has shown that 90% of those hospitalised with COPD were known to have the diagnosis before admission and to suffer exacerbations frequently.[267] However, a population survey from Spain showed a high rate of undiagnosed COPD, and significant impairment in health-related quality of life and activities of daily living in undiagnosed individuals.[268] Patients who have frequent exacerbations irrespective of disease severity may represent a distinct COPD phenotype, implying that earlier diagnosis and intervention may improve outcomes.[269] The validity of the model, and the cost-effectiveness of targeted COPD case-finding strategies in practices with high estimates of undiagnosed disease, should therefore be evaluated.

The new QOF Patient Experience indicators were introduced, as their title suggests, to improve patient satisfaction with access to primary healthcare. However the strong association between these indicators and admission rates suggests that they also reflect clinically important aspects of access. The National COPD Audit showed that 74% of

admitted patients make contact with their general practice in the month before admission and 31% have three or more contacts in those four weeks.[267] Given that the ability to access primary care rapidly during a COPD exacerbation may well reduce admission risks, the association with PE07, the ease of obtaining a consultation, is plausible. That there is also an association with PE08, the ease of making non-urgent appointments, and with GPs per 100,000 practice population, suggests that access barriers may also affect ongoing COPD management.

At practice level, only QOF COPD indicator 8, uptake of influenza immunisation, was associated with hospital admission rates. This is clinically plausible. The other QOF COPD indicators are measures of clinical processes and standards which are less likely to affect outcomes. Alternatives which are more likely to affect outcomes should be considered. For example, practices could be incentivised to deliver more intensive smoking cessation services to COPD patients. The updated NICE COPD guidelines currently out for consultation recommend that pulmonary rehabilitation should be made available to all appropriate patients with COPD, and that newer long-acting drugs be prescribed to appropriate patients.[89] The interventions in the Chronic Care Model reduce admission rates, although they are difficult to measure.[263] Finally, other factors which could affect admission rates include the quality of out of hours services, the presence of specialist nurse services, and the supply and quality of hospital services.

10.7 Exploring the utility of mapping and spatial statistics in the context of health & social care regulation

10.7.1 Introduction

This section illustrates some of the analyses carried out for the CQC as part of a project we were commissioned to carry out to explore the utility of mapping and spatial statistics in the context of primary healthcare regulation.

The role of the CQC includes the registration and regulatory assessment of primary health care. Providers of NHS primary care medical services that provide regulated activities must be registered from 1 April 2012. This includes general practices and out-of-hours services. Research carried out by the CQC has demonstrated the importance of conducting analyses at sub-PCT/practice level to disseminate salient messages to local commissioning and public health teams. In addition, methodological advances in quantitative geography have led to the development of GWR and related spatial statistical techniques. These examine spatial relationships and could add another dimension to CQC practice-level analyses. The use of mapping facilities in modern GIS also provides helpful methods of displaying geo-referenced data. The CQC has therefore commissioned a small project to investigate the application of GIS and GWR to CQC healthcare data, and to explore its utility for local monitoring purposes within the context of CQC's functions to register and regulate primary healthcare services delivered by GP practices.

The specific objectives were:

- to use the LMI method for detecting outlying performance of individual PCTs and GP practices, including statistically significant geographical clustering
- to investigate the relationship between cross-, intra- and inter-PCT performance variation with and without adjustment for confounding factors using classical regression and GWR, to compare the two methods
- to demonstrate options for displaying and mapping this variation using GIS for purposes of local monitoring and regulation.
-

To explore spatial relationships in this context, the two scenarios chosen as a "proof-of-concept" were (Table 66):

Scenario 1/disease detection: associations with practice-registered ("observed") coronary heart disease (CHD) and stroke prevalence and undiagnosed CHD and stroke prevalence as outcome variables. This can be viewed as a metric of the success of practices in identifying and registering cases of disease in their registered populations.

Scenario 2/disease management: associations with PCT and practice level unscheduled/emergency admission rates for CHD and stroke as outcomes of disease management by general practices. There are therefore analyses for two scenarios and two disease groups for each, as shown in Table 1.

Table 66: Analytic scenarios and data sources

	Scenario 1/disease detection	Scenario 2/disease management
Outcome (y) variables	Registered and undiagnosed CHD and stroke prevalence	Admission rates for CHD (MI+angina) and stroke
PCT/practice population contextual exposure variables	Age, sex, deprivation score, ethnicity, smoking	Age, sex, deprivation, ethnicity, smoking, registered and undiagnosed prevalence
Healthcare exposure variables	GP supply, CQC PCT ranking	GP supply, CQC PCT ranking, quality of care

A characteristic of most of the ordinary least squares (OLS) or classical regression analyses is IRRs/ORs mainly very close to unity, suggesting that effect sizes are relatively small. This is somewhat surprising as several other published PCT-level analyses show, for example, relatively strong positive associations of prevalence with deprivation,[270] and it is included in the model used to estimate overall prevalence, which is based on HSfE patient-reported doctor-diagnosed CHD (OR for fifth IMD quintile 2.4). It does not appear that this is mainly due to multicollinearity.

Nevertheless, because of the 8,000 practices included in the analyses, a number of the IRRs are statistically significant. For unregistered prevalence of CHD (Scenario 1), IMD score, GP supply/100,000 population, and the QOF Patient Experience Indicator PE08 (can book an appointment with a GP > 2 days ahead) are marginally positively associated in a multivariate OLS model. Smoking prevalence and the QOF Clinical Indicator CHD08 (Cholesterol<5Mmol/L) were negatively associated. Other variables were not retained in the model. The directions of some of these associations are unexpected: one might expect GP supply to be negatively associated and smoking prevalence to be positively associated. One explanation for this could be multi-collinearity. However only IMD score and estimated smoking prevalence (which uses IMD score as an input) showed evidence of multicollinearity.

Simple practice-level mapping of registered and expected prevalence rates shows the typical North-South divide, with higher prevalence in South coastal areas because of population age structure. However unregistered (expected minus registered) CHD prevalence rate, shows evidence of possible spatial clustering, with the highest rates clustered in London and West Midlands conurbations.

For stroke, OLS multivariate stepwise analysis results are similar to CHD. As for CHD, practice-level mapping of the unregistered stroke prevalence rate, shows evidence of possible spatial clustering, with higher rates in the London and West Midlands conurbations, but there is a more dispersed pattern than for CHD. We also performed Local Moran's I (LMI) tests on unrecorded prevalence of stroke. There is marked clustering of high levels of unrecorded prevalence evident in London and other conurbations, which is even more prominent in the 5 km map. There is some, but much less, clustering of low levels of

unrecorded prevalence which tends to be in the hinterlands of urban areas. The results of the spatial non-stationarity test for stroke prevalence revealed little difference between the AIC values for the non-spatial and GWR models.

For CHD (MI+angina) admission rate as the outcome variable, observed prevalence, IMD score, smoking prevalence and Asian population are as expected all risk factors for hospital admission- the IRR for the latter is comparatively high. The IRR for unregistered prevalence is less than unity, which is unexpected, as patients with undiagnosed disease might be expected to be more likely to be admitted. The directions of other associations are generally as expected, with higher scores on QOF Clinical Indicators as protective, apart from CHD06 (BP<150/90mm Hg). We carried out LMI tests on stroke only.

As for MI+angina, observed prevalence, IMD score and smoking prevalence are as expected all risk factors for hospital admission for stroke. The IRR for unregistered prevalence is less than unity, which is unexpected, as patients with undiagnosed disease might be expected to be more likely to be admitted. In comparison with a similar analysis on COPD admissions we have carried out recently, IRRs were generally smaller for CHD and stroke, and the results for COPD were more consistent with our prior hypotheses.[250] We also have some concerns about the provenance of the CQC HES data, as we have been unable to ascertain how it was grouped into spells. Only one QOF Clinical Indicator, Stroke08 (Last cholesterol $\leq$5mmol/l), remains in the model and is as expected, with higher scores protective.

LMI tests at 2km and 5km for stroke admission rates showed marked clustering of high-high admission rate values evident in London and other conurbations, which is even more prominent in the 5 km map. However there is also clustering of low-low levels, especially in West London and in the West Midlands. The results of the spatial non-stationarity test reveal little difference between the AIC values for the non-spatial and GWR models, with marginal improvement with the spatial model. Nevertheless, the AIC values confirmed that it was statistically valid to use spatial regression as a basis for exploratory mapping.

Practice-level local R-squared maps were produced for stroke admission rates to demonstrate the utility of this analytic approach. We used a split-variable analysis (grouping population factors and healthcare factors together) to illustrate the differing contributions of these variables on admission rates. We have demonstrated, for example, that in one region's more rural PCTs there are associations of practices for whom service delivery factors predict admissions more strongly than population factors: these associations do not appear to cross boundaries into the more urban PCTs in the centre of the region. In NW England, the socio-demographic model shows a general south to north statistical incline and the healthcare model the reverse. On the other hand, along the South Coast, the maps show remarkably strong agreement between the healthcare/institutional and population models in terms of directionality.

In summary, although the analysis did not demonstrate that GWR is definitely superior to OLS analysis in this situation, it is at least as useful, and in addition provides the opportunity to visualise and investigate spatial relationships more easily, as we have shown. Performance-related practice associations identified by the use of GWR provides a basis for further investigation of potential 'overachievers' and 'underachievers'. Mapping of results

enables visualisation readily interpretable by CQC statisticians, and with adaptation cartographic output can be understood by lay members of the public, a facet important in a Freedom of Information age. CQC should be actively promoting GIS approaches as part of its remit.

Recommendations
We recommended that CQC:

1. Make a senior manager within CQC responsible for the implementation of mapping and GIS techniques for the general public and healthcare professionals, and spatial analytic techniques (either in the form of GWR or Bayesian methods[271] using WinBUGS) for internal and external analytic staff;
2. Use one or more GIS packages to display national data on its website e.g. InstantAtlas or GapMinder
3. Procure training and guidance in the use of relevant analytic methods and technologies from an external body; a two-day course should cover the requisite material. Core topics covered must include an introduction to GIS, effective map communication, spatial cluster/outlier analysis, spatial regression and software implementation;
4. After the training of key individuals, hold a brainstorming workshop to discuss how mapping and GIS can be used in practice within CQC;
5. Develop data exchange agreements between the CQC and other institutions such as the Ordnance Survey, and ensure that a core database of key variables is developed and maintained.

10.7.2 Background/Introduction

The role of the Care Quality Commission
The CQC regulates health and adult social care services in England, including monitoring and inspection of all health and adult social care, and improving health and social care services by undertaking regular reviews of how well those who arrange and provide services locally are performing and special reviews on particular care services, pathways of care, or themes where there are particular concerns about quality.

A significant division of its responsibilities includes the registration and regulatory assessment of primary health care. Providers of NHS primary medical services that provide regulated activities must be registered from 1 April 2012. This includes general practices and out-of-hours services. The new registration system will bring all providers of health and adult social care under a single set of essential standards of quality and safety for the first time.[249] Different methods will be used for different providers depending on the regulated activities they provide, and information that is available to CQC about their activities and performance. CQC will constantly update information about providers from people who use services, organisations and other regulators.

This will be managed through a Quality and Risk Profile for each registered provider that will gather all CQC knows about a provider in one place. They will enable CQC to assess where risks lie and prompt front line regulatory activity, such as inspection, and will support teams

to make robust judgments about the quality of services. These Profiles will develop over time as the information about a provider increases.

Currently, the CQC also has a responsibility to ensure that primary care trusts (PCTs) are monitoring and assisting practices and achieving the standards for commissioning laid out in World Class Commissioning. If there are differences in PCT performance, this could be reflected in performance of practices within the PCT. The NHS White Paper *Equity and excellence: Liberating the NHS* states that PCTs' role will be taken over by consortia of GP practices, working with other health and care professionals, and in partnership with local communities and local authorities.[272] To avoid double jeopardy and duplication, the NHS Commissioning Board will take over the current CQC responsibility of assessing NHS commissioners and will hold GP consortia to account for their performance and quality. However, there will still be a need to assess the impact consortia are making on health outcomes.

Spatial Analytic Methods

Research carried out by the CQC has demonstrated significant deprivation-related variations in the quality of primary care as measured by key outcome measures in the Quality and Outcomes Framework (QOF). Furthermore, a recent national report by the CQC also highlights primary medical care inequalities in the assessment and treatment of people with cardiovascular disease (CVD).[249] These findings highlight the importance of conducting analyses at sub-PCT level to disseminate salient messages to local commissioning and public health teams.

Methodological advances in quantitative geography have led to the development of GWR)and related spatial statistical techniques. These examine spatial relationships and can add another dimension to health-related analyses. The use of mapping facilities in modern GIS also provides helpful methods of displaying geo-referenced data.

A further consideration is that, while mapping of data is becoming very pervasive because of the internet and computer software development, understanding of GWR and spatial statistical methods is not widespread in commissioners or providers. A recent report from Canada illustrated how these methods could be used by local public health departments to investigate cancer risks,[273] using the LMI test and subsequently Bayesian hierarchical models to smooth small area incidence rates. However because of lack of familiarity with these methods, local use is rare even for environmental health risk data. There is an obvious need to accompany the use of any GIS techniques for regulatory purposes with appropriate raising of awareness and training, through multiple channels, and to use a stepwise approach in doing so. The LMI test provides an excellent stepping stone to GWR initially, and then to the use of Bayesian techniques where they add understanding and value.

For example, using local data on the observed and expected prevalence of COPD, we recently compared classical regression results and those of GWR, using analysis of variance and AIC methods, and confirmed that GWR modelled the dataset more accurately than classical regression.[274] The supplementary employment of cluster detection methods, such as the LMI test, additionally enabled hotspots and outliers to be detected. The CQC therefore commissioned a small project to investigate the application of GIS, GWR and LMI

to CQC healthcare data, and to explore its utility for local monitoring purposes within the context of CQC's functions to register and regulate primary healthcare services delivered by GP practices.

10.7.3 Project Description/Methods

Aims and Objectives

The overall aim of the project was to explore the utility of GIS and GWR in the context of healthcare regulation of PCTs as commissioners and developers of primary care services and general practices as providers. The specific objectives were:

- to use the LMI method for detecting outlying performance of individual PCTs and GP practices, including statistically significant geographical clustering
- to investigate the relationship between cross-, intra- and inter-PCT performance variation with and without adjustment for confounding factors using classical regression and GWR
- to demonstrate options for displaying and mapping this variation using GIS for purposes of local monitoring and regulation.

To explore spatial relationships in this context, the two scenarios chosen as a "proof-of-concept" were:

Scenario 1/disease detection: associations with practice-registered ("observed") coronary heart disease (CHD) and stroke prevalence and undiagnosed CHD and stroke prevalence as outcome variables. This can be viewed as a metric of the success of practices in identifying and registering cases of disease in their registered populations. The latter was calculated by subtracting observed prevalence from the expected prevalence of disease estimated by the APHO/ICL CHD and stroke prevalence models.[116]

Scenario 2/disease management: associations with PCT and practice level hospital admission rates for CHD and stroke as outcomes of disease management by general practices. While unscheduled/emergency admissions for CHD could be viewed as an adverse outcome of primary care management, elective CHD admissions, for example for revascularisation, should be viewed as appropriate care. We therefore agreed with CQC that the appropriate admission diagnostic groups for CHD were MI and angina.
There are therefore analyses for two scenarios and two disease groups for each.

Business Case/Benefits Expected

The main benefits for the CQC were expected to be:

- a better understanding of the specific associations between local variables and the detected prevalence of CVD at PCT and practice levels
- identification of geographical associations and clusters of care indicators within, between and across PCT boundaries
- an example of an approach to mapping and geographic analysis of primary care and associated datasets which could be applied across a range of primary care indicators
- options for further application of available data and methods developed.

Data sources

The data used fell into three main categories:

1. contextual/population data (independent/predictor variables in both scenarios, apart from registered and undiagnosed disease prevalence, which became independent variables in Scenario 2)
2. primary healthcare service access data
3. healthcare activity and quality data

Contextual/population data

Data on practice populations by sex and five-year age band of registered patients, as of April 2008, was obtained from the Practice-Based Commissioning Toolkit produced by the DH. The Department of Communities & Local Government IMD 2004 data for England was used to measure deprivation. A deprivation weighting for each practice (aggregated from patient-level post codes) was provided by the CQC- high scores indicate a high level of deprivation. Estimates of the proportions of practice populations in major Census ethnic groups and the proportion who were smokers were provided by CQC.

Undiagnosed prevalence data

We used practice-level predictions of CHD and stroke prevalence from the APHO website. Further information about model development is provided on the APHO website.[116] Practice-registered prevalence was then subtracted to estimate undiagnosed prevalence.

Healthcare access/supply data

The IC provided data on GP and nurse supply, at practice and PCT level respectively, as of September 2007. To measure broad access to care from the patient perspective we used two new patient experience (PE) indicators introduced in the 2008-9 QOF, which were derived from the results of the national GP Patient Survey, and which measure ability to access a GP consultation or appointment (see Table 2).[220]

Quality & Outcomes Framework (QOF) Data

QOF includes standardised information on the quality of care provided by 8.229 general practices in 152 English PCTs. We used QOF data from the IC website for April 2008-March 2009.[220] QOF indicators are split between four quality domains (clinical, organisational, patient experience and additional services). The QOF clinical domain included 129 evidence-based indicators of the clinical quality of care, of which 10 cover CHD and eight stroke (Table 67). We selected three CHD and three stroke indicators which we considered to be good intermediate outcomes of primary healthcare services (as opposed to indicators purely of the process of care). We also included blood pressure control (from the hypertension domain) as it is a strong risk factor for stroke, and two indicators of the accessibility of primary care. The CHD indicators were CHD6 (blood pressure $\leq$150/90), CHD8 (cholesterol $\leq$5mmol/l), CHD11 (treated with an ACE inhibitor or Angiotensin II antagonist). The stroke indicators were STROKE6 (blood pressure $\leq$150/90), STROKE8 (cholesterol $\leq$5mmol/l), STROKE12 (on an anti-platelet agent or an anti-coagulant), and BP05 (blood pressure $\leq$150/90 in patients with hypertension only). The access indicators were PE7 (able to obtain

a consultation with a GP within 2 working days) and PE8 (able to book an appointment with a GP more than 2 days ahead).

Table 67: 2008-9 Quality & Outcomes Framework indicators for CHD & stroke/TIA

Indicator	Points	Payment thresholds*
CHD quality indicators		
Records		
CHD 1: The practice can produce a register of patients with coronary heart disease.	4	
Initial diagnosis		
CHD 2: The percentage of patients with newly diagnosed angina (diagnosed after 1 April 2003) who are referred for exercise testing and/or specialist assessment.	7	40-90%
Ongoing management		
CHD 5: The percentage of patients with coronary heart disease whose notes have a record of blood pressure in the previous 15 months.	7	40-90%
CHD 6: The percentage of patients with coronary heart disease in whom the last blood pressure reading (measured in the previous 15 months) is 150/90 or less.	17	40-70%
CHD 7: The percentage of patients with coronary heart disease whose notes have a record of total cholesterol in the previous 15 months.	7	40-90%
CHD 8: The percentage of patients with coronary heart disease whose last measured total cholesterol (measured in the previous 15 months) is 5mmol/l or less.	17	40-70%
CHD 9: The percentage of patients with coronary heart disease with a record in the previous 15 months that aspirin, an alternative anti-platelet therapy, or an anti-coagulant is being taken (unless a contraindication or side-effects are recorded).	7	40-90%
CHD 10: The percentage of patients with coronary heart disease who are currently treated with a beta blocker (unless a contraindication or side-effects are recorded).	7	40-60%
CHD 11: The percentage of patients with a history of myocardial infarction (diagnosed after 1 April 2003) who are currently treated with an ACE inhibitor or Angiotensin II antagonist.	7	40-80%
CHD 12: The percentage of patients with coronary heart disease who have a record of influenza immunisation in the	7	40-90%

Indicator	Points	Payment thresholds*
preceding 1 September to 31 March.		
Records		
STROKE 1. The practice can produce a register of patients with stroke or TIA	2	
Initial diagnosis		
STROKE 13. The percentage of new patients with a stroke or TIA who have been referred for further investigation	2	40-80%
Ongoing management		
STROKE 5. The percentage of patients with TIA or stroke who have a record of blood pressure in the notes in the preceding 15 months 2	2	40-90%
STROKE 6. The percentage of patients with a history of TIA or stroke in whom the last blood pressure reading (measured in the previous 15 months) is 150/90 or less	5	40-70%
STROKE 7. The percentage of patients with TIA or stroke who have a record of total cholesterol in the last 15 months	2	40-90%
STROKE 8. The percentage of patients with TIA or stroke whose last measured total cholesterol (measured in the previous 15 months) is 5mmol/l or less	5	40-60%
STROKE 12. The percentage of patients with a stroke shown to be non-haemorrhagic, or a history of TIA, who have a record that an anti-platelet agent (aspirin, clopidogrel, dipyridamole or a combination), or an anti-coagulant is being taken (unless a contraindication or side-effects are recorded)	4	40-90%
STROKE 10. The percentage of patients with TIA or stroke who have had influenza immunisation in the preceding 1 September to 31 March	2	40-85%
Smoking quality indicator		
Smoking 4 The percentage of patients with coronary heart disease, stroke or TIA, hypertension, diabetes, CHD, CKD, asthma, schizophrenia, bipolar affective disorder or other psychoses who smoke whose notes contain a record that smoking cessation advice or referral to a specialist service, where available, has been offered within the previous xx months	30	40-90%
Patient Experience indicators		
PE 7 Patient experience of access (1) The percentage of patients who, in the appropriate national survey, indicate that they were able to obtain a consultation with a GP (in England)	23.5	70-90%
PE 8 Patient experience of access (2)	35	60-90%

Indicator	Points	Payment thresholds*
The percentage of patients who, in the appropriate national survey, indicate that they were able to book an appointment with a GP more than x days ahead		

Hospital Episodes Statistics Data

We were given HES data by CQC for this analysis, apparently grouped into spells to avoid multi-counting of admissions/Finished Consultant Episodes (FCEs). Admission events were selected where the primary diagnosis i.e. main reason for admission met ICD version 10 J44.0-44.9, for all patients over the age of 15 admitted during financial years 2006-7, 2007-8, 2008-9 to maximise the number of admissions per practice.

Statistical analyses

Analytic scenarios are shown in Table 68.

Table 68: analytic scenarios and data sources

	Scenario 1/disease detection	Scenario 2/disease management
Outcome (y) variables	Registered and undiagnosed CHD and stroke prevalence	Admission rates for CHD (MI+angina) and stroke
PCT/practice population contextual exposure variables	Deprivation score, ethnicity, smoking	Age, sex, deprivation, ethnicity, smoking, registered and undiagnosed prevalence
Healthcare exposure variables	GP supply, CQC PCT ranking,	GP supply, CQC PCT ranking, quality of care

Scenario 1/ disease detection:

The basic unit of analysis was the general practice population. Counts of QOF registered CHD and stroke prevalence for 2008-9 were subtracted from counts of estimated prevalence for each practice to calculate undiagnosed prevalence. Independent variables were age, sex, IMD score, practice smoking prevalence (%), Asian population (%), Black population (%), GP supply/100.000 population, CQC PCT ranking, and QOF Clinical and Patient Experience Indicators as in the table above).

Indirect standardisation is generally accepted as more robust when counts are relatively small, but prevalence counts were generally much higher than for admissions e.g. mean 233 and median 190 per practice for registered CHD prevalence. Baseline PCT characteristics were then calculated including registered population (the sum of registers of general practices located within PCTs' geographic boundaries), age and sex distribution, IMD scores, practices per PCT and practice list size.

After producing summary statistics (mean, standard deviations, minima/maxima, median and interquartile ranges) we used OLS linear regression to perform univariate analyses for registered and unregistered CHD and/or stroke prevalence separately. We then performed multiple linear regression (using the GLM command in Stata) with reverse stepwise variable selection to fit the most parsimonious model for each population outcome variable, initially separating healthcare and institution-type variables. Practice characteristics were aggregated to obtain PCT-level results, and the same tests were performed.

For the spatial analyses, we applied two methods commonly used in geographical studies: LMI and GWR. The LMI technique is used to identify geographic clusters and outliers in data by testing for randomness in spatial distribution across a dataset, localities with significance scores (Z scores) greater than two standard deviations being considered to be either clusters or outliers.[197] Strongly positive Z scores indicate statistically significant similar values in close geographic proximity hence the presence of a cluster; a strongly negative Z score demonstrates a locality with a significantly dissimilar value in relation to its neighbouring localities thus indicating an outlier. It is very useful as a means of identifying and introducing, for a non-specialist audience, the concept of geographical "hot" and "cold" spots where spatial autocorrelation may be occurring. Like temporal autocorrelation, spatial autocorrelation means that adjacent observations of the same phenomenon are correlated. However, temporal autocorrelation is about proximity in time. Spatial autocorrelation is about proximity in (two-dimensional) space. Spatial autocorrelation is more complex than autocorrelation because the correlation is two-dimensional and hence bi-directional unlike in time-based studies.

GWR is a form of spatial statistics which disaggregates geographic data into spatial blocks using a probability distribution kernel and Monte Carlo simulations, which moves from location to location across the dataset to test for geographic variation in regression relationships. In situations where there is geographic variation in the strength of a regression relationship, a phenomenon referred to as spatial non-stationarity, GWR will produce higher correlation coefficients, lower residuals and higher degrees of freedom than traditional OLS regression.[198;199;275]

For this project, we used GWR to assess whether linear regression relationships existed between observed and expected prevalence in Scenario 1 (and admission rates in Scenario 2), and if so whether they varied in strength over space. This purpose should be viewed as distinct from that of mapping, for example, observed to expected prevalence ratios, as the latter aims to measure equality between two variables rather than to assess predictability of an association. Both OLS and GWR models were run in the software package GWR3 to test for spatial non-stationarity. The optimal bandwidth for the kernel was estimated using the AIC.[202] Two rounds of regression were performed, the first a univariate regression involving expected prevalence as the independent variable and observed prevalence as the dependent variable; the second bivariate including whole-time equivalent GP supply as an additional variable.

For the GWR, we used the sub-regional groupings or analysis blocks shown in Table 69 below. These were chosen because on the one hand, we wished to show spatial relationships which might overlap PCT boundaries, while on the other hand, Government Regions were too large to visualise practices and were beyond the software capacity in terms of practice numbers.

Microsoft Excel and Access were used to manage the data and the software package Stata10 was used for OLS statistical analysis. The geographical activity was also divided into 2 phases. Phase 1 involved basic mapping of observed, expected and unrecorded disease prevalence, as well as of admission rates. This was followed by the use of LMI for unrecorded prevalence and admission rates to facilitate the Phase 2 use of GWR. Disease

prevalence was mapped using ESRI shapefiles. GeoDa was used for the LMI tests, which required a spatial weights matrix that defines a local neighbourhood around each geographic unit. The value at each unit is compared with the weighted average of the values of its neighbours. A weights file identifies the neighbours; 2km and 5km weight matrices were created for sensitivity analyses. The LMI results were mapped in QGIS.

Table 69: Rationale behind analysis blocks used in the split-variable spatial regression models

ANALYSIS BLOCK	MEAN IMD SCORE	DOMINANT OAC GROUPS	ANALYSIS BLOCK DESCRIPTION
Home Counties	12.72	New and Growing Towns, Prospering Southern England	Affluent, mixed urban-rural
North West London	25.37	London Centre, London Suburbs	Mixed SES, metropolitan
North East London	31.82	London Cosmopolitan	Deprived, metropolitan
South West London	14.32	London Suburbs, Thriving London Periphery	Affluent, metropolitan
South East London	26.45	London Cosmopolitan, London Suburbs	Deprived, metropolitan
South Coast	18.91	Coastal and Countryside, Regional Centres	Mixed SES, mixed urban-rural
South West England	16.91	Coastal and Countryside, Prospering Small Towns	Mixed SES, mixed urban-rural
West Midlands	23.00	Cities and Services	Mixed SES, metropolitan
East Midlands	19.49	Prospering Smaller Towns, New and Growing Towns	Mixed SES, urban
East of England	17.45	Prospering Smaller Towns, Coastal and Countryside	Mixed SES, mixed urban-rural
Yorkshire and the Humber	23.39	Prospering Smaller Towns, Manufacturing Towns	Mixed SES, urban
Greater Manchester	28.79	Centres with Industry	Deprived, metropolitan
Merseyside	34.61	Regional Centres, Industrial Hinterlands	Deprived, urban
North West England	26.05	Centres with Industry, Prospering Smaller Towns, Coastal and Countryside	Deprived, mixed urban-rural
North East England	26.61	Regional Centres, Industrial Hinterlands	Deprived, urban

Scenario 2/disease management

Again the basic unit of analysis was the general practice population. Indirectly standardised admission rates were calculated at the practice level, to produce expected admission rates

based on the England level, for comparisons with observed counts. Indirect standardisation is generally accepted as more robust when counts are relatively small.

The same baseline practice and PCT characteristics were used as for Scenario 1, including registered population (the sum of registers of general practices located within PCTs' geographic boundaries), age and sex distribution, IMD scores, practices per PCT and practice list size. Practice level characteristics (QOF points and indicator breakdown) were aggregated to PCT level, as required. As counts were much larger, PCT directly standardised (age and sex) admission rates were calculated per 100.000 GP registered population, using the total England population as the standard. The same counts of registered, estimated and undiagnosed CHD and stroke prevalence used for Scenario 1 were used as independent variables in this analysis.

Univariate associations between admission rates and possible explanatory factors at practice and PCT levels were first assessed using Spearman's correlation coefficient. OLS linear multivariate regression models were then fitted to examine the effects of all independent variables. At the practice level, because of low admission numbers we applied Poisson regression, where standard errors have been scaled using the square root of the deviance-based dispersion. The dependent variable was observed count of hospital admission and the offset was the expected number of admissions.

Phase 2 GWR involved statistical mapping results for stroke and CHD. The outputs of GWR models were stored in .csv format to facilitate import in any GIS package. QGIS can only access .csv files via "add delimited text layer" plug-in to add a point layer (general practices) to the map. Local R-squared correlation coefficients were mapped using quantile classifications and drawn using diverging red to blue colour schemes analysis region. PCT (source: CQC) and England boundaries (source: http://www.edina.ac.uk/digimap) were overlain to the final maps to aid visualisation of results. CQC data for stroke and CHD was manipulated using Microsoft Excel and GWR 3 so that it could be mapped in QGIS 1.3.0 (source: http://www.qgis.org). The datasets were projected to the British National Grid, which is based on a Transverse Mercator projection.

Two batches of GWR analysis were performed to reflect analysis 1 and analysis 2 in the foundation non-spatial regression modelling. The GWR 3.0 software permits the choice of different probability distribution models, Gaussian, logistic and Poisson. For analysis 1, a Gaussian model was chosen as unrecorded prevalence is a continuous variable. The situation proved to be more complex for analysis 2, for which a Poisson model would strictly speaking be appropriate, as admission rates are treated as discrete variables in epidemiology. However, we were obliged to use a Gaussian GWR model for analysis 2 as the Poisson functionality in the software was found not to operate correctly (this problem may have been remedied in more recent versions of the package). The spatial kernel options offered are fixed distance and adaptive distance; adaptive was selected to reflect the heterogeneous density of GP locations across the country and even with analysis regions. Statistical significance testing and optimal bandwidth selection was performed using Monte Carlo simulations. The GWR software generates a dichotomous output – spreadsheet and text file. The spreadsheet holds standardised residuals, Cook's D and Local R square values for every regression point (in our case general practice). The text file contains diagnostics for

the global model fitted to the data and GWR estimates. AIC values are compared to test for spatial non-stationarity.

10.7.4 Results

This section contains a sample of results from each scenario, disease and type of analysis (OLS, GWR).

Scenario 1/ disease detection: outcome=registered and undiagnosed prevalence

Scenario 1 aims to describe the effectiveness of general practices in detecting disease and registering cases on QOF registers so that patients are offered regular evidence-based management. We therefore used as the outcome variable the estimated prevalence of undiagnosed CHD and stroke in each practice. We undertook an OLS analysis and then LMI/GWR.

We have used selected QOF clinical indicators as independent variables in these analyses because, although they reflect the standard of care after diagnosis, it is possible that quality of care post-diagnosis is also a predictor of successful case-finding by practices (for example, by interrogating practice information systems to detect new cases, or by opportunistic screening).

CHD

Table 70 provides a descriptive analysis of the variables used. Data from 7,969 practices was analysed. Nationally the overall prevalence of CHD is estimated to be about 4.5 per cent, with about 20 per cent of cases unregistered as such by practices. Rather than using estimated proportion of ethnic minority population per practice, we have used estimated proportion South Asian because it is known to be a risk factor for CHD.

Table 70: Descriptive analysis

	Mean	(SD)	Min - Max	Median	IQR
Observed prevalence (%)	3.50	1.19	0.00 - 9.81	3.47	2.70 - 4.25
Unregistered prevalence (%)	0.95	0.87	0.00 - 5.78	0.87	0.40 - 1.44
IMD score	23.63	12.61	2.24 - 71.85	21.31	13.32 - 31.89
Smoking prevalence (%)	24.87	6.43	13.18 - 51.73	23.68	19.71 - 28.81
Asian population (%)	0.06	0.15	0.00 - 0.97	0.01	0.00 - 0.05
GP supply/100.000 population	56.17	17.98	0.00 - 337.08	55.10	45.80 - 65.28
QOF Clinical Indicators					
CHD06 (BP<150/90mm Hg)	89.82	5.03	0.00 - 100	90.29	87.16 - 93.05
CHD08 (Cholesterol<5Mmol/L)	81.71	6.78	0.00 - 100	82.07	77.83 - 86.15
CHD11 (treated with an ACE inhibitor/ Angiotensin II antagonist)	90.23	7.40	0.00 - 100	90.00	85.71 - 95.24
QOF Patient Care Experience Indicators					
PE07 (can obtain a consultation with GP in 2 days)	84.19	14.07	0.00 - 100	87.28	79.39 - 92.97
PE08 (can book an appointment with a GP > 2 days ahead)	76.97	17.57	0.00 - 100	80.67	67.23 - 90.28

For Scenario 1, we were interested to see whether there were associations between the quality of primary care for diagnosed and registered disease and its initial identification. We therefore used the same individual QOF indicators which we used for Scenario 2, that is, those we considered to be more likely to affect patient outcomes or intermediate outcome indicators. Total QOF achievement for a domain reflects a combination of intermediate outcome indicators and process of care indicators. The former may be better indicators of quality of care. In addition, achievement of process indicators is almost uniformly very high, creating a "ceiling effect" with little dispersion of practice values. As shown in Table 70, the chosen QOF indicators have mean achievements of 80-90 per cent.

We initially fitted models using observed/registered prevalence as the outcome variable to gain an understanding of associations with population and healthcare factors. A characteristic of most of the OLS analyses is incidence rate/odds ratios (IRRs/ORs) mainly very close to unity, suggesting that effect sizes are relatively small. This is somewhat surprising as several other published PCT-level analyses show, for example, relatively strong positive associations of prevalence with deprivation,[270] and it is included in the model used to estimate overall prevalence, which is based on HSfE patient-reported doctor-diagnosed CHD (OR for fifth IMD quintile 2.4).

Table 71 shows a univariate analysis with unregistered CHD as the outcome variable. Deprivation, smoking and South Asian ethnicity are risk factors i.e. predictors of unregistered disease, whereas higher scores for QOF indicators CHD06 and CHD08 are protective.

Table 71: OLS univariate analysis. Dependent variable: unregistered CHD count 2008-9

	IRR	Std. Err.	z	P>|z|	[95% Conf. Interval]	
IMD score	1.025	0.001	40.280	0.000	1.024	1.027
Smoking prevalence (%)	1.045	0.001	34.840	0.000	1.042	1.047
Asian population (%)	2.098	0.127	12.240	0.000	1.863	2.362
GP supply/100,000 population	1.000	0.001	0.560	0.575	0.999	1.002
QOF Clinical Indicators						
CHD06 (BP<150/90mm Hg)	0.993	0.002	-3.940	0.000	0.989	0.996
CHD08 (Cholesterol<5Mmol/L)	0.993	0.001	-5.110	0.000	0.990	0.996
CHD11 (treated with an ACE inhibitor/ Angiotensin II antagonist)	1.000	0.001	0.200	0.840	0.998	1.003
QOF Patient Experience Indicators						
PE07 (can obtain a consultation with GP in 2 days)	0.998	0.001	-2.930	0.003	0.997	0.999
PE08 (can book an appointment with a GP > 2 days ahead)	0.999	0.001	-1.790	0.073	0.998	1.000

Table 72 shows a multivariate analysis with unregistered CHD count as the outcome variable, using the practice population as the offset variable in the Poisson regression. The sign for smoking prevalence is now negative, which is unexpected, and suggests that multicollinearity may be present. Other variables were not retained in the model. The directions of some of these associations are unexpected: one might expect GP supply to be negatively associated and smoking prevalence to be positively associated. In the presence of

multicollinearity, the estimate of one variable's impact on y while controlling for the others tends to be less precise than if predictors were uncorrelated with one another. If, for example, smoking is highly correlated with deprivation, we have many observations for which these two variables have a particular relationship, and fewer observations for which smoking changes independently of deprivation, or vice versa, so we have an imprecise estimate of the effect of independent changes in each variable. However, in the presence of multicollinearity we expect larger standard errors, and this is not the case in our dataset.

Table 72: OLS multivariate stepwise analysis. Dependent variable: unregistered CHD count 2008-9 Independent variables: population + institutional factors

	IRR	Std. Err.	z	P>|z|	[95% Conf. Interval]	
IMD score	1.048	0.002	20.680	0.000	1.043	1.053
Smoking prevalence (%)	0.957	0.004	-9.900	0.000	0.948	0.965
GP supply/100.000 population	1.001	0.001	2.590	0.010	1.000	1.002
QOF Clinical Indicators						
CHD08 (Cholesterol<5Mmol/L)	0.998	0.001	-1.910	0.056	0.995	1.000
QOF Patient Experience Indicators						
PE08 (can book an appointment with a GP > 2 days ahead)	1.002	0.000	3.470	0.001	1.001	1.003

The variance inflation factor (VIF) can be estimated to quantify the severity of multicollinearity in OLS regression analysis. It provides an index that measures how much the variance of an estimated regression coefficient (the square of the estimate's standard deviation) is increased because of collinearity. A VIF>10 and a tolerance<0.1 indicates multicollinearity (see Table 73). Only IMD score and estimated smoking prevalence (which uses IMD score as an input) showed evidence of multicollinearity.

Table 73: Collinearity Diagnostics

Variable	VIF	SQRT VIF	Tolerance	Local R-squared
Observed CHD prevalence	1.44	1.20	0.6923	0.3077
Unregistered CHD prevalence	1.39	1.18	0.7198	0.2802
IMD score	14.74	3.84	0.0678	0.9322
Smoking prevalence	13.46	3.67	0.0743	0.9257
Asian population (%)	1.23	1.11	0.8101	0.1899

The next section illustrates practice-level mapping of prevalence data. Figure 24 shows expected CHD prevalence rate by practice, for England in 2008 (so that it is comparable with the latest available QOF prevalence data), using estimates from the ICL/APHO prevalence models. Higher rates are evident in the North and in coastal (retirement) areas. Lower rates are seen in London (because of its relatively young population age structure) and the Home Counties (because they are less deprived). Figure 25 shows QOF registered CHD prevalence rate by practice in England in 2008. At first sight there are many similarities with expected prevalence. However, Figure 26, unregistered (expected minus registered) CHD prevalence

rate, shows evidence of possible spatial clustering, with higher rates in urban areas and vice versa. Figure 27, observed divided by expected CHD prevalence rate by practice, is an alternative means of displaying this. The lowest proportions are clearly clustered in London and West Midlands conurbations.

The next step would be to visualise this data using the LMI test. To avoid making the report excessively detailed, yet to demonstrate utility, we have carried out LMI tests for stroke only.

Figure 24: expected CHD prevalence rate by practice, England 2008

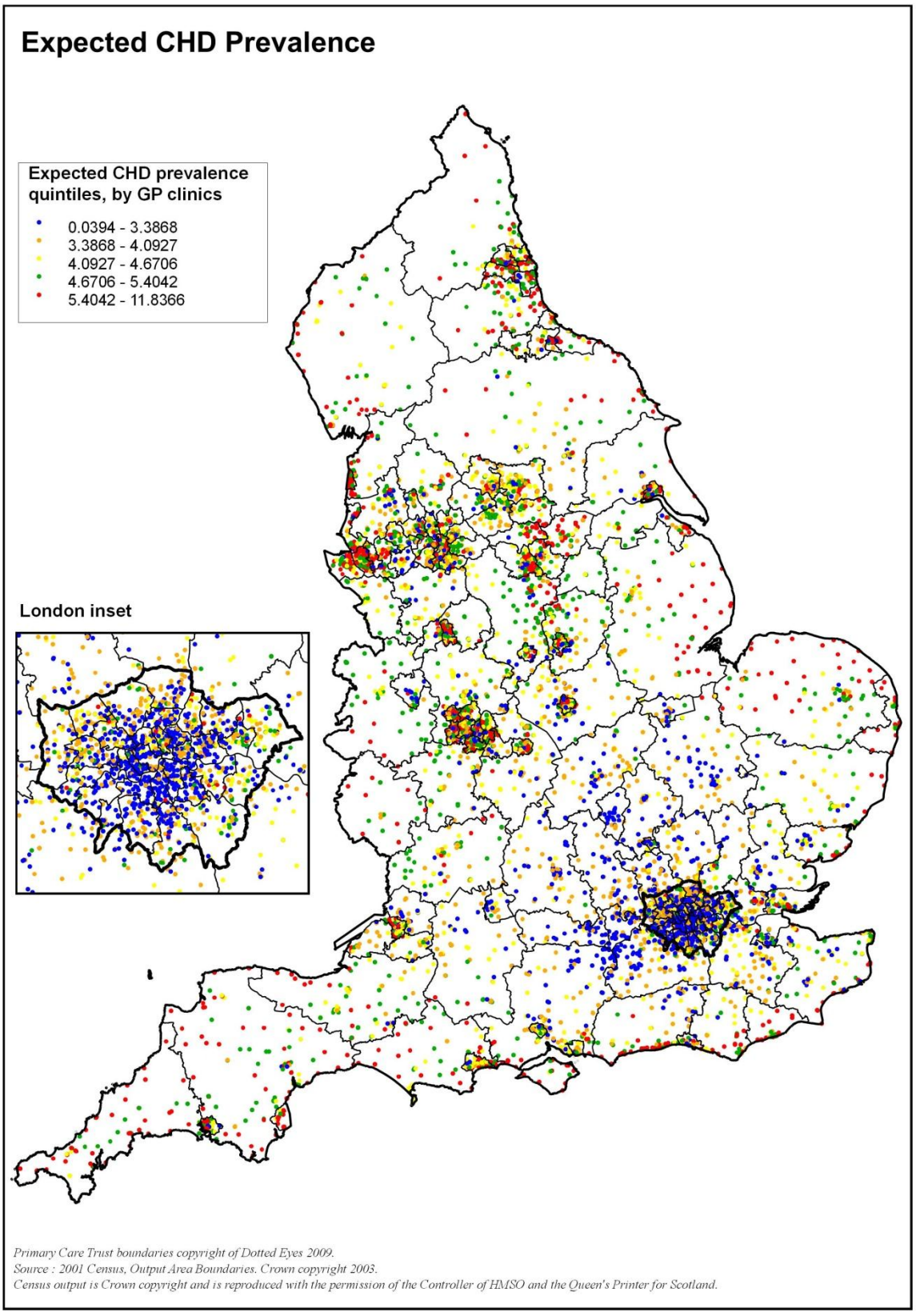


Figure 25: QOF registered CHD prevalence rate by practice, England 2008

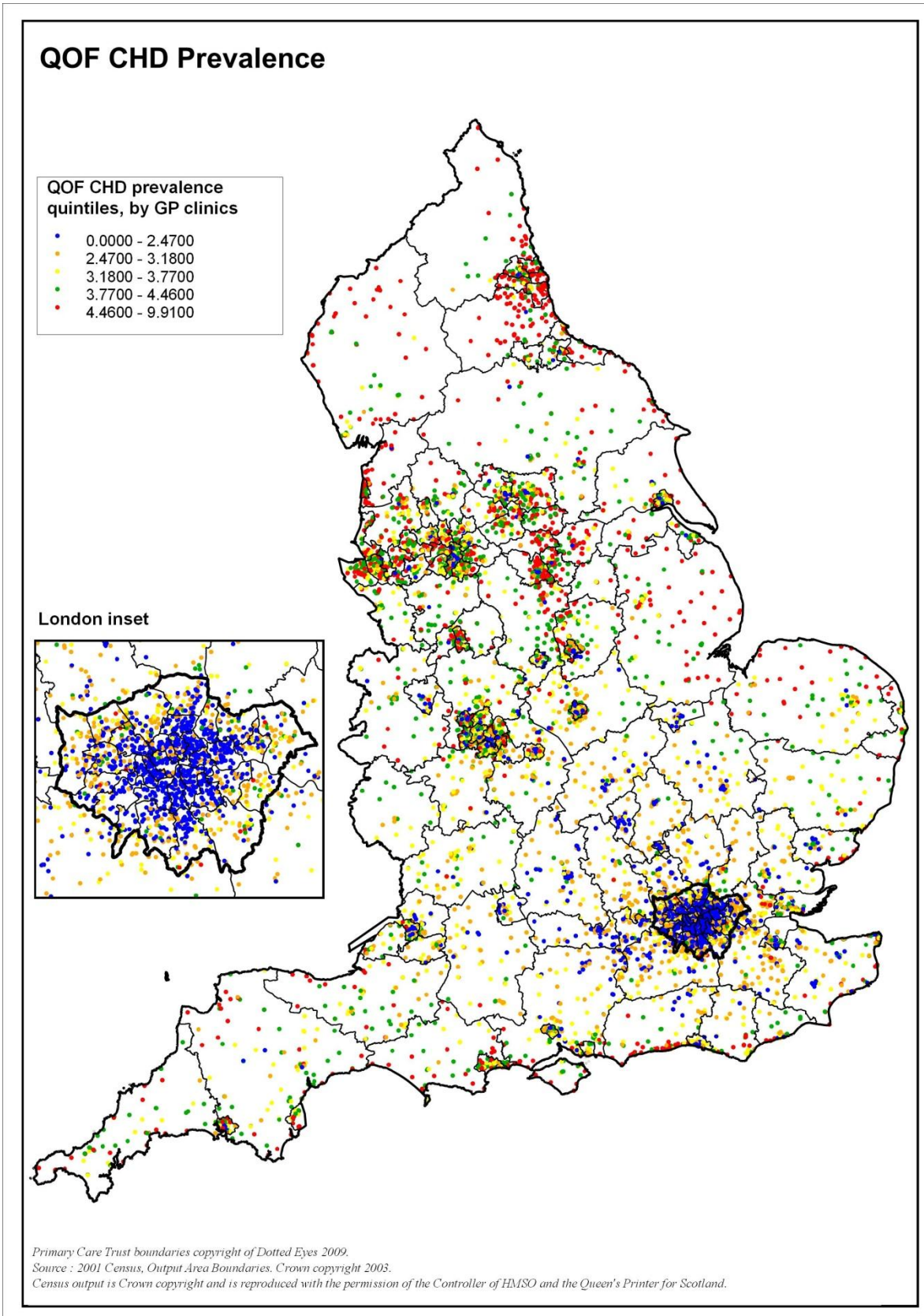


Figure 26: unregistered CHD prevalence rate by practice, England 2008

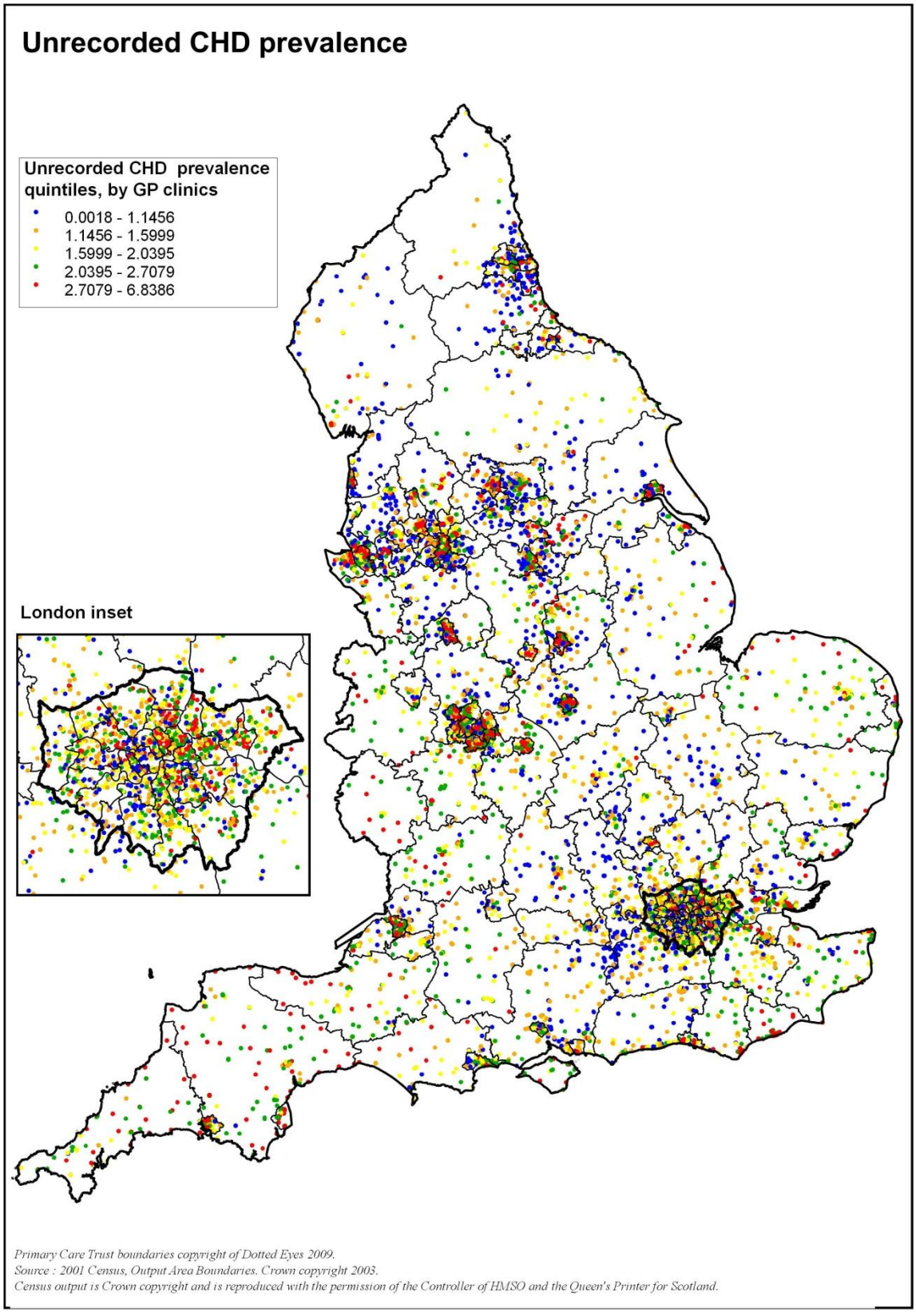
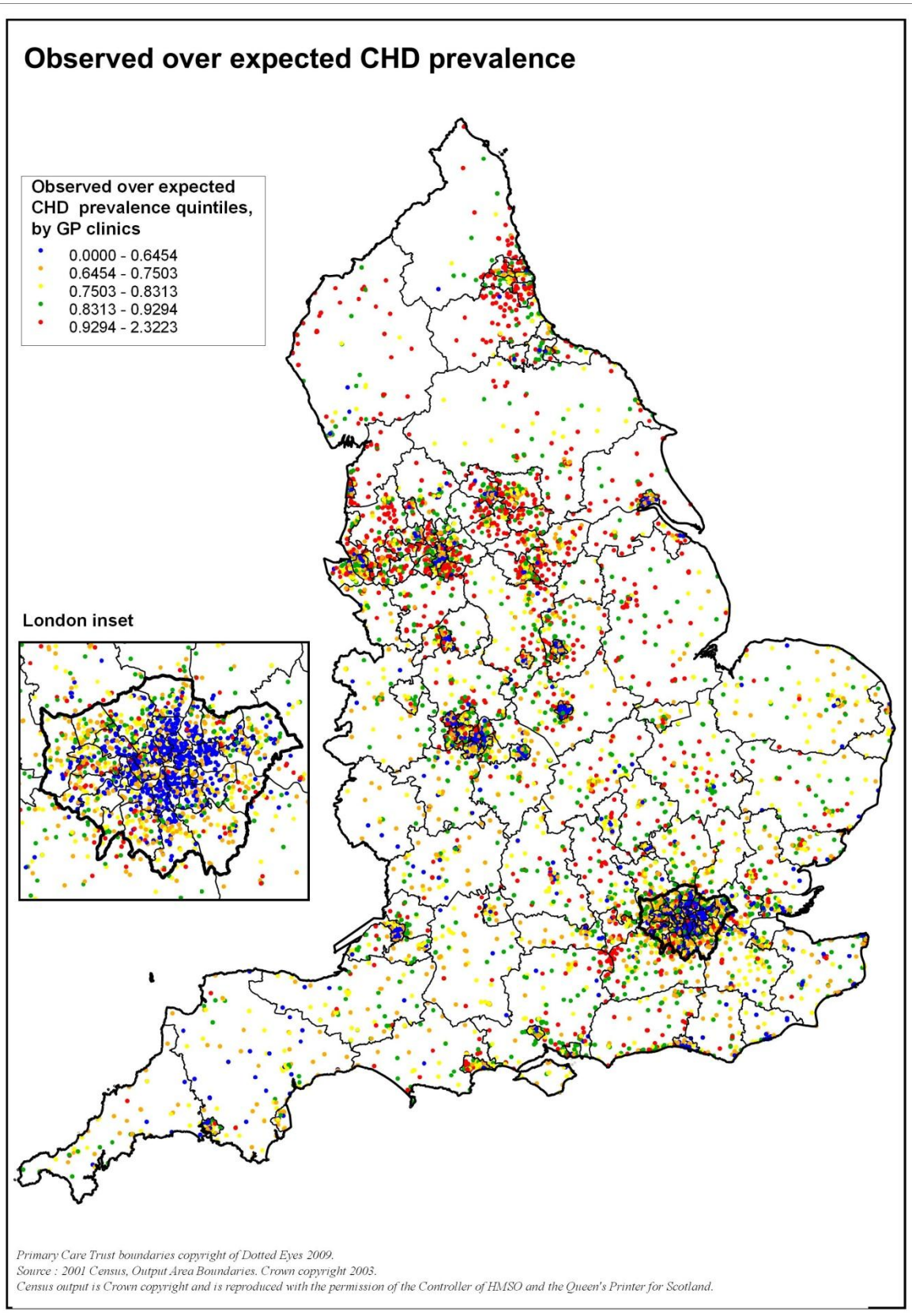


Figure 27: observed: expected CHD prevalence rate by practice, England 2008



Stroke

Table 74 provides a descriptive analysis of the stroke variables used. Data from 7,969 practices was analysed. Nationally the overall prevalence of stroke is estimated to be about 1.6 per cent, with about 25 per cent of cases apparently unregistered as such by practices. As a general rule, the proportion of expected cases undiagnosed is lower for conditions like stroke which have a sudden as opposed to gradual or insidious onset. Rather than using estimated proportion of ethnic minority population per practice, we have used estimated proportion of Black/Black Caribbean because it is known to be a risk factor for stroke.

Table 74: descriptive analysis

	Mean	(SD)	Min - Max	Median	IQR
Observed prevalence (%)	1.61	0.67	0.00 - 18.55	1.61	1.15 - 2.02
Unregistered prevalence (%)	0.42	0.44	0.00 - 3.83	0.31	0.01 - 0.65
IMD score	23.67	12.62	2.24 - 71.85	21.34	13.35 - 31.99
Smoking prevalence (%)	24.90	6.44	13.18 - 51.73	23.71	19.73 - 28.89
Black population (%)	3.43	7.35	0.00 - 72.39	0.52	0.15 - 2.58
GP supply/100.000 population	56.10	18.06	0.00 - 337.08	55.05	45.72 - 65.26
QOF Clinical Indicators					
Stroke06 (Last BP ≤ 150/90)	88.09	6.73	0.00 - 100	88.64	84.67 - 92.12
Stroke08 (Last cholesterol ≤5mmol/l)	76.82	9.31	0.00 - 100	77.55	71.72 - 82.86
Stroke12 (on anti-platelet agent/anti-coagulant)	94.56	5.53	0.00 - 100	94.74	92.45 - 97.53
BP05 (In all hypertensives,BP<150/90)	78.90	6.35	22.22 - 100	79.04	75.10 - 83.00
QOF Patient Care Experience Indicators					
QOF Patient Experience Indicators					
PE07 (can obtain a consultation with GP in 2 days)	84.21	14.05	0.00 - 100	87.30	79.39 - 92.98
PE08 percentage of patients who were able to book an appointment with a GP > 2 days ahead	77.07	17.54	0.00 - 100	80.79	67.32 - 90.36

Table 75 shows the results of a univariate analysis. Once again effect sizes and standard errors are small.

Table 75: OLS univariate analysis. Dependent variable: unregistered stroke count 2008-9

	IRR	Std. Err.	z	P>|z|	[95% Conf. Interval]	
IMD score	1.037	0.001	45.550	0.000	1.036	1.039
Smoking prevalence (%)	1.066	0.002	39.220	0.000	1.062	1.069
Black population (%)	1.000	0.000	24.640	0.000	1.000	1.000
GP supply/100.000 population	0.990	0.001	-12.710	0.000	0.988	0.991
QOF Clinical Indicators						
Stroke06 (Last BP ≤ 150/90)	0.991	0.002	-4.430	0.000	0.987	0.995
Stroke08 (Last cholesterol ≤5mmol/l)	0.990	0.001	-7.100	0.000	0.987	0.993
Stroke12 (on anti-platelet	1.011	0.003	3.560	0.000	1.005	1.017

	IRR	Std. Err.	z	P>|z|	[95% Conf. Interval]	
agent/anti-coagulant)						
QOF Patient Experience Indicators						
PE07 (can obtain a consultation with GP in 2 days)	0.992	0.001	-10.790	0.000	0.991	0.994
PE08 (can book an appointment with a GP > 2 days ahead)	0.995	0.001	-6.930	0.000	0.994	0.997
Blood pressure 05	0.997	0.002	-1.490	0.136	0.992	1.001
CQC PCT rating = fair vs. poor	0.594	0.044	-7.050	0.000	0.514	0.686
CQC PCT rating = good vs. poor	0.534	0.040	-8.470	0.000	0.462	0.617
CQC PCT rating = excellent vs. poor	0.527	0.071	-4.730	0.000	0.404	0.687

Table 76 shows a multivariate model with reverse stepwise variable selection. Results are generally similar to the CHD analysis.

Table 76: OLS multivariate stepwise analysis. Dependent variable: unregistered stroke count 2008-9 Independent variables: population + institutional factors

	IRR	Std. Err.	z	P>|z|	[95% Conf. Interval]	
IMD score	1.071	0.003	22.580	0.000	1.065	1.078
Smoking prevalence (%)	0.931	0.006	-12.010	0.000	0.921	0.942
Black population (%)	1.000	0.000	5.310	0.000	1.000	1.000
GP supply/100.000 population	0.994	0.001	-9.020	0.000	0.992	0.995
QOF Clinical Indicators						
BP05 (In all hypertensives, BP<150/90)	1.007	0.002	3.560	0.000	1.003	1.011
Stroke08 (Last cholesterol ≤5mmol/l)	0.989	0.001	-8.260	0.000	0.986	0.991
Stroke12 (on anti-platelet agent/anti-coagulant)	1.012	0.003	4.770	0.000	1.007	1.017
QOF Patient Experience Indicators						
PE07 (can obtain a consultation with GP in 2 days)	0.998	0.001	-3.010	0.003	0.996	0.999
CQC PCT rating = good vs. poor	0.931	0.021	-3.230	0.001	0.891	0.972

The next section illustrates practice-level mapping of stroke prevalence data. Figure 28 shows expected stroke prevalence rate by practice, for England in 2008 (so that it is comparable with the latest available QOF prevalence data), using estimates from the ICL/APHO prevalence models. As for CHD, higher rates are evident in the North and in coastal (retirement) areas, and lower rates are seen in London (because of its relatively young population age structure) and the Home Counties (because they are less deprived). Figure 29 shows QOF registered stroke prevalence rate by practice in England 2008. Again, there are many similarities with expected prevalence. Figure 30, unregistered (expected minus registered) stroke prevalence rate, shows evidence of possible spatial clustering, with higher rates in urban areas and vice versa. Figure 31, observed divided by expected stroke

prevalence rate by practice, is an alternative means of displaying this. The lowest proportions are clustered in London and West Midlands conurbations, but there is a more dispersed pattern than for CHD.

Figure 28: expected stroke prevalence rate by practice, England 2008

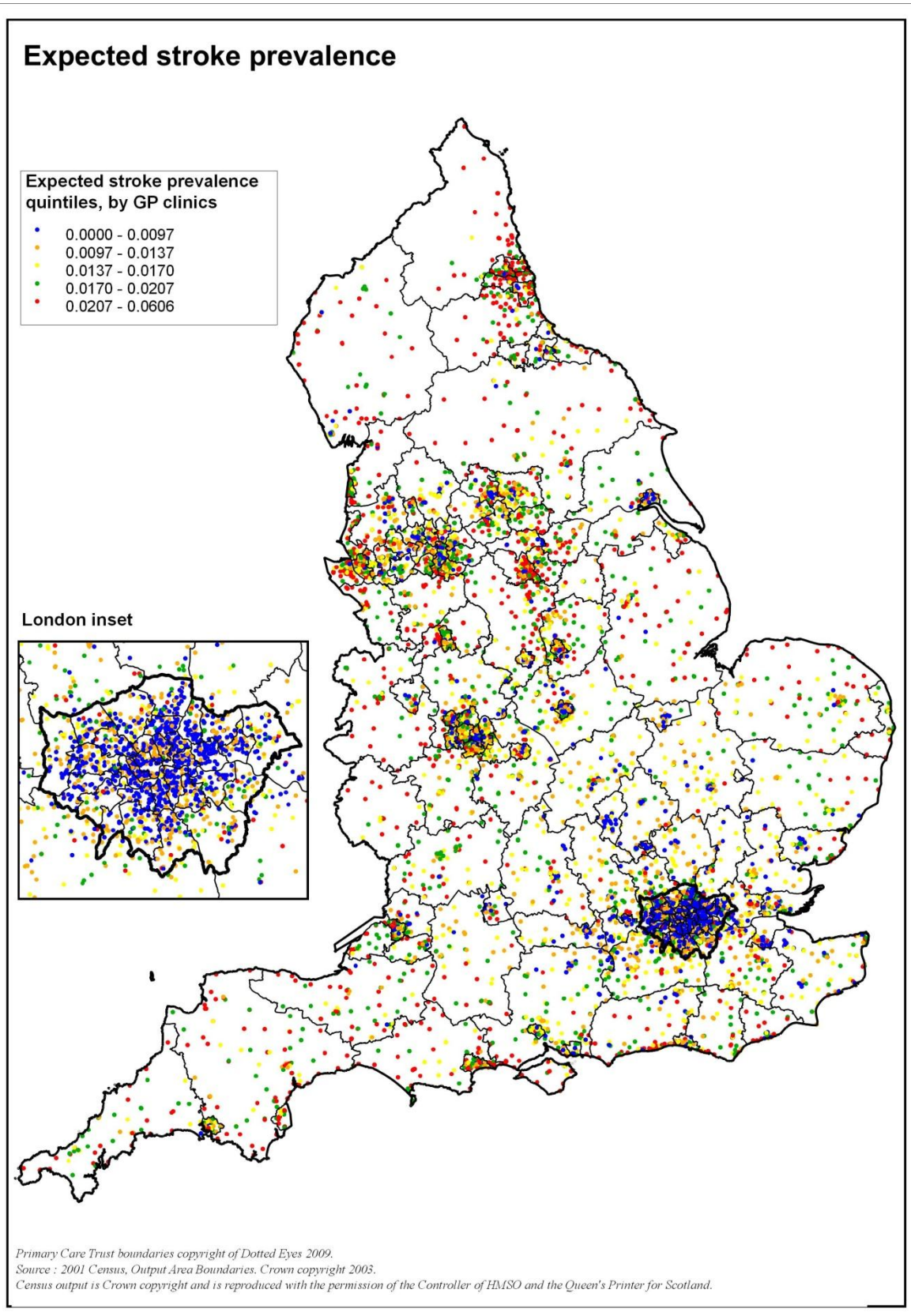


Figure 29: QOF registered stroke prevalence rate by practice, England 2008

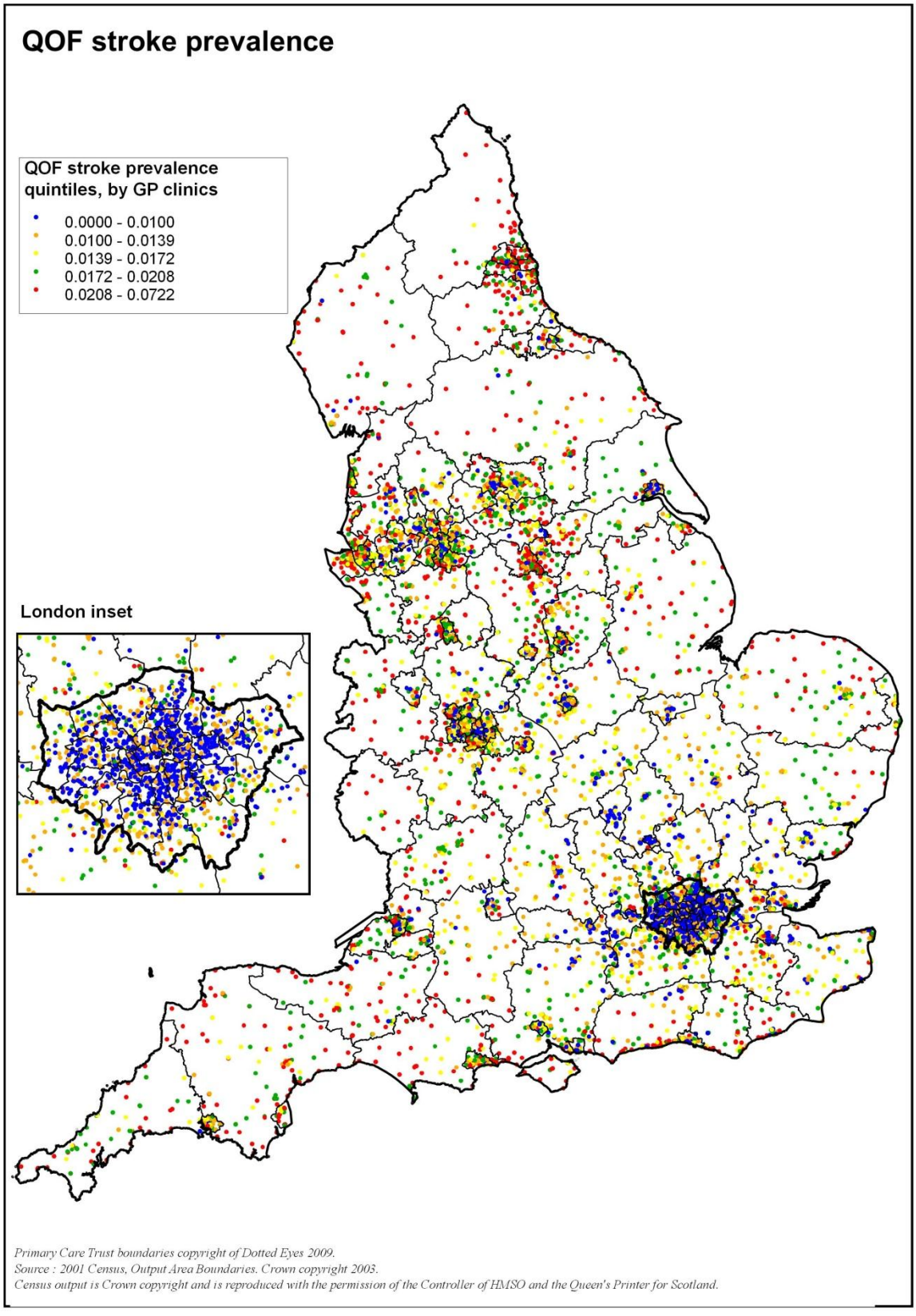


Figure 30: unregistered CHD prevalence rate by practice, England 2008

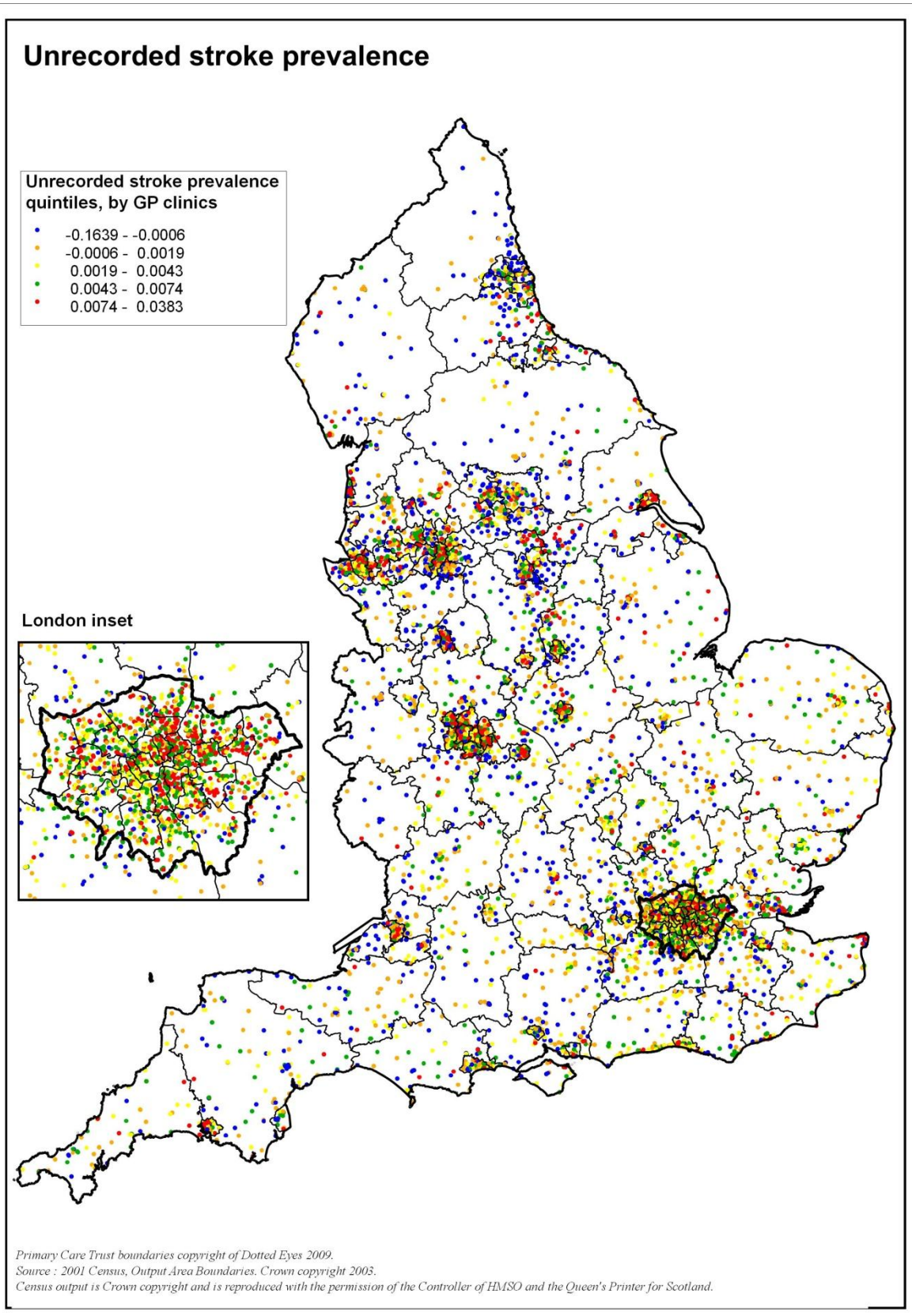
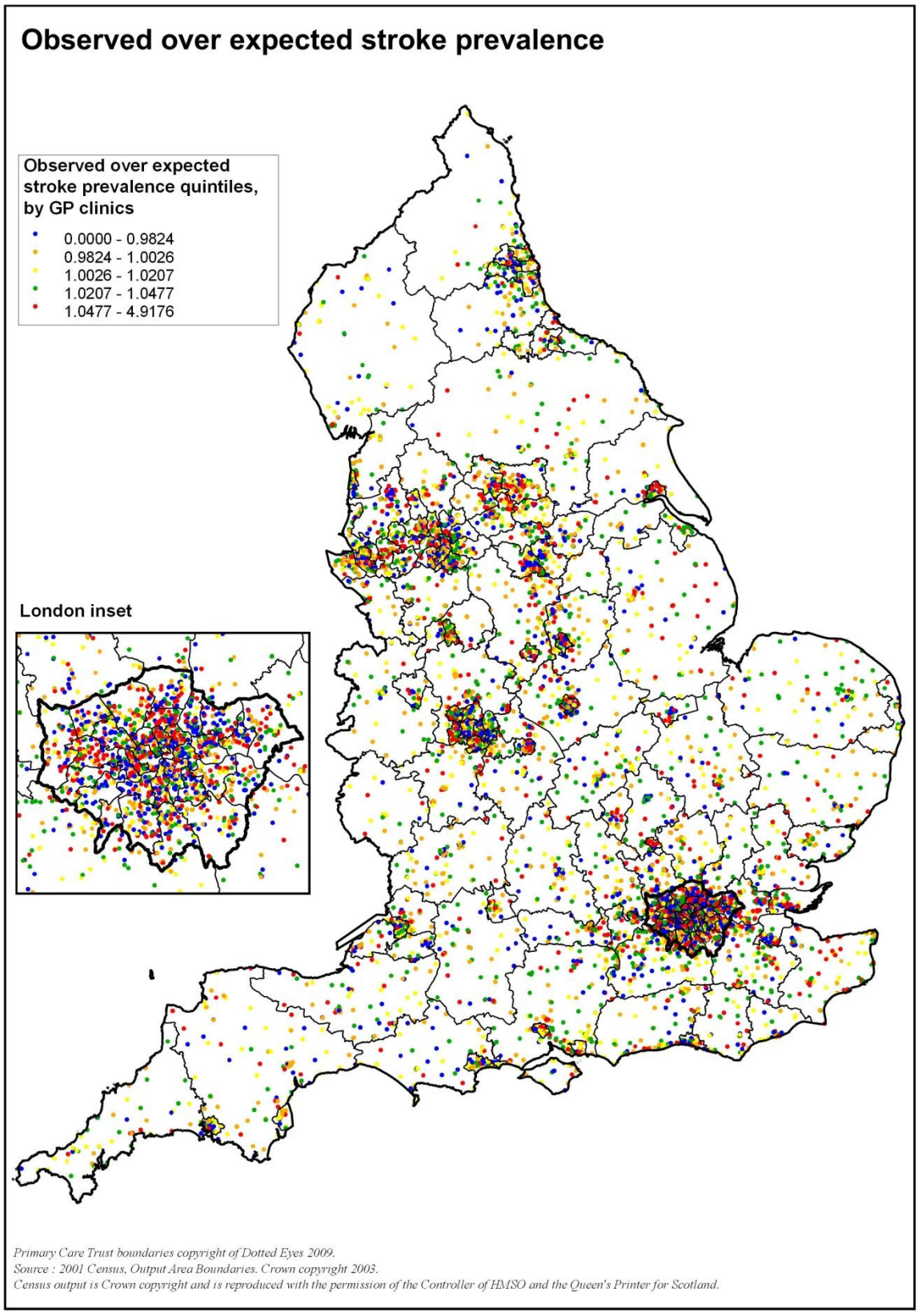


Figure 31: observed: expected CHD prevalence rate by practice, England 2008



The next two practice level maps (Figure 32 and Figure 33) show Local Moran's I tests at 2km and 5km for unrecorded prevalence of stroke. There is marked clustering of high levels of unrecorded prevalence evident in London and other conurbations, which is even more prominent in the 5 km map. There is some, but much less, clustering of low levels of unrecorded prevalence which tends to be in the hinterlands of urban areas.

Figure 32: Local Moran's I (2km) of stroke unrecorded prevalence

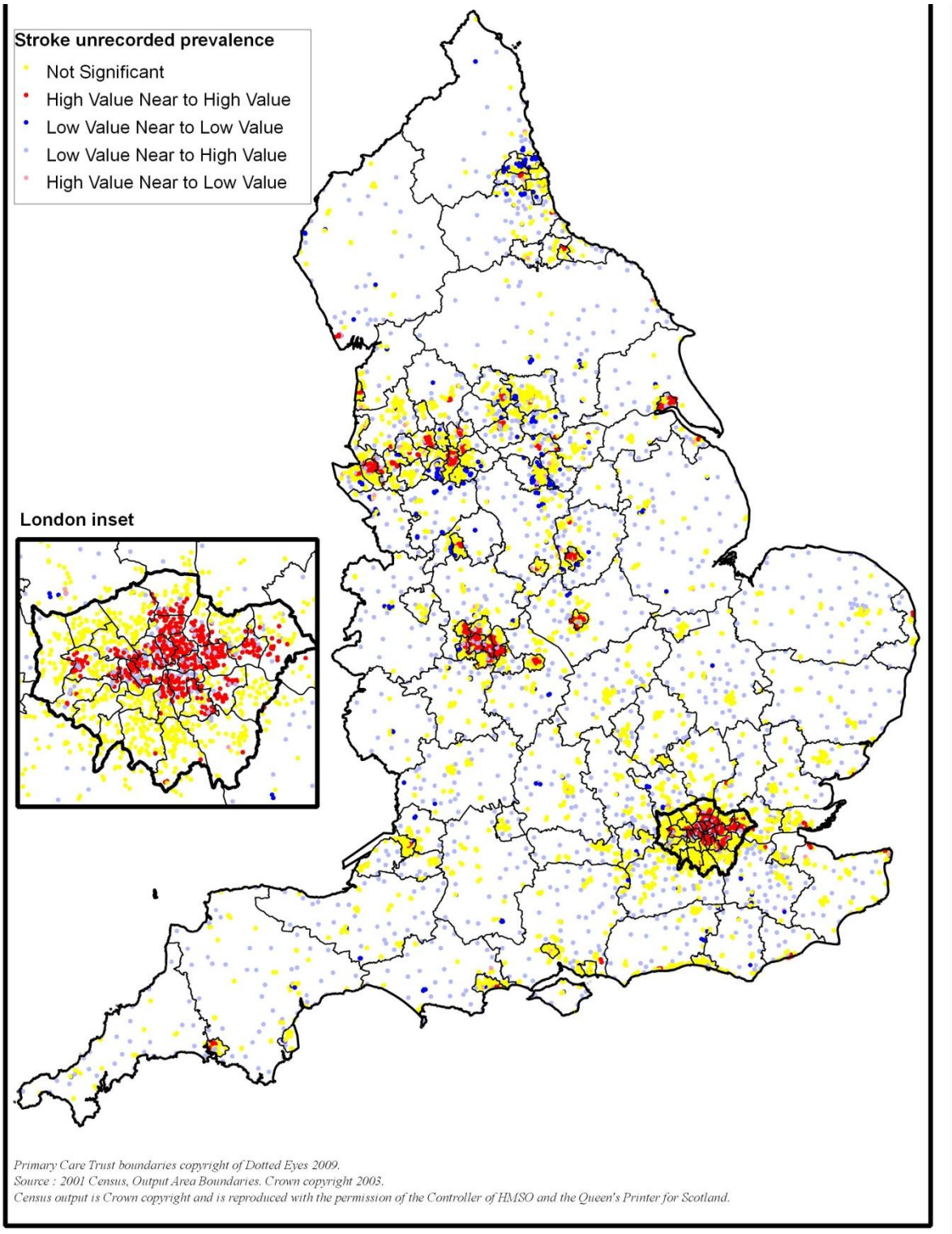


Figure 33: Local Moran's I (5 km) of stroke unrecorded prevalence

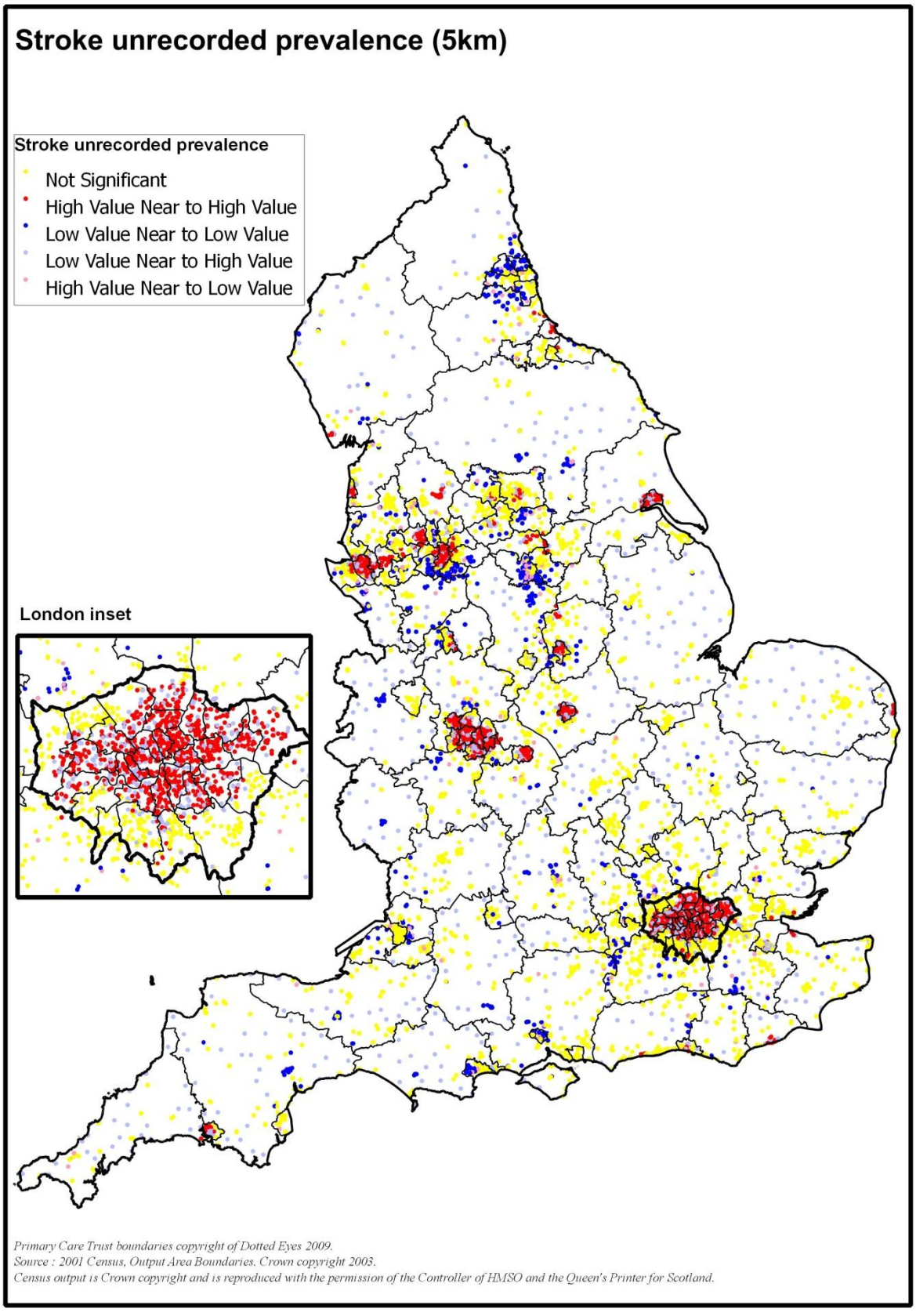


Table 77 compares OLS and geographical analyses for stroke prevalence. Comparison with classical regression outcomes was made using the AIC[202] to test for spatial non-stationarity.[276] Identification of the optimal bandwidth for each analysis block was made using Monte Carlo simulations. The results of the spatial non-stationarity test revealed little difference between the AIC values for the non-spatial and GWR models, with marginal improvement with the spatial model. For example AIC values ranged from -1567 to -7032 between analysis blocks for the non-spatial model using primary care quality variables; the comparable range for the GWR model was -1565 to -7062. Whilst non-stationarity was not significant nevertheless the AIC values confirmed that it was statistically valid to use spatial regression as a basis for exploratory mapping.

Table 77: regression diagnostics for stroke prevalence

Sub regions	Residual Sum of Squares	Number of parameters	Sigma	AIC	Coefficient of Determ-ination	Sum of Squares	Degrees of Freedom	Adjusted r-square
	Classical Regression							
East Midlands	0.02	12.00	0.01	-6258.58	0.22	0.00	12.00	0.21
East of England	0.01	12.00	0.00	-2588.62	0.19	0.00	12.00	0.16
Greater Manchester	0.01	12.00	0.01	-2645.27	0.25	0.00	12.00	0.22
Home Counties	0.05	12.00	0.01	-7032.45	0.06	0.10	12.00	0.04
Merseyside	0.01	12.00	0.01	-2079.41	0.25	0.00	12.00	0.22
North East England	0.01	12.00	0.01	-2855.90	0.29	0.00	12.00	0.26
North East London	0.01	12.00	0.01	-2864.06	0.13	0.00	12.00	0.10
North West England	0.02	12.00	0.01	-4126.28	0.19	0.00	12.00	0.17
North West London	0.02	12.00	0.01	-4589.65	0.17	0.00	12.00	0.15
South Coast	0.01	12.00	0.00	-4184.88	0.20	0.00	12.00	0.18
South East London	0.01	12.00	0.01	-2147.34	0.20	0.00	12.00	0.16
South West England	0.01	12.00	0.00	-4800.19	0.13	0.00	12.00	0.12
South West London	0.00	12.00	0.00	-1567.75	0.17	0.00	12.00	0.12
West Midlands	0.03	12.00	0.01	-6996.77	0.27	0.00	12.00	0.26
Yorks & Humberside	0.03	12.00	0.01	-5342.75	0.24	0.00	12.00	0.23
	GWR Estimation							
East Midlands	0.02	21.79	0.01	-6280.93	0.26	0.00	9.79	0.24
East of England	0.01	26.60	0.00	-2585.45	0.26	0.00	14.60	0.20
Greater Manchester	0.01	22.63	0.01	-2649.56	0.30	0.00	10.63	0.26
Home Counties	0.05	49.27	0.01	-7062.32	0.15	0.00	37.27	0.11
Merseyside	0.01	19.28	0.01	-2074.25	0.28	0.00	7.28	0.23
North East England	0.01	25.18	0.01	-2864.45	0.36	0.00	13.18	0.31
North East London	0.01	20.11	0.01	-2856.82	0.15	0.00	8.11	0.10

Sub regions	Residual Sum of Squares	Number of parameters	Sigma	AIC	Coefficient of Determ-ination	Sum of Squares	Degrees of Freedom	Adjusted r-square
North West England	0.02	23.12	0.01	-4129.83	0.23	0.00	11.12	0.19
North West London	0.02	24.98	0.01	-4595.40	0.21	0.00	12.98	0.18
South Coast	0.01	18.84	0.00	-4185.28	0.23	0.00	6.84	0.20
South East London	0.01	23.49	0.00	-2152.77	0.28	0.00	11.49	0.22
South West England	0.01	21.05	0.00	-4823.66	0.19	0.00	9.05	0.16
South West London	0.00	20.51	0.00	-1565.15	0.24	0.00	8.51	0.16
West Midlands	0.03	42.72	0.01	-7009.74	0.33	0.00	30.72	0.29
Yorks & Humberside	0.03	21.37	0.01	-5362.36	0.28	0.00	9.37	0.26

Scenario 2/disease management: outcome=admission rates
CHD (MI+angina)

Table 78 shows descriptive data for Scenario 2. As noted in the Objectives (Section 4.1), while unscheduled/emergency admissions for CHD could be viewed as an adverse outcome of primary care management, elective CHD admissions, for example for revascularisation, could be viewed as appropriate care. We therefore agreed with CQC that the appropriate admission diagnostic groups for CHD were MI and angina, which are mainly unscheduled/emergency admissions.

Table 78: descriptive analysis

	Mean	(SD)	Min - Max	Median	IQR
Admissions/100.000 population	782.56	340.66	0.00 - 3966.02	720.02	559.92 - 933.88
Observed prevalence (%)	3.50	1.19	0.00 - 9.81	3.47	2.70 - 4.25
Unregistered prevalence (%)	0.95	0.87	0.00 - 5.78	0.87	0.40 - 1.44
IMD score	23.63	12.61	2.24 - 71.85	21.31	13.32 - 31.89
Smoking prevalence (%)	24.87	6.43	13.18 - 51.73	23.68	19.71 - 28.81
Asian population (%)	0.06	0.15	0.00 - 0.97	0.01	0.00 - 0.05
GP supply/100.000 population	56.17	17.98	0.00 - 337.08	55.10	45.80 - 65.28
QOF Clinical Indicators					
CHD06 (BP<150/90mm Hg)	89.82	5.03	0.00 - 100	90.29	87.16 - 93.05
CHD08 (Cholesterol<5mmol/l)	81.71	6.78	0.00 - 100	82.07	77.83 - 86.15
CHD11 (treated with ACE inhibitor/angiotensin antagonist)	90.23	7.40	0.00 - 100	90.00	85.71 - 95.24
QOF Patient Care Experience Indicators					
PE07 (can obtain a consultation with GP in 2 days)	84.19	14.07	0.00 - 100	87.28	79.39 - 92.97
PE08 (can book an appointment with GP more than 2 days ahead)	76.97	17.57	0.00 - 100	80.67	67.23 - 90.28

Table 79 shows a univariate analysis with practice level CHD (MI+angina) admission rate as the outcome variable. Observed prevalence, unregistered prevalence, IMD score, smoking prevalence and Asian population are all risk factors, whereas as expected higher scores for GP supply and QOF indicators are protective.

Table 79: OLS univariate analysis. Dependent variable: observed admission count.

	IRR	Std. Err.	z	P>|z|	[95% Conf. Interval]	
Observed prevalence (%)	1.089	0.004	24.750	0.000	1.081	1.096
Unregistered prevalence (%)	1.008	0.005	1.670	0.094	0.999	1.019
IMD score	1.015	0.000	54.090	0.000	1.015	1.016
Smoking prevalence (%)	1.029	0.001	50.240	0.000	1.028	1.030
Asian population (%)	2.221	0.072	24.530	0.000	2.083	2.367
GP supply/100.000 population	0.999	0.000	-5.340	0.000	0.998	0.999

	IRR	Std. Err.	z	P>|z|	[95% Conf. Interval]	
QOF Clinical Indicators						
CHD06 (BP<150/90mm Hg)	0.998	0.001	-1.810	0.071	0.996	1.000
CHD08 (Cholesterol<5Mmol/L)	0.996	0.001	-5.510	0.000	0.995	0.998
CHD11 (treated with an ACE inhibitor/ Angiotensin II antagonist)	0.998	0.001	-3.360	0.001	0.996	0.999
QOF Patient Experience Indicators						
PE07 (can obtain a consultation with GP in 2 days)	0.996	0.000	13.770	0.000	0.996	0.997
PE08 (can book an appointment with GP more than 2 days ahead)	0.996	0.000	15.860	0.000	0.996	0.997

Table 80 shows a multivariate analysis with MI+angina admission rate as the outcome variable, with the expected practice rate (as we used indirect standardisation here) as the offset variable in the Poisson regression. Other variables not shown were eliminated from the model by reverse stepwise variable selection. Observed prevalence, IMD score, smoking prevalence and Asian population are as expected all risk factors for hospital admission- the IRR for the latter is comparatively high. The IRR for unregistered prevalence is less than unity, which is unexpected, as patients with undiagnosed disease might be expected to be more likely to be admitted. The directions of other associations are generally as expected, with higher scores on QOF Clinical Indicators as protective, apart from CHD06 (BP<150/90mm Hg). GP supply/100.000 population is not retained in the model.

Table 80: OLS multivariate stepwise analysis. Dependent variable: observed admission count. Independent variables: population + institutional factors

	IRR	Std. Err.	z	P>|z|	[95% Conf. Interval]	
Observed prevalence (%)	1.086	0.004	23.560	0.000	1.078	1.093
Unregistered prevalence (%)	0.954	0.004	10.400	0.000	0.945	0.962
IMD score	1.007	0.001	6.660	0.000	1.005	1.009
Smoking prevalence (%)	1.014	0.002	7.030	0.000	1.010	1.018
Asian population (%)	1.835	0.054	20.600	0.000	1.732	1.944
QOF Clinical Indicators						
CHD06 (BP<150/90mm Hg)	1.002	0.001	2.520	0.012	1.000	1.004
CHD08 (Cholesterol<5mmol/l)	0.998	0.001	-2.600	0.009	0.997	1.000
CHD11 (treated with ACE inhibitor/angiotensin antagonist)	0.998	0.001	-3.500	0.000	0.997	0.999
QOF Patient Experience Indicators						
PE07 (can obtain a consultation with GP in 2 days)	0.999	0.000	-4.320	0.000	0.998	0.999
PE08 (can book an appointment with GP more than 2 days ahead)	0.999	0.000	-5.950	0.000	0.998	0.999

Other results are shown in Section 9 Appendix C: detailed results. The next step would be to visualise this data using the LMI test. To avoid making the report excessively detailed, yet to demonstrate utility, we have carried out LMI tests for stroke only.

Stroke

Table 81 provides a descriptive analysis of the stroke variables used. The mean admission rate is 104/100,000 practice population, with very wide variation. Rather than using estimated proportion of ethnic minority population per practice, we have used estimated proportion Black/Black Caribbean because it is known to be a risk factor for stroke.

Table 81: descriptive analysis

	Mean	(SD)	Min - Max	Median	IQR
Admissions/100.000 population	103.74	39.13	0.00 - 476.51	99.28	79.02 - 123.77
Observed prevalence (%)	1.61	0.67	0.00 - 18.55	1.61	1.15 - 2.02
Unregistered prevalence (%)	0.42	0.44	0.00 - 3.83	0.31	0.01 - 0.65
IMD score	23.67	12.62	2.24 - 71.85	21.34	13.35 - 31.99
Smoking prevalence (%)	24.90	6.44	13.18 - 51.73	23.71	19.73 - 28.89
Black population (%)	3.43	7.35	0.00 - 72.39	0.52	0.15 - 2.58
GP supply/100.000 population	56.10	18.06	0.00 - 337.08	55.05	45.72 - 65.26
QOF Clinical Indicators					
Stroke06 (Last BP $\leq$ 150/90)	88.09	6.73	0.00 - 100	88.64	84.67 - 92.12
Stroke08 (Last cholesterol $\leq$5mmol/l)	76.82	9.31	0.00 - 100	77.55	71.72 - 82.86
Stroke12 (on anti-platelet agent/anti-coagulant)	94.56	5.53	0.00 - 100	94.74	92.45 - 97.53
BP05 (In all hypertensives,BP<150/90)	78.90	6.35	22.22 - 100	79.04	75.10 - 83.00
QOF Patient Care Experience Indicators					
PE07 (can obtain a consultation with GP in 2 days)	84.21	14.05	0.00 - 100	87.30	79.39 - 92.98
PE08 (can book an appointment with GP more than 2 days ahead)	77.07	17.54	0.00 - 100	80.79	67.32 - 90.36

Table 82 shows a univariate analysis with practice level stroke admission rate as the outcome variable. Observed prevalence, unregistered prevalence, IMD score, smoking prevalence and Asian population are all risk factors, whereas as expected higher scores for GP supply and QOF indicators are protective.

Table 82: univariate analysis. Dependent variable: observed admission count

	IRR	SE	z	P>|z|	[95% Conf. Interval]	
Observed prevalence (%)	1.026	0.006	4.650	0.000	1.015	1.037
Unregistered prevalence (%)	1.066	0.009	7.580	0.000	1.049	1.084
IMD score	1.009	0.000	33.890	0.000	1.009	1.010
Smoking prevalence (%)	1.017	0.001	32.340	0.000	1.016	1.018
Black population (%)	1.004	0.001	6.500	0.000	1.003	1.005
GP supply/100.000 population	1.000	0.000	-0.170	0.868	1.000	1.000
QOF Clinical Indicators						

	IRR	SE	z	P>|z|	[95% Conf. Interval]	
Stroke06 (Last BP ≤ 150/90)	0.999	0.001	-1.900	0.057	0.998	1.000
Stroke08 (Last cholesterol ≤5mmol/l)	0.999	0.000	-1.500	0.134	0.999	1.000
Stroke12 (on anti-platelet agent/anti-coagulant)	1.000	0.001	-0.520	0.604	0.998	1.001
BP05 (In all hypertensives,BP<150/90)	0.999	0.001	-1.390	0.163	0.998	1.000
QOF Patient Experience Indicators						
PE07 (can obtain a consultation with GP in 2 days)	0.999	0.000	-3.950	0.000	0.998	0.999
PE08 (can book an appointment with GP more than 2 days ahead)	0.999	0.000	-5.890	0.000	0.998	0.999
CQC PCT rating = fair vs. poor	0.964	0.027	-1.270	0.202	0.912	1.020
CQC PCT rating = good vs. poor	0.993	0.028	-0.250	0.800	0.939	1.050
CQC PCT rating = excellent vs poor	0.939	0.037	-1.590	0.111	0.869	1.015

Table 83 shows a multivariate analysis with stroke admission rate as the outcome variable, with the expected practice rate (as we used indirect standardisation here) as the offset variable in the Poisson regression. Other variables not shown were eliminated from the model by reverse stepwise variable selection. As for MI+angina, observed prevalence, IMD score and smoking prevalence are as expected all risk factors for hospital admission. The IRR for unregistered prevalence is less than unity, which is unexpected, as patients with undiagnosed disease might be expected to be more likely to be admitted. Only one QOF Clinical Indicator, Stroke08 (Last cholesterol ≤5mmol/l), remains in the model and is as expected, with higher scores protective. GP supply/100.000 population is not retained in the model.

Table 83: OLS multivariate stepwise analysis. Dependent variable: observed admission count. Independent variables: population + institutional factors

	IRR	SE	z	P>|z|	[95% Conf. Interval]	
Observed prevalence (%)	1.033	0.006	5.120	0.000	1.020	1.045
Unregistered prevalence (%)	0.957	0.010	-4.290	0.000	0.937	0.976
IMD score	1.008	0.001	7.600	0.000	1.006	1.010
Smoking prevalence (%)	1.005	0.002	2.370	0.018	1.001	1.008
QOF Clinical Indicators						
Stroke08 (Last cholesterol ≤5mmol/l)	0.998	0.000	-3.750	0.000	0.998	0.999
CQC PCT rating = fair vs. poor	0.983	0.006	-2.680	0.007	0.971	0.995
CQC PCT rating = excellent vs. poor	0.954	0.025	-1.780	0.075	0.906	1.005

The next two practice level maps (Figure 34 and Figure 35) show Local Moran's I tests at 2km and 5km for stroke admission rates. Once again there is marked clustering of high-high admission rate values evident in London and other conurbations, which is even more prominent in the 5 km map. However there is also clustering of low-low levels, especially in West London and in the West Midlands.

Figure 34: Local Moran's I (2km) of stroke admission rates

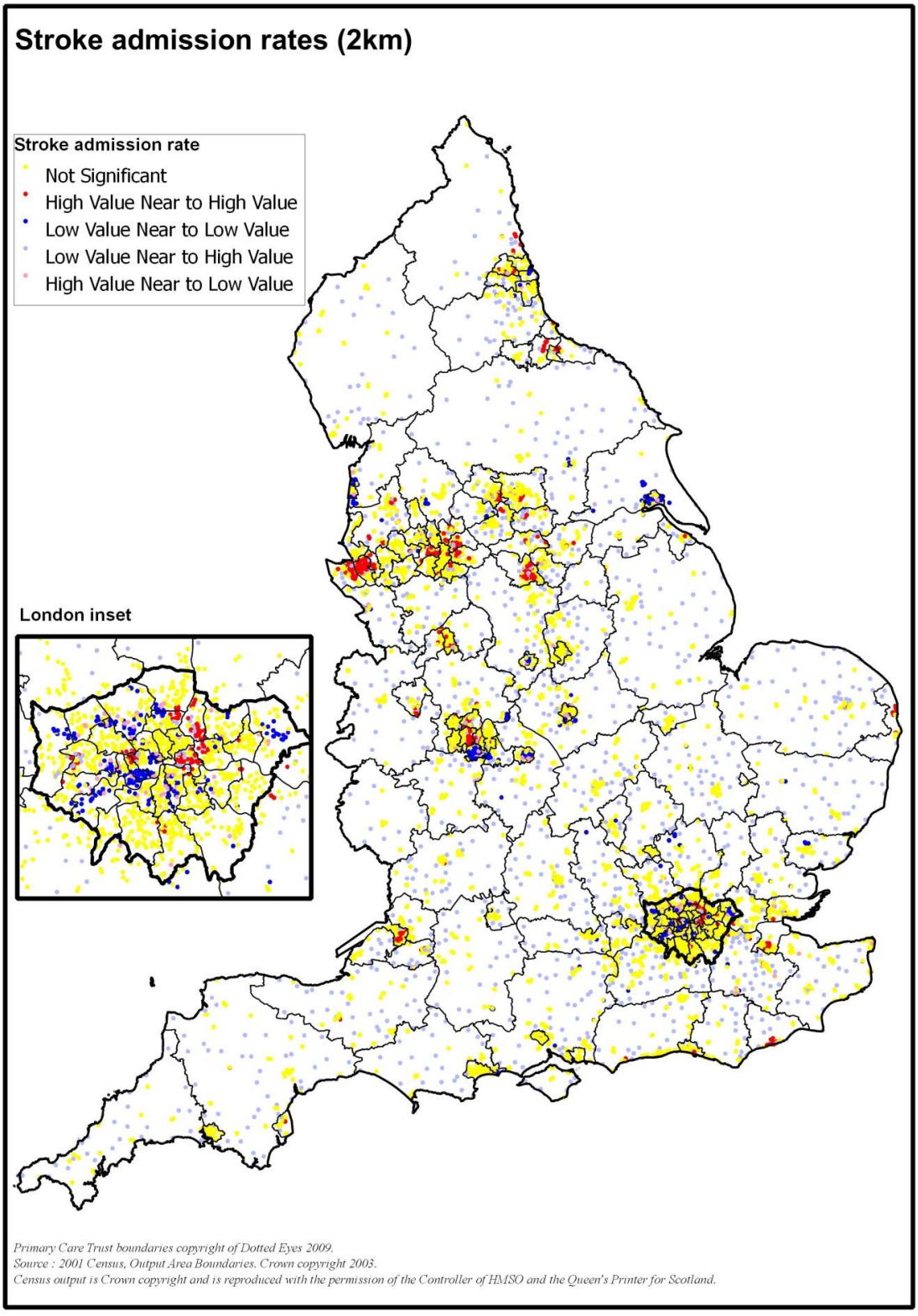


Figure 35: Local Moran's I (5km) of stroke admission rates

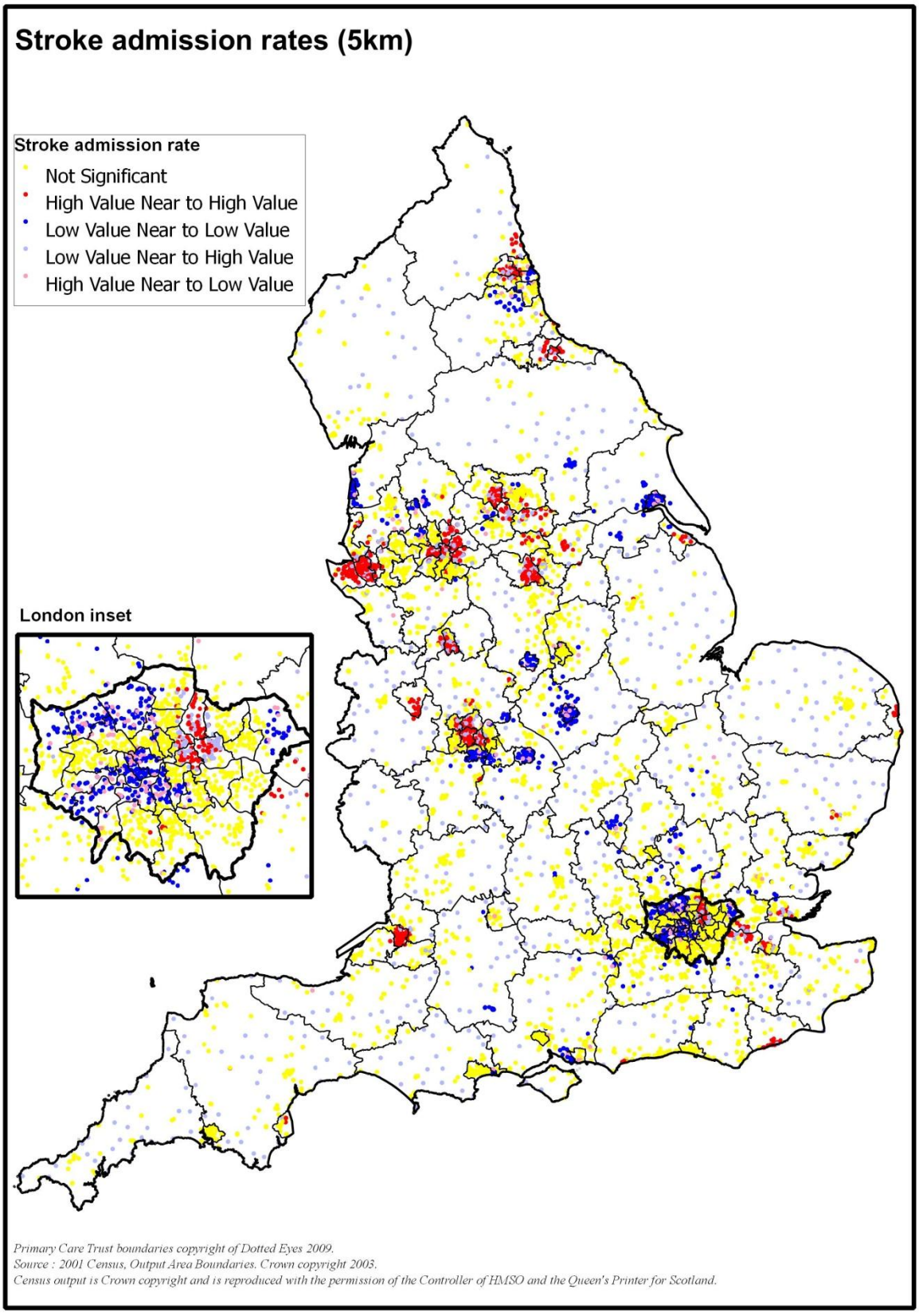


Table 84 compares OLS and geographical analyses for stroke admissions, including all variables except PCT CQC scores. Comparison with classical regression outcomes was made using the Akaike Information Criterion (AIC)[202] to test for spatial non-stationarity.[276] Identification of the optimal bandwidth for each analysis block was made using Monte Carlo simulations. The results of the spatial non-stationarity test reveal little difference between the AIC values for the non-spatial and GWR models, with marginal improvement with the spatial model. For example AIC values range from 2024 to 9899 between analysis blocks for the non-spatial model; the comparable range for the GWR model was 2024 to 9897. Thus while non-stationarity was not significant, nevertheless the AIC values confirmed that it was statistically valid to use spatial regression as a basis for exploratory mapping.

Table 84: regression diagnostics for stroke admissions- all variables (PCT Scores not included)

Sub regions	Residual Sum of Squares	Number of para-meters	Sigma	AIC	Coefficient of Deter-mination	Sum of Squares	Degrees of Freedom	Adjusted r-square
	Classical Regression							
East Midlands	794350.19	13.00	31.49	7941.58	0.13	794350.20	13.00	0.12
East of England	233667.98	13.00	27.41	3081.05	0.22	233668.00	13.00	0.19
Greater Manchester	600900.79	13.00	41.98	3666.51	0.17	600900.80	13.00	0.14
Home Counties	1084418.71	13.00	33.06	9899.22	0.14	1084418.70	13.00	0.13
Merseyside	414106.86	14.00	39.16	2906.66	0.24	414106.90	14.00	0.20
North East England	573430.79	13.00	39.69	3861.37	0.13	573430.80	13.00	0.10
North East London	831133.98	14.00	48.18	3955.77	0.10	831134.00	14.00	0.06
North West England	597069.82	13.00	33.31	5442.87	0.11	597069.80	13.00	0.09
North West London	852586.67	13.00	38.08	6096.00	0.14	852586.70	13.00	0.12
South Coast	396893.62	13.00	27.55	5091.42	0.19	396893.60	13.00	0.17
South East London	332856.05	14.00	35.44	2800.10	0.15	332856.10	14.00	0.11
South West England	491973.60	13.00	28.83	5799.70	0.13	491973.60	13.00	0.11
South West London	239149.68	13.00	35.67	2024.06	0.04	239149.70	13.00	-0.03
West Midlands	1459898.48	14.00	39.75	9586.87	0.12	1459898.50	14.00	0.10
Yorks& Humberside	1051371.77	13.00	38.37	7381.87	0.22	1051371.80	13.00	0.20
	GWR Estimation							
East Midlands	698472.52	45.81	30.15	7907.82	0.24	95877.70	32.81	0.19
East of England	208918.95	28.72	26.60	3081.08	0.30	24749.00	15.72	0.24
Greater Manchester	478405.44	42.50	39.19	3656.06	0.34	122495.30	29.50	0.25
Home Counties	1065104.83	20.76	32.90	9897.26	0.15	19313.90	7.76	0.14

Sub regions	Residual Sum of Squares	Number of para-meters	Sigma	AIC	Coefficient of Deter-mination	Sum of Squares	Degrees of Freedom	Adjusted r-square
Merseyside	374593.21	25.90	38.10	2906.06	0.31	39513.70	11.90	0.24
North East England	219164.36	119.79	29.19	3826.46	0.67	354266.40	106.79	0.51
North East London	782636.84	23.74	47.41	3955.21	0.15	48497.10	9.74	0.09
North West England	509909.11	47.84	31.83	5434.53	0.24	87160.70	34.84	0.17
North West London	832693.22	22.26	37.93	6101.56	0.16	19893.40	9.26	0.13
South Coast	348797.59	32.25	26.31	5064.41	0.29	48096.00	19.25	0.24
South East London	310591.12	24.12	34.91	2804.39	0.21	22264.90	10.12	0.13
South West England	426437.00	47.38	27.65	5789.87	0.25	65536.60	34.38	0.18
South West London	210289.72	23.59	34.42	2024.31	0.16	28860.00	10.59	0.04
West Midlands	1186389.38	75.62	37.09	9528.81	0.28	273509.10	61.62	0.22
Yorks & Humberside	974670.19	23.58	37.22	7349.17	0.27	76701.60	10.58	0.25

Table 85 summarises AIC scores for remaining stroke admission analyses. Looking down the columns, the results of the spatial non-stationarity test revealed little difference between the maximum and minimum AIC values (highlighted in yellow) for the non-spatial and GWR models, with marginal improvement with the spatial model.

Table 85: Stroke admissions: split variable Akaike Information Criterion scores

Sub regions	Service Delivery Variables (PCT Scores not included)	Service Delivery Variables (PCT Scores included)	Demographic Variables
	Classical regression		
East Midlands	8007.91	7657.21	7958.83
East of England	3120.28	3120.59	3072.83
Greater Manchester	3697.11	3695.51	3667.71
Home Counties	10016.27	10018.30	9897.83
Merseyside	2964.94	2953.66	2889.12
North East England	3887.69	3416.61	3870.43
North East London	3963.55	3963.55	3953.08
North West England	5484.68	5486.48	5443.39
North West London	6173.01	6172.90	6100.53
South Coast	5168.44	5167.20	5092.57
South East London	2798.94	2798.94	2816.06
South West England	5844.26	5646.88	5814.22
South West London	2018.14	2020.14	2010.45
West Midlands	9645.68	9345.65	9607.95

Sub regions	Service Delivery Variables (PCT Scores not included)	Service Delivery Variables (PCT Scores included)	Demographic Variables
Yorks & Humberside	7447.38	7449.41	9620.35
	GWR Estimation		
East Midlands	7950.02	7618.49	7876.03
East of England	3107.67	3099.55	3068.79
Greater Manchester	3655.69	3654.28	3619.75
Home Counties	9985.54	9990.27	9888.13
Merseyside	2954.32	2952.80	2882.15
North East England	3822.19	3374.74	3798.02
North East London	3955.60	3955.60	3944.95
North West England	5462.33	5454.94	5408.82
North West London	6177.15	6177.83	6099.34
South Coast	5150.98	5091.03	5055.81
South East London	2799.94	2799.94	2819.65
South West England	5828.68	5626.68	5758.97
South West London	2015.22	2017.51	2006.66
West Midlands	9609.42	9291.41	9467.91
Yorks & Humberside	7404.12	7401.65	9479.87

The following figures show a small sample of the the GWR analysis carried out for CQC, which covered all the English sub-regions we defined pragmatically for mapping purposes- specifically, local R-squared results- to demonstrate practice-level mapping and the utility of GWR regression diagnostics. We realised that there was little point in having a London inset when there was no London data, and that the inset could instead be used to show a higher detail map of practices. We have therefore shown three sub-regions- Greater Manchester, Home Counties and Merseyside- with an inset for the sub-region itself.

Figure 36 and Figure 37 show local R-squared for stroke admission rates for **Greater Manchester**, considering demographic and service delivery variables separately. There is broad agreement in local R-squared values between the population and service delivery maps, though not with the consistency observed for some other metropolitan contexts such as South West London.

Figure 36: local R-squared for stroke admission rates, Greater Manchester, population factors only

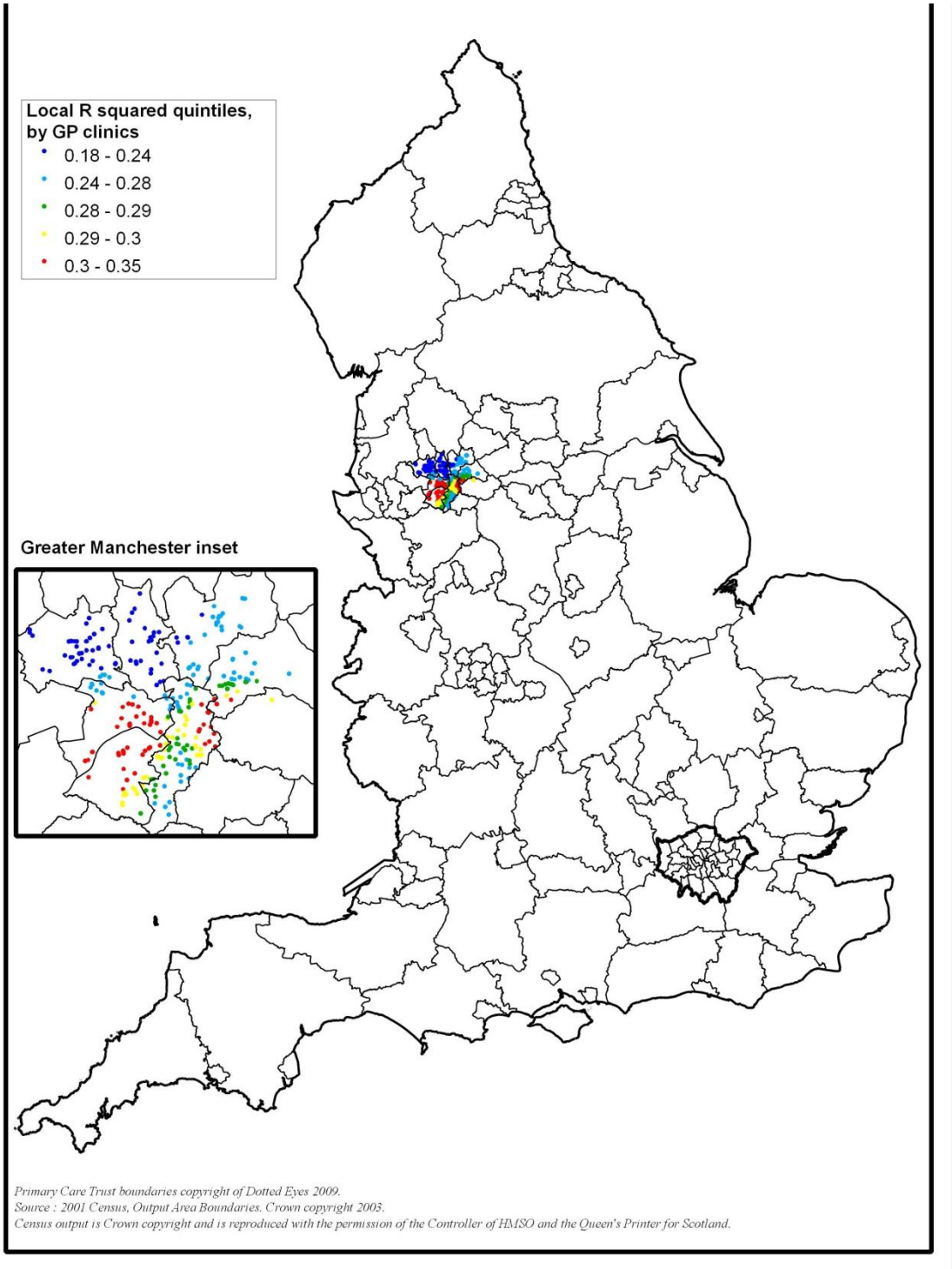
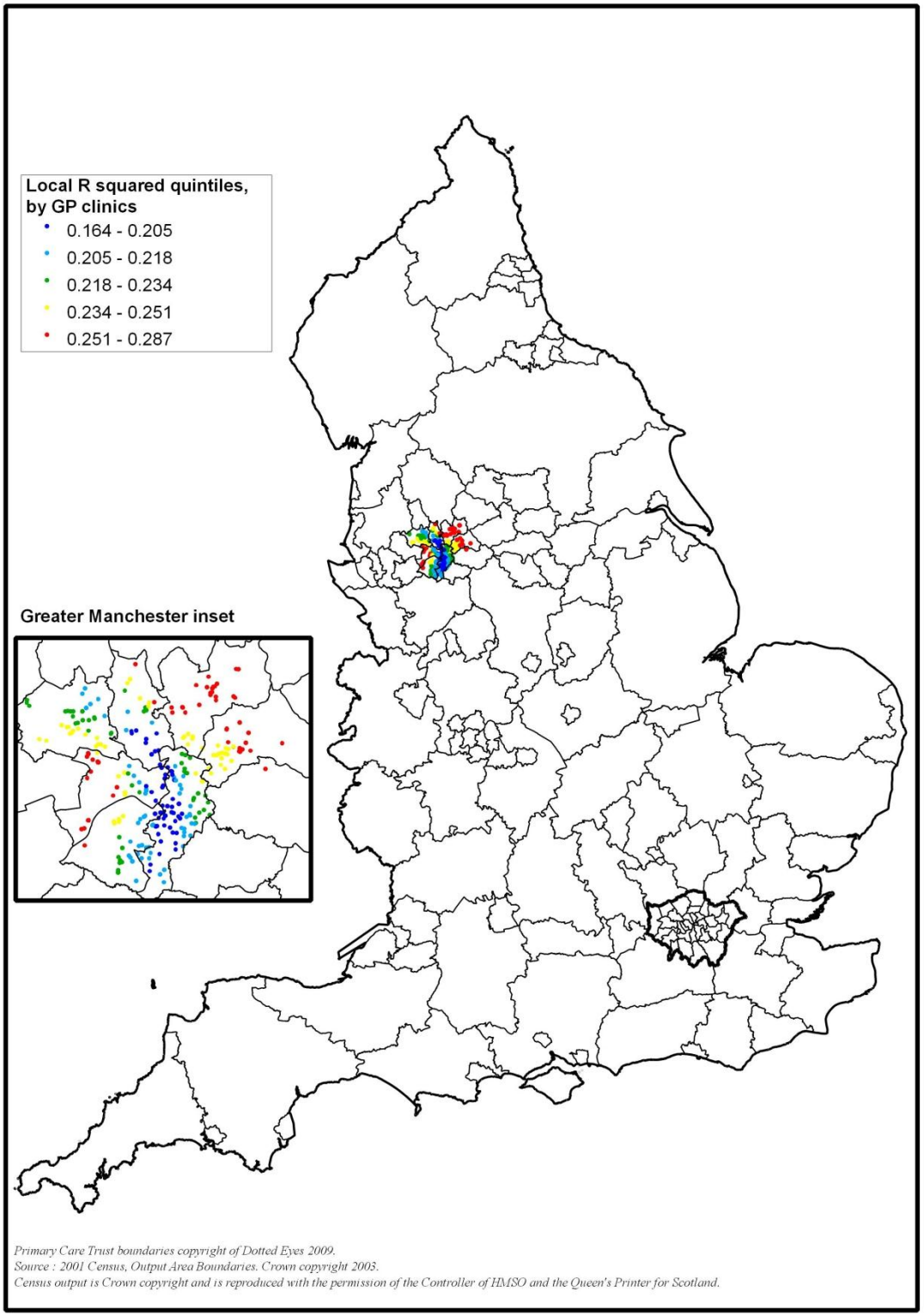


Figure 37: local R-squared for stroke admission rates, Greater Manchester, service delivery variables only



F

Figure 38 and Figure 39 show local R-squared for stroke admission rates for **Merseyside**, using demographic and service delivery variables. Demographic factors associated more strongly with admissions than did care-related factors suggesting that practices in this region are managing populations better than might be expected considering population context.

Figure 38: local R-squared for stroke admission rates, Merseyside, population factors only

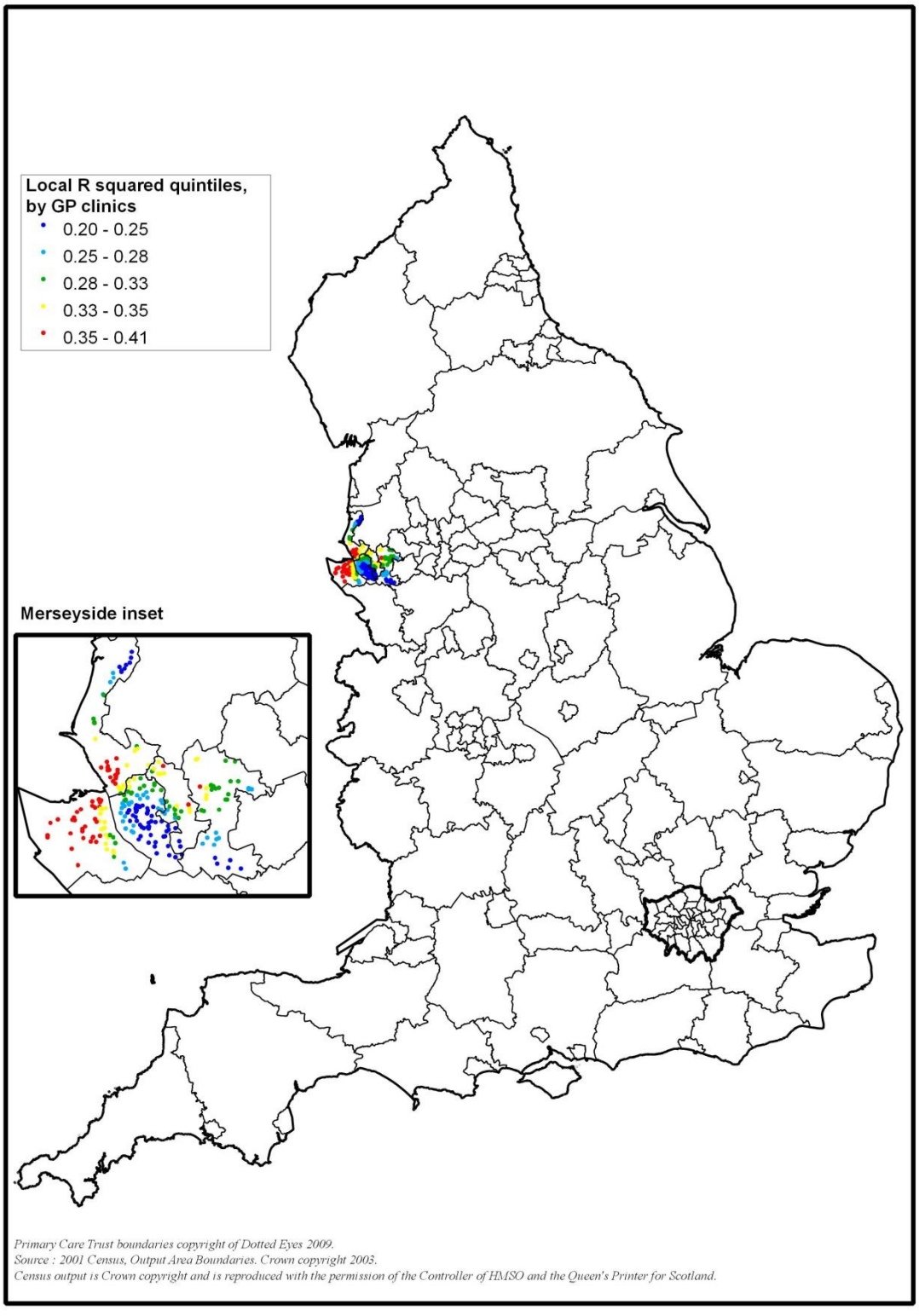


Figure 39: local R-squared for stroke admission rates, Merseyside, service delivery variables only

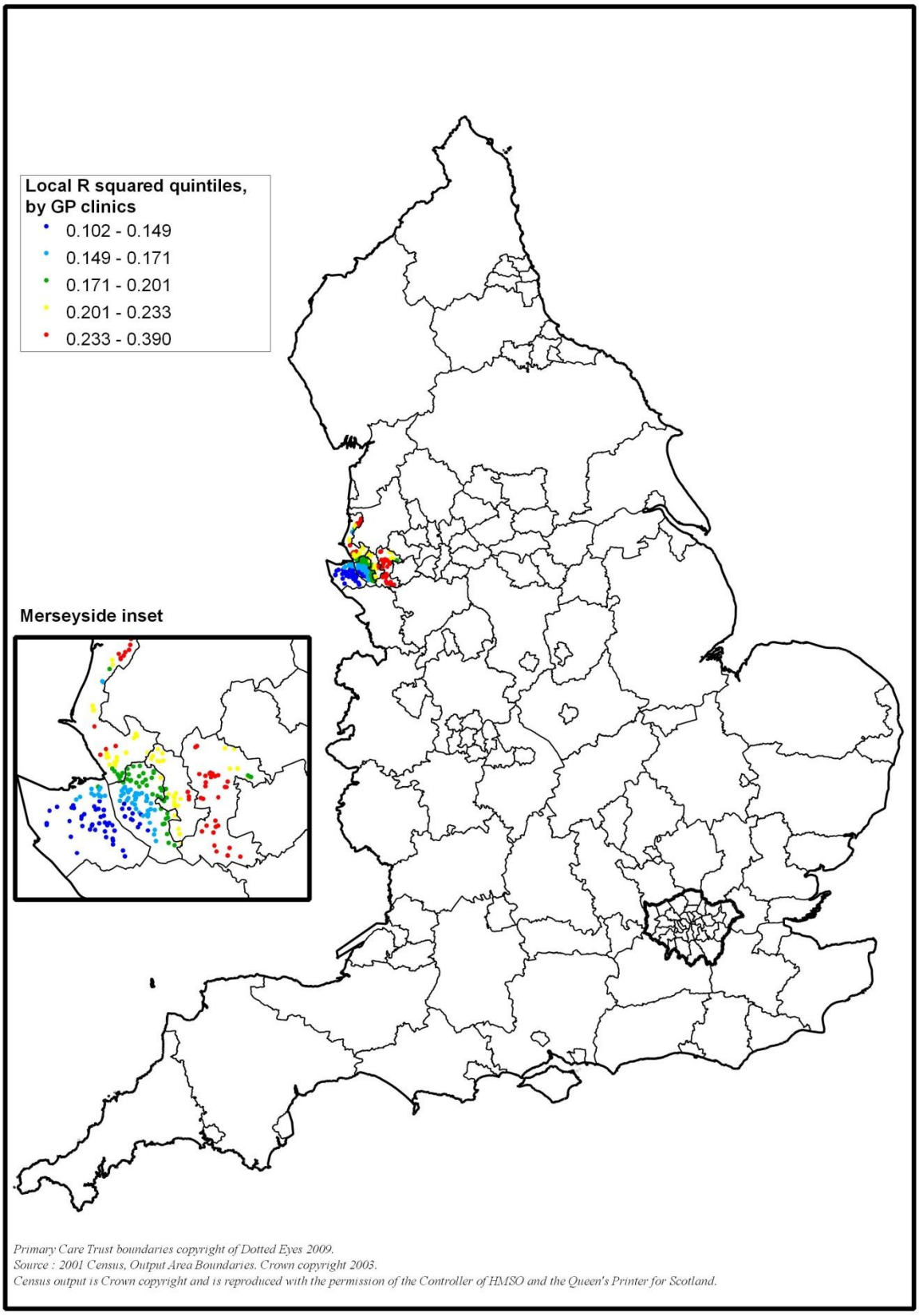
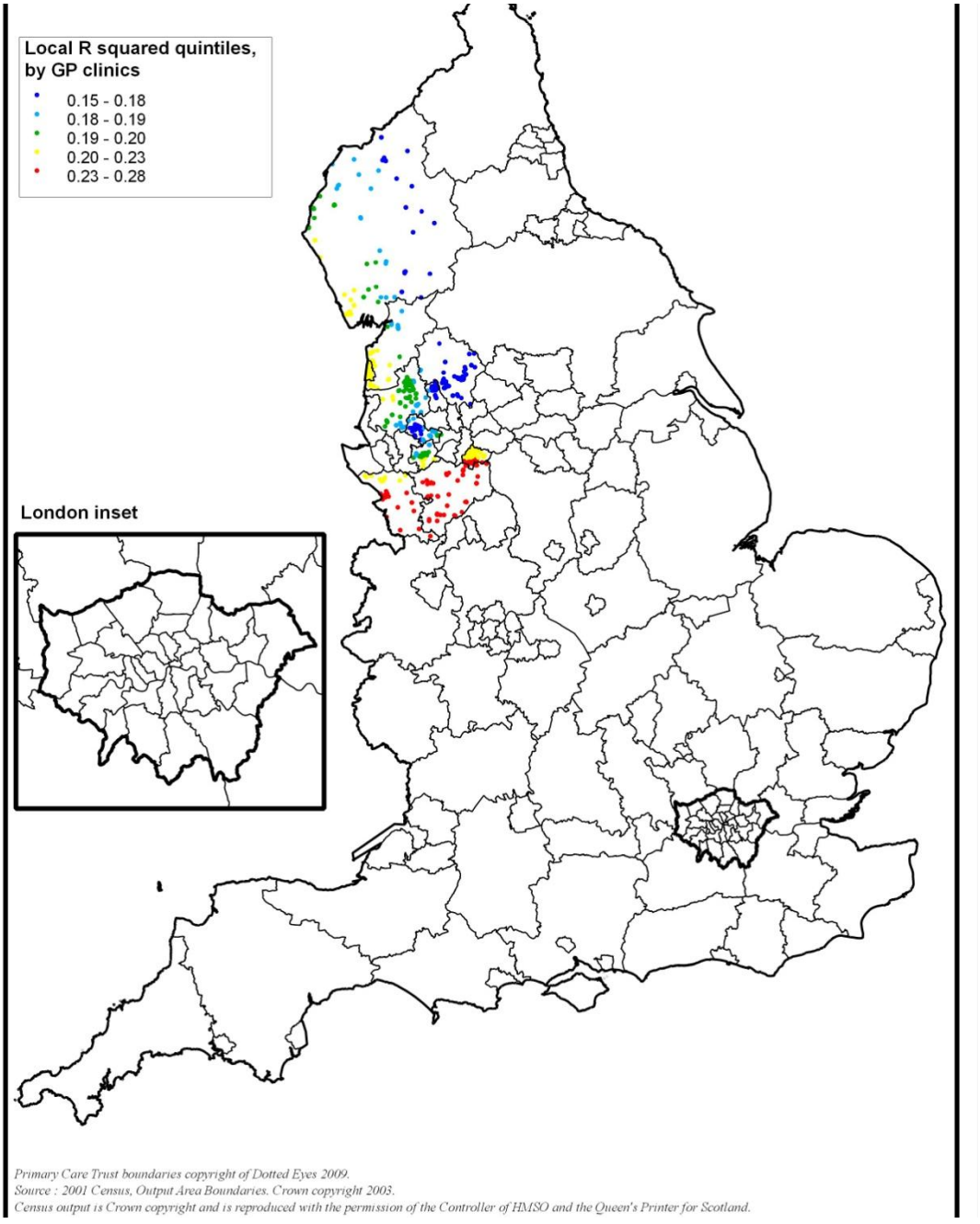


Figure 40 and Figure 41 show local R-squared for stroke admission rates for North West England, using demographic and service delivery variables. There is a clear reversal in the directionality of the geographic trend in local R-squared values between the split-variable models across the region, with the socio-demographic model showing a general south to north statistical incline and the institutional model the reverse. There was wider spatial variability in the predictive power of the service delivery model than for the demographic model.

Figure 40: local R-squared for stroke admission rates, NW England, population factors only



Within the more northerly, more rural PCTs there are associations of practices for whom service delivery factors predict admissions more strongly than population factors: these

associations do not appear to cross boundaries into the more urban PCTs in the centre of the region. In the southernmost, more affluent PCTs of the study area the local R-squared values are similar for both models.

Figure 41: local R-squared for stroke admission rates, NW England, service delivery variables only

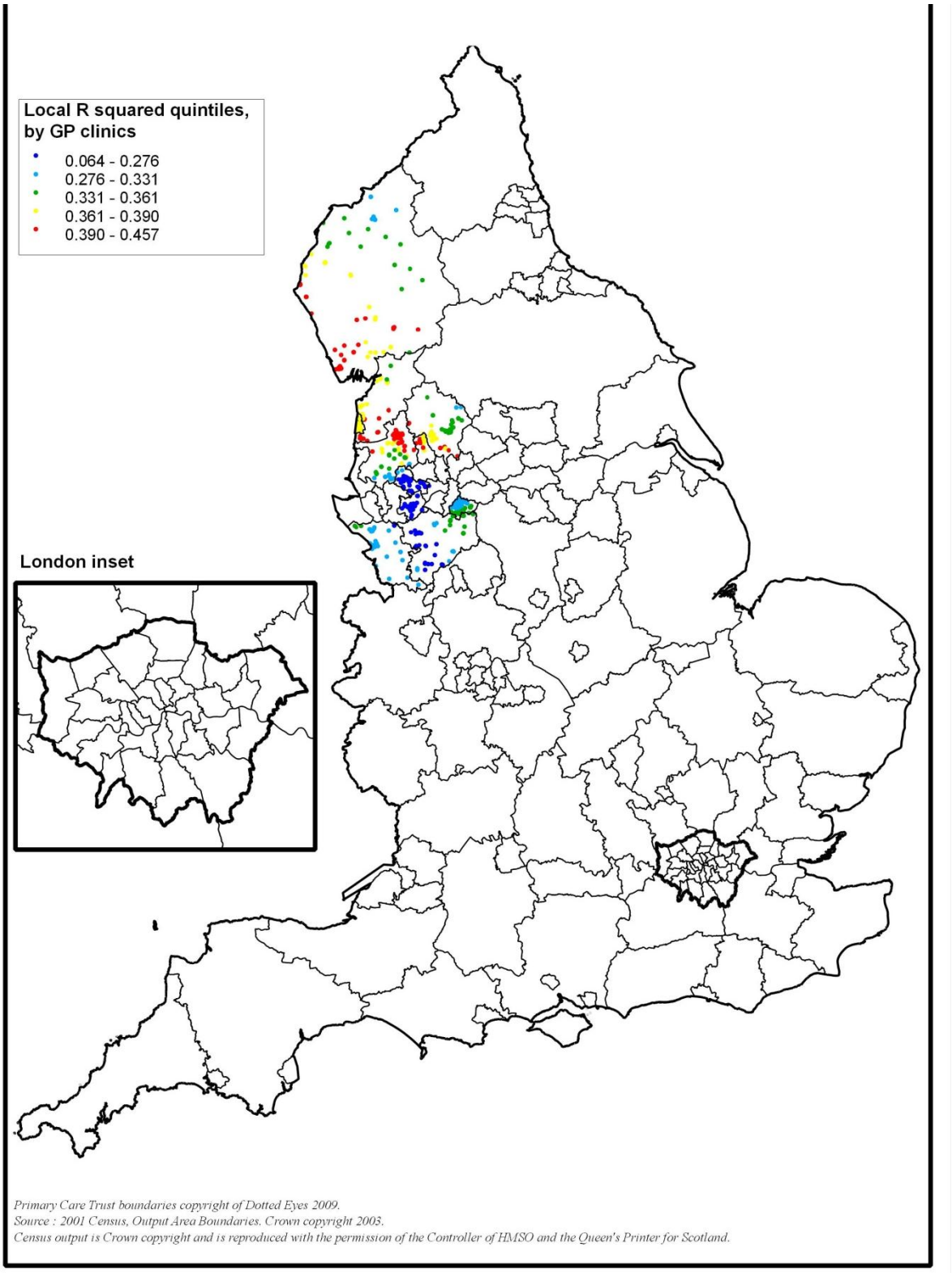
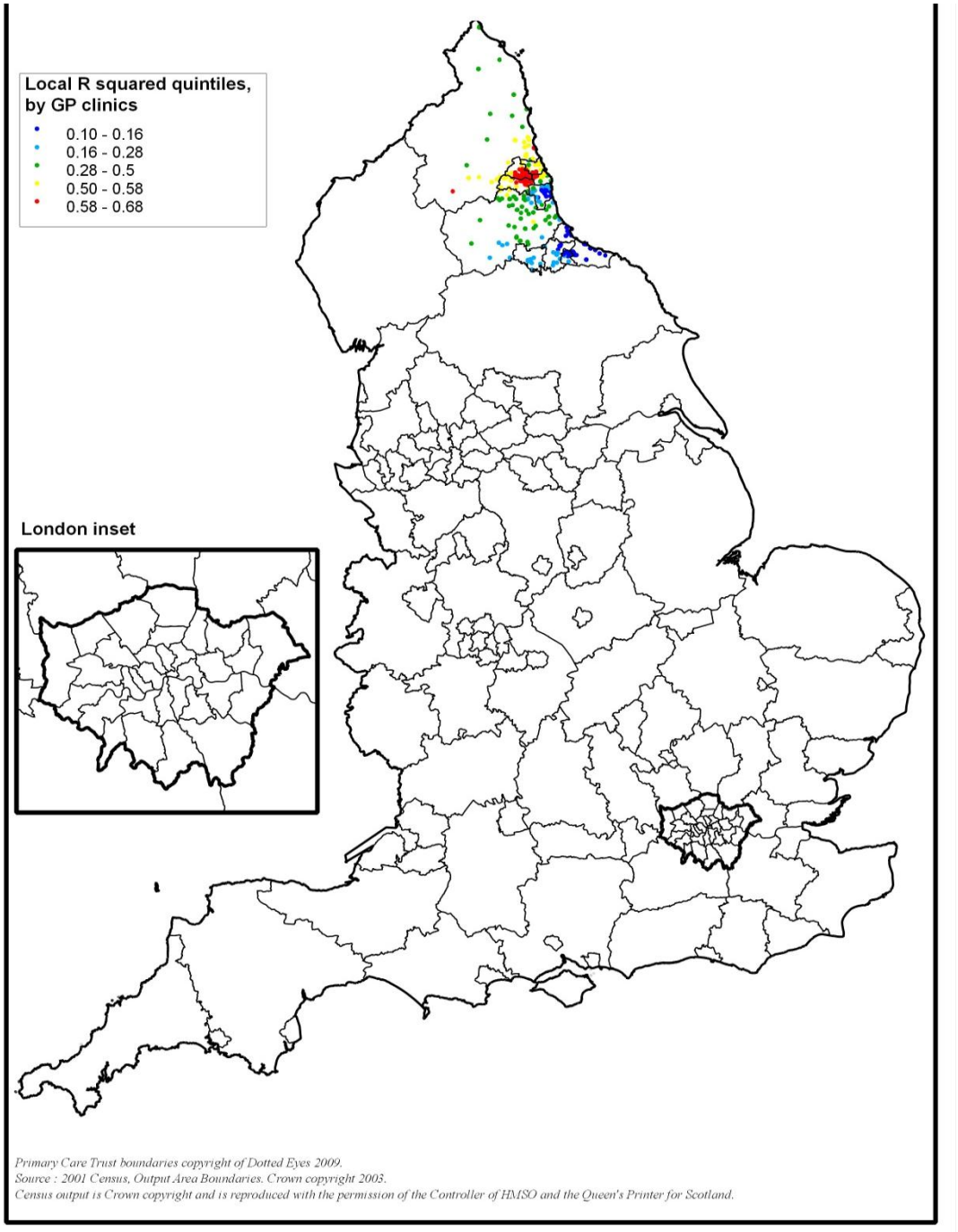


Figure 42 and Figure 43 show R-squared for stroke admission rates for North East England, using demographic and service delivery variables. The highest correlations between admissions and population factors are observed in parts of the Tyne and Wear locality of this area (up to 0.68 for one association of practices) and much stronger than institutional factors, indicating deep underlying socio-demographic problems and little local population diversity.

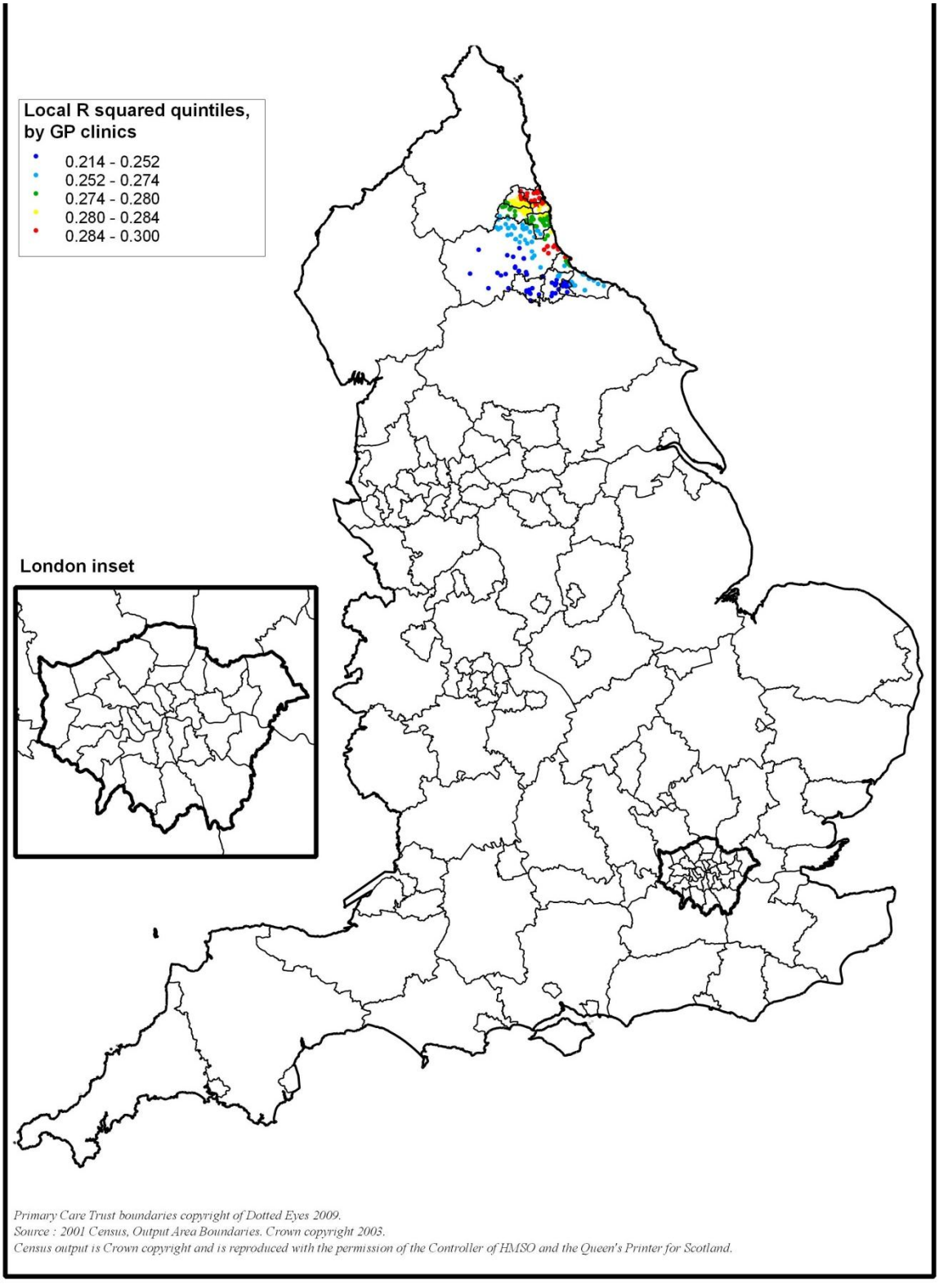
Figure 42: local R-squared for stroke admission rates, NE England, population factors only



This corresponds to known health promotion problems in the area relating to lifestyles.[218] However, by contrast the South of Tyne and Wear PCTs (e.g. Sunderland) indicate admission

correlations with primary care factors over and above that predicted by socio-demographic factors, whereas the values are low for Newcastle.

Figure 43: local R-squared for stroke admission rates, NE England, service delivery variables only



In the **West Midlands** (Figure 44 and Figure 45), as with Greater Manchester, general consistency was observed in the geographic patterning of local R-squared values between the two models. Socio-demographic factors were stronger predictors of admissions than institutional variables.

Figure 44: local R-squared for stroke admission rates, West Midlands, population factors only

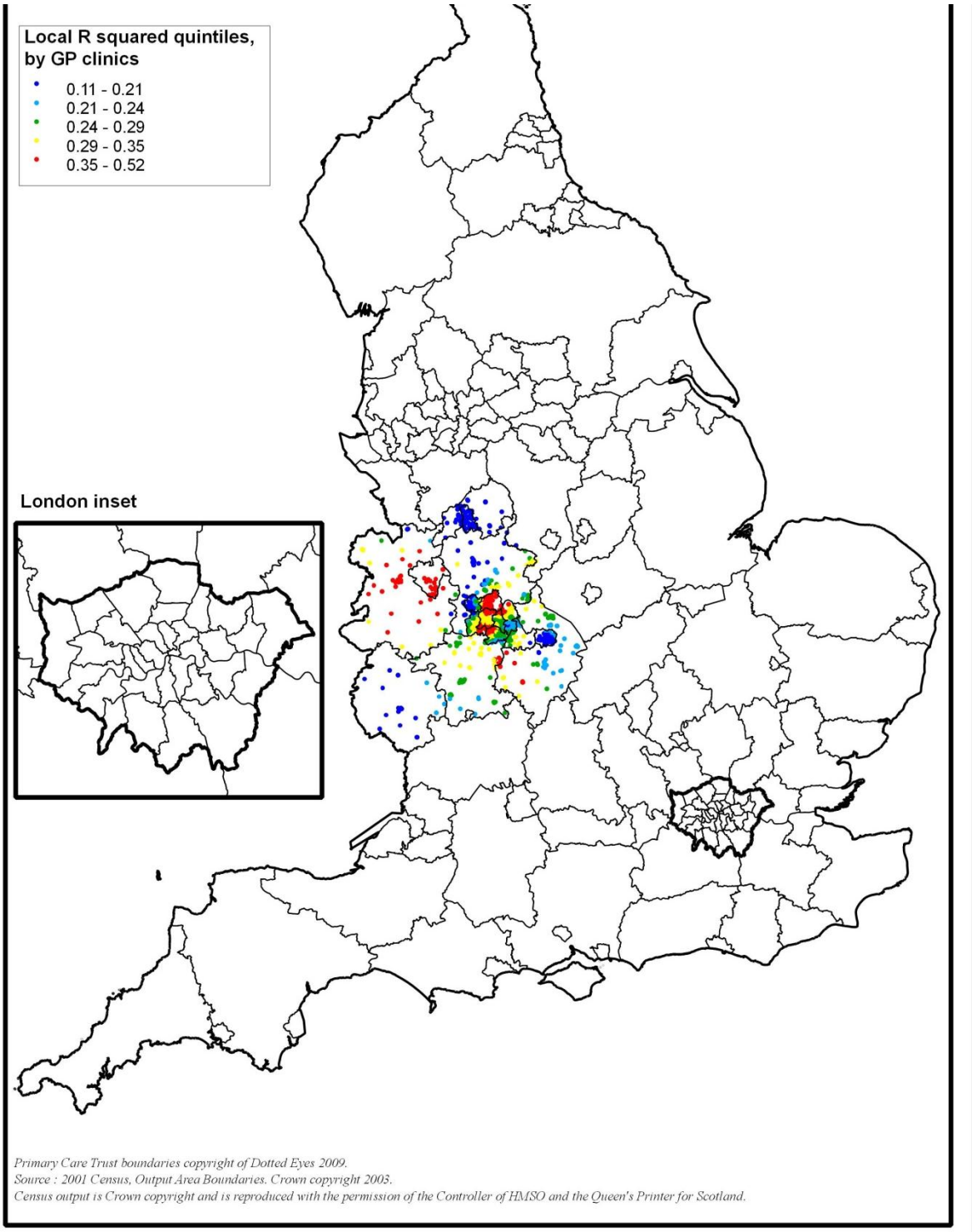
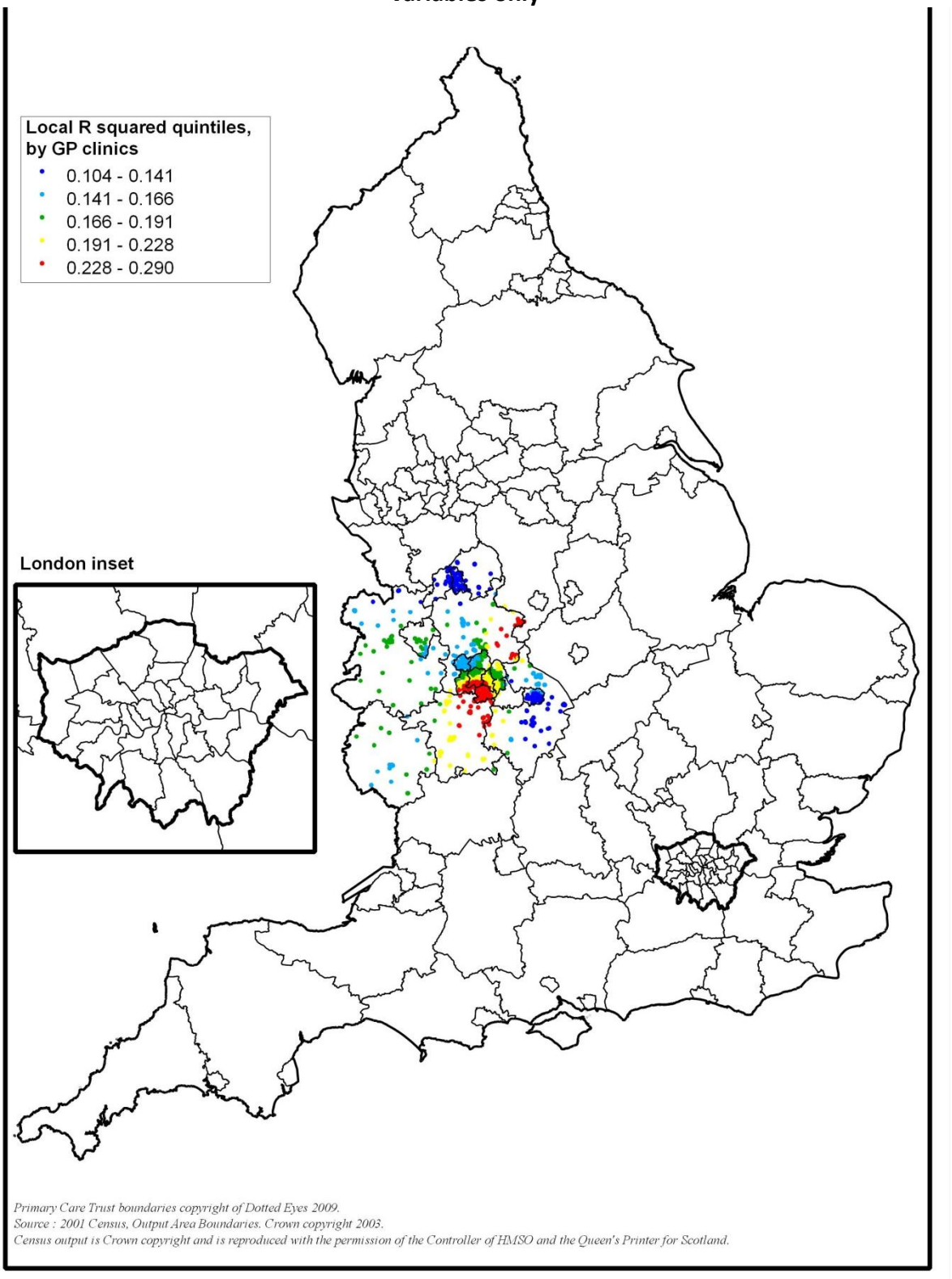
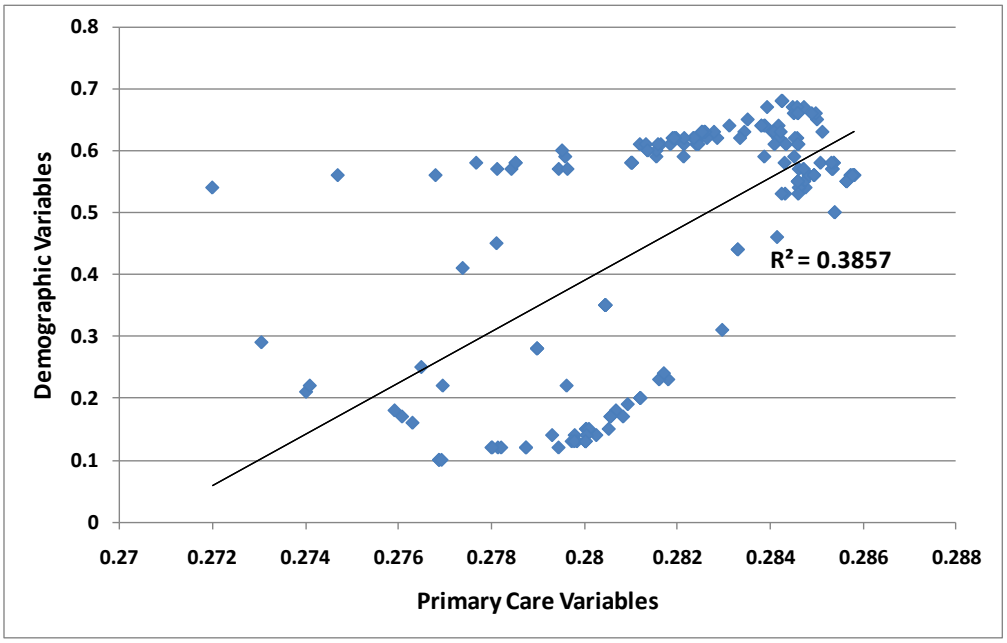


Figure 45: local R-squared for stroke admission rates, West Midlands, service delivery variables only



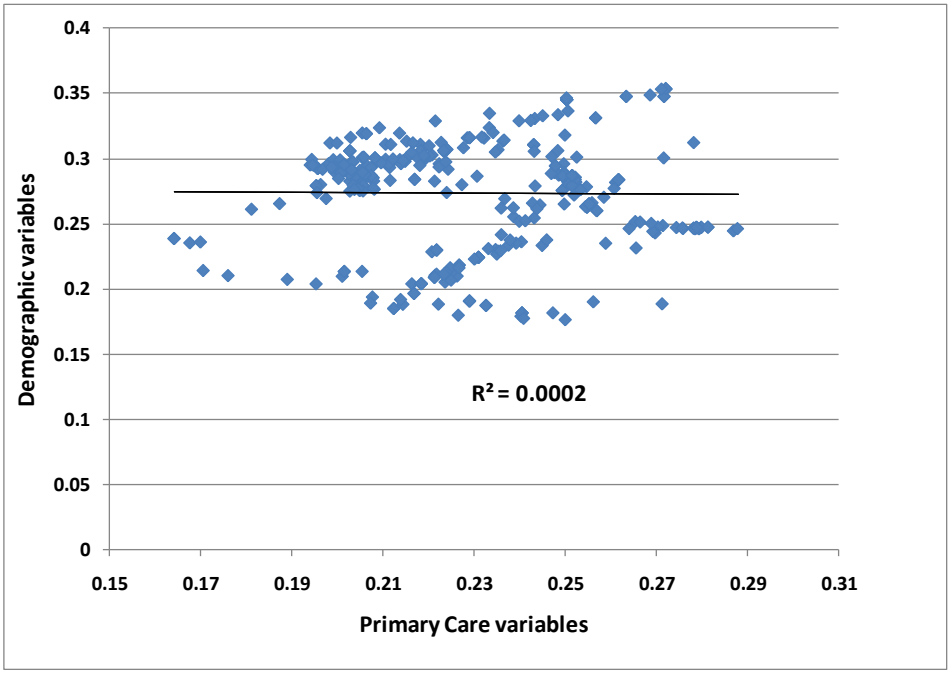
We explored these relationships further using scatterplots of R squared values for practices in two sub-regions. The first chart (Figure 46) shows a scatterplot of R squared demographic vs primary healthcare variables for Tyne & Wear practices. There is a fairly strong correlation between these in this area.

Figure 46: scatterplot of demographic vs primary healthcare R squareds for Tyne & Wear practices



The second chart (Figure 47) shows a scatterplot of demographic vs primary healthcare variables for Greater Manchester practices. There is no correlation whatsoever, so apparently different relationships hold for these variables between the two areas.

Figure 47: scatterplot of demographic vs primary healthcare R squareds for Greater Manchester practices



In North West London (Figure 48 and Figure 49) whilst population factors are stronger predictors of admission rates than service-related factors, there is nevertheless consistent

anisotropy (directionality in a pattern) across the area without the suggestion of spatial anomalies. The decline in R squared declines from the West, whereas primary care factors do not correlate at all, with very low R squareds.

Figure 48: local R-squared for stroke admission rates, NW London, population factors only

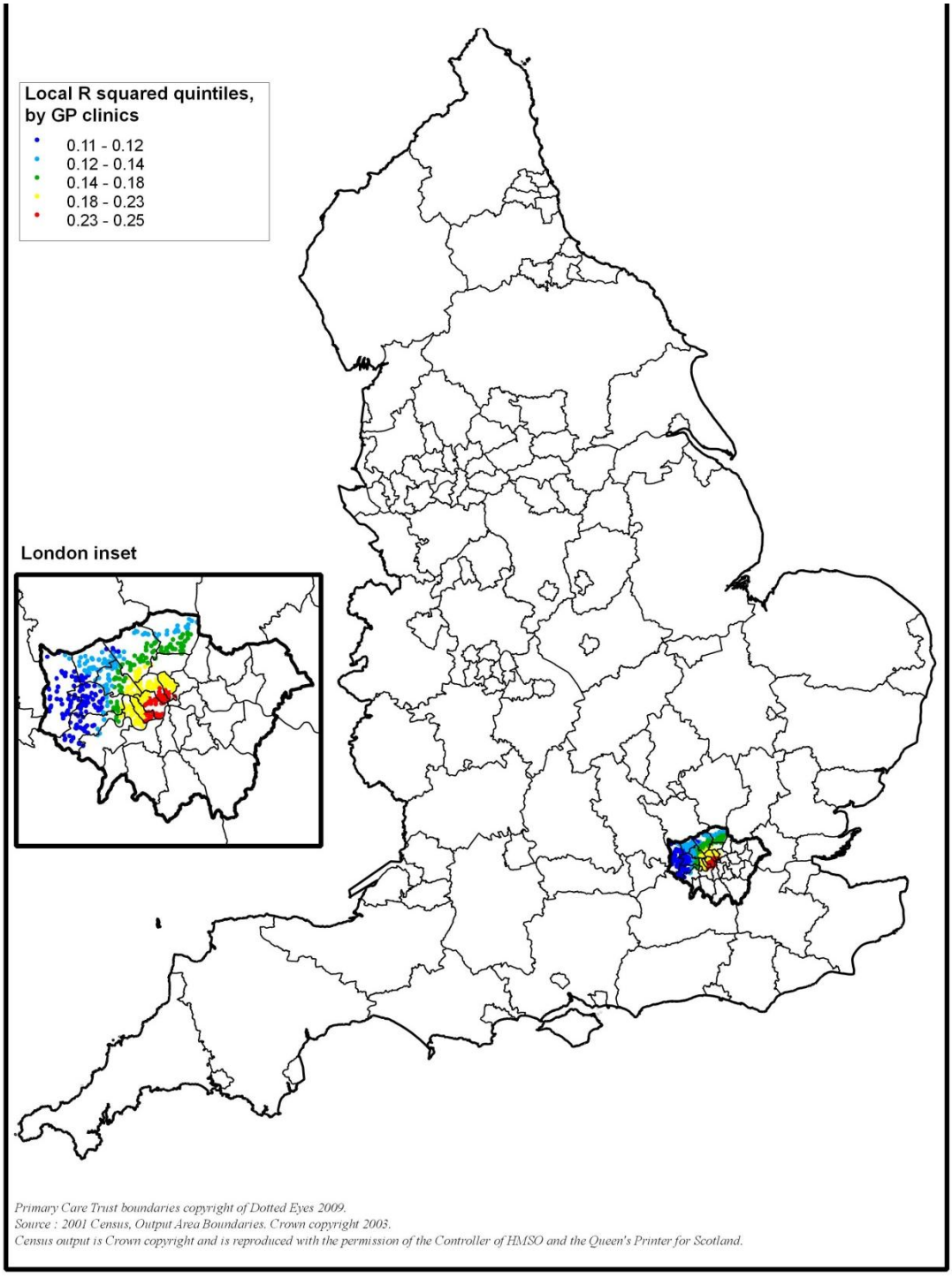
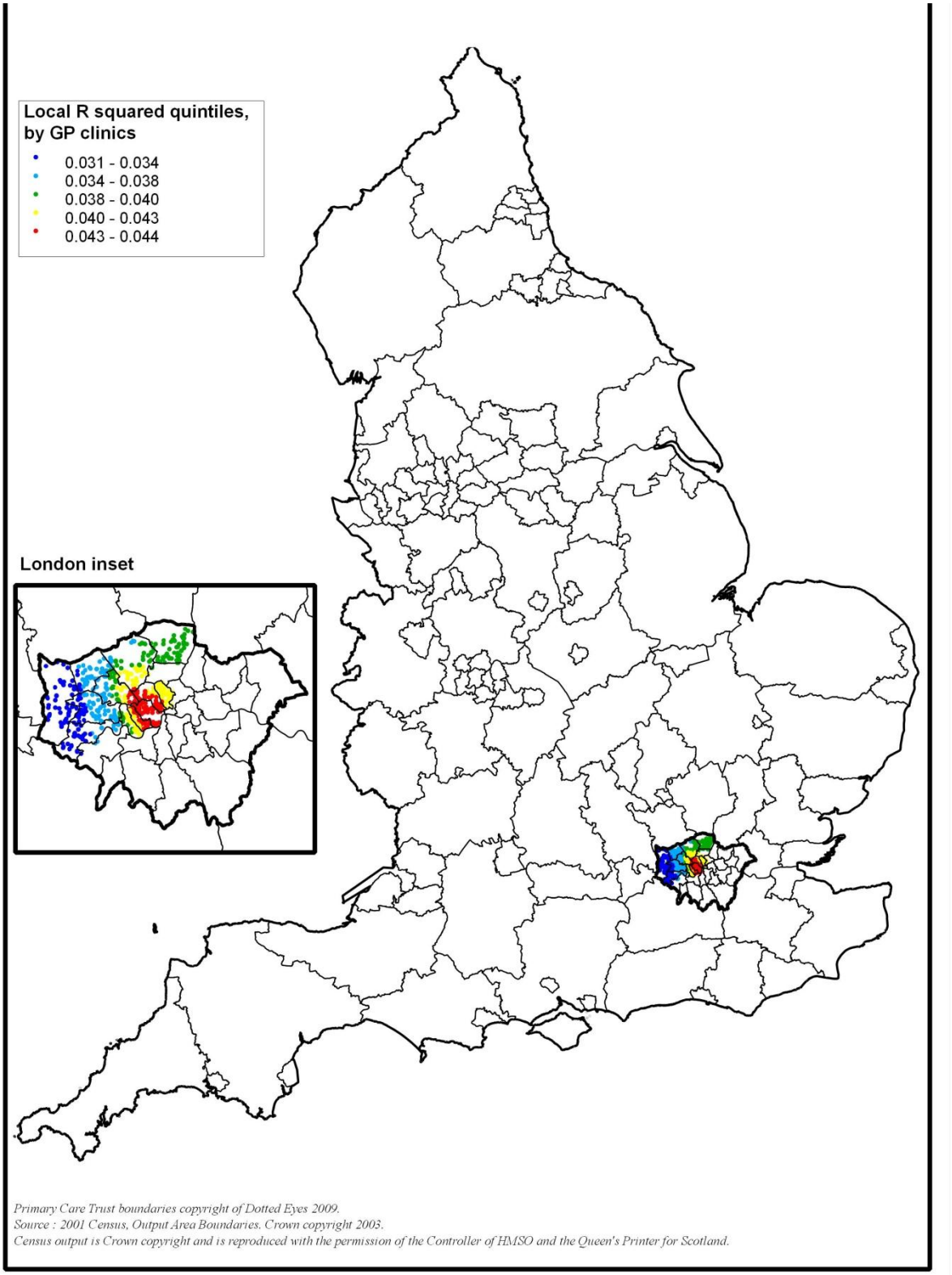


Figure 49: local R-squared for stroke admission rates, NW London, service delivery variables only



In North East London (Figure 50 and Figure 51) the same scenario is evident for Hackney and Tower Hamlets.

Figure 50: local R-squared for stroke admission rates, NE London, population factors only

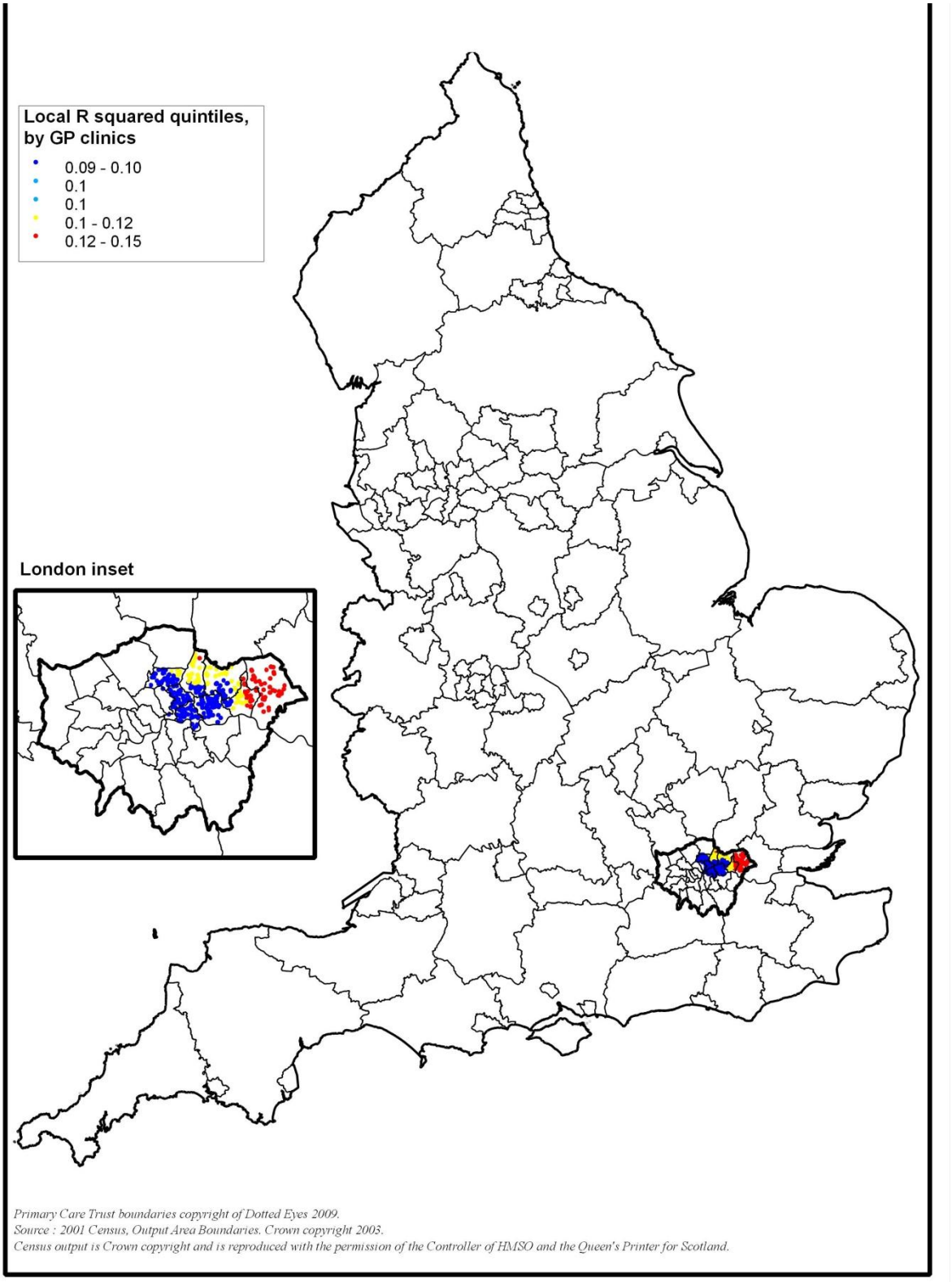
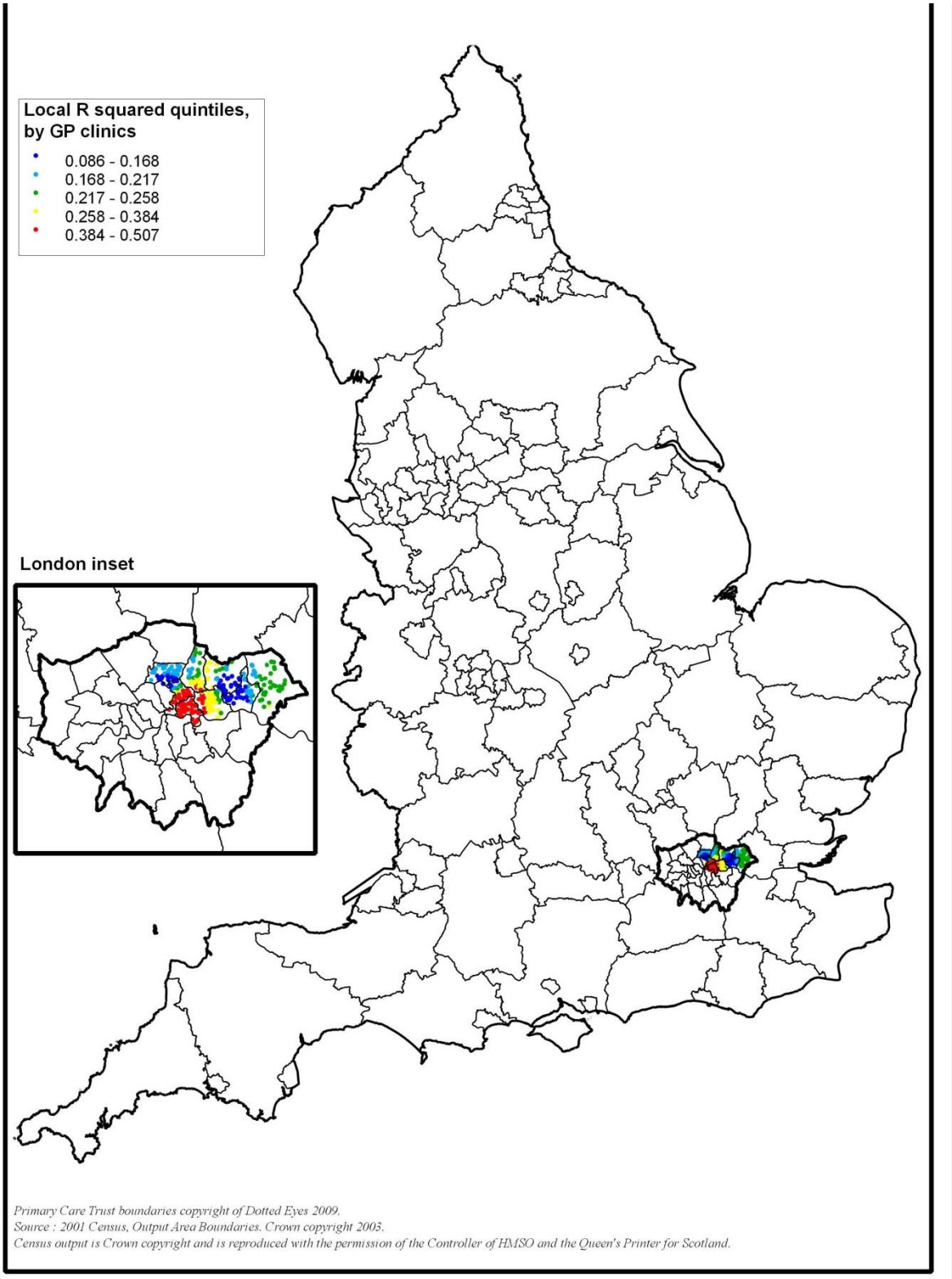


Figure 51: local R-squared for stroke admission rates, NE London, service delivery variables only



Some significant associations were observed in **South Coast** (Figure 52 and Figure 53). The maps show remarkably strong agreement between the institutional and population models in terms of directionality, though greater variability in R-squared values can be seen in the healthcare-related model map. There is no suggestion of performance anomalies in this analysis block.

Figure 52: local R-squared for stroke admission rates, South Coast, population factors only

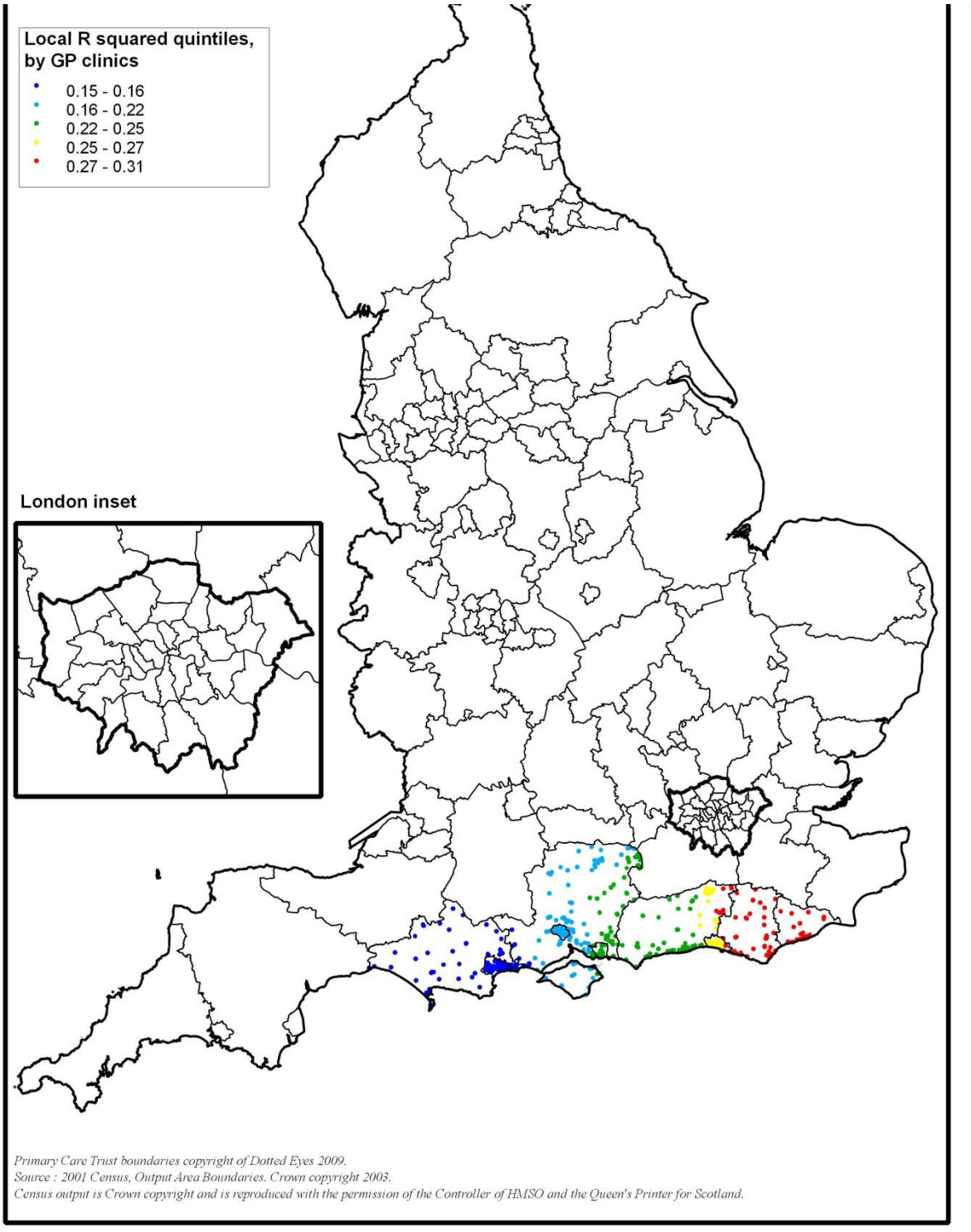
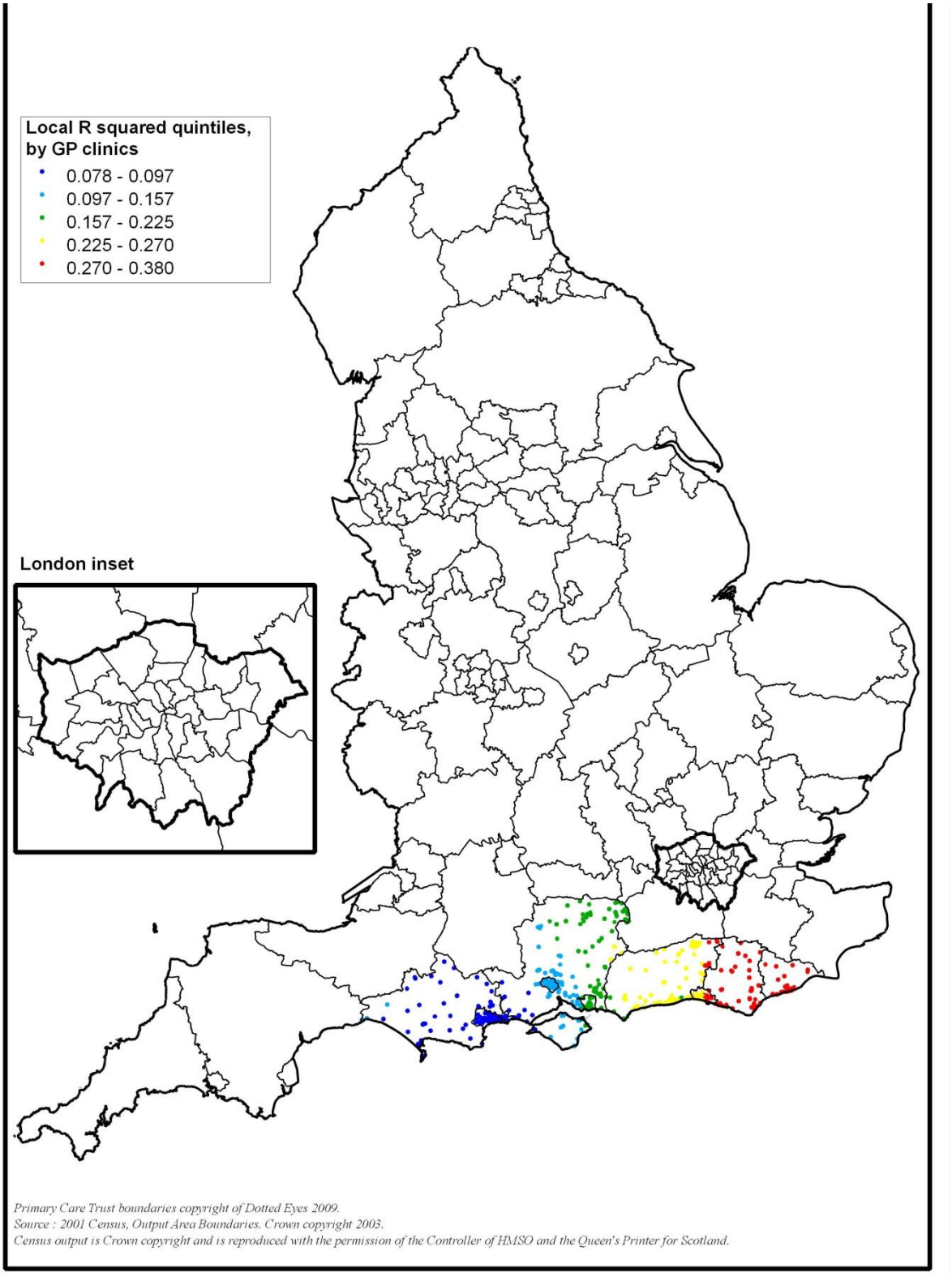


Figure 53: local R-squared for stroke admission rates, South Coast, service delivery variables only



General patterns

There is little evidence that urbanity/rurality consistently correlates with the relationship between primary care variables and admissions, though there is evidence that more affluent with the exception of the Home Counties and locality-specific contexts appear to be relevant. Deprivation has an inconsistent relationship with the effectiveness of care delivery regarding admissions with suggested variation in primary care achievement between Greater Manchester, Merseyside and central Tyne and Wear on the one hand, and inner east London and south Tyne and Wear on the other.

10.7.5 Discussion & Conclusions

Scenario 1/ disease detection: outcome=registered and undiagnosed prevalence

An obvious initial question is: why was there such a discrepancy between expected and practice-registered disease prevalence, when the model-based CHD and stroke prevalence estimates were based on patient reports of doctor-diagnosed disease? Reasons are likely to include inadequate searching of practice records (which are now mainly electronic) for previous diagnoses or CVD-related prescriptions, lack of linkage of practice and hospital records (practices still generally enter codes for hospital admissions manually), and high population mobility in urban areas with concomitant changes (or delays) in registration with general practices. In the case of modelled hypertension (not included in this report), for which the modelled estimates used a combination of doctor-diagnosed disease and blood pressure measurements, the gap is greatest in younger and middle age groups, when males in particular are less likely to be seen by their practices. In contrast, diagnosis appears to be more complete in the north of England.

For the Scenario 1 stroke and CHD OLS analysis, neither the registered nor the unregistered prevalence were standardised by age/sex. For registered prevalence, we applied a Poisson regression where we included registered count of cases as the dependent variable and the practice population as the offset. For unregistered prevalence, we applied a Poisson regression where we used unregistered (estimated-registered) count of cases as the dependent variable and the practice population as the offset. This fails to take into account the age and sex structure of the practice population. There are two possible solutions:

1. for registered prevalence, an OLS regression introducing age/sex standardized prevalence (observed count as the dependent variable and the expected count the offset) could be run. For unregistered prevalence, the problem is less of an issue since the estimated prevalence from the APHO prevalence models includes age, sex and a number of other variables for estimation. The subtraction "estimated-registered" and the ratio "registered/estimated" are therefore comparable between practices
2. if we wish to see the “otherwise adjusted” effects of age and sex distribution of practices on the outcome variable, we would need to include the latter among the independent variables of the OLS regressions. We intend to follow up both these options in developing papers for publication.

The multivariate analysis with unregistered CHD count as the outcome variable, using the practice population as the offset variable in the Poisson regression, gave an IRR less than 1, which is unexpected. One reason would be the presence of multicollinearity, but only IMD

score and estimated smoking prevalence (which uses IMD score as an input) showed evidence of multicollinearity (Table 6). We cannot explain this anomaly.

From a spatial analytic perspective, there are significant geographic discrepancies between QOF prevalence and modelled prevalence for all three conditions in much of London and its hinterland. The inclusion of a measure of GP supply in the GWR analyses suggests that, despite needs-based healthcare resource allocation in the UK, persistent differences in availability of primary care services is an important limiting factor in diagnosis.[231] An analysis of Gini coefficients to measure geographical equity in GPs per 100.000 population in England and Scotland showed that equity in England rose between 1974 and 1994, but then decreased, and in 2006 it was below the 1974 level.

Finally, the Scenario 1 analyses presented here should be used to guide practical case-finding initiatives in practices. If they are, the results should be audited and the prevalence models validated. Birmingham University has recently received a Programme Grant from the National Institute for Health Research for COPD which has as one of its objectives COPD case-finding in practices and through that, validation of the COPD prevalence model and of practice smoking status data. Similar audits are needed for CHD and stroke. The NHS Health Checks programme provides an opportunity to do this, but only if practices ask patients about previous diagnoses of and admissions for CHD/stroke. This is not part of the current Health Checks protocol.

Scenario 2/disease management: outcome=admission rates

For the OLS analysis with MI+angina admission rate as the outcome variable, observed prevalence, IMD score, smoking prevalence and Asian population are as expected all risk factors for hospital admission- the IRR for the latter is very high. The IRR for unregistered prevalence is less than unity, which (as for Scenario 1) is unexpected, as patients with undiagnosed disease might be expected to be more likely to be admitted. The findings for stroke were broadly similar.

In comparison with a similar analysis on COPD admissions we have carried out recently, IRRs were generally smaller for CHD and stroke, and the results for COPD were more consistent with our prior hypotheses.[250] Further analyses could be undertaken with the current aggregate HES data, but the questions about the impact of unregistered CHD may not be answered until analyses of linked person-specific data are examined to determine how many patients who are admitted with second MIs or angina have not been on practice registers. We also have some concerns about the provenance of the CQC HES data, as we have been unable to ascertain how it was grouped into spells. There does not appear to be a consistent relationship between the spells or admissions in the CQC data and episode-based data from HESOnline (the HES team have also been unsure how the data was grouped). The CQC could also consider using a scatterplot of R squareds, with primary healthcare/population axes, and cropping as an alternative to remapping.

We are aware that York University and Professor Peter Smith in ICL Business School have recently undertaken an analysis for DH and the Health Foundation using person-specific data, which could be used to examine these additional questions. We also had some remaining questions about the processing of the HES data within CQC, as we have not been

able to determine how it was spell grouped, i.e. how FCEs were linked into spells. We would be able to provide the expertise to advise and audit this process.

Conclusions

Surveillance for care regulation, both for health and social well-being, can be aided by the routine use of mapping and spatial statistics, as this report demonstrates. Although the analysis did not demonstrate that GWR is definitely superior to OLS analysis in this situation, it is at least as useful, and in addition provides the opportunity to visualise and investigate relationships more easily. Performance-related practice associations identified by the use of GWR provides a basis for further investigation of potential 'overachievers' and 'underachievers'. Mapping of results enables visualisation readily interpretable by CQC statisticians, and with adaptation cartographic output can be understood by lay members of the public, a facet important in a Freedom of Information age.

The key questions for discussion are (a) how the CQC can implement GIS and spatial analysis, and (b) incorporate the relevant methods and technologies as routine parts of Information Policy and Research.

Perhaps a starting point should be GIS experiences in the NHS. Research carried out covering the period 1991-2001 indicates that whilst the use of GIS for mapping had become well-established during this period as a result of the general mainstreaming of spatial technologies throughout the economy, the take-up of GIS for analytic purposes remained very much a minority activity. In 1991 71.9% of Health Authorities used such systems and by 2001 the percentage had grown to 85.7%, accompanying a shift from cartographic software usage to fully-functional GIS packages. Nevertheless the figure for those organisations actually using such packages for statistical analysis amounted to only 33.5% even by 2001.

Even with the use of GIS now being pervasive across the NHS, its primary application remains one of core mapping as evidenced by the NHS SHAPE facility (http://shape.dh.gov.uk/public/mapping.asp), Public Health Observatory interactive mapping websites (e.g. http://www.nwpho.org.uk/healthgis/) and the custom package iQ HealthMaps (http://www.iq-medical.co.uk/products-healthmaps.php), alongside more traditional GIS proprietary software such as MapInfo and ESRI ArcGIS. This can be largely put down to a lack of spatial analysis knowledge within healthcare organisations. This is not an obstacle for the CQC, which could also encourage and disseminate spatial statistical understanding. The use of core mapping with a software package such as InstantAtlas or GapMinder is a good starting point, and will generally raise awareness amongst healthcare providers and the public. At the same time the CQC should be developing its spatial analytic capacity, and should be expecting NHS organisations to be doing the same.

Whilst the CQC does not have widespread GIS and spatial statistics knowledge, the organisation does have broad statistical expertise which can be used as a foundation for developing a geographical analysis capability. A lack of GIS knowledge in the organisation is not a fundamental issue as spatial statistical analysis can be performed in specialist statistical packages with GIS simply used for map production, and a MapInfo licence is already held by the organisation. The CQC uses R and SAS which could be used for spatial statistics; currently CQC uses general linear modelling, and it can then add spatial factors.

CQC should also consider GIS software- GeoDA have just brought out an R version which is a stepping stone to more specialised software. The main alternative is Luc Anselin's analysis blocks these may be suitable for some models.

Costs, frequently an inhibiting factor in implementing new information technology in organisations including GIS, is also not an obstacle, as the last five or so years have seen a boom in reliable freeware mapping packages with basic mapping functionality aimed at users who primarily need a basic mapping tool with the ability to convert commercial file formats from one to the other. Packages such as GeoVista (http://www.geovistastudio.psu.edu/jsp/index.jsp) and QGIS (http://www.qgis.org/) offer cost-free alternatives with user support groups. Spatial statistics packages, as they are typically outside mainstream needs, are often free and developed by academics with research and development grants from academic funding councils.[277;278]

More significant issues will be data sharing with other organisations and arranging spatial statistics training for staff. Data sharing within the NHS has previously been identified as a problem for implementing spatial information systems and a core obstacle for routine analytic work. The importance of data provider service level agreements/requirements cannot be stressed enough for any attempt by the CQC to make geographical analyses an integrated part of its regulatory work. Staff training in spatial statistics, the relevant software and the basics of GIS, whilst inevitably at some initial cost from an external source, is realistic for the CQC considering the statistical expertise in the organisation. However, a 'champion' from within the Commission will be necessary. With the likely eventual mainstreaming of spatial analysis in health services, development of a strategic plan by the CQC will be important to maximise the potential of spatial analytics for health/social care regulation. With this aim in focus we recommend that the model suggested by the Higgs et al study already referred to in this report be followed due to the overlap between GIS and spatial analysis. The suggested stages in a systematic plan for capacity building in GIS are: (1) training and guidance, (2) assessment of scope and remit of methods and technologies and (3) development of data exchange agreements.

In the context of CQC, training and guidance in the use of relevant methods and technologies from an external body will be necessary; a two-day course should cover the requisite material. Core topics covered must include an introduction to GIS, effective map communication, spatial cluster/outlier analysis, spatial regression and software implementation. Whilst there are many other aspects to spatial analysis, such as geostatistics which is frequently used in environmental health, we do not envisage these other methods as being essential to mapping and spatial statistics for health/social care regulation. A possible exception to this could be spatio-temporal analysis for examining past geographic changes in quality of care delivery over time and predicting future change. Internal assessment of the scope and remit of mapping and spatial statistics within the Commission may be preferable using domain knowledge and an internal 'champion' rather than being tendered out to an external body. Development of data exchange agreements between the CQC and other institutions such as the Ordnance Survey can be developed if necessary.

11 Discussion/Conclusions

11.1 Discussion

The Discussion focuses on the extent to which the research I have undertaken can answer the research questions originally posed in late 2007 (or the modified questions).

11.1.1 How should the prevalence of chronic diseases be estimated? How do different chronic disease prevalence estimation methods compare in terms of their validity?

These questions are obviously closely related. The answer depends on the criteria which should be used to compare prevalence modelling methodologies, their "fitness for purpose" in providing reasonably robust local, small population/area estimates which can be used by PCTs, practices and in future GP commissioning groups and LA public health departments. It is important to distinguish between decision-analysis "cost-effectiveness" models and population-based "surveillance" models.[279] A surveillance or prevalence model differs from decision-analysis models in that "rather than representing a hypothetical cohort, it models the population, that is, a collection of birth cohorts, over a specified period of time"

We can start by stating that, currently, the use of a highly complex agent-based, discrete event simulation models to provide local estimates as well would be a major undertaking. Developers of such models have until now focused on model construction- the iterative process of scope selection, hypothesis generation, causal diagramming, quantification, and reliability testing in large populations, rather than combining simulation models with small area covariates. However it is feasible to do this and it will certainly be done in the future.

It is likely, but by no means certain, that more sophisticated methods of model development provide better predictions, but a notable feature of the literature is that very little comparative validation has been carried out, and most have compared similar methods.[280] Only one paper which I could find has done this for different types of model- it compared Markov and discrete event simulation (DES) models i.e. agent-based models.[41] The DES model predicted the course of a disease naturally, with few restrictions. This may give these models superior face validity with decision makers. Furthermore, this model automatically provided a probabilistic sensitivity analysis, which is cumbersome to perform with a Markov model. DES models also allow inclusion of more variables without aggregation, which may improve model precision. However the differences were not great in terms of the actual predictions. Moreover these comparisons are only one aspect of model validation.

I could find no other published comparisons of, for example, models produced by MLM, multilevel, Markov, Bayesian or simulation methods using criteria which might be important to non-expert model users. On our part, we were unable to demonstrate, in our practice-level CHD and stroke modelling, the superiority of GWR compared to MLM methods, using the AIC as a criterion. Given the breadth and complexity of modelling methodology, few academics have in-depth knowledge of two let alone all of these methods. Geographers understandably favour spatial methodologies (GWR or Bayesian), and classically-trained statisticians non spatial methods such as MLM with clustering options. It appears to me that there is a need for methods to compare the fit of various types of models (such as AIC), and some comparisons of their application across several diseases.

For example, two or more modelling methodologies could be compared to a validation gold standard of local prevalence. This data is difficult to obtain; one option would be using data from populations in which extensive case-finding has apparently detected nearly all cases of disease. NHS Health Checks data might serve this purpose, once a large enough proportion of the population has been screened. Unfortunately, although the present national policy excludes cases currently on QOF registers, it does not specify that questions about patient-reported doctor-diagnosed disease be asked to ensure that cases not on practice registers are found.

In an important paper just published by the Institute for Health Metrics and Evaluation (IHME), Srebotnjak et al propose a validation framework to cope with the problem of validating estimates in data-sparse domains.[289] The framework combines the three conventional approaches to small area measurement, all of which we have used:

- pooling data across time by combining multiple survey years
- exploiting spatial correlation by including a spatial component; and
- utilizing structured relationships between the outcome variable and domain-specific covariates to define four increasingly complex model types - coined the ***Naïve*** i.e. age/race(ethnicity) for each sex***, Geospatial, Covariate, and Full models***.

Their Covariate and Full model types draw leverage from the relationships between outcome variable and additional domain-level covariates, as in the disease prevalence models. Spatial relationships are accomplished through an additional covariate in the Geospatial and Full models. Thus, their basic model is a generalized linear mixed-effects MLM with binomial outcome and logit link function. The combinations of the four model families and four model specifications generate a total of 16 models for each sex (for the English disease prevalence models the equivalent number is 25, because we used more covariates).

The gold standard serves as a benchmark judged to be the best available direct estimate for the small area domain. The IHME used Behavioural Risk Factor Surveillance Survey data (BRFSS, a US equivalent of the HSfE). Their gold standard is counties with sufficiently large sample sizes, which can be obtained by choosing small domains with large sample sizes in a single survey year, pooling multiple survey years, or increasing domain size. IHME pooled the 2000-2008 BRFSS and calculated the direct age-standardized, sex-specific BRFSS estimates for self-reported Type 2 diabetes prevalence, taking post-stratification weights into account, and weighting each survey year equally. They used as their gold standard the direct, age-standardized, sex-specific estimates for counties that had more than 900 observations (by sex) in both periods 1996-2004 and 2000-2008 (the validation sets).

The second step of the validation framework involves determining the minimum sample size needed to achieve sufficient correspondence with the gold standard. Such a validation environment allows the selection of a modelling strategy that optimally mixes the three approaches of pooling data across time, harnessing spatial patterns in the distribution of the outcome of interest, and adjusting for estimates for local area characteristics. This approach is analogous to that used in the development of risk predictions models, where derivation and validation cohorts are often initially derived from the same population: for example, in

the case of QRISK2 derivation, random sampling was used to assign practices to the derivation or validation cohort, and then their population data was used.[111] However, validation should still be carried out in other populations if possible.[188;281]

The IHME validation framework compares model estimates against these using all available observations for the large domains, and systematically reduced sample sizes obtained through random sampling with replacement. At each sampling level, the model is rerun, and the validity of the model estimates from the four model types is then determined by calculating the (average) concordance correlation coefficient (CCC) and (average) root mean squared error (RMSE) against the gold standard. The CCC is closely related to the intra-class correlation coefficient and can be used when the units are organized in groups, and when it is of interest to measure the agreement between units in the same group. The RMSE is often used to measure the differences between values predicted by a model or an estimator and the actually observed values- it is a useful measure to capture the precision of the model or estimator.

All the IHME model types had substantially higher CCC and lower RMSE than the direct, single-year estimates from the BRFSS. In addition, the inclusion of relevant domain-specific covariates generally improved predictive validity, especially at small sample sizes, and their leverage can be equivalent to a five- to tenfold increase in sample size in the Naïve and Geospatial models.

The International Society for Pharmacoeconomics and Outcomes Research (ISPOR) has published principles of good practice for decision analytic modelling in health-care evaluation.[282] A Canadian group has also just published developed a comprehensive framework for validating population-based chronic disease simulation models, and used it in a review of published model validation guidelines.[283] Based on the review, they formulated a set of recommendations for gathering evidence of model credibility. Evidence of model credibility derives from examining:

- **Evidence from examining model development process**
    - Conceptual model: "determining that the theories and assumptions underlying the conceptual model are correct and the model representation of the problem entity and the model's structure, logic, and mathematical and causal relationships are reasonable for the intended purpose of the model".
    - Parameters/ definition of variables: the conceptual and operational definitions of the variables in the model should be justified. Evidence that the model is sufficiently complete and that the relationships between the variables are specified correctly should come from both theory and empirical data evidence generally comes from examining the process used to derive a value for the parameter (primary source, method of derivation), and comparisons with data from other sources.
    - Computer implementation: "Published validation guidelines do not specify how the appropriateness of model type should be determined". This harks back to the methodological discussion above.
- **Evidence from examining model performance**

  - Plausibility (face validity): assessment of output plausibility (face validity), involving comparisons of model output with expectations
  - Internal consistency: assessed by considering functional and logical relationships between different output variables
  - Parameter sensitivity: "a technique that measures the effect on model output of alternative choices about model structure". In the context of the disease prevalence models, the best approach would be to produce credible intervals for the estimates using MLWin
  - Between-model comparisons: "modellers should cooperate in comparing results and articulating the reasons for discrepancies". This could be done by comparing the APHO prevalence models with, for example, Congdon's prevalence models,[46] or the South London Stroke Register[126]
  - Comparisons with external data: "all available data should be used in model development and data should not be withheld for the purpose of external validation". If, however, one of the objectives of the model(s) is to encourage case-finding in practices, this suggests that QOF register data should not be used for model development
- **Evidence from examining the consequences of model-based decisions**
  - The impact of models on the quality of decisions could be evaluated directly in a variety of ways, including subjective and objective measures.

The YHEC evaluation is useful in meeting the decision-making criterion, by providing feedback from model users. The additional criteria they suggest include:

- **flexibility:** including enabling users to input their own local covariate data if it is more accurate than that already in the models
- **transparency:** ensuring that the models are parsimonious, and provision of good methodological detail and clarity of assumptions within accompanying documents is important. Provision of credible intervals for estimates, possibly using Monte Carlo methods and MLWin software
- **accessibility: t**he models should be easy to locate, both within appropriate websites and through internet search engines. A single interface, with avoidance of competing models which produce different results
- **version control and updates:** user registration and email notifications of any problems with the models or upcoming work
- **format:** useful easily downloadable formats. Any such estimates should be provided for both resident i.e. LSOA and GP-registered i.e. practice populations. The additional effort involved in applying modelled disease odds to resident-based population data is minimal. The data could and should be used in both formats.
- **training and support:** an obvious criterion.

In summary, although there are gaps in establishing model credibility e.g. credible interval estimates, the majority of the above criteria have been satisfied.

11.1.2 What is the best available methodology given the requirements for prevalence models?

The IHME and Canadian frameworks described above appears to provide very useful technical criteria for assessing internal validity of prevalence models, and could be applied in the UK using HSfE data. They could also be used to compare not only geospatial, covariate, and full models but also other modelling methodologies described in the Introduction. Until this is done, hopefully with some agreement between those involved in developing UK models, it is not possible to identify the best methodology.

11.1.3 How valid are the chronic disease model prevalence estimates?

As discussed in 11.1.2, the internal validity of the prevalence models described here remains to be evaluated using the IHME framework or similar. The external validations described provide some assurances, the most obvious being the correlations with QOF-registered prevalence described in the analyses in Section 9. For example, for COPD, it seems that the model provides very conservative prevalence estimates which rarely (in only 3.5 per cent of practices) exceed QOF-registered prevalence. In the COPD admission rate analysis, undiagnosed COPD prevalence remains a risk factor for higher admission rates after adjustment for other population covariates. This is not the case for the equivalent CHD and stroke analyses, but the provenance of the HES data is uncertain.

Moreover, the unregistered/undiagnosed prevalence of COPD and hypertension is greater than that for CHD and stroke. This is plausible, on the basis that the onset of COPD is insidious, whereas that for CHD and stroke is often marked by a sentinel event resulting in hospital admission and definitive diagnosis (see Table 86). The COPD prevalence model is based on objective spirometry and not patient reports.

Table 86: Comparison of registered and unregistered prevalence for COPD, CHD and stroke

		Mean	(SD)	Min - Max	Median
COPD	Observed prevalence (%)	1.56	0.82	0.00-7.47	1.45
	Unregistered prevalence (%)	1.50	0.89	000-5.30	1.46
CHD	Observed prevalence (%)	3.50	1.19	0.00 - 9.81	3.47
	Unregistered prevalence (%)	0.95	0.87	0.00 - 5.78	0.87
Stroke	Observed prevalence (%)	1.61	0.67	0.00 - 18.55	1.61
	Unregistered prevalence (%)	0.42	0.44	0.00 - 3.83	0.31

Finally, there are three other sources of data for external validation:

- comparisons with disease-specific population-based surveys using objective criteria e.g. ECG evidence for CHD prevalence
- validity of patient-reported doctor-diagnosed disease as a diagnostic criterion
- studies of individual practices like the limited validation provided by the Westminster COPD case-finding pilot, which can be extended when a major cluster RCT of COPD screening led by Birmingham University is completed.[284] A good validation data source would be a number of practices where extensive interrogation and updating of practice registers had been carried out, using searches for prescription items and test results for specific diagnoses.

11.1.4 How useful are chronic disease prevalence estimates for case-finding at GP practice (and LSOA) level? Can chronic disease prevalence estimates be used with practice disease register prevalence to estimate the prevalence of undiagnosed disease?

As an example of how local estimates can be generated and used, a local Canadian public health agency (population 250,000) recently published a "textbook" for local disease incidence modelling.[273] Age standardized incidence ratios for cancer and the prevalence of Census covariates were calculated for each of 331 dissemination areas. The standardized incidence ratios for cancer varied dramatically across these areas. However, application of the LMI statistic as used by ourselves suggested significant spatial patterns for only two cancers, lung and prostate, both in males. Employing Bayesian hierarchical models, areas in the urban core were found to have significantly higher SIRs for male lung cancer than the remainder; and, neighbourhoods in some urban and surrounding rural areas exhibited significantly higher SIRs for prostate cancer. After adjustment for age and spatial dependence, average household income attenuated much of the spatial pattern of lung cancer, but not of prostate cancer.

This paper demonstrates the feasibility and utility of a systematic approach to identifying neighbourhoods, within the area served by a public health unit, that have significantly higher risks of cancer. It suggests several hypotheses for these spatial patterns that warrant further investigations. However few public health teams have the skills and knowledge to carry out this complex analysis. It makes much more sense to provide them with local estimates and the training to use them properly.

This is what the *Informing Healthier Choices* project aimed to do with the APHO/ICL prevalence models.[3] The best way to answer these questions is to assess the impact of the models at local level. An early assessment has been provided in 10.1 Evaluation of the uptake of the models by York University Health Economics Consortium and 10.3 Case-finding pilots and case studies.

11.2 The future

11.2.1 linkages with risk factor prevalence data from GP systems

I have already illustrated how the current models can be linked, both with each other (10.4 Modelling risk factor scenarios and impact on disease prevalence) and with other modelled data such as smoking prevalence (8.1 Chronic Obstructive Pulmonary Disease Model). As noted above, local estimates of prevalence or population means (for continuous variables) of disease risk factors and lifestyle behaviours e.g. current smoking, obesity, blood pressure and physical activity, are of great interest to policy-makers, and increasingly to local organisations for planning and performance monitoring purposes e.g. in Local Area Agreements. They are also of interest to researchers examining the relationships between health determinants, such as income and education, and health outcomes such as disease prevalence, SRHS and death, because risk factor prevalence and means provide a linkage between determinants and outcomes, and may be used to predict the latter.[243] Risk factor estimates may be obtained and used for geographically-based populations such as Regions, Counties, LAs, and Wards or SOAs, or populations enrolled for healthcare e.g. PCTs, or GP commissioning groups.

Until now, risk factor prevalence estimates have come mainly from the HSfE, from local surveys either run independently from the HSfE, or from synthetic estimates. The traditional ("direct") approach to estimation used for the HSfE uses classical design-based survey sampling. However, sample sizes are typically small within small areas, so the direct estimators have large sampling errors (and hence large confidence intervals), even when pooling years of survey data (which prevents trend estimates). When there are no sample observations in the small areas of interest, direct estimators cannot even be calculated.

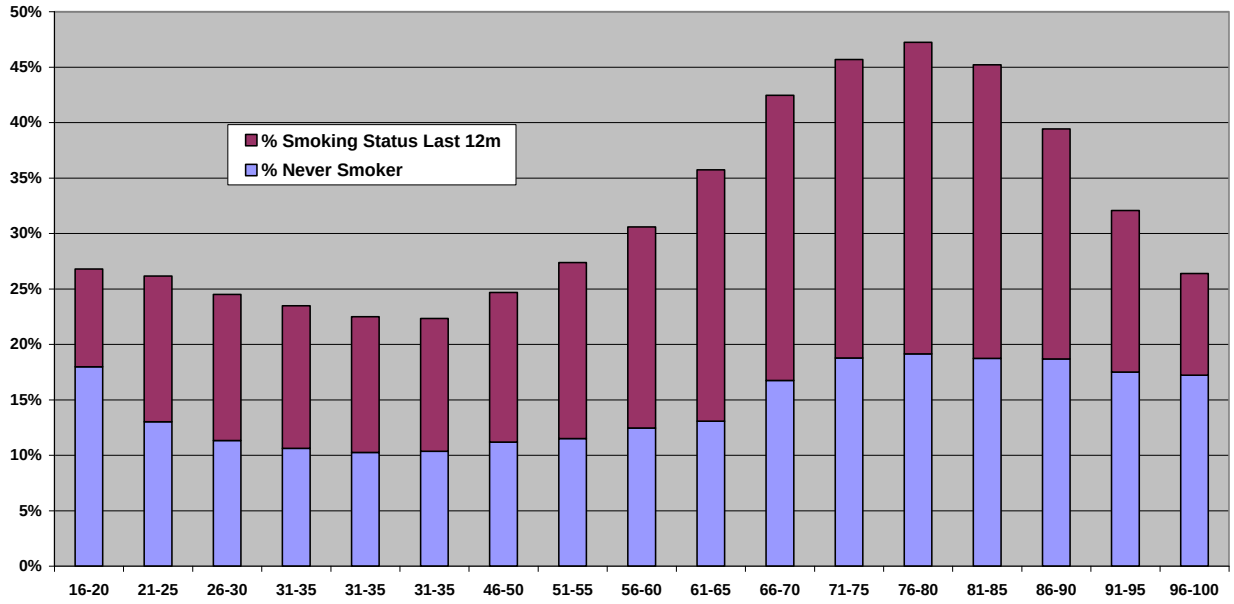
Even major boosts result in a wide margin of error at the PCO or LA level, of the order of five per cent for risk factors like smoking and obesity. In 2004, DH costed a national boost to provide this margin at PCO level at £14m, and so did not proceed with it. The new ONS Integrated Household Survey will provide a much bigger sample size, and will provide more reliable local estimates of smoking prevalence, but not for other risk factors.

Potential use of primary care data for risk factor prevalence/mean estimation

The quality and quantity of population-based data collected by primary healthcare teams has improved dramatically in recent years, as a result of overall quality improvement, specific data quality initiatives, and the impact of pay for performance through the QOF.[220;221;285] The data is also becoming more accessible through national GP databases, local IT projects and, eventually, aggregation through the Secondary Uses Service via GPES, which could cover over 50% of English practices.[4] This would allow near to real-time monitoring of risk factor prevalence data.

There is a large volume of risk factor and lifestyle data in GP systems. For example, our analysis of the 2005-6 IMS national GP database (about 1m active patients) shows that overall about 30% of patients have their smoking status recorded in the last 12 months, and 17% their BMI recorded in that period. There is therefore the potential to use this data, either alone or in combination with other data, to improve local prevalence estimation. It is also easier to measure risk factor prevalence through population-based programmes such as NHS Health Checks than it is to ensure QOF registers reflect actual disease prevalence. However, there are a number of methodological problems to resolve. For example, the data is non-random- sicker and older people and females are more likely to be sampled. Figure 54 below shows the percentage of patients with a valid smoking status recorded in the last 12 months in the IMS database. The result is that this data source appears to underestimate smoking prevalence in younger age groups.

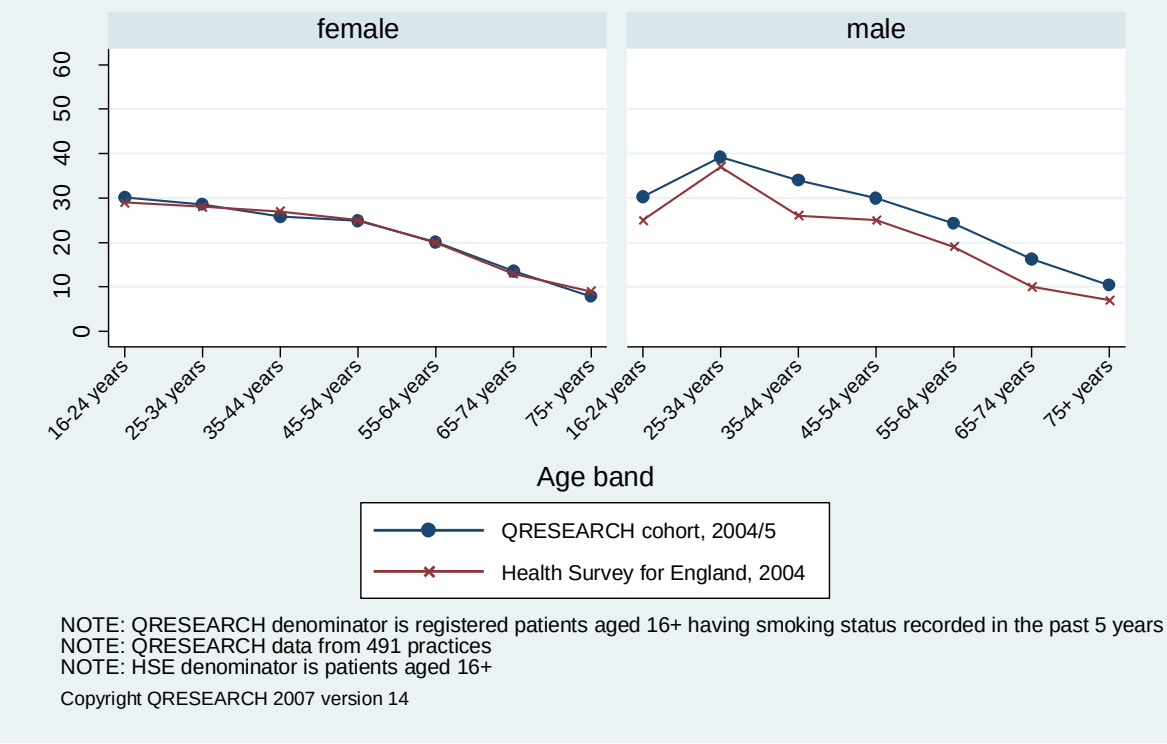
Figure 54: Percent of patients with smoking status recorded in last 12 months



Source: IMS GP database, Information Centre for Health & Social Care

Figure 55 compares smoking prevalence by age and sex between the QRESEARCH database and the HSfE. The proportions of males recorded as smokers in QRESEARCH (2004/5) were higher in each age band than the proportions reported in the HSfE (2004) data, but for females the proportions are very similar in the two sets of data. Insofar as disease developing in middle and old age is likely to produce sampling bias, differences might be expected in this age group. In addition, data has been captured over longer periods than is the case in surveys. In conclusion, once local risk factor prevalence data becomes available it is and obvious step to link it to disease prevalence models.

Figure 55: proportion of smokers by age and sex, comparison of HSfE and QRESEARCH data



NOTE: QRESEARCH denominator is registered patients aged 16+ having smoking status recorded in the past 5 years
NOTE: QRESEARCH data from 491 practices
NOTE: HSE denominator is patients aged 16+

Copyright QRESEARCH 2007 version 14

11.2.2 A simulation model of the UK population?

A major disadvantage of the current prevalence models is their static, cross-sectional structure. In the US, the Centers for Disease Control are already well advanced in the development of dynamic models.[54] In San Francisco, Archimedes is continuing to expand both in its scope and functionality.[57] The UK should be developing equally dynamic epidemiologic models. This requires at the very least a multidisciplinary team with ongoing support, not a lone researcher and a single injection of funding.

In the UK, in addition to the prevalence models described here, there is already a great deal of academic expertise in mathematical modelling, particularly for CVD. Examples include the Foresight Obesity Report, for which quantitative modelling included cross-sectional regression and longitudinal simulation analyses, and the IMPACT CHD model developed by Liverpool University.[286] A systematic review of policy models covering CHD found 75 articles describing 42 models: 12 were micro-simulation, 8 spreadsheet or cell-based, and 8 life table analyses, while 14 used other methods.[58] This review includes the CHD model developed for DH, which has never been used, and the PREVENT cell-based model.[287] Healthcare provider capacity models have also been developed.[288]

Yet there is currently no synergy in UK modelling research or development because of the lack of coordination, and no single organisation has the resources to produce a large scale model. A recent survey of academic epidemiologic modellers proposed that the development of this discipline is likely to require a framework providing recommendations and guidance at various steps of the studies, from design to report. A possible way forward would be the development of a consensus statement about methods and steps similar to other statements related to reporting of clinical trials.

12 List of related publications

(1) Nacul LC, Soljak M, Meade T. Model for estimating the population prevalence of chronic obstructive pulmonary disease: cross sectional data from the Health Survey for England. *Popul Health Metr* 2007; 5:8-16.

(2) Soljak M, Flowers J. Closing the Gap: Using Prevalence Models for Long-term Conditions in the United Kingdom. *J Ambul Care Manage* 2008; 31(3): 211-215.

(3) Calderon-Larranaga A, Carney L, Soljak M, Bottle A, Partridge M, Bell D et al. Association of population and primary healthcare factors with hospital admission rates for chronic obstructive pulmonary disease in England: national cross-sectional study. Thorax 2011;66(3):191-96.

(4) Dalton A, Soljak M, Samarasundera E, Millett C, Majeed A. The prevalence of CVD risk amongst the population eligible for the NHS Health Checks in England; a modelling study using the QRISK2 and JBS2 scores. *BMJ*, Submitted. 2010.

(5) Nacul L, Soljak M, Samarasundera E, Hopkinson NS, Lacerda E, Indulkar T et al. COPD in England: a comparison of expected, model-based prevalence and observed prevalence from general practice data. *J Publ Health 2011;33(1):108-16.*.

(6) Soljak M, Samarasundera E, Indulkar T, Walford H, Majeed A. Variations in cardiovascular disease under-diagnosis in England: national cross-sectional spatial analysis. *BMC Cardiovasc Disorders*; accepted for publication..

(7) Stafford M, Soljak M, Pledge V, Mindell J. Health related quality of life and long-term conditions: findings from a national survey in England. *Eur J Publ Health* ; accepted for publication.

13 Technical Appendix

This appendix provides some further explanation of the less well-known analytic techniques used, particularly those related to spatial statistics.

Concordance correlation coefficient
The concordance correlation coefficient (CCC) measures agreement between two variables and correspondence within groups, i.e., how strongly do units within the same area resemble each other. The CCC is closely related to the intra-class correlation coefficient and can be used when the units are organized in groups and when it is of interest to measure the agreement between units in the same group e.g. practices.

Root Mean Squared Error
Root Mean Squared Error (RMSE) is a measure for the average squared difference between model estimates and the gold standard. The RMSE is often used to measure the differences between values predicted by a model or an estimator and the actually observed values. It is a useful measure to capture the precision of the model or estimator.

In the validation framework proposed by Srebotnjak et al,[289] the best-performing models are those with the highest CCC and lowest RMSE.

The techniques used in quantitative geography and spatial statistics include cluster/outlier detection, smoothing, interpolation, GIS, spatial databases, mapping, querying, analysis and modelling software. Some GIS software examples include relatively simple tools for mapping and visualisation such as InstantAtlas, GapMinder, ArcGIS, GRASS, and MapInfo.

Ordinary least squares or OLS regression is the classical or 'global' method of regression modelling. There are limitations to using OLS models, including omission of spatial adjacency and spatial non-stationarity. Geographically weighted regression (GWR) is a 'local' method which:

- accounts for edge effects, spatial influence, and non-random spatial variability in strength and outcome
- allows disaggregation → local estimates of R^2, error terms, etc.
- groups together of proximate, similar outcomes
- uses customised rather than predefined units (horses for courses)

Spatial statistics software examples include GeoDa, GWR (used for this thesis), R extensions, and S-PLUS extensions. Stata can also be used for mapping purposes. Some of the most important techniques are outline below.

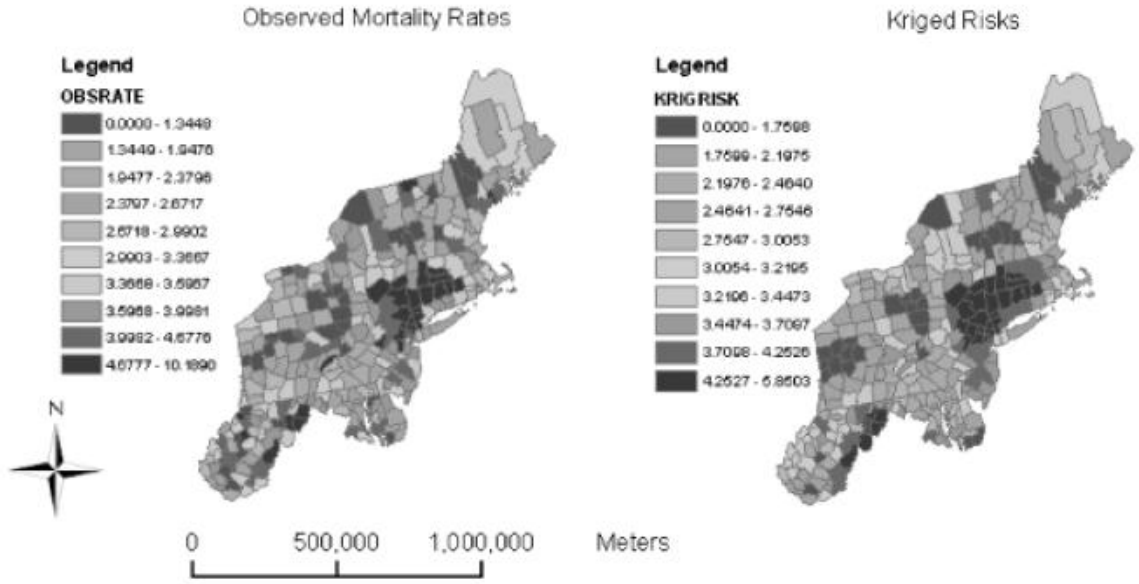
Local Moran's I (LMI) analysis
This is a useful and intuitive cluster/outlier detection technique, which is one of many similar techniques; but LMI is more related to mainstream statistical concepts. How it works:

- it examines each data point in relation to whole dataset
- but gives more weighting to nearer data points
- there is a threshold distance e.g. 2km, 5km

- the null hypothesis that there is no spatial autocorrelation
- scores between +1 and -1 are generated
- the significance test is a Z scores i.e. >1.96 standard deviations from the mean

"Kriging"

An example is kriging of cancer rates (see *Goovaerts et al, Geographical Analysis 2005;37:152-182).*



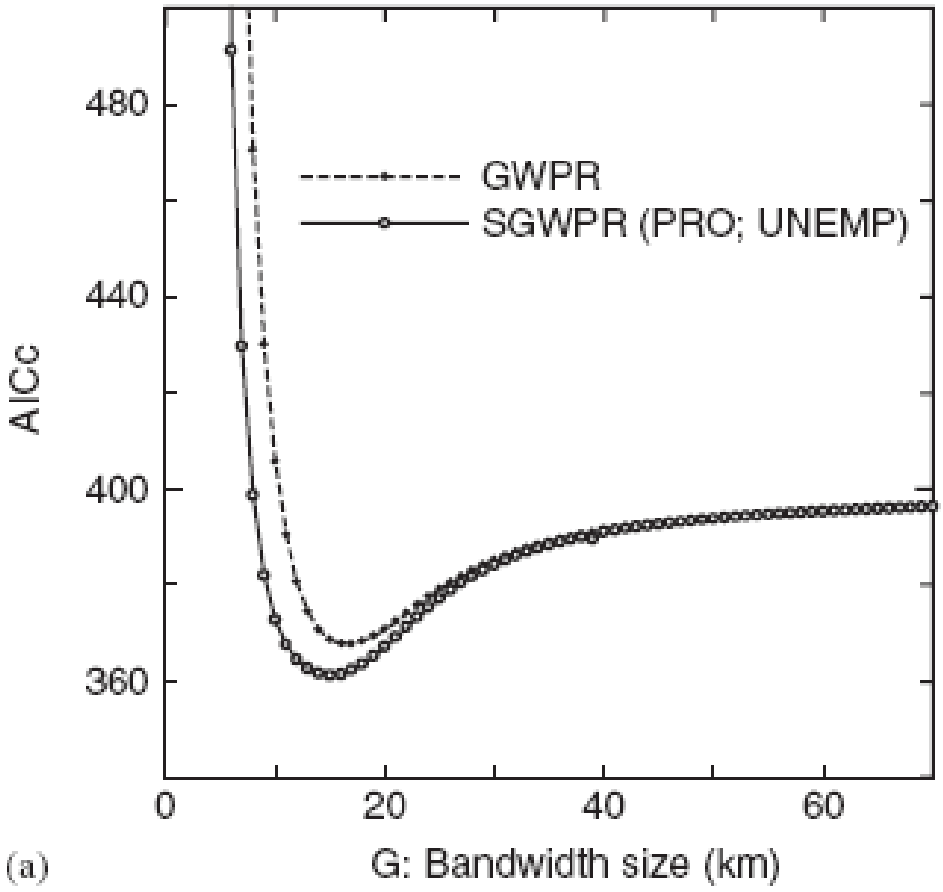
Spatial regression diagnostics: we compared OLS and GWR using Monte Carlo simulations to model outcomes for different bandwidths: a 'microscope' technique. We used the Akaike Information Criterion (AIC), ANOVA, R^2 increases, standardised residuals, and Cook's D.

Analysis of variance (ANOVA): the null hypothesis is that there is random variation in OLS (global) regression results leading to chance spatial patterns.

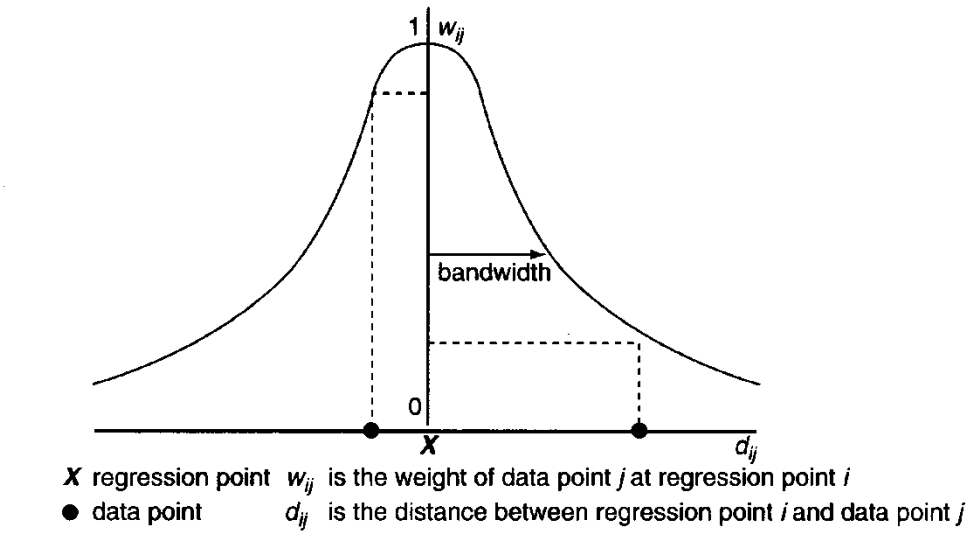
- degrees of freedom
- F test (Gauss-Jordan method)
- standardised residuals
- correlation coefficient

Akaike Information Criterion (AIC):

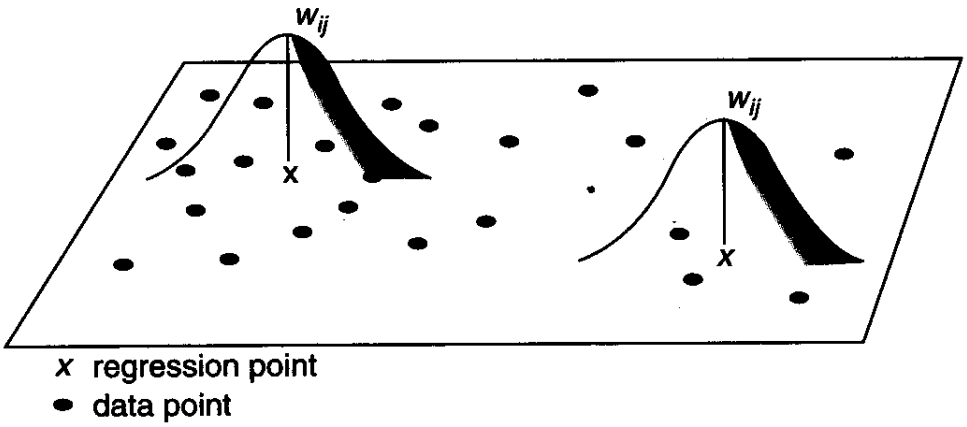
- instead of a confidence interval approach, the AIC is a trade-off between R^2 and degrees of freedom
- lower AIC → model describes data better
- different bandwidths will yield different results; an optimal bandwidth is determined



Source: Nakaya et al, Statistics in Medicine 2005;24:2695-2717



Source: Fotheringham et al, Geographically Weighted Regression 2002:44


w_{ij}
x
w_{ij}
x
x regression point
data point

14 Pragmatic methods for estimating prevalence in small populations

An excellent guide to prevalence model development can be found in APHO Technical Briefing no 8: Prevalence Modelling, which is drawn upon here.[290]

Prevalence modelling is a technique used to estimate the number of people with a particular condition or risk factor in a population when direct evidence is not available. Direct evidence may be lacking because surveys or data collection have not been undertaken, are technically impractical, or are unreliable. Methods for generating synthetic or modelled estimates range in complexity from simple to highly sophisticated. Crude estimates of the number of cases can be generated by applying known prevalence rates to a different population, for example applying national rates measured in a large survey to a local population; or applying local rates for a recent year to a projected future population. Very simple models typically apply age/sex specific rates to local population data. However, many factors in addition to age and gender influence prevalence, for example, deprivation smoking and ethnicity can influence the prevalence of a behaviour, risk factor or disease, and more complex epidemiological modelling techniques are required in order to take such factors into account.

Many methods exist for creating synthetic estimates of prevalence, and in many cases methodologies are combined and adapted to make best use of the information and data available. There is often a balance to be struck between increasing the complexity of the model by incorporating more contributory factors and the availability of good quality data at local level to populate the model. These input requirements of a model are often restricted by what information is available. Complex models can also suffer from difficulty of interpretation, which negates the benefit of increased accuracy. All models are based on assumptions. Good models clearly state the assumptions that have been made and good interpretation of modelled estimates takes into account the limitations of the assumptions. The two simplest methods are:

14.1 Prevalence estimates from studies applied to local populations

Although this can be a quick and simple method, its usefulness depends on the size of the studies and hence the precision of the prevalence estimates, and whether or not they include prevalence estimates for population subgroups. Incorporating different prevalence rates for subgroups by deprivation, gender and/or age is usually useful and technically straightforward. However, increasing the specificity of the model by using prevalence estimates in many subgroups (for example gender, age, ethnicity, smoking status and employment status) can limit the range of local levels to which it can be applied, because there is seldom sufficient alignment between data routinely available at local level and all the variables used in the source study.

14.2 Regression models using large surveys

Multiple variables from large surveys can be used to model the risk factors for a behaviour or disease, using techniques such as multinomial logistic regression. However, it is important to limit the factors considered to those for which data are available in the population of interest. For example, cholesterol level or family history of disease may be important risk factors which were recorded in the source survey, but such information is not usually available at population level and therefore these are not appropriate variables to be included in a disease prevalence model.

National surveys are usually limited to people living in private households and omit populations such as the homeless, those living in institutional care, 'special populations' (armed forces and prisoners) and people particularly likely to decline to participate. For some disease areas, notably some types of mental illness, these omitted populations can be particularly important. Despite this limitation, national surveys are often the best source of prevalence information available, but where possible should be used in conjunction with other evidence about the likely extent to which they miss cases. Models can then be adjusted to take account of the resultant under-estimation of prevalence. Although regression models most commonly use survey data, other data sources, for example information recorded in general practices, can also be used to create this type of prevalence model.

14.3 Validation, confidence intervals and robustness

The accuracy of model outputs depends on the predictive power of the model and on the accuracy of the input data. Models should be subjected to validation checks to ascertain their robustness and general applicability. Estimates of the accuracy of prevalence estimates based on simple models can be generated by combining the uncertainty in prevalence rates from the source study or trial with the stochastic variation expected given the size of the local population. This approach results in a range estimate for the prevalence, rather than confidence intervals. The range estimates are calculated using the same methods as those used to derive the control limits for funnel plots. However, if there is uncertainty around both the population data and the input data, the calculation of confidence intervals can be complex. Bootstrapping methods i.e. repeated sampling such as MCMC are commonly used in such situations. Four ways of validating models are described here (14.4-14.7).

14.4 Sensitivity testing

Sensitivity testing can be useful in assessing how the uncertainty in input data affects prevalence estimates. For some models, very small variations in the input data will have a large effect on the results. Other models may be relatively insensitive to variability in input data. For example, different sources of practice level smoking prevalence data were input into the APHO chronic obstructive pulmonary disease (COPD) models. Estimated COPD prevalence in general practices ranges from 1% to 7%. In 92% of cases, changing the source of smoking prevalence data made an absolute difference of less than 0.3% in the COPD prevalence estimate. Estimates were strongly correlated with each other ($r^2 > 0.95$). One-way sensitivity analysis such as this evaluates the impact of a change in one variable on the model results. Multi-way sensitivity analysis is more powerful and can be used to assess the impact of changing two or more variables simultaneously.

14.5 Internal validation

One method of assessing the performance of a model is to use it to predict the response for each subject in the source data (e.g. a large survey). These predictions are called fitted values. The differences between the fitted and the observed values are called residuals. Residual analysis can be used to check the adequacy of any assumptions used when creating the model. It can also be used to identify whether any additional factors should be included.

To check the accuracy of the model, the predicted 'classification' of each individual (i.e. whether or not they have the disease or behaviour that is being modelled) can be compared with their actual classification. This will result in a 'misclassification' (also known as a 'contingency' or 'confusion') table

14.6 Receiver-Operating Characteristic (ROC) Analysis

ROC analysis is a useful way of assessing the accuracy of a model by understanding the trade-off between the sensitivity (in this sense referring to the true positive rate; Table 2) and the specificity (the true negative rate).[12] The method was developed to assess the accuracy of distinguishing signal from noise in radar systems and has since been applied in many other settings, including clinical diagnostic testing and the evaluation of regression models that classify cases into two categories, for example diseased and non-diseased. Sensitivity is plotted against 1-specificity (specificity subtracted from one) over a range of values and the area under the curve (AUC or AUROC) is used as a summary of the predictive or diagnostic accuracy. A 'perfect' model that accurately predicts every case has AUROC = 1. Typically, models have a convex ROC curve and an AUROC between 0.5 (equivalent to random chance) and 1. A model with AUROC < 0.5 is less accurate than random chance.

14.7 External validation

Modelled estimates can be compared with observed prevalence where such measures exist. For example, modelled estimates for small areas can be aggregated to regional or national level and compared with survey measures of prevalence. In some models, local values are adjusted so that regional or national aggregates are consistent with observed prevalence.

14.8 Strengths and limitations of prevalence models

Understanding the strengths and limitations of any model is crucial if the results are to be appropriately used. However, this understanding is often hampered by the limited amount of information published about the methodology, development and testing of models. The following checklist of ten areas that should be reported and discussed in a modelling paper.

1. Aims of the project
2. Structure and methods of the model
3. Data quality (data availability, how up to date, comprehensive, any gaps in certain population groups or interventions, reasons for selecting or excluding specific data sources)
4. Methodological limitations
5. The assumptions used to address these deficiencies Sensitivity analyses (one-way or preferably multi-way)

6. Whether the validity of the model was checked (with real observational data or with other models)
7. Replication of the model in different populations
8. Model results and comparisons with other studies
9. Social and economic policy implications of model outcomes

References

(1) Information Centre for Health & Social Care. Health Survey for England 2006 Latest Trends. 2008 http://www.ic.nhs.uk/statistics-and-data-collections/healthand-lifestyles-related-surveys/health-survey-for-england/health-survey-for-england-2006-latest-trends

(2) Wanless D. Wanless Report: Securing Good Health for the Whole *Pop*. National Archives 2008 [cited 2010 Sept. 23]; http://webarchive.nationalarchives.gov.uk/+/http://www.hm-treasury.gov.uk/consultations_and_legislation/wanless/consult_wanless04_final.cfm treasury.gov.uk/consultations_and_legislation/wanless/consult_wanless04_final.cfm

(3) Public Health Development Unit. Informing Healthier Choices: information and intelligence for healthy populations. Department of Health 2007 [cited 2008 Aug. 29]; http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_075488

(4) Information Centre. GP Extraction Service. The Information Centre website 2010 [cited 2010 Sept. 21]; http://www.ic.nhs.uk/services/in-development/general-practice-extraction-service

(5) Soljak M. Faculty of Public Health, NHS CFH and NHS IC joint workshop: The Future of Public Health Information and Intelligence: needs and applications: the GP Extraction Service. Connecting for Health events online 2009 http://etdevents.connectingforhealth.nhs.uk/events/uploads/michael_soljak.ppt

(6) Hospital Episode Statistics Online. Information Centre website 2010 http://www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937&categoryID=537

(7) Soljak M, Flowers J. Closing the Gap: Using Prevalence Models for Long-term Conditions in the United Kingdom. *J Ambulatory Care Management* 2008; 31(3):211-215.

(8) Billings J, Dixon J, Mijanovich T, Wennberg D. Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. *BMJ* 2006; 333(7563):327.

(9) Dr Foster Intelligence. High-impact User Manager (HUM) - Reducing avoidable emergency admissions. Dr Foster Intelligence website 2010 [cited 2010 Sept. 23]; http://www.drfosterintelligence.co.uk/managementInformation/longTermConditions.asp

(10) McGauran A. Over 55s in US are less healthy than their English peers but live longer. *BMJ* 2010; 341.

(11) Banks J, Muriel A, Smith J. Disease Prevalence, Disease Incidence, and Mortality in the U.S. and England. Demography 47[Supplement], S211-S231. 2010.

(12) Forouhi NG, Merrick E, Goyder E, Ferguson BA, Abbas J, Lachowycz K et al. Diabetes prevalence in England, 2001-estimates from an epidemiological model. *Diabetic Medicine* 2006; 23(2):189-197.

(13) Information Centre for Health & Social Care. Quality and Outcomes Framework. Information Centre website 2010 http://www.ic.nhs.uk/statistics-and-data-collections/audits-and-performance/the-quality-and-outcomes-framework

(14) Darzi A. Healthcare for London: a Framework for Action. Technical Paper. Healthcare for London website 2007 [cited 2010 Sept. 23]; http://www.healthcareforlondon.nhs.uk/assets/Publications/A-Framework-for-Action/FFA-Technical-Paper.pdf

(15) Yorkshire & Humber Public Health Observatory. PBS Phase 3 Diabetes *Pop* Prevalence Model. Yorkshire & Humber Public Health Observatory 2008 [cited 2008 Sept. 2]; http://www.yhpho.org.uk/PBS_diabetes.aspx

(16) Information Centre for Health & Social Care. QOF 2006/07 data tables. 2007 [cited 2008 July 8]; http://www.ic.nhs.uk/our-services/improving-patient-care/the-quality-and-outcomes-framework-qof-2006/07/qof-2006-07-data-tables

(17) NHS Information Centre for health and social care. National Diabetes Audit. Information Centre website 2010 [cited 2010 Sept. 27]; http://www.ic.nhs.uk/services/national-clinical-audit-support-programme-ncasp/audit-reports/diabetes

(18) National Statistics Methodology Advisory Committee. Model-Based Small Area Estimation Series No 2. Office for National Statistics. National Statistics website 2003 [cited 2010 Sept. 23]; http://www.statistics.gov.uk/methods_quality/downloads/small_area_est_report/saep1_Prelims&ch1&2_v2.pdf

(19) Goubar A, Ades AE, De Angelis D, McGarrigle CA, Mercer CH, Tookey PA et al. Estimates of HIV prevalence and proportion diagnosed based on Bayesian multi-parameter synthesis of surveillance data. Health Protection Agency 2008 [cited 2008 Sept. 2]; http://www.hpa.org.uk/web/HPAwebFile/HPAweb_C/1194947329693

(20) Alkema L, Raftery AE, Clark SJ. Probabilistic projections of HIV prevalence using Bayesian melding. Annals of Applied Statistics 1[1], 229-248. 2007.

(21) NICE/*Br* Thoracic Society. Appendix B: The cost effectiveness of opportunistic case finding in primary care. *Thorax* 2004; 59(90001):i175-i190.

(22) Gillies CL, Lambert PC, Abrams KR, Sutton AJ, Cooper NJ, Hsu RT et al. Different strategies for screening and prevention of type 2 diabetes in adults: cost effectiveness analysis. *BMJ* 2008; 336(7654):1180-1185.

(23) Thorpe KE, Florence CS, Howard DH, Joski P. The Rising Prevalence Of Treated Disease: Effects On Private Health Insurance Spending. *Health Aff* 2005;hlthaff.

(24) *Br* Cardiac Society BHS. JBS 2: Joint *Br* Societies' guidelines on prevention of cardiovascular disease in clinical practice. *Heart* 2005; 91(Suppl V):1-52.

(25) Gravelle H, Sutton M. Inequality in the geographical distribution of general practitioners in England and Wales 1974-1995. *J Health Services Research and Policy* 2001; 6(1):6-13.

(26) Stern M, Williams K, Eddy D, Kahn R. Validation of Prediction of Diabetes by the Archimedes Model and Comparison With Other Predicting Models. *Diabetes Care* 2008; 31(8):1670-1671.

(27) Danaei G, Friedman A, Oza S, Murray C, Ezzati M. Diabetes prevalence and diagnosis in US states: analysis of health surveys. *Pop Health Metrics* 2009; 7(1):16.

(28) Health Development Agency. The smoking epidemic in England. NICE website 2004 [cited 2010 Sept. 23]; http://www.nice.org.uk/niceMedia/documents/smoking_epidemic.pdf

(29) Pickering K, Scholes S, Bajekal M. Synthetic estimation of healthy lifestyles indicators: stage 3 report. 2005. London, National Centre for Social Research.

(30) EURAREA Consortium. Enhancing Small Area Estimation Techniques to meet *Eur* Needs. National Statistics website 2004 [cited 2010 Sept. 23]; http://www.statistics.gov.uk/eurarea/downloads/EURAREA_PRV_1.pdf

(31) Jackson C, Best N, Richardson S. Improving ecological inference using individual-level data. *Statistics in Medicine* 25[12], 2136-2159. 11-10-2005. 1-9-2008.

(32) Jackson C, Best N, Richardson S. Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. *J* Royal Statistical Society 171[1], 159-178. 18-9-2007.

(33) Diez-Roux AV. Multilevel analysis in public health research. *Annual Review of Public Health* 2000; 21(1):171-192.

(34) Markov models. *SO: Statistics in Medicine* 1993; 12(22):2127-2130.

(35) Honeycutt AA, Boyle JP, Broglio KR, Thompson TJ, Hoerger TJ, Geiss LS et al. A Dynamic Markov Model for Forecasting Diabetes Prevalence in the United States through 2050. *Health Care Management Science* 2003; 6(3):155-164.

(36) Martin C. Description and validation of a Markov model of survival for individuals free of cardiovascular disease that uses Framingham risk factors. *BMC Medical Informatics and Decision Making* 2004; 4(1):6.

(37) Moran A, Gu D, Zhao D, Coxson P, Wang YC, Chen CS et al. Future Cardiovascular Disease in China: Markov Model and Risk Factor Scenario Projections From the Coronary Heart Disease Policy Model-China. *Circ Cardiovasc Qual Outcomes* 2010; 3(3):243-252.

(38) Nijhuis RL, Stijnen T, Peeters A, Witteman JCM, Hofman A, Hunink MGM. Apparent and Internal Validity of a Monte Carlo-Markov Model for Cardiovascular Disease in a Cohort Follow-up Study. *Med Decis Making* 2006; 26(2):134-144.

(39) SB P. Refining estimates of major depression incidence and episode duration in Canada using a Monte Carlo Markov model. *Med Decis Making* 2004; 24(4):351.

(40) Sesso HD. Blood pressure lowering and life expectancy based on a Markov model of cardiovascular events. *Hypertension* 2003; 42(5):885.

(41) Simpson KN, Strassburger A, Jones WJ, Dietz B, Rajagopalan R. Comparison of Markov Model and Discrete-Event Simulation Techniques for HIV. *Pharmacoeconomics* 2009; 27(2):159-165.

(42) Congdon P. Health status and healthy life measures for population health need assessment: modelling variability and uncertainty. *Health & place* 2001; 7(1):13-25.

(43) Congdon P, Almog M, Curtis S, Ellerman R. A spatial structural equation modelling framework for health count responses. Statistics in Medicine 26[29], 5267-5284. 11-5-2007.

(44) Congdon P. Estimating CHD prevalence by small area: integrating information from health surveys and area mortality. *Health & place* 2008; 14(1):59-75.

(45) Congdon P. Estimating population prevalence of psychiatric conditions by small area with applications to analysing outcome and referral variations. *Health & place* 2006; 12(4):465-478.

(46) Congdon P. Estimating diabetes prevalence by small area in England. *J Public Health* 2006; 28(1):71-81.

(47) Cortes C, Vapnik V. Support-vector networks. *Machine learning* 1995; 20:273-297.

(48) Yu W, Liu T, Valdez R, Gwinn M, Khoury M. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Medical Informatics and Decision Making* 2010; 10(1):16.

(49) Homer JB, Hirsch GB. System Dynamics Modeling for Public Health: Background and Opportunities. *Am J Public Health* 2006; 96(3):452-458.

(50) Gunning-Schepers LJ. The health benefits of prevention - a simulation approach. *Health Policy* 1989; 12:1-256.

(51) Cooper K, Davies R, Roderick P, Chase D, Raftery J. The development of a simulation model of the treatment of coronary heart disease. *Health Care Manag Sci* 2002; 5:259-267.

(52) Davies R, Roderick P, Raftery J. The evaluation of disease prevention and treatment using simulation models. *Eur J Operational Research* 2003; 150:53-66.

(53) Marma AK, Berry JD, Ning H, Persell SD, Lloyd-Jones DM. Distribution of 10-Year and Lifetime Predicted Risks for Cardiovascular Disease in US Adults: Findings From the National Health and Nutrition Examination Survey 2003 to 2006. *Circ Cardiovasc Qual Outcomes* 2009.

(54) Boyle J, Thompson T, Gregg E, Barker L, Williamson D. Projection of the year 2050 burden of diabetes in the US adult population: dynamic modeling of incidence, mortality, and prediabetes prevalence. *Pop Health Metrics* 2010; 8(1):29.

(55) Schlessinger L, Eddy DM. Archimedes: a new model for simulating health care systems--the mathematical formulation. *J Biomedical Informatics* 2002; 35(1):37-50.

(56) Eddy D. Bringing Health Economic Modeling to the 21st Century. *Value in Health* 2006; 9(3):168.

(57) Eddy DM. Linking Electronic Medical Records To Large-Scale Simulation Models: Can We Put Rapid Learning On Turbo? *Health Aff* 2007; 26(2):w125-w136.

(58) Unal B, Capewell S, Critchley J. Coronary heart disease policy models: a systematic review. *BMC Public Health* 2006; 6(1):213.

(59) Nacul L, Soljak M, Meade T. Model for estimating the population prevalence of chronic obstructive pulmonary disease: cross sectional data from the Health Survey for England. *Pop Health Metrics* 2007; 5(1):8.

(60) National collaborating centre for chronic conditions. Chronic Obstructive Pulmonary Disease. National clinical guideline on management of chronic obstructive pulmonary disease in adults in primary and secondary care. Appendix B: The cost effectiveness of opportunistic case finding in primary care. *Thorax* 2004; 59(90001):i175-i190.

(61) Chief Medical Officer. CMO Report 2004: It takes your breath away: the impact of chronic obstructive pulmonary disease. UK Department of Health 2005 [cited 2008 July 7];

(62) Sorlie PD, Kannel WB, O'Connor G. Mortality associated with respiratory function and symptoms in advanced age. The Framingham Study. *Am Rev Respir Dis* 1989; 140:379-384.

(63) Devereux G. ABC of chronic obstructive pulmonary disease: Definition, epidemiology, and risk factors. *BMJ* 2006; 332(7550):1142-1144.

(64) Coultas DB, Mapel D, Gagnon RC, Lydick E. The Health Impact of Undiagnosed Airflow Obstruction in a National Sample of United States Adults. *Am J Respir Crit Care Med* 2001; 164(3):372-377.

(65) Hansell AL. The epidemiology of chronic obstructive pulmonary disease in the UK: spatial and temporal variations. *PhD Thesis* 2004.

(66) Rushton L. Review of the literature on chronic bronchitis and emphysema and occupational exposure. MRC Institute of Environment and Health, Leicester 2005 [cited 8 A.D. July 7]; http://www.silsoe.cranfield.ac.uk/ieh/links/publishedoffsite.html

(67) Barnes PJ. Molecular genetics of chronic obstructive pulmonary disease. *Thorax* 1999; 54:245-252.

(68) Chatila WM, Wynkoop WA, Vance G, Criner GJ. Smoking patterns in African *Am*s and Whites with advanced COPD. *Chest* 2004; 125:15-21.

(69) Godtfredsen NS, Vestbo J, Osler M, Prescott E. Risk of hospital admission for COPD following smoking cessation and reduction: a Danish population study. *Thorax* 2002; 57(11):967-972.

(70) Anthonisen NR, Skeans MA, Wise RA, Manfreda J, Kanner RE, Connett JE et al. The effects of smoking cessation intervention on 14.5-year mortality. A randomized clinical trial. *Ann Intern Med* 2005; 142:233-239.

(71) Gorecka D, Bednarek M, Nowinski A, Puscinska E, Goljan-Geremek A, Zielinski J. Diagnosis of airflow limitation combined with smoking cessation advice increases stop-smoking rate. *Chest* 2003; 123:1916-1923.

(72) Enright PL, Kronmal RA, Higgins M, Schenker M, Haponik EF. Spirometry reference values for women and men 65 to 85 years of age. *Am Rev Respir Dis* 1993; 147:125-133.

(73) Talitha L, Feenstra TL, van Genugten MLL, Hoogenveen RT, Wouters EF, van Molken MPMH. The impact of aging and smoking on the future burden of chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2001; 164:590-596.

(74) Stang P, Lydick E, Silberman C, Kempel A, Keating ET. The Prevalence of COPD. *Chest* 2000; 117(5 suppl 2):354S-359S.

(75) Lopez AD, Shibuya K, Rao C, Mathers CD, Hansell AL, Held LS et al. Chronic obstructive pulmonary disease: current burden and future projections. *Eur Respir J* 2006; 27(2):397-412.

(76) Primatesta P, Stamatakis E. Health Survey for England 2001: Respiratory symptoms, atopic conditions and lung function. Health Survey for England 2001: Respiratory symptoms, atopic conditions and lung function 2003 http://www.archive2.official-documents.co.uk/document/deps/doh/survey01/rsac/rsac01.htm

(77) *Br* Thoracic Society. BTS Guidelines for the Management of Chronic Obstructive Pulmonary Disease. *Thorax* 1997; 52(90005):S1-28.

(78) Falaschetti E, Laiho J, Primatesta P, Purdon S. Reference equations for normal and low lung function from the Health Survey for England. *Eur Respir J* 2004; 23:456-463.

(79) Department of Communities & Local Government. The English Indices of Deprivation 2004 (revised). Department for Communities & Local Government 2008 [cited 2008 July 8]; http://www.communities.gov.uk/publications/communities/englishindices

(80) Office for National Statistics. *Pop* Estimates by Ethnic Group (experimental). National Statistics Online 2007 [cited 2008 July 8]; http://www.statistics.gov.uk/statbase/Product.asp?vlnk=14238

(81) Goddard E. Smoking and drinking among adults, 2005. ONS website 2006 [cited 2010 Sept. 23]; http://www.statistics.gov.uk/statbase/product.asp?vlnk=5756

(82) Halbert RJ, Isonaka S, George D, Iqbal A. Interpreting COPD prevalence est imates. What is the true burden of disease? *Chest* 2003; 123(5):1684-1692.

(83) Walters JA, Hansen EC, Johns DP, Blizzard EL, Walters EH, Wood-Baker R. A mixed methods study to compare models of spirometry delivery in primary care for patients at risk of COPD. *Thorax* 2008; 63:408-414.

(84) Shahab L, Jarvis MJ, Britton J, West R. Prevalence, diagnosis and relation to tobacco dependence of chronic obstructive pulmonary disease in a nationally representative population sample. *Thorax* 2006; 61:1043-1047.

(85) Bakke PS, Baste V, Hanoa R, Gulsvik A. Prevalence of obstructive lung disease in a general population: relation to occupational title and exposure to some airborne agents. *Thorax* 1991; 46:863-870.

(86) Mannino DM, Gagnon RC, Petty TL, Lydick E. Obstructive lung disease and low lung function in adults in the United States. *Arch Intern Med* 2000; 160:1683-1689.

(87) Jordan LM, Berra JCM, Inigo MC, Diez L, Garmendia Z. Enfermedad pulmonary obstructive cronica en la poblacion general. Estudio epidemiologico realizado em Guipuzcoa. *Arch Bronchoneumol* 1998; 34(1):23-27.

(88) Murtagh E, Heaney L, Gingles J, Shepherd R, Kee F, Patterson C et al. The prevalence of obstructive lung disease in a general population sample: The NICECOPD study. *Eur J Epidemiology* 2005; 20(5):443-453.

(89) National Institute for Clinical Excellence. CG101 Chronic obstructive pulmonary disease (update): full guideline. National Institute for Clinical Excellence website 2010 [cited 2010 Sept. 23]; http://guidance.nice.org.uk/CG101/Guidance/pdf/English

(90) Department of Health. Lung Function. The Health Survey for England 2001. The Stationery Office; 2010.

(91) Pauwels R, Buist AS, Calverley PMA, JENKINS CR, HURD SS. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease . NHLBI/WHO Global Initiative for Chronic Obstructive Lung Disease (GOLD) Workshop Summary. *Am J Respir Crit Care Med* 2001; 163(5):1256-1276.

(92) Department of Health. Equitable Access to Primary Medical Care (EAPMC). Department of Health website 2010 http://www.dh.gov.uk/en/Procurementandproposals/Procurement/ProcurementatPCTs/index.htm

(93) Information Centre. Health Survey for England 2006 Latest Trends, Table 12: Prevalence of IHD, stroke, IHD or stroke (ever), by survey year, age and sex. The Information Centre 2008 [cited 2008 Apr. 25]; http://www.ic.nhs.uk/webfiles/publications/HSE06/ADULT%20TREND%20TABLES%202006.xls

(94) Information Centre for Health & Social Care. The Quality & Outcomes Framework. Information Centre website 2009 [cited 2010 Jan. 24]; http://www.ic.nhs.uk/statistics-and-data-collections/audits-and-performance/the-quality-and-outcomes-framework

(95) De Bacquer D, De Backer G, Kornitzer M. Prevalences of ECG findings in large population based samples of men and women. *Heart* 2000; 84(6):625-633.

(96) Lampe FC, Walker M, Lennon LT, Whincup PH, Ebrahim S. Validity of a Self-reported History of Doctor-diagnosed Angina. *J Clin Epidemiol* 1999; 52(1):73-81.

(97) Walker MK, Whincup PH, Shaper AG, Lennon LT, Thomson AG. Validation of Patient Recall of Doctor-diagnosed Heart Attack and Stroke: A Postal Questionnaire and Record Review Comparison. *Am J Epidemiol* 1998; 148(4):355-361.

(98) Shaper AG, Cook DG, Walker M, Macfarlane PW. Prevalence of ischaemic heart disease in middle aged *Br* men. *Br Heart J* 1984; 51(6):595-605.

(99) Shaper AG, Cook DG, Walker M, Macfarlane PW. Recall of diagnosis by men with ischaemic heart disease. *Br Heart J* 1984; 51(6):606-611.

(100) Lampe FC, Morris RW, Whincup PH, Walker M, Ebrahim S, Shaper AG. Is the prevalence of coronary heart disease falling in *Br* men? *Heart* 2001; 86(5):499-505.

(101) Alexander CM, Landsman PB, Teutsch SM, Haffner SM. NCEP-defined metabolic syndrome, diabetes, and prevalence of coronary heart disease among NHANES III participants age 50 years and older. *Diabetes* 2003; 52(5):1210.

(102) Ford ES, Giles WH, Croft JB. Prevalence of nonfatal coronary heart disease among *Am* adults. *The Am heart journal* 2000; 139(3):371-377.

(103) Goldman L, Phillips KA, Coxson P, Goldman PA, Williams L, Hunink MGM et al. The effect of risk factor reductions between 1981 and 1990 on coronary heart disease incidence, prevalence, mortality and cost. *J Am College of Cardiology* 2001; 38(4):1012-1017.

(104) Tunstall-Pedoe H. Contribution of trends in survival and coronary-event rates to changes in coronary heart disease mortality: 10-year results from 37 WHO MONICA project populations. Monitoring trends and determinants in cardiovascular disease. *Lancet, The* 1999; 353(9164):1547-1557.

(105) Gupta R, Prakash H, Gupta VP, Gupta KD. Prevalence and determinants of coronary heart disease in a rural population of India. *J Clin Epidemiol* 1997; 50(2):203-209.

(106) Ahto M, Isoaho R, Puolijoki H, Laippala P, Romo M, KIVELA SL. Prevalence of coronary heart disease, associated manifestations and electrocardiographic findings in elderly Finns. *Age Ageing* 1998; 27(6):729-738.

(107) Hemingway H, Langenberg C, Damant J, Frost C, Pyorala K, Barrett-Connor E. Prevalence of Angina in Women Versus Men: A Systematic Review and Meta-Analysis of International Variations Across 31 Countries. *Circulation* 2008; 117(12):1526-1536.

(108) Carroll K, Majeed A, Firth C, Gray J. Prevalence and management of coronary heart disease in primary care: population-based cross-sectional study using a disease register. *J Public Health* 2003; 25(1):29-35.

(109) Davies AR, Smeeth L, Grundy EMD. Contribution of changes in incidence and mortality to trends in the prevalence of coronary heart disease in the UK: 1996 2005. *Eur Heart J* 2007; 28(17):2142-2147.

(110) Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ* 2007; 335(7611):136.

(111) Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* 2008; 336(7659):1475-1482.

(112) Mozaffarian D, Wilson PWF, Kannel WB. Beyond Established and Novel Risk Factors: Lifestyle Risk Factors for Cardiovascular Disease. *Circulation* 2008; 117(23):3031-3038.

(113) Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143(1):29-36.

(114) Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 1983; 148(3):839-843.

(115) Zou KH, O'Malley AJ, Mauri L. Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. *Circulation* 2007; 115(5):654-657.

(116) Association of Public Health Observatories. Browsing Disease prevalence models. 2010 [cited 3/5/ 2010]; http://www.apho.org.uk/resource/view.aspx?RID=48308

(117) Truelsen T, Piechowski J, Bonita R, Mathers C, Bogousslavsky J, Boysen G. Stroke incidence and prevalence in Europe: a review of available data. *Eur J Neurology* 2006; 13(6):581-598.

(118) Feigin VL, Lawes CM, Bennett DA, Anderson CS. Stroke epidemiology: a review of population-based studies of incidence, prevalence, and case-fatality in the late 20th century. *The Lancet Neurology* 2003; 2(1):43-53.

(119) Saposnik G, Del Brutto OH. Stroke in South America: A Systematic Review of Incidence, Prevalence, and Stroke Subtypes. *Stroke* 2003; 34(9):2103-2107.

(120) Jungehulsing GJ. Prevalence of stroke and stroke symptoms: a population-based survey of 28,090 participants. *Neuroepidemiology* 2008; 30(1):51.

(121) Geddes JM, Fear J, Tennant A, Pickering A, Hillman M, Chamberlain MA. Prevalence of self reported stroke in a population in northern England. *J Epidemiol Community Health* 1996; 50(2):140-143.

(122) MacDonald BK, Cockerell OC, Sander JWAS, Shorvon SD. The incidence and lifetime prevalence of neurological disorders in a prospective community-based study in the UK. *Brain* 2000; 123(4):665-676.

(123) O'Mahony PG, Thomson RG, Dobson R, Rodgers H, James OFW. The prevalence of stroke and associated disability. *J Public Health* 1999; 21(2):166-171.

(124) Sproston K, Primatesta P. Health Survey for England 2003: Cardiovascular Disease. The Stationery Office 2008 http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsStatistics/DH_4098712

(125) Heuschmann PU, Grieve AP, Toschke AM, Rudd AG, Wolfe CDA. Ethnic Group Disparities in 10-Year Trends in Stroke Incidence and Vascular Risk Factors: The South London Stroke Register (SLSR). *Stroke* 2008; 39(8):2204-2210.

(126) Smeeton NC, Heuschmann PU, Rudd AG, McEvoy AW, Kitchen ND, Sarker SJ et al. Incidence of Hemorrhagic Stroke in Black Caribbean, Black African, and White *Pops*: The South London Stroke Register, 1995 2004. *Stroke* 2007; 38(12):3133-3138.

(127) Tilling K, Sterne JA, Wolfe CD. Estimation of the incidence of stroke using a capture-recapture model including covariates. *Int J Epidemiol* 2001; 30(6):1351-1359.

(128) National Institute for Health and Clinical Excellence. Hypertension: Management of hypertension in adults in primary care. NICE clinical guideline 34; National Institute for Health and Clinical Excellence 2006 [cited 2008 May 13]; http://www.nice.org.uk/nicemedia/pdf/CG034NICEguideline.doc

(129) Kearney PM, Whelton M, Reynolds K, Muntner P, Whelton PK, He J. Global burden of hypertension: analysis of worldwide data. *The Lancet* 2005; 365(9455):217-223.

(130) Lawes CM, Hoorn SV, Rodgers A. Global burden of blood-pressure-related disease, 2001. *The Lancet* 2008; 371(9623):1513-1518.

(131) Lawes CMM, Bennett DA, Feigin VL, Rodgers A. Blood Pressure and Stroke: An Overview of Published Reviews. *Stroke* 2004; 35(3):776-785.

(132) Blood Pressure Lowering Treatment Trialists' Collaboration. Effects of different regimens to lower blood pressure on major cardiovascular events in older and younger adults: meta-analysis of randomised trials. *BMJ* 2008; 336(7653):1121-1123.

(133) Hajat C, Tilling K, Stewart JA, Lemic-Stojcevic N, Wolfe CDA. Ethnic Differences in Risk Factors for Ischemic Stroke: A *Eur* Case-Control Study. *Stroke* 2004; 35(7):1562-1567.

(134) Whelton PK, He J, Muntner P. Prevalence, awareness, treatment and control of hypertension in North America, North Africa and Asia. *J Hum Hypertens* 2004; 18(8):545-551.

(135) Wolf-Maier K, Cooper RS, Banegas JR, Giampaoli S, Hense HW, Joffres M et al. Hypertension Prevalence and Blood Pressure Levels in 6 *Eur* Countries, Canada, and the United States. *JAMA* 2003; 289(18):2363-2369.

(136) Singh RB. Hypertension and stroke in Asia: prevalence, control and strategies in developing countries for prevention. *J human hypertension* 2000; 14(10-11):749.

(137) Chen X, Wang Y. Tracking of Blood Pressure From Childhood to Adulthood. A Systematic Review and Meta-Regression Analysis. *Circulation* 2008;CIRCULATIONAHA.

(138) Droyvold WB, Midthjell K, Nilsen TIL, Holmen J. Change in body mass index and its impact on blood pressure: a prospective population study. *Int J Obes Relat Metab Disord* 2005; 29(6):650-655.

(139) Thorpe J, Brandon DT, LaVeist TA. Social context as an explanation for race disparities in hypertension: Findings from the Exploring Health Disparities in Integrated Communities (EHDIC) Study. *Social Science & Medicine* 2008; 67(10):1604-1611.

(140) Hardy R. Age at puberty and adult blood pressure and body size in a *Br* birth cohort study. *J hypertension* 2006; 24(1):59.

(141) Beckett NS, Peters R, Fletcher AE, Staessen JA, Liu L, Dumitrascu D et al. Treatment of Hypertension in Patients 80 Years of Age or Older. *N Engl J Med* 2008; 358(18):1887-1898.

(142) Strong K, Mathers C, Bonita R. Preventing stroke: saving lives around the world. *The Lancet Neurology* 2007; 6(2):182-187.

(143) Joffres MR, Ghadirian P, Fodor JG, Petrasovits A, Chockalingam A, Hamet P. Awareness, Treatment, and Control of Hypertension in Canada. *Am J Hypertension* 1997; 10(10):1097-1102.

(144) Brindel P. Prevalence, awareness, treatment, and control of hypertension in the elderly: the Three City study. *J hypertension* 2006; 24(1):51.

(145) Angell SY, Garg RK, Gwynn RC, Bash L, Thorpe LE, Frieden TR. Prevalence, Awareness, Treatment, and Predictors of Control of Hypertension in New York City. *Circ Cardiovasc Qual Outcomes* 2008; 1(1):46-53.

(146) Wang YR, Alexander GC, Stafford RS. Outpatient Hypertension Treatment, Treatment Intensification, and Control in Western Europe and the United States. *Arch Intern Med* 2007; 167(2):141-147.

(147) Wang Y, Wang QJ. The Prevalence of Prehypertension and Hypertension Among US Adults According to the New Joint National Committee Guidelines: New Challenges of the Old Problem. *Arch Intern Med* 2004; 164(19):2126-2134.

(148) Wald NJ, Law MR. A strategy to reduce cardiovascular disease by more than 80%. *BMJ* 2003; 326(7404):1419.

(149) Ezzati M, Oza S, Danaei G, Murray CJL. Trends and Cardiovascular Mortality Effects of State-Level Blood Pressure and Uncontrolled Hypertension in the United States. *Circulation* 2008; 117(7):905-914.

(150) Craig R, Mindell J. Health Survey for England 2006 Latest Trends. The Information Centre for Health & Social Care 2008 [cited 2008 Nov. 14]; http://www.ic.nhs.uk/statistics-and-data-collections/health-and-lifestyles-related-surveys/health-survey-for-england/health-survey-for-england-2006-latest-trends

(151) National Center for Health Statistics. National Health and Nutrition Examination Survey (NHANES). Centers for Disease Control 2010 [cited 2010 Oct. 4]; http://www.cdc.gov/nchs/nhanes/nhanes2001-2002/nhanes01_02.htm

(152) Gillum RF, Feinleib M, Margolis JR, Fabsitz RR, Brasch RC. Community surveillance for cardiovascular disease: The Framingham Cardiovascular Disease Survey : Some methodological problems in the community study of cardiovascular disease. *J Chronic Diseases* 1976; 29(5):289-299.

(153) Anderson KM, Wilson PW, Odell PM, Kannel WB. An updated coronary risk profile. A statement for health professionals. *Circulation* 1991; 83(1):356-362.

(154) Conroy RM, Pyorala K, Fitzgerald AP, Sans S, Menotti A, De Backer G et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *Eur Heart J* 2003; 24(11):987-1003.

(155) Tunstall-Pedoe H, Kuulasmaa K, Amouyel P, Arveiler D, Rajakangas AM, Pajak A. Myocardial infarction and coronary deaths in the World Health Organization MONICA

Project. Registration procedures, event rates, and case-fatality rates in 38 populations from 21 countries in four continents. *Circulation* 1994; 90(1):583-612.

(156) Chan WC, Wright C, Riddell T, Wells S, Kerr AJ, Jackson R. Ethnic and socioeconomic disparities in the prevalence of cardiovascular disease in New Zealand. *New Zealand Medical Journal* 2008; 121(1285):11.

(157) Wells S. Estimated prevalence of cardiovascular disease and distribution of cardiovascular risk in New Zealanders: data for healthcare planners, funders, and providers. *New Zealand medical journal* 2006; 119(1232):U1935.

(158) Kirkwood B, Sterne JAC. Essential Medical Statistics. Second ed. Oxford: Blackwell Science; 2003.

(159) Department of Health. Putting prevention first- vascular checks: risk assessment and management - next steps guidance for primary care trusts. [cited 13-11-2009]; http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_097489

(160) McFadden E, Luben R, Wareham N, Bingham S, Khaw KT. Occupational social class, risk factors and cardiovascular disease incidence in men and women: a prospective study in the *Eur* Prospective Investigation of Cancer and Nutrition in Norfolk (EPIC-Norfolk) cohort. *Eur J Epidemiology* 2008; 23(7):449-458.

(161) Department of Health. Putting prevention first- vascular checks: risk assessment and management - next steps guidance for primary care trusts. [cited 13-11-2009]; http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_097489

(162) National Institute for Clinical Excellence. Cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease. NICE website 2010 [cited 2010 Sept. 30]; http://guidance.nice.org.uk/CG67

(163) Jackson R, Lawes CM, Bennett DA, Milne RJ, Rodgers A. Treatment with drugs to lower blood pressure and blood cholesterol based on an individual's absolute cardiovascular risk. *The Lancet* 2005; 365(9457):434-441.

(164) Manuel DG, Lim J, Tanuseputro P, Anderson GM, Alter DA, Laupacis A et al. Revisiting Rose: strategies for reducing coronary heart disease. *BMJ* 2006; 332(7542):659-662.

(165) Marshall T, Westerby P, Chen J, Fairfield M, Harding J, Westerby R et al. The Sandwell Project: A controlled evaluation of a programme of targeted screening for prevention of cardiovascular disease in primary care. *BMC Public Health* 2008; 8(1):73.

(166) DH/Medical/Vascular Policy Team. Economic modelling for vascular checks. Department of Health 2008 [cited 2009 Mar. 12]; http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_085869

(167) Chamnan P, Simmons RK, Khaw KT, Wareham NJ, Griffin SJ. Estimating the population impact of screening strategies for identifying and treating people at high risk of cardiovascular disease: modelling study. *BMJ* 2010; 340(apr23_2):c1693.

(168) Lawson KD, Fenwick EAL, Pell ACH, Pell JP. Comparison of mass and targeted screening strategies for cardiovascular risk: simulation of the effectiveness, cost-effectiveness and coverage using a cross-sectional survey of 3921 people. *Heart* 2010; 96(3):208-212.

(169) Marshall T. The value of information: modelling the efficiency of strategies for identifying high-risk patients for primary prevention of cardiovascular disease. *Informatics in Primary Care* 2006; 14:85-92.

(170) Donders AR, van der Heijden GJMG, Stijnen T, Moons KGM. Review: A gentle introduction to imputation of missing values. *J Clin Epidemiol* 2006; 59(10):1087-1091.

(171) Jefrey LH, Amylou D. Handling Missing Data. Current problems in cancer 29[6], 317-325. 1-11-2005.

(172) Royston P. Multiple imputation of missing values: update. *Stata Journal* 2005; 5(2):188-201.

(173) Clark TG, Altman DG. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. *J Clin Epidemiol* 2003; 56(1):28-37.

(174) Hippisley-Cox J. QRISK2 algorithm. http://www qresearch org/PowerPointpresentations/QRISK2%20algorithm pdf 2009

(175) Department of Health. Practice Baced Commisioning Toolkit. http://www dh gov uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_094364 2010

(176) Batsis J, Lopez-Jimenez F. Cardiovascular risk assessment - From individual risk prediction to estimation of global risk and change in risk in the population. *BMC Medicine* 2010; 8(1):29.

(177) Dalton ARH, Bottle A, Okoro C, Majeed A, Millett C. Implementation of the NHS Health Checks programme: baseline assessment of risk factor recording in an urban culturally diverse setting. *Fam Pract* online 2010 10.1093/fampra/cmq068.

(178) Waugh N, Scotland G, McNamee P, Gillett M, Brennan A, Goyder E et al. Screening for type 2 diabetes: literature review and economic modelling. *Health Technology Assessment* 2007; 11(17):1-+.

(179) Gillies CL, Abrams KR, Lambert PC, Cooper NJ, Sutton AJ, Hsu RT et al. Pharmacological and lifestyle interventions to prevent or delay type 2 diabetes in people with impaired glucose tolerance: systematic review and meta-analysis. *BMJ* 2007; 334(7588):299.

(180) Lindstrom J, Ilanne-Parikka P, Peltonen M, Aunola S, Eriksson JG, Hemio K et al. Sustained reduction in the incidence of type 2 diabetes by lifestyle intervention: follow-up of the Finnish Diabetes Prevention Study. *The Lancet* 2006; 368(9548):1673-1679.

(181) Alberti KGMM, Zimmet P, Shaw J. International Diabetes Federation: a consensus on Type 2 diabetes prevention. *Diabetic Medicine* 2007; 24(5):451-463.

(182) Reynolds TM, Twomey PJ, Wierzbicki AS. Concordance evaluation of coronary risk scores: implications for cardiovascular risk screening. *Current medical research and opinion* 2004; 20(6):811-818.

(183) Pierce JP. International comparisons of trends in cigarette smoking prevalence. *Am J Public Health* 1989; 79(2):152-157.

(184) Craig R, Mindell J. Health Survey for England 2006: Volume 1 CVD and risk factors adults, obesity and risk factors children. Information Centre for Health & Social Care 2008 [cited 2008 Nov. 26]; http://www.ic.nhs.uk/webfiles/publications/HSE06/HSE%2006%20report%20VOL%201%20v2.pdf

(185) Marshall T. Identification of patients for clinical risk assessment by prediction of cardiovascular risk using default risk factor values. *BMC Public Health* 2008; 8(1):25.

(186) Horgan JMP, Blenkinsopp A, McManus RJ. Evaluation of a cardiovascular disease opportunistic risk assessment pilot ("Heart MOT service") in community pharmacies. *J Public Health* 2010; 32(1):110-116.

(187) Brindle P, Beswick A, Fahey T, Ebrahim S. Accuracy and impact of risk assessment in the primary prevention of cardiovascular disease: a systematic review. *Heart* 2006; 92(12):1752-1759.

(188) Collins GS, Altman DG. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. *BMJ* 2010; 340(may13_2):c2442.

(189) Campbell I. The obesity epidemic: can we turn the tide?. *Heart* 2003; 89(suppl 2):ii22-ii24.

(190) Marshall T. The use of cardiovascular risk factor information in practice databases: making the best of patient data. *Br J General Practice* 2006; 8:600-605.

(191) Ray KK, Seshasai SRK, Erqou S, Sever P, Jukema JW, Ford I et al. Statins and All-Cause Mortality in High-Risk Primary Prevention: A Meta-analysis of 11 Randomized Controlled Trials Involving 65 229 Participants. *Arch Intern Med* 2010; 170(12):1024-1031.

(192) Jackson R. Cardiovascular risk prediction: are we there yet? *Heart* 2008; 94(1):1-3.

(193) NICE. National clinical guideline on management of COPD in primary & secondary care: Diagnosing COPD. Thorax 2004 59(90001):[i27-i38] http://www.nice.org.uk/nicemedia/pdf/CG012_niceguideline.pdf

(194) Stevens A, Raftery J, Mant J. Health Care Needs Assessment. 2nd ed. Abingdon: Radcliffe Publishing; 2004.

(195) Lakhani A, Olearnik H, Eayres D. Clinical and Health Outcomes Knowledge Base: Mortality from bronchitis, emphysema and other chronic obstructive pulmonary disease. The Information Centre for health and social care / National Centre for Health Outcomes Development 2008 [cited 2008 July 8]; http://www.nchod.nhs.uk/

(196) NHS Strategic Tracing Service. NHS Connecting for Health 2008 [cited 2008 July 8]; http://www.connectingforhealth.nhs.uk/systemsandservices/nsts?searchterm=NHS+Strategic+Tracing+Service+

(197) Anselin L. Local indicators of spatial association: LISA. *Geographical Analysis* 2009; 27:93-115.

(198) Brunsdon C, Fotheringham S, Charlton M. Geographically weighted regression--modelling spatial non-stationarity. *J Royal Statistical Society: Series D (The Statistician)* 1998; 47(3):431.

(199) Fotheringham AS, Brunsdon C, Charlton M. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Chichester: Wiley; 2002.

(200) Nakaya T, Fotheringham AS, Brunsdon C, Charlton M. Geographically weighted Poisson regression for disease association mapping. *Statistics in Medicine* 2005; 24(17):2695-2717.

(201) Ricketts TC, Holmes GM. Mortality and Physician Supply: Does Region Hold the Key to the Paradox? *Health Services Research* 2007; 42(6p1):2233-2251.

(202) Akaike H. A new look at statistical model identification. *Automatic Control, IEEE Transactions on* 1974; 19(6):716-723.

(203) Information Centre for Health & Social Care. NHS Staff 1996 - 2006 (General Practice). Information Centre for health & social care website 2010 [cited 2010 Sept. 30]; http://www.ic.nhs.uk/statistics-and-data-collections/workforce/nhs-staff-numbers/nhs-staff-1996--2006-general-practice

(204) Mohangoo AD, van der Linden MW, Schellevis FG, Raat H. Prevalence estimates of asthma or COPD from a health interview survey and from general practitioner registration: what's the difference? *Eur J Public Health* 2006; 16(1):101-105.

(205) Vandevoorde J, Verbanck S, Gijssels L, Schuermans D, Devroey D, De Backer J et al. Early detection of COPD: A case finding study in general practice. *Respiratory Medicine* 2007; 101(3):525-530.

(206) Tinkelman DG, Price D, Nordyke RJ, Halbert RJ. COPD screening efforts in primary care: what is the yield? *Primary Care Resp J* 2007; 16(1):41-48.

(207) National Clinical Guideline Centre. Management of adults with chronic obstructive pulmonary disease in primary and secondary care: update guideline (DRAFT). NICE website 2009 [cited 2010 Jan. 28]; http://www.nice.org.uk/nicemedia/pdf/COPDDraftConsultationGuideline.pdf

(208) Calverley PMA, Anderson JA, Celli B, Ferguson GT, Jenkins C, Jones PW et al. Salmeterol and Fluticasone Propionate and Survival in Chronic Obstructive Pulmonary Disease. *N Engl J Med* 2007; 356(8):775-789.

(209) Sutherland ER, Allmers H, Ayas NT, Venn AJ, Martin RJ. Inhaled corticosteroids reduce the progression of airflow limitation in chronic obstructive pulmonary disease: a meta-analysis. *Thorax* 2003; 58(11):937-941.

(210) Sridhar M, Taylor R, Dawson S, Roberts NJ, Partridge MR. A nurse led intermediate care package in patients who have been hospitalised with an acute exacerbation of chronic obstructive pulmonary disease. *Thorax* 2008; 63(3):194-200.

(211) Vollmer WM, Gislason T, Burney P, Enright PL, Gulsvik A, Kocabas A et al. Comparison of spirometry criteria for the diagnosis of COPD: results from the BOLD study. *Eur Respir J* 2009; 34(3):588-597.

(212) Cox AM, McKevitt C, Rudd AG, Wolfe CD. Socioeconomic status and stroke. *The Lancet Neurology* 2006; 5(2):181-188.

(213) McFadden E, Luben R, Wareham N, Bingham S, Khaw KT. Social Class, Risk Factors, and Stroke Incidence in Men and Women. A Prospective Study in the *Eur* Prospective Investigation Into Cancer in Norfolk Cohort. *Stroke* 2009; 40(4):1070-1077.

(214) Morris RW, Whincup PH, Lampe FC, Walker M, Wannamethee SG, Shaper AG. Geographic variation in incidence of coronary heart disease in Britain: the contribution of established risk factors. *Heart* 2001; 86(3):277-283.

(215) Doran T, Drever F, Whitehead M. Health underachievement and overachievement in English local authorities. *J Epidemiol Community Health* 2006; 60(8):686-693.

(216) Levene LS, Baker R, Bankart MJ, Khunti K. Association of Features of Primary Health Care With Coronary Heart Disease Mortality. *JAMA* 2010; 304(18):2028-2034.

(217) Clark AM, Hartling L, Vandermeer B, McAlister FA. Meta-Analysis: Secondary Prevention Programs for Patients with Coronary Artery Disease. *Annals of Internal Medicine* 2005; 143(9):659-672.

(218) Ashworth M, Medina J, Morgan M. Effect of social deprivation on blood pressure monitoring and control in England: a survey of data from the quality and outcomes framework. *BMJ* 2008; 337(oct28_2):a2030.

(219) Ward P, Noyce P, Leger A. Are GP practice prescribing rates for coronary heart disease drugs equitable? A cross sectional analysis in four primary care trusts in England. *J Epidemiol Community Health* 2004; 58:89-96.

(220) Information Centre. Quality and Outcomes Framework. The Information Centre 2009 [cited 2008 Apr. 25]; http://www.ic.nhs.uk/statistics-and-data-collections/audits-and-performance/the-quality-and-outcomes-framework

(221) Campbell SM, Reeves D, Kontopantelis E, Sibbald B, Roland M. Effects of Pay for Performance on the Quality of Primary Care in England. *N Engl J Med* 2009; 361(4):368-378.

(222) Informing Healthier Choices Programme. Browsing Disease prevalence models. Association of Public Health Observatories website 2009 [cited 2010 Mar. 10]; http://www.apho.org.uk/resource/view.aspx?RID=48308

(223) Connecting for Health. The Personal Demographics Service. Connecting for Health 2010 [cited 2010 Sept. 30]; http://www.connectingforhealth.nhs.uk/systemsandservices/demographics/pds

(224) Glymour MM, Avendano M. Can Self-Reported Strokes Be Used to Study Stroke Incidence and Risk Factors?: Evidence From the Health and Retirement Study. *Stroke* 2009; 40(3):873-879.

(225) Kazumasa Y, Ai I, Hiroyasu I, Manami I, Shoichiro T. Self-reported stroke and myocardial infarction had adequate sensitivity in a population-based prospective study. J clinical epidemiology 62[6], 667-673. 1-6-2009.

(226) O'Mahony PG, Dobson R, Rodgers H, James OFW, Thomson RG. Validation of a *Pop* Screening Questionnaire to Assess Prevalence of Stroke. *Stroke* 1995; 26(8):1334-1337.

(227) Office for National Statistics. Model Based Estimate for Smoking in Healthy Lifestyle Behaviours: Model Based Estimates, 2003-2005. Neighbourhood Statistics 2007 http://neighbourhood.statistics.gov.uk/dissemination/viewFullDataset.do?$ph=60_61_62&step=4&productId=969&instanceSelection=022951&timeId=242&containerAreaId=6059484&startColumn=1&numberOfColumns=8&viewAction=parent

(228) Gray J, Majeed A, Kerry S, Rowlands G. Identifying patients with ischaemic heart disease in general practice: cross sectional study of paper and computerised medical records. *BMJ* 2000; 321(7260):548-550.

(229) Donnan PT, Dougall HT, Sullivan FM. Optimal strategies for identifying patients with myocardial infarction in general practice. *Fam Pract* 2003; 20(6):706-710.

(230) Millett C, Zelenyanszki C, Binysh K, Lancaster J, Majeed A. *Pop* mobility: characteristics of people registering with general practices. *Public Health* 2005; 119(7):632-638.

(231) Goddard M, Gravelle H, Hole A, Marini G. Where did all the GPs go? Increasing supply and geographical equity in England and Scotland. *J Health Serv Res Policy* 2010; 15(1):28-35.

(232) Morris RW, Whincup PH, Papacosta O, Walker M, Thomson A. Inequalities in coronary revascularisation during the 1990s: evidence from the *Br* regional heart study. *Heart* 2005; 91(5):635-640.

(233) *Br* Lung Foundation. Invisible Lives: Chronic Obstructive Pulmonary Disease (COPD)-finding the missing millions. *Br* Lung Foundation website 2007 [cited 2010 Oct. 8]; http://www.lunguk.org/Resources/*Br*%20Lung%20Foundation/Migrated%20Resources/Documents/I/Invisible%20Lives%20report.pdf

(234) Bednarek M, Maciejewski J, Wozniak M, Kuca P, Zielinski J. Prevalence, severity and underdiagnosis of COPD in the primary care setting. *Thorax* 2008; 63(5):402-407.

(235) Stratelis G, Jakobsson P, Molstad S, Zetterstrom O. Early detection of COPD in primary care: screening by invitation of smokers aged 40 to 55 years. *Br J General Practice* 2004; 54:201-206.

(236) Falzon C, Elkin SL, Kelly JL, Lynch F, Blake ID, Hopkinson NS. Can financial incentives for improvements in healthcare quality enhance identification of COPD in primary care? *Thorax* online 2010. *thx.2010.140913*

(237) Information Centre for Health & Social Care. NHS Comparators. Information Centre for health & social care website 2010 [cited 2010 Oct. 6]; https://www.nhscomparators.nhs.uk/NHSComparators/Login.aspx

(238) West P, Tatlock S. Evaluating the Use of APHO Disease Prevalence Models in PCTs. York Health Economics Consortium 2009.

(239) Vascular Programme. Vascular Programme briefing packs. Department of Health website 2009 [cited 2010 Oct. 8]; http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_098627

(240) Walter SD. Local estimates of population attributable risk. *J Clin Epidemiol* 2010; 63(1):85-93.

(241) Uter W, Pfahlberg A. The application of methods to quantify attributable risk in medical practice. *Statistical Methods in Medical Research* 2001; 10(3):231-237.

(242) Graubard BI, Flegal KM, Williamson DF, Gail MH. Estimation of attributable number of deaths and standard errors from simple and complex sampled cohorts. *Statist Med* 2007; 26(13):2639-2649.

(243) Rodgers A, Ezzati M, Vander Hoorn S, Lopez AD, Lin RB, Murray CJL et al. Distribution of Major Health Risks: Findings from the Global Burden of Disease Study. *PLoS Med* 2004; 1(1):e27.

(244) Rodgers A. Reducing the global burden of blood pressure-related cardiovascular disease. *J hypertension Supplement* 2000; 18(1):S3.

(245) Eide GE, Gefeller O. Sequential and average attributable fractions as aids in the selection of preventive strategies. *J Clin Epidemiol* 1995; 48(5):645-655.

(246) Eide GE, Heuch I. Attributable fractions: fundamental concepts and their visualization. *Statistical Methods in Medical Research* 2001; 10(3):159-193.

(247) Iso H, Date C, Yamamoto A, Toyoshima H, Watanabe Y, Kikuchi S et al. Smoking Cessation and Mortality from Cardiovascular Disease among Japanese Men and Women: The JACC Study. *Am J Epidemiol* 2005; 161(2):170-179.

(248) Kenfield SA, Stampfer MJ, Rosner BA, Colditz GA. Smoking and Smoking Cessation in Relation to Mortality in Women. *JAMA* 2008; 299(17):2037-2047.

(249) Care Quality Commission. Closing the gap: Tackling cardiovascular disease and health inequalities by prescribing statins and stop smoking services. Care Quality Commission website 2009 [cited 2010 Feb. 2]; http://www.cqc.org.uk/_db/_documents/Closing_the_gap.pdf

(250) Calderon A, Carney L, Soljak M, Bottle A, Partridge M, Bell D et al. Association of population and primary healthcare factors with hospital admission rates for chronic obstructive pulmonary disease in England: national cross-sectional study. *Thorax* 2010; 10.1136/thx.2010.147058 .

(251) Healthcare Commission. Clearing the air: A national study of chronic obstructive pulmonary disease. Healthcare Commission website 2006 [cited 2010 Jan. 20]; http://www.cqc.org.uk/_db/_documents/COPD_factsheet.pdf

(252) Department of Health. Strategy for services for Chronic Obstructive Pulmonary Disease (COPD) in England - consultation document. Department of Health website 2010 [cited 2010 Sept. 29]; http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_113279.pdf

(253) Chancellor of the Exchequer. Pre-Budget Report 2009: Securing the recovery: growth and opportunity. HM Treasury website 2009 [cited 10 A.D. Mar. 20]; http://www.hm-treasury.gov.uk/prebud_pbr09_repindex.htm

(254) Health Intelligence Unit. Victorian ambulatory care sensitive conditions study. Victorian Government health information website 2009 [cited 2010 Jan. 8]; http://www.health.vic.gov.au/healthstatus/downloads/prelimanalyses.pdf

(255) NHS Institute for Improvement & Innovation. NHS Better Care, Better Value Indicators: Clinical productivity- managing variation in emergency admissions. NHS Institute website 2010 [cited 2010 Jan. 24]; http://www.productivity.nhs.uk/Def_ManagingVariationInEmergencyAdmissions.aspx

(256) Prevention Quality Indicators Composite Measure Workgroup. Guide to Prevention Quality Indicators: Hospital Admission for Ambulatory Care Sensitive Conditions. AHRQ Quality Indicators website 2006 [cited 2010 Jan. 8]; http://www.qualityindicators.ahrq.gov/downloads/pqi/pqi_guide_v31.pdf

(257) Purdy S, Griffin T, Salisbury C, Sharp D. Ambulatory care sensitive conditions: terminology and disease coding need to be more specific to aid policy makers and clinicians. *Public Health* 2009; 123(2):169-173.

(258) Sanderson C, Dixon J. Conditions for which onset or hospital admission is potentially preventable by timely and effective ambulatory care. *J Health Services Research and Policy* 2000; 5(4):222-230.

(259) Ansari Z, Laditka JN, Laditka SB. Access to Health Care and Hospitalization for Ambulatory Care Sensitive Conditions. *Med Care Res Rev* 2006; 63(6):719-741.

(260) Basu J, Friedman B, Burstin H. Primary care, HMO enrollment, and hospitalization for ambulatory care sensitive conditions. *Med Care* 2002; 40(12):1260-1269.

(261) Griffiths C, Sturdy P, Naish J, Omar R, Dolan S, Feder G. Hospital admissions for asthma in east london: associations with characteristics of local general practices, prescribing, and population. *BMJ* 1997; 314(7079):482.

(262) Saxena S, George J, Barber J, Fitzpatrick J, Majeed A. Association of population and practice factors with potentially avoidable admission rates for chronic diseases in London: cross sectional analysis. *J R Soc Med* 2006; 99(2):81-89.

(263) Adams SG, Smith PK, Allan PF, Anzueto A, Pugh JA, Cornell JE. Systematic Review of the Chronic Care Model in Chronic Obstructive Pulmonary Disease Prevention and Management. *Arch Intern Med* 2007; 167(6):551-561.

(264) Majeed A, Bardsley M, Morgan D, O'Sullivan C, Bindman AB. Cross sectional study of primary care groups in London: association of measures of socioeconomic and health status with hospital admission rates. *BMJ* 2000; 321(7268):1057-1060.

(265) Reid FDA, Cook DG, Majeed A. Explaining variation in hospital admission rates between general practices: cross sectional study. *BMJ* 1999; 319(7202):98-103.

(266) Audit Commission. PbR Data Assurance Framework. Audit Commission website [Internet] 2009 [cited 2010 May 3]; http://www.audit-commission.gov.uk/SiteCollectionDocuments/AuditCommissionReports/NationalStudies/20090827pbrdataassuranceframework0809rep.pdf

(267) Clinical Effectiveness and Evaluation Unit. National COPD Audit 2008. The Royal College of Physicians 2010 [cited 2010 Jan. 28]; http://www.rcplondon.ac.uk/clinical-standards/ceeu/Current-work/ncrop/Pages/audit.aspx

(268) Miravitlles M, Soriano JB, Garc+¡a-R+¡o F, Mu+¦oz L, Duran-Tauleria E, Sanchez G et al. Prevalence of COPD in Spain: impact of undiagnosed COPD on quality of life and daily life activities. *Thorax* 2009; 64(10):863-868.

(269) Hurst JR, Vestbo J+, Anzueto A, Locantore N, M++llerova H, Tal-Singer R et al. Susceptibility to Exacerbation in Chronic Obstructive Pulmonary Disease. *New England J Medicine* 2010; 363(12):1128-1138.

(270) Bottle A, Gnani S, Saxena S, Aylin P, Mainous A, Majeed A. Association Between Quality of Primary Care and Hospitalization for Coronary Heart Disease in England: National Cross-sectional Study. J General Int Med 23[2], 135-141. 1-2-2008. 8-1-2010.

(271) Wakefield JC, Best NG, Waller L. Bayesian approaches to disease mapping. In: Elliott P, Wakefield J, Best N, Briggs D, editors. Spatial Epidemiology: Methods and Application. Oxford University Press; 2000. 104-127.

(272) Department of Health. Lung Function. The Health Survey for England 2001. The Stationery Office; 2010. http://www.archive2.official-documents.co.uk/document/deps/doh/survey01/rsac/rsac01.htm

(273) Holowaty E, Norwood T, Wanigaratne S, Abellan J, Beale L. Feasibility and utility of mapping disease risk at the neighbourhood level within a Canadian public health unit: an ecological study. *Int J Health Geog* 2010; 9(1):21.

(274) Nacul L, Soljak M, Samarasundera E, Hopkinson NS, Lacerda E, Indulkar T et al. COPD in England: a comparison of expected, model-based prevalence and observed prevalence from general practice data. *J Public Health* 2010;fdq031.

(275) Anselin L. Spatial econometrics. *Palgrave Handbook of Econometrics* 2006; 1.

(276) Leung Y, Mei CL, Zhang WX. Statistical tests for spatial nonstationarity based on the geographically weighted regression model. *Environment and Planning A* 2000; 32(1):9-32.

(277) GeoDa Center for Geospatial Analysis and Computation. GeoDa Center & Affiliated Software. GeoDa Center website 2010 [cited 2010 Nov. 25]; http://geodacenter.asu.edu/software

(278) Fotheringham AS, Charlton M, Brunsdon C. Geographically Weighted Regression: software for GWR. GWR website 2010 [cited 2010 Nov. 25]; http://ncg.nuim.ie/ncg/GWR/index.htm

(279) Feuer EJ, Etzioni R, Cronin KA, Mariotto A. The use of modeling to understand the impact of screening on US mortality: examples from mammography and PSA testing. *Statistical Methods in Medical Research* 2004; 13(6):421-442.

(280) Best N, Richardson S, Thomson A. A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research* 2005; 14(1):35-59.

(281) Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Brindle P. Performance of the QRISK cardiovascular risk prediction algorithm in an independent UK sample of patients from general practice: a validation study. *Heart* 2008; 94(1):34-39.

(282) Weinstein MC, O'Brien B, Hornberger J, Jackson J, Johannesson M, McCabe C et al. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices-Modeling Studies. *Value in Health* 2003; 6:9-17.

(283) Kopec J, Fines P, Manuel D, Buckeridge D, Flanagan W, Oderkirk J et al. Validation of population-based disease simulation models: a review of concepts and methods. *BMC Public Health* 2010; 10(1):710.

(284) Jordan RE, Lam KbH, Cheng KK, Miller MR, Marsh JL, Ayres JG et al. Case finding for chronic obstructive pulmonary disease: a model for optimising a targeted approach. *Thorax* 2010; 65:492-498.

(285) Campbell S, Reeves D, Kontopantelis E, Middleton E, Sibbald B, Roland M. Quality of primary care in England with the introduction of pay for performance. *New England J Medicine 357(2):181-90,* 2007.

(286) Unal B, Critchley JA, Capewell S. Modelling the decline in coronary heart disease deaths in England and Wales, 1981-2000: comparing contributions from primary prevention and secondary prevention. *BMJ* 2005; 331(7517):614.

(287) Naidoo B, J C. Using the PREVENT model in Birmingham (Society for Social Medicine abstract). *J Epidemiol Community Health* 1998; 52:683.

(288) Harper PR, Yates J, Wilderspin H, Shahani AK, Bowie C, Duff C. *Pop* Needs Assessment and Access Modelling for Cardiac Services. *J Operational Research Society* 2004; 55(5):529-534.

(289) Srebotnjak T, Mokdad A, Murray C. A novel framework for validating and applying standardized small area measurement strategies. *Population Health Metrics* 2010; 8(1):26.

(290) Walford H, Kearns B, Barron S. APHO Technical Briefing 8: Prevalence Modelling. In: Bolton H, Cunningham A, Dancox M, Flowers J, Fryers P, Glover G, et al., editors. Online: Association of Public Health Observatories, 2011. http://www.apho.org.uk/resource/item.aspx?RID=100181.

(291) Hejblum G, Setbon M, Temime L, Lesieur S, Valleron A-J. Modelers' Perception of Mathematical Modeling in Epidemiology: A Web-Based Survey. *PLoS ONE* 2011;6(1):e16531.