UNIVERSITY OF SOUTHAMPTON

Faculty of Social, Human
and Mathematical Sciences

Statistical Models for Small Area Public Health Intelligence on Chronic Morbidity

by

Peter Dutey-Magni

Thesis submitted for the degree of Doctor of Philosophy

February 2017

University of Southampton

Abstract

Faculty of Social, Human and Mathematical Sciences

Thesis for the degree of Doctor of Philosophy

Statistical Models for Small Area Public Health Intelligence on Chronic Morbidity

Peter Dutey-Magni

Local indicators of chronic morbidity are needed to conduct needs assessments, plan health care services, allocate funds and monitor health inequalities. Model-based estimation is increasingly perceived as a possible avenue to enhance future local population health statistics methodology. In the UK, model-based small area estimation attracts particular interest both as a possible alternative to traditional population census enumeration and as a way to expand the range of indicators currently available. These methods however remain complex and still neglected in official statistics production. The present thesis brings applied contributions to this field by examining the potential of model-based estimation in England and Wales. First, a systematic literature review identifies the latest statistical developments and key methodological weak points. This informs the designs of three empirical academic papers designed around 2011 census health outputs. The first study builds two models predicting the crude prevalence of long-term limiting illness and self-rated health, and examines their reliability compared with 2011 census estimates. Secondly, an observational study analyses the spatial structure of morbidity for twenty age by ethnic groups in 2011 census long-term limiting illness data. This assesses the potential to borrow strength across space and demographic groups, and to improve prediction efficiency. The final study proposes a survey design approach determining sample size requirements to achieve a desired level of statistical reliability. It is tested in a simulation study on 2011 census long-term limiting illness data. Together, these contributions provide applied testing work on well-established European population health indicators which inform the reliability of model-based estimation methods in a UK context.

Contents

List of Figures

List of Tables

List of Algorithms

Declarations of Authorship

I, Peter Dutey-Magni, declare that this thesis and the work presented in it are my own and have been generated by me as the result of my own original research. **Statistical Models for Small Area Public Health Intelligence on Chronic Morbidity**

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Chapter 6 of this thesis has been published as:
   Dutey-Magni, P. F. & Moon, G. (2016). 'The Spatial Structure of Chronic Morbidity: Evidence from UK Census Returns', *International Journal of Health Geographics*, 15(1). doi: 10.1186/s12942-016-0057-5

Signed: .....

Date: .....

Declarations of Authorship

I, undersigned Graham Moon, confirm that the cited paper (included as Chapter 6) is co-authored by me. The paper was shaped by initial discussions in supervisions over an idea initiated by Peter. I subsequently advised on the structure and presentation of the paper. I read and commented on early drafts and advised on how to respond to reviews. Peter led the submission, undertaking all empirical work and writing all drafts including the final accepted version.

Signed: .....

Date: .....

Acknowledgements

The research produced in this thesis has been conducted in relation to the Population, Health and Well-being research group at the school of Geography and Environment, with support from the Department of Social Statistics & Demography, both at the University of Southampton. It is funded by a +3 Advanced Quantitative Methods studentship awarded by the Economic and Social Research Council as part of the Southampton Doctoral Training Centre [grant reference number 1223155].

I would like to thank the Economic and Social Research Council, the Doctoral Training Centre and my supervisors Professor Graham Moon and Professor Nikos Tzavidis for their support in conducting this work. Producing this thesis has been an exercise of perseverance and humility but, across this journey, I have been privileged to cross paths with very special people who all inspired me in their own ways. Among them, Doctor Andrew Hinde, Professor Peter van der Heijden, Professor Dankmar Böhning, Professor Tarani Chandola, Professor Natalie Shlomo, Doctor Alison Rodger, Doctor Charlotte Warren-Gash, Doctor Marta Blangiardo, Professor Sujit Sahu and Doctor Julia Sinclair have been role models who kept my intellectual curiosity fed. I feel particularly lucky to have worked with Angela Bartley and Professor Alison Rodger during my time at the Royal Free London. They were never short of enthusiasm, encouragements and constructive feedback, and made me a much more competent researcher. Finally, I owe a debt of gratitude to the Cathie Marsh Centre for Census and Survey Research in Manchester for the incredible encouragement, moral and financial support, and inspiration I received from its people and the wider university. They played a determining role in getting me to do research.

Special thanks go to my mother, Bertalan, Dorothée and Dylan, Radhika, Caio, Joanna, Becky, Sarah, Thierry, Brette, Chet, Pauline, Jonathan, Pete & Steph who taught me that jazz and a good swingout make trouble go away.

„Temu, kto szukając przyczyny, nie godzi się z żadną hipotezą rozmysłu, ani w jej postaci opatrznościowej, ani diabelskiej, pozostaje tylko racjonalnyurogat demonologii – statystyka."
Stanisław Lem, *Głos Pana*

'To him who seeks a reason but cannot abide any hypothesis of a design, whether in the form of Providence or of the Diabolical, there remains only the rationalist's substitute for demonology—statistics.'
Lem, 1999, p. 10

Disclaimer

Abbreviations

AIC Akaike information criterion
AR1 first-order autoregressive
ARB absolute relative bias
AUC area under the curve
BMI body mass index
CAR conditional autoregressive
CCG Clinical Commissioning Group
CE communal establishment
CSEW Crime Survey for England and Wales
CV coefficient of variation
DEFF design effect
DHS Demographic and Health Survey
DIC deviance information criterion
EB empirical best
EBLUP empirical best linear unbiased predictor
EHS English Housing Survey
EPP empirical plug-in predictor
ESS effective sample size
FRS Family Resources Survey
GIS geographical information system
HB hierarchical Bayes
HSCIC Health & Social Care Information Centre
HWB Health and Wellbeing Board
ICAR intrinsic conditional autoregressive
ICD International Classification of Diseases
IHS Integrated Household Survey
INLA integrated nested Laplace approximation

ISAR indirectly standardised emergency hospital admissions rate
JSNA joint strategic needs assessment
LAD local authority district
LFS Labour Force Survey
LISA local indicator of autocorrelation
LLB Leroux-Lei-Breslow
LLTI limiting long-term illness
LOS Life Opportunities Survey
MSE mean squared error
MSOA middle layer super output area
MYPE mid-year population estimate
NEX normal exchangeable
NHS National Health Service
ONS UK Office for National Statistics
PCT Primary Care Trust
QLFS Quarterly Labour Force Survey
Q–Q quantile-quantile
RMSE root mean squared error
RSE relative standard error
RW1 first-order random walk
SAE small area estimation
SAR simultaneous autoregressive
SILC Survey on Income and Living Conditions
SMR directly age-standardised mortality rate
SRH self-rated health
SSD sample size determination
SYN synthetic predictor
UPSEM unequal probability of sampling method

Chapter 1

Introduction

Public health agencies and health care organisations demand increasingly comprehensive baseline data on local risk factors and health outcomes in order to fulfil a range of analytical and strategic needs. To this end, they traditionally rely on sources of public health information such as government surveys; locally commissioned surveys; death registrations; hospital activity statistics and school measurement programmes. Yet, many of these sources become unreliable when it comes to making inferences for very small geographical areas such as wards. In particular, survey estimates are often subject to unreasonable sampling error. While locally commissioned health and lifestyle surveys have in the past produced more reliable estimates, rising costs of fieldwork and nonresponse rates have gradually discouraged their use. As a result, data from decennial censuses are sometimes the only source that can be relied upon to produce official statistics with high levels of granularity, that is, for very small population groups defined as a cross-classification of demographic traits and area of residence. Yet, census information comes with three key limitations. First, decennial UK censuses collect a very narrow range of health indicators: limiting long-term illness (LLTI), self-rated health (SRH), provision of unpaid care, and long-term sickness/disability. More focused aspects of specific diseases and behavioural risk factors are absent from census questionnaires. Second, outputs are only produced every ten years, which is often insufficient to monitor progress against public health objectives. Third, censuses and large population surveys are costly exercises. The UK government is seeking more cost-effective ways to produce population statistics, following the example of other nations that successfully phased out their own censuses. There is potential to use new methods and sources of data to respond to the range of challenges posed by this context.

1.1 Existing research

Work is needed to ascertain the potential of developing new techniques to address the need for small area information and the shortcomings of existing data. A range of statistical techniques collectively known as model-based small area estimation (SAE) have been put forward which share a reliance on statistical modelling to predict a parameter of interest across a set of small geographical areas or some other population subdivisions. Substantial research effort has been dedicated to them internationally. The UK Office for National Statistics (ONS) has conducted successive model-based SAE trials (Heady and Ruddock, 1996; Heady, 2003; Heady et al., 2003). Three successive European Framework Programmes have encouraged international collaboration on the development of statistical estimators, published software implementing these methods, and considered feasibility constraints for national statistics production, particularly around poverty and social exclusion indicators (AMELI, 2011; Heady and Ralphs, 2005; SAMPLE, 2011). With regard to health statistics, the UK Department of Health has prompted research on small area indicators of lifestyle and behavioural risk factors (Pickering et al., 2004) and government departments regularly commission similar outputs (UK Department for Business Innovation & Skills, 2012; Ipsos Mori, 2007; Public Health England, 2014c, 2015). We also note concomitant work on consolidating transnational population health information with the European Core Health Indicators tool as well as the Life and Health Expectancy Information System (Kilpeläinen et al., 2008; Public Health Evaluation and Impact Assessment Consortium, 2013; Robine et al., 2013). Comparable work is under way at the US Centers for Disease Control and Prevention and elsewhere globally.

1.2 The problem

This international research and development effort bears testament to the continued interest in making health indicators available and comparable across areas, whether on a regional scale (countries and states), or on a more local scale (political and administrative localities). A recent impetus for this interest has been the UK government's ambition to abolish traditional decennial censuses (Maude, 2014). Planning work has been conducted by the ONS under the Beyond 2011 methodology task force between 2010 and 2015 (ONS, 2014b), and thereafter as part of the Census Transformation Programme (ONS, 2016b). The task force has reached the conclusion that 'different and more complex methods than those currently applied

by ONS will be required to apply SAE methods and there are likely to be significant challenges to their development including identification of suitable data sources to provide the covariates required' (ONS, 2013f, p. 24). There remain significant methodological unknowns in relation to:

- the existence of reliable models meeting, for a range of census topics, a set of relevant quality and reliability criteria to be determined—a consensus remains that 'the lack of model validation has been a significant impediment to developing robust small area measures on which policy and resource allocation decisions can be reliably based, potentially impeding the acceptance of estimation models by policy-makers' (Smith et al., 2011, p. 618);
- the sampling design requirements of each of these models, which presupposes extending the range of survey planning methods;
- the existence and reliability of sources of data supporting model-based SAE—the ONS (2013e) has noted that the discontinuation of the census would decrease 'the amount of detailed, small area data available for use as auxiliary information in the models. Census variables are used widely in existing SAE models, both at ONS and internationally and the impact on modelling outcomes and estimator viability of replacing these variables with covariates from administrative or survey sources is being investigated.'

The motivation for this doctoral research is driven by these three methodological gaps. An original call for proposals entitled 'Estimating the 2011 Census Geography for Health Care' was issued by Prof Graham Moon during Spring 2012. This call aimed to promote alternative ways to produce population health information, taking into account the uncertainty that model-based methodologies were capable of meeting the requirements of UK census outputs in the future. It was issued with an emphasis on validity and feasibility assessment:

> 'We envisage a five stage approach that will involve simulating a robust, accurate and valid synthetic small area baseline population with age, sex, ethnicity and marital status characteristics, comparing this synthetic population to 2011 census (available in autumn 2012), development of multilevel models using routine census data (e.g. Labour Force Survey, General Lifestyle Survey, Health Survey for England) on census 'health' variables (disability, carers, limiting long term illness), identification of small area synthetic estimates of the selected variables, and validation against 2011 census output.'

A proposal was developed for this call on the basis that this validation work should be completed with due consideration for recent developments in model-based

SAE, particularly robust estimation, spatial modelling, and the reliance on administrative data. It also incorporated the need to review model-based methods and develop survey design and planning tools in addition to this validation work.

1.3 Thesis scope and structure

The present thesis aims to contribute towards a better understanding of model-based SAE techniques, a consolidation of approaches to validating their statistical designs, and the demonstration of effective implementations for official statistics on morbidity. Within these aims, we focus on the bias and efficiency of applying model-based estimation to existing census health indicators.

The body of this thesis commences with a chapter examining the need for small area population health statistics in the public and academic sectors. This highlights substantive gaps to be addressed with new methodology. Chapter 3 introduces the basic concepts and terminology of model-based disease mapping and SAE. It appraises both the methodological and applied literatures around techniques applicable to population health indicators. This informs the design of three empirical papers, which are introduced from the viewpoint of their research objectives and designs in chapter 4. The papers themselves are included in chapters 5 to 7. A final chapter 8 sets out the overall contributions of the thesis and discusses its limitations and implications for future research and policy.

Chapter 2

The Case for Small Area Population Health Statistics

2.1 Concepts of public health intelligence

2.1.1 Data, statistics and information

Public health is defined by the Faculty of Public Health (2016, p. 2) as 'the science and art of promoting and protecting health and well-being, preventing ill-health and prolonging life through the organised efforts of society.' To understand the role of statistical modelling in supporting public health action, this first chapter begins with an examination of how data are turned into information. Doing so requires to draw out the range of functions assumed by public health information.

In the US, the concept of 'public health surveillance' appeared in the 1950s in infectious disease medicine, only to attain academic and institutional recognition twenty years later (Thacker and Berkelman, 1988). Its current definition 'encompasses the continuous collection of health information; evaluation, analysis, and translation of data into knowledge about the health of communities; and communication of that knowledge to the public and to public health staff, policy makers, and others positioned to take action' (Smith et al., 2013, p. 231). Prominent examples include syndromic surveillance systems, outbreak detection, behavioural risk factor surveillance, health inequalities monitoring, and studies of the burden of health of certain risk factors. The UK public health profession interchangeably uses the equivalent phrase 'public health intelligence' to designate the functions of 'surveillance, monitoring, and assessment of health and the determinants of health, plus the development of the public health evidence base and knowledge' (Faculty of Public Health, 2014, p. 1). This wider definition places further emphasis on interventional

research evidence (including cost-effectiveness evaluation), as well as two activities which have acquired particular importance in the UK: health needs assessments and tackling health inequalities. A more governmental and holistic view is that public health intelligence is 'a core public health function that ensures the right information is available at the right time and in the right place to inform public health decisions and actions' (UK Department of Health, 2012, p. 5).

'Surveillance', 'intelligence' or 'information' thus refer to data which have been gathered, processed, validated and analysed in a way that can be made sense of in terms of (a) understanding the underpinning mechanisms and the significance of health problems (needs assessment), or (b) making policy decisions (public health action). The role of the intelligence function is to convey information with the help of definitions, units of measurement, standardised indicators and user guides, allowing for intelligibility, comparability, auditability and accountability.

A further specificity of the public health intelligence function when compared to medical statistics is its focus on population representativeness. Public health is characterised as a 'population-based approach' by the UK Faculty of Public Health, in other words 'a framework characterised by a broad definition of health, that recognised determinants outside the health care system, and explicitly acknowledged trade-offs between investing in health care and investing in other social goods' (Diez Roux, 2016). This focus has a continued justification due to the existing tension between prevention and health care, which has been highlighted by economists (Evans and Stoddart, 1990). This affects to a very large degree the set of indicators, data sources and methodologies privileged in public health intelligence, which cannot only rely on clinical sources of information.

With these lexical considerations in mind, it is worth considering what public health information is made up of, that is, what tools, constructs and definitions make it possible to extract information from data. The remainder of this first section introduces formal concepts of public health information with specific relevance to local analytical needs. Section 2.2 then reviews uses of this information by health organisations, highlighting drivers of an increasing demand for local population health information available with very high levels of granularity. Section 2.3 reviews modern developments of local public health information in a UK context and particularly in England. Public health intelligence has been influenced not just by reorganisation of government and health care organisations, but also by changes in the way data can be collected and accessed. Particular emphasis is placed on existing methodological limitations and consequences of the review led by the ONS into the future of the decennial census. Finally, a section 2.4 summarises key aspects and

considers the potential role of statistical modelling in addressing current challenges.

2.1.2 Morbidity, mortality and summary health indicators

Concepts of health are usually framed according to the definition of health by the World Health Organisation (1948) as 'a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity'. This corresponds to a wider approach comparatively to strict clinical and biomedical diagnoses nomenclatures. A wide perspective on health also implies definitions and measurements across the whole natural history of disease; taking into account the successive stages of an illness, including risk factor exposure (social, behavioural, environmental, genetic and clinical determinants of health), biological onset of disease, onset of clinical symptoms, patient presenting complaints, medical diagnoses, therapeutic decisions (prescription, operations and other care interventions), health outcomes, vital and health events (including death), overall burden of disease (quality of life, functional measurements, degree of limitation). These concepts demand closer examination in order to understand information needs of health agencies.

In clinical terms, health and illness are usually qualified at the diagnosis phase as medical conditions listed under classifications of diseases (World Health Organisation, 2011), mental health disorders (American Psychiatric Association, 2013) or functional ailments (World Health Organisation, 2001). In accounting terms, this can be measured from health care records or from self-reports of doctor diagnoses by individuals themselves.

In public health and epidemiology, however, morbidity is studied under a wider variety of 'health indicators' which can relate to a health status determined by an appropriate tool such as a quality of life questionnaire, a screening form detecting hazardous drinking, a census or a household survey questions on disability or subjective/perceived health.

Disease definitions are complex to operationalise for the production of population-level official statistics. Performing clinical examinations and diagnoses on a wider population than the one encountered in health care settings is not only costly, but the construct validity of such a measurement is not evident. The International Classification of Diseases (ICD) is largely designed for diagnoses made in a clinical setting; it is determined by a patients' presenting complaints and access to health care. In some diseases, however, simple physiological measurements can be taken which are sufficient to predict an ICD diagnosis even if the condition has not yet been diagnosed by a medical professional. Examples frequently encountered in

Figure 2.1 | Diagram of life history and summary population health indicators

health examination surveys include chronic kidney disease or chronic obstructive pulmonary disease.

These measurements remain more costly and cannot be implemented in large household surveys or in a census. This is why alternative self-reported indicators have been developed overtime. Some of them have a clinical origin, for instance the EuroQOL five dimensions quality of life questionnaire (EuroQol Group, 2009) or the short-form health survey (Ware and Sherbourne, 1992). Simpler indicators have been developed for population investigations; a very common one is SRH, which has become ubiquitous in European surveys, patient experience and health improvement initiatives. Respondents are asked to rate their present health state on a scale of one to three or five. It is used widely in epidemiology, demography and economics to study health needs in different populations and across the life course. Another common indicator is the chronic illness question or LLTI. Both of these indicators are the object of a more detailed overview in sections 4.3.1 and 4.3.2.

Many population indicators can be combined with mortality rates into summary population health indicators (Gold et al., 2002; Murray et al., 2000). Such indicators have emerged as a necessity due to the rise in burden of disease concomitant of the extension in life expectancy. Life expectancies such as the healthy life expectancy (see Figure 2.1), the disability-free life expectancy, or the handicap-free life expectancy rely on Sullivan's (1971) method to collapse mortality and disability indicators into a single measurement of quality-adjusted life expectancy at birth. More complex summary population health indicators known as 'potential years of life lost' focus on either a specific cause of illness present on death certificates, or on exposures to a given behavioural or environmental risk factor which has to be estimated to produce an attributable fraction estimate (Murray et al., 2003; World Health Organisation, 2009). Healthy and disability-free life expectancy estimates

are produced on a regular basis in England using SRH and LLTI survey questions (Smith et al., 2010), and there is interest in publishing them for small geographical areas in order to measure inequalities (ONS, 2013g).

To summarise, health is measured under a variety of constructs, at different stages of the history of disease, and in relation to various stages of life and time periods. These constructs correspond to the specific information needs of a variety of public health functions.

To conclude this brief overview of measurements of population, the emphasis on assessing need at a very small geographic level is a distinct characteristic of research carried out since the eighties. The interest in morbidity and disability prevalence from other sources than hospital and primary care has led to the establishment of uniform definitions and methodologies. In addition, there has been a rising interest in the measurement of burden of disease in terms of healthy life years lost, severity of limitation, care provided by relatives, as well as its contextual determinants. Such indicators have proved to be more relevant to assess the health care need and use, not only in terms of hospitalisation, primary and social care, but also in terms of avoidable mortality, activity limitation, dependence.

2.2 The functions of local public health intelligence

2.2.1 Needs assessments and health care planning

The process by which a doctor examines his or her patient, reviews the existing biomedical evidence base, and determines an appropriate treatment can be metaphorically extrapolated to the population level. Key questions that occur in local public health departments are: what are the community's health problems; are there effective interventions; what is the cost of such interventions; and are there gaps in information on the health of the community?

'Medical need' has been defined as 'the medically modifiable morbidity (illness) burden of a defined population' (Fries et al., 1998, p. 71). Wright et al. (1998) have argued that health needs assessments are required to circumvent a 'top-down approach to providing health services, which relies too heavily on what a few people perceive to be the needs of the population rather than what they actually are' (*ibid.*, p. 1310). This widely accepted view is motivated by seminal research on health inequalities highlighting the unequal access to health care and the absence of straightforward relationship between health needs and care utilisation (Last, 1963; Tudor Hart, 1971). This was further evidenced in subsequent investigations of socioeco-

nomic gaps in health outcomes (Whitehead et al., 1992). As a result, community health assessments require a very strong grasp not just of the most effective prevention and treatment strategies, but also the specific determinants driving the health of any given local community. This presupposes access to information on key risk factors in the local community as well as their significance in terms of quality of life and economic burden.

Health needs assessment, despite a long history in the UK, have been the object of a renewed attention in the 1990s, when they came to be regarded as 'an objective and valid method of tailoring health services—an evidence-based approach to commissioning and planning health services' (Wright et al., 1998, p. 1310)). The purchaser-provider split acted in 1989 vested statutory obligations in needs assessments, with commissioning defined as 'the process of assessing the needs of a local population and putting in place services to meet those needs' (UK Department of Health, 2010, p. 54).

In the aftermath of the Health and Social Care Act 2012 (c. 7), Clinical Commissioning Groups (CCGs) were established and given the charge of arranging for the provision of medical services or facilities as they considered appropriate in relation to the National Health Service (NHS) Constitution and other national policy. The Health and Social Care Act 2012 (c. 7, s. 14) provides that CCGs fulfil general duties to secure improvements in:

(a) the physical and mental health of the persons for whom it has responsibility, or

(b) the prevention, diagnosis and treatment of illness in those persons.

Simultaneously, the Act handed some of public health responsibilities previously controlled by primary care organisations and the Department of Health over to upper tier local authorities and their Health and Wellbeing Boards (HWBs) (UK Department of Health, 2011b). A population-centred approach to commissioning at the local level was to be strengthened thanks to the co-operation of boards, local government and CCGs and formalised into Joint Health and Wellbeing Strategies adopted by local authorities following a Joint Strategic Needs Assessment.

These developments explain current challenges in terms of providing the detailed intelligence and skilful planning needed to determine priorities and organise services at a local scale. Richer evidence is needed on the health of local populations, and the conditions in which such evidence can be collated and published have evolved considerably. Local public health information gaps have been one of the key motivations behind the World Class Commissioning programme launched in 2007. The Department of Health and the Health & Social Care Information

Centre (HSCIC) subsequently took initiative to develop a basket of indicators available to commissioners, either by

- aggregating and updating the output of existing statistics such as census data, the HSCIC's Compendium of Population Health Indicators, the Local Basket of Inequalities Indicators (Fitzpatrick and Jacobson, 2003), or the joint strategic needs assessment (JSNA) core dataset;
- or by developing new statistical outputs, for example indicators on healthy behaviours (NHS Information Centre, 2005), or more recently NHS England's regularly updated CCG outcomes indicator sets (see for instance NHS England 2012).

> Local populations vary considerably in their mental health needs and commissioning should respond to that. But there is a lack of good quality local information on population needs, including unmet needs, which is important because those who are in most need may be less able to seek help. Poor use of evidence by commissioners may be the reason why the fivefold variation per person in NHS budget for mental health services does not seem to follow known patterns of prevalence or need but appears to be almost random. (*ibid.*, p. 30)

Overall, the gradual devolution of commissioning and public health in recent years justifies the development of a strong and disaggregated evidence based to support specialised prevention and care plans likely to meet local need. Despite the government's World Class Commissioning programme and efforts by the NHS to provide relevant indicators for local strategic planning, there remains scope to improve intelligence supporting commissioning, especially in areas where prevention and specialised commissioning are required.

2.2.2 Policy evaluation and monitoring

Population health data have a long history of informing public health action in domains as varied as sexual and reproductive health, smoking cessation or medical case finding initiatives. Many of the health service performance objectives are monitored using administrative data, as is for instance the case with the English Quality and Outcomes Framework (NHS Digital, 2016b), a pay-for-performance scheme under which General Practitioners commit to recording conditions, obesity or smoking status of patients they encounter during the year. Other policies focus on reducing emergency readmissions, reduce the length of inpatient stays, or the use of antibiotics. For many aspects of policy making, it is desirable to have statistics

covering the entire population, rather than individuals attending health care services. In that case, it is necessary to design population investigations.

A prime example can be found in tackling health inequalities, which shows how initial evidence can shape policy and, in turn, further evidence and monitoring. Successive censuses have provided the necessary denominator data to calculate standardised mortality ratios across items of industrial or occupational classifications. These have mostly been produced in Decennial Supplements by the Register General (see for instance Griffiths and Fitzpatrick, 2001). Health inequalities in mortality have been a recurrent policy topic in the UK since the publication of the very influential report entitled *Inequalities in Health* by Sir Douglas Black (UK Department of Health and Social Security, 1980). This publication prompted a renewed interest in examining inequalities in life expectancy and other indicators, and was followed by similar investigations (Acheson, 1998; Marmot, 2010). This body of evidence spurred government to develop policy, the most tangible being formulated around the funding, planning and delivery of health care (UK Department of Health, 1995). By 2000, reducing health inequalities had become an official governing objective of the Department of Health and the NHS, with an emphasis on improving funding allocation and focusing health improvement interventions on the most deprived and ill populations (UK Department of Health, 1999; of Health, 2000). This goal was later operationalised into a range of targets, starting with the 2002 Spending Review Public Service Agreements aiming 'by 2010 to reduce by at least 10 per cent the gap between the fifth of areas with the lowest life expectancy at birth and the population as a whole' (UK Department of Health, 2003). Since the Health and Social Care Act 2012 (c. 7), reducing health inequalities in the planning and delivery of health care has a statutory basis. Its section 13G provides that the national commissioning board (NHS England) must 'in the exercise of its functions, have regard to the need to:

(a) reduce inequalities between patients with respect to their ability to access health services; and

(b) reduce inequalities between patients with respect to the outcomes achieved for them by the provision of health services'.

The establishment of population data collection through the Integrated Household Survey and the Annual Population Survey have made it possible to monitor the temporal trend and identify that gaps in healthy life expectancy across districts have not narrowed across a ten-year period (ONS, 2014d, 2013g).

Population health indicators remain an important part of policy making in England. The Public Health Outcomes Framework (UK Department of Health,

2016) contains a wide range of such indicators focusing on social determinants on health, which are reviewed on a regular basis. Currently, the vast majority of these indicators stem from registers and health care return systems (e.g. cancer screening coverage, children poverty and school readiness, rates of hospital admission and mortality indicators, chlamydia detection rates, state school children overweight). Few indicators beyond healthy life expectancies are based on population investigations. Wellbeing, unemployment, sickness and crude smoking prevalence are obtained from the Annual Population Survey using traditional survey sampling methods. Estimates of the population exposed to road, rail and air transport noise is derived from the decennial census. Healthy eating, smoking and exercise statistics are produced by a random-digit telephone survey, Sport England's Active People Survey and a school survey, the What About Youth Study. Other specialist population surveys exist (English Housing Survey, Health Survey for England, Adult Dental Health Survey, Adult Psychiatric Morbidity Survey), but their design does not permit to produce estimates beyond the regional level, for small areas such as local authority districts (LADs) or CCGs.

In addition to information on epidemiological frequency, key public health priority are determined by other criteria: severity, cost, preventability, communicability and public interest (Teutsch, 2010). Severity and cost any given health problems are indicators of burden and are often operationalised into constructs such as years of life lost. Burden of health studies are thus conducted to monitor progress in tackling health problems and identifying priorities for health improvement. Yet, they remain very demanding in terms of data. In England, such analysis has been produced down to the regional level (Newton et al., 2015) and by area deprivation quintile. Data exist for smaller geographical levels but are less reliable owing to a range of factors, namely: stochastic variability in mortality counts; sampling variance in disease prevalence estimates; distorting factors such as the presence of nursing homes in a given area; error in population denominators. This justifies current efforts to develop new methodology to produce small area estimates of burden of health (Williams et al., 2005).

2.2.3 Epidemiological surveillance

In the UK, surveillance is defined as 'the systematic regular collection, analysis, interpretation and dissemination of data for a given population to detect changes on patterns of disease or disease determinant with action taken if a predefined criteria or thresholds are met' (UK Department of Health, 2012, p. 49). In other words,

surveillance is the continuous and protocol-driven component of public health intelligence, which focuses on hazards, exposures, or behaviours determining health with reference to a geographical or temporal baseline. Amongst the goals of surveillance reviewed by Meriwether (1996), we highlight two important epidemiological ones:

1. to detect cases or clusters so as to trigger appropriate interventions preventing transmission or reducing morbidity and mortality;
2. to develop scientific hypotheses leading to analytic studies about risk factors for disease causation, propagation or progression.

Traditionally, the first goal is achieved by syndromic surveillance systems in Emergency Departments, vital statistics monitoring, and mandatory case reporting systems on a range of conditions and syndromes (both communicable and non-communicable). Permanent examples include influenza-like illness weekly returns, meticillin-resistant *Staphylococcus aureus* post infection reviews, or variant Creutz-feldt-Jakob disease monitoring. Resulting data is analysed to detect anomalies and trigger an immediate response.

But epidemiological surveillance also encompasses systems monitoring population health without automatic triggers: for instance drug use and treatment statistics, cancer registries and cancer screening databases, teenage conception statistics, rough sleeping statistics, hospital and death statistics. For example, increases in mortality and hospitalisations attributed to respiratory diseases in Tower Hamlets in the decade leading to 2011 justified a policy response in the form of a local enhanced public health action to improve outcomes (Tower Hamlets JSNA Reference Group, 2011). Health behaviours is an area of particular interest currently receiving continuous funding for data collection in the Annual Population Survey, the Active People Survey, the What About Youth Study and the Health Survey for England. These enable the production of time series for behaviours such as smoking, hazardous drinking, healthy eating and physical exercise. Additional statistics are produced on overweight by the Health Survey for England and the National Child Measurement Programme.

There is a second and more scientific purpose in monitoring small area levels of exposures to a range of risk factors. Geographically disaggregated information fulfils important research goals in observational epidemiology. The dose-response curve between mortality/morbidity and many risk factors cannot be estimated under experimental conditions. Risk factors such as air pollution, weather conditions, noise, or processes of social segregation cannot be administered to groups under a randomised scheme, as their manipulation is not only unethical, but impractical. Alongside

cohort and time series studies, ecological cross-sectional or spatio-temporal observational designs are an affordable and simple approach that has been used to estimate dose-response relationships. Spatial variation in a risk factor of interest is used as a quasi-experimental allocation mechanism. The confounding of this non-random allocation by other risk must yet be controlled through appropriate confounding adjustments. For example, Rushworth et al. (2014) examined the association between the concentration in particulate matter and rates of hospitalisation for respiratory illness across London wards, adjusting for socioeconomic confounders such as the proportion of residents claiming unemployment benefit and median house prices. They demonstrate significant dose-responses with some air pollutants. Another example is recent research intended to study the confounding between air and noise pollution by Halonen et al. (2015). In this design, authors use aggregate small area health outcomes for age by sex subgroups. Such investigations are important to formulate appropriate scientific hypotheses needed to design more conclusive research designs, for instance cohort investigations.

Conducting ecological studies such as the above presupposes not only health outcome data, but also good quality data on spatial and demographic confounders such as population age and ethnic make-up, smoking habits, deprivation or social segregation. Such confounders are not always readily available, particularly around behavioural risk factors such as smoking (Rushworth et al., 2014). Covariate measurement error in most cases leads to the underestimation of relative risks or dose-response coefficients, as per the well-documented phenomenon of regression dilution (Clarke et al., 1999). Population health indicators thus play an important role in improving the validity of ecological research into health risks.

2.2.4 Funding allocation

Resource allocation mechanisms have come under review in the nineties with the provider-purchaser split, and later when government developed policy to tackle health inequalities. They have become more evidence-based and have come to rely on local public health intelligence to great extent. During the first thirty years of its existence, the NHS did not have an explicit funding allocation formula. Regional Health Authorities used to channel funding locally without established, nationally consistent methodology. By the early seventies, unbalances in *per capita* expenditure had become flagrant and mostly dictated by the number of already existing hospital beds, to the detriment of regions suffering from insufficient care provision (Mays, 1995).

From the late seventies onwards, a substantial body of research attempted to capture care needs using population health statistics and infer a fair capitation formula. The first such formula was proposed by the Resource Allocation Working Party in 1975 and implemented gradually from 1977–78 onwards. The formula predicted hospital bed utilisation using regional all-age years standardised mortality ratios for a range of conditions (UK Department for Health and Social Services, 1976). This formula nevertheless came under criticism in that it assumed that mortality was an accurate and current measurement of need for health care. Until the 1991 census, the only other measurement of need available at the local level was hospital utilisation, which was also criticised for being mostly determined by the *provision* or *availability* of care services rather than need. By conducting analysis of variance of hospital admission rates, mean length of stay, and total bed-days by catchment or resident population, Kirkup and Forster (1990) found that bed supply was the most powerful predictor of use, rather than local mortality, deprivation or percentage of elderly in the local population. Overall, concerns around strictly activity-based funding were that a population with very poor health could display low levels of need owing to well-documented obstacles on access to health care. Such a funding mechanism has widely been regarded as a case of inverse care law (Tudor Hart, 1971), in the sense that indexing funding on utilisation can exacerbate geographical distortions in the amount spent on each individual.

This observation was a fundamental motive behind funding a question on health for the very first time in the 1991 census. The Department of Health and Social Services wished to obtain higher quality evidence regarding the health of populations across communities, so as to allocate funding based on *need* rather than just the size of the resident population or some other criteria based on existing staff headcounts. The health indicator introduced in the 1991 census was developed precisely with the ambition to operationalise a simple proxy that would make it possible to measure need, as explained in more detail in sections 4.3.1–4.3.2 (see p. 88). It was used to produce what soon became known as the 'York formula' of capitation Carr-Hill et al. (1994, 1997) articulated around more refined analysis disentangling need, demand, supply and utilisation of health care. Based on the assumption that need in itself was a latent unobservable variable, Carr-Hill et al. (1994) took into account a range of measurements of local population health status in addition to standardised mortality ratios, namely: standardised illness ratios from the 1991 census LLTI question; standardised permanent sickness ratios from the 1991 census question on economic activity; and local low birth weight statistics. The availability of this information at a very small level (electoral wards) from the 1991 census accommod-

ated, to some extent, limitations of modelling need at the coarse regional level, since it is in many ways ecologically spurious.

More recently, research on resource allocation has tended to challenge the relevance of area-level statistics to measure care need. Carr-Hill et al. (1994) invoke the ecological fallacy by explaining that 'summary measures of need in an area do not necessarily reflect the sum of the needs of individuals living in the area' (*ibid.*, p. 15). Even though the weighted capitation formula has experienced only minor adjustments since the nineties, resource allocation remains an area of debate.

Firstly, research has attempted to refine the range of determinants of need under consideration. Gravelle et al. (2003) have presented a model specified to adjust for supply-side characteristics so as to compute an adjusted estimate of need. In particular, they account for determinants such as waiting times, distance, capacity and the availability of private health care. On the other hand, they regard mortality, self-reported morbidity, and socioeconomic deprivation in terms of education and income as 'legitimate' determinants of need. As for the census LLTI indicator, it results from the search for a simpler, yet appropriate proxy capturing health complaints leading to attending health services. Yet, this a powerful and effective proxy does not identify the nature of health issues that contribute to the overall prevalence of long-term activity limitation, and it is arguably insufficient on its own to capture all determinants of health expenditure. In particular, it does not reflect specialist care needs, or their severity, or their cost.

Secondly, research has revealed the importance of producing such indicators for very small populations rather than regional ensembles. Aggregation at a high level such as a region can lead to inappropriate resource estimates at more local levels, and there is a risk that cost function parameters for individuals (that is to say, parameters of weighted capitation) may not be adequately deduced from regional relationships between sociodemographic makeup and aggregate measurements of need or utilisation. There is considerable variance in needs at small levels which justify estimating and implementing weighted capitation formulæ at a those same levels, based on more detailed population health information than is currently available.

The concern over potential ecological fallacy in estimating funding formulæ led to a shift in research away from area-level determinants of need to patient-level predictors (from care records) in models predicting individual-level health care expenditure. This began with an attempt to determine budgets for General Practitioner fundholding experimentations conducted since 1991. Those have illustrated difficulties in fitting cost functions with sufficient explanatory power to hold year after year even in small practices. Martin et al. (1998) gave evidence regarding the

explanatory model of simple expenditure models on the basis of information available from registers and concluded that capitation was not likely to be an efficient practice-level allocation system. Nevertheless, more recent research by the Nuffield Trust (Person-Based Resource Allocation Team, 2010; Dixon et al., 2011) has led to the development of longitudinal models taking into account known information from patient records, in particular existing diagnostics. These models are specified with a patient time-invariant random effect capturing an individual's unobservable needs.

These models are nevertheless very complex, and they cannot fully disregard ecological determinants. Models presented by Dixon et al. (2011) include covariates such as the individual's ward-level Index of Multiple Deprivation and proportion of residents without higher qualifications, or the General Practice's register prevalence of asthma (Quality and Outcomes Framework). These covariates are yet not enough to explain all the variance in expenditure and random General Practice and Primary Care Trust (PCT) effects are also included which capture other unobservable differences.

This shows that ecological determinants remain important even in patient-level resource allocation mechanisms. In response to this work, public health specialist have emphasised the potential limitations of disconnecting resource allocation (and the associated incentives) from community needs assessments (Pollock et al., 2012). A critical characteristic of work led by the Nuffield Trust is the shift towards predicting current patterns in care utilisation as opposed to need or, more specifically, 'legitimate need' as intended by Morgan et al. (1987). Forecasting patterns from commissioning datasets is bound to reproduce existing patterns of supply rather than gradually allocate resources to reach an equitable optimum, shifting funding in such a way as to achieve marginal gains in health while reducing avoidable health inequalities across demographics and parts of the territory.

Although the Advisory Committee on Resource Allocation recommended adopting a patient-based allocation mechanism (Fillingham, 2012) in the 2013–14 CCG budget allocation, it emphasised that risks of perpetuating existing inequitable repartition of funds were not addressed.

> [The Advisory Committee on Resource Allocation] believe that rather than 'unmet need' the problem is one of 'sub-optimal access' where an individual's care costs more or less than it would if it had been accessed at the most beneficial point. We believe that the best way forward is to focus future research on a selection of specific diseases/conditions with sufficient coverage to be representative. (Fillingham, 2012, p. 6-7)

This formula was ultimately not retained by NHS England over concerns that it would lead to a radical geographical shift in allocation that could disrupt services. Only the mental health person-based formula was integrated in the overall calculation (NHS England, 2013).

Overall, research in funding allocation mechanisms has tended to emphasise the disconnect between epidemiological prevalence and hospital utilisation. In an observational study, Payne et al. (1994) found no relationship between the prevalence of digestive disease, musculoskeletal disease, and obesity on the one hand, and hospitalisation and mortality rates on the other hand. This is in contrast with respiratory disease and depression, where a significant association does exist. Yet, association does not presume a strong relationship with funding needs, and other observational investigations by Asthana et al. (2004) and Vallejo-Torres et al. (2009) concluded that epidemiological prevalence is a poor predictor of need for health care, particularly due to differences in populations' disease stage, age, comorbidities, access to carers and lifestyles. This prompts substantial policy attention into what is often referred to as 'unwarranted variations' across regions and social groups in health care, ethnicity being a well-documented predictor of need (Gravelle et al., 2003; Schneider et al., 2009). The interpretation of such variations has been an object of academic debate for decades in the small area analysis literature (Wennberg and Gittelsohn, 1973; Wennberg, 1987), highlighting many different drivers beyond differences in epidemiological prevalence. This has influenced health care strategies in several ways internationally. For instance, health care quality agencies are increasingly paying attention to auditing and harmonisation of care pathways across different providers, and efforts to address unequal access to healthcare have been made with case finding exercise and population screening. Chronic obstructive pulmonary disease is a typical example of conditions where spatial inequalities have driven a change in clinical practice (UK Department of Health, 2011a). We note a strengthening of the clinical audit strategy in respiratory medicine, with an emphasis on national multicentre audits (Stone et al., 2015). In parallel, targeted case-finding exercises have been proposed which prioritise areas where strong gaps are found between the expected epidemiological prevalence and the observed primary care register prevalence (Jordan et al., 2010; Nacul et al., 2007, 2011; Peña et al., 2000).

Difficulties around funding allocation highlight challenges around the geographical distribution of the burden of illness expressed not simply in terms of prevalence, but rather in terms of risks of developing conditions, access to health care and quality of self-care. Vallejo-Torres et al. (2009) emphasised that needs assessments require 'rich epidemiological data on the prevalence of different diseases

at an appropriate level of geographical disaggregation for allocating health care resources. If needs are considered to vary by population groups within areas (e.g., by age and gender) then prevalence data is required for each population group within every area. Crucially this requires reliable data for all geographical areas at the units of disaggregation at which allocations are to be made' (*ibid.*, p. 1635).

The outcome of research into resource allocation acts as a reminder that a funding allocation is no replacement for a public health strategy. Person-based allocation mechanisms do not undermine the case for more advanced local public health conditions-based indicators of prevalence and incidence, so as to allocate funds not just on the basis of care utilisation, but also in such a way as to support a strong prevention policy in primary care.

2.3 The future sources of population health information: Potentials and constraints

Previous sections have illustrated how local intelligence plays an increasingly determinant role in setting public health priorities. The devolution of healthcare and public health responsibilities has augmented user requirements on information, whether in terms of quality, timeliness and disaggregation into geographical areas, demographic subgroups and clinical specialities.

Meanwhile, sources of public health intelligence are becoming more technical. Recent developments in official statistics highlight cost constraints on surveys and the decennial census, which make increasingly apparent the need to maximise the information obtained from surveys and administrative data through most sophisticated methodology. Yet, in order to maintain trust an support local public engagement, public health intelligence must remain transparent and accessible. While traditional household surveys and population or school censuses are well understood by a range of audiences with minimal statistical literacy, recent innovations are introducing complexity involving statistical modelling and administrative sources of data. In this section, we give an overview of ongoing developments in the way population health statistics are produced.

2.3.1 Disease and lifestyle prevalence models

In recent decades, a new source of local information has become common in public health which is referred to as 'model-based small area estimation', 'synthetic estimation' or 'prevalence modelling', around the basic concept of interpolating na-

tional prevalence information from surveys to local populations, based on their known characteristics. These techniques share a reliance on statistical modelling to 'borrow strength' across some dimension of estimation (time, space, sociodemographic group) and predict a parameter of interest across a set of small geographical areas or some other population subdivisions. In particular, cross-sectional SAE borrows strength across small areas (or neighbouring areas), while time series SAE and state-space models tend to borrow strength across time to smooth out sampling error across successive waves of data collection. The methodological appraisal of these techniques being left to chapter 3, we focus here on their impact in public health practice.

The earliest instance of synthetic estimation in public health is the production of state-level estimates of prevalence of disability by the US National Center for Health Statistics (1968), by triangulating information from the National Health Interview Survey and a decennial census. This work was later extended by validations published in (Levy and French, 1977). In the UK, synthetic estimation has been investigated by the ONS in the first place. A first project investigated the prevalence of anxiety and depression (Heady and Ruddock, 1996). Other projects ensued on other topics (Heady, 2003; Heady et al., 2003; Saei and Chambers, 2003b; Saei and Taylor, 2012; Silva and Clarke, 2008). The Department of Health's strategy to tackle social and behavioural determinants of health inequalities led to a first project on small area prevalence estimates for current smoking, obesity, binge drinking, and healthy eating (Bajekal et al., 2006; Pickering et al., 2004; Scholes et al., 2008), building on research by Twigg et al. (2000); Twigg and Moon (2002); Twigg et al. (2004). A technical briefing by the Association of Public Health Observatories (2005, p. 2) provided official guidance with regard to four main utilisations that can be made of various sources of public health intelligence:

(a) Comparison with national and regional benchmarks and with other LADs or PCTs
(b) Analysis of within-area inequalities by age, gender, ethnicity and area of residence.
(c) Monitoring trends over time and progress towards local targets.
(d) Measuring/auditing outcomes of particular services/initiatives.

The briefing emphasises that synthetic estimates 'do not take account of any additional local factors, e.g. local health improvement initiatives, which may impact on the true prevalence rate' and conclude that they should not be used for the above goals (c) and (d).

Beyond lifestyle characteristics, synthetic estimates have been commissioned

by public health observatories for the prevalence of chronic conditions such as heart disease, chronic obstructive pulmonary disease (Nacul et al., 2007; Walford and Ramsay, 2011), chronic kidney disease (Public Health England, 2014c) or coronary vascular disease (Walford et al., 2011).

2.3.2 Design of future censuses and population estimates

Good quality small area population estimates are crucial to produce accurate estimates of disease prevalence and burden; they are usually used as denominators in mortality rates and as grossing factors to calculate population headcounts from survey estimates. Uncertainty in the population estimates is usually unknown and typically ignored. Since this uncertainty is not taken into account, it is especially important that error remains small. In the case of disease prevalence models, population estimates are often used as auxiliary data. Here again, uncertainty in the composition of the domain, as opposed to uncertainty regarding risk ratios, is typically not represented.

Since 1991, every UK census has delivered methodological innovations, most of which have been made possible by expansion of administrative data sharing. For instance, information on subnational migrations has been used since 1999 to improve the mid-year population estimate (MYPE) methodology. Such migrations are estimated from annual ratio changes in primary care registrations recorded from the NHS Central Register down to local authority level (Jefferies and Fulton, 2005). A variety of other sources are employed to strengthen estimates using ratio change, apportionment and cohort component methods. These include data from the Defence Analytical Service Agency regarding armed forces, the Ministry of Justice regarding prisoners populations, Patient Registers regarding residents of private households as well as communal establishment, the Department for Work and Pensions regarding Older Persons Datasets, as well as Her Majesty's Revenue and Customs regarding Child Benefit data (ONS, 2011d).

Such incremental improvements will continue with the 2021 census, which will rely on administrative sources to optimise field operations and help with estimation of certain characteristics, for instance income or population residing in communal establishments (ONS, 2016a). Yet, the ONS has been working since 2003 on a more fundamental shift away from decennial full enumerations towards statistical population database (ONS, 2003). This redesign is envisaged to result in a continuous 'administrative data census', producing population estimates and characteristics more frequently by combining administrative records linkage with government sur-

veys. In 2014, the UK cabinet declared its ambition that 'censuses after 2021 will be conducted using other sources of data and providing more timely statistical information' (Maude, 2014). Preliminary assessment work conducted by the ONS has led to narrow down indicative quality standards to be achieved by any future census. For population characteristics, the basic standard is the publication of middle layer super output area (MSOA) statistics for a characteristic applying to 3 per cent of the total MSOA population with a coefficient of variation (CV) of up to 20 per cent, or a margin of error (width of interval estimates) of up to 40 per cent (ONS, 2014b). Internationally, several countries have successfully completed transitions, including the Netherlands, Austria and Sweden (Ralphs and Tutton, 2011). These countries tend to have some form of municipal registration which provides very reliable population registers. In the case of the UK and, to a more limited extend, the Netherlands, the challenge is that no single reference register exists, and that it must be created in some way or another.

In 2016, the ONS' new Census Transformation Programme completed a first assessment of progress in building a Statistical Population Dataset, and set out a range of objectives to have the methodology in place by 2023 to begin official statistics production (ONS, 2016b). Early work is focusing on establishing efficient record matching and data cleaning rules, before coverage adjustment based on capture-recapture methods. As of November 2016, the second version of the Statistical Population Dataset is produced using extracts from the NHS Patient Register, the Department for Work and Pensions' Customer Information System, school censuses and data from the Higher Education Statistics Agency. 2011 population estimates produced with this Statistical Population Database were compared with 2011 census population estimates for small geographical areas totalling between 1,000 and 3,000 usual residents known as lower layer super output areas. For 80 per cent of these areas, all-age population estimates were within ±5% of the 2011 census estimate (ONS, 2016d). The ONS is working on refining the methodology so as to produce official population estimates for sex by quinary age cross-classification using a Population Coverage Survey and a capture-recapture estimation methodology. The question of producing population estimates for ethnic cross-classifications is not yet addressed. It is intended that a 4 per cent annual population survey could produce estimates of population ethnicity breakdowns (ONS, 2013f), but the ONS has already identified the potential to use ethnicity attributes available from healthcare and school databases (ONS, 2016b). It is thus reasonable to assume that MYPEs for demographic cross-classifications of sex, age and ethnic group should be available in the medium term.

The main challenge for the ONS is to develop methodology to produce population attribute statistics after the 2021 census. Such attributes, including health attributes, being mostly absent from administrative databases, new methodology must be developed. The Beyond 2011 task force has noted that while a 4 per cent annual population survey could help produce such outputs, 'surveys alone might not be able to produce direct estimates for the cross-tabulated outputs at small area levels that users require [...], administrative data can be used in model-based approaches to improve the precision of survey estimates, particularly where the survey sample size is too small for direct estimates, for example certain minority ethnic groups. There are different approaches to doing this, but a common feature of these methods is that they use relationships between the data and the target characteristic to produce estimates. This could be a direct relationship (for example, using ethnicity information from both administrative and survey sources) or an indirect one (for example, using ethnicity information from administrative sources to produce estimates on main language or religion from the survey because there may be a relationship between these variables and ethnicity)' ONS (2013f, p. 9).

It is not clear yet how and when such methodology will be developed, and decennial census data of interest to UK public health intelligence users may no longer be available after 2021. This is likely to also negatively impact estimates produced by disease and lifestyle models discussed in section 2.3.1. As noted by the Beyond 2011 task force, these model-based estimates tend to rely on auxiliary data from decennial census, particularly with regard to population estimates for socioeconomic cross-classification (ONS, 2013f, p. 8).

2.3.3 Administrative data for official statistics

Systematic administrative data collections are very important to public health surveillance and disease prevalence modelling. They include the oldest source of epidemiological data (death and birth registrations) as well as data collected routinely by NHS trusts, primary care and community services. Mortality statistics are used in the field of public health for one very important indicator. Life tables and life expectancy statistics destined to describe life course epidemiology characteristics represented in Figure 2.1 are some of the key indicators used to monitor health inequalities and generally evaluate the performance of a national health and public health system. There has also been interest in monitoring these at the subnational level, as was the case with the evaluation of interventions to improve health determinants in the 20 per cent local areas with the worst health and deprivation characteristics in

England (known as 'spearhead areas') (UK Department of Health, 2003). Methods to monitor progress in life expectancies were developed, with proposals for ward-level estimation methodologies by public health observatories (Williams et al., 2005) and the ONS (Toson and Baker, 2003). Academic research is also active in the area of healthy life expectations (Congdon, 2009a; Eayres, 2004; Hennessy et al., 2015; Jonker et al., 2012, 2013a).

In recent years, considerable efforts have been made to strengthen other sources of information for official statistics, by creating a legal basis for accessing government departments' administrative registers. The Statistics and Registration Service Act 2007 (c. 18) sets an incomplete legal framework under which some of these data are communicated to the UK Statistics Authority, for the purposes of population estimation. Disclosure control measures are regulated by section 39. Data sharing covers:

— NHS patient registration data: date of birth, sex, address and any previous address, NHS number, history of registration
— Birth and death-related registrations provided by the Births and Deaths Registration Act 1953 (c. 20).

The National Statistician noted that policy makers expect the Government Statistical Service 'to make better use of data, much of which is already held on behalf of the citizen within government departments and public bodies, but also within rich datasets held within the private sector' (UK Statistics Authority, 2016a, p. 1). The UK government completed a consultation in 2016 on public attitudes around giving official statisticians statutory access to their personal data (UK Cabinet Office, 2016). A bill (UK House of Commons, 2016) was introduced to Parliament in July 2016 which grants the UK Statistics Authority powers to require disclosure of a wide variety of information held by public organisations, including the Health and Social Care Information Centre known as NHS Digital.

Amongst these sources of information, one particular set of records is likely to become very useful to produce population health statistics in the future. National Commissioning Data Sets (Information Standards Board for Health and Social Care, 2012) are harmonised data warehouses involved in recording and charging for secondary care hospital services (Emergency Care, Admitted Patient Care, Out-Patient Care, Elective Admission List). Commissioning Data Sets are returned centrally by hospital providers to NHS to be collated into a single warehouse known as Hospital Episode Statistics accessible for analytical and research purposes. Records are directly linked to patients and include their personal NHS number, address of residence, demographic information, a complete breakdown of care episodes, attend-

ances, clinical procedures and operations as well as medical diagnoses. The introduction of activity-based funding (Payment by Results) as well as audits of clinical coding in recent decades have provided incentives for substantial improvements in the quality of these data. Hospital Episode Statistics is produced with respect to all secondary care offered under the NHS in England, regardless of its provider. Its coverage is therefore extremely wide and the geographic information on patients' places of residence makes it particularly suitable to compute a wide range of meaningful population indicators, for instance hospitalisation rates, re-admission rates, hospital mortality summary indicators, cancer treatment and survival. A linkage study matching Hospital Episodes Statistics and records from a primary care records research sample, the Clinical Practice Research Datalink, was found to produce sensible estimates of demographic breakdowns by ethnicity as compared to the 2011 census at the national level (Mathur et al., 2014). No evidence was published regarding the quality of estimates for smaller geographies, but data quality audits carried out by the HSCIC (2016) are showing that data quality is improving. Key challenges currently faced by official researchers remain questions of safe access and legal framework, which are being addressed.

In contrast, there is no equivalent data extraction system to maintain a nationwide primary care records warehouse. Efforts in this direction have been hampered by the technological failure of the General Practice Extraction System (UK National Audit Office, 2015). Subsequently, the care.data project aimed at extracting all primary and secondary care records into a single data warehouse caused a loss of public trust and had to be halted (Freeman, 2016). This has resulted in many patients opting out of their records being shared, to the detriment of research and possibly statistical production. As of December 2016, NHS Digital (2016a) estimated that 1,310,798 patients in England (2.3%) opted out of their personal information being disseminated or published by NHS Digital for purposes beyond direct care (type 2 objections), with some areas such as Oldham recording proportions as high as 13 per cent. An undetermined number of patients opted out of their general practice records being shared with NHS Digital at all (type 1 objections). Type 2 objections have an important consequence on Hospital Episode Statistics: as of 29 April 2016, NHS Digital stopped sharing personal records for patients with type 2 objections with the exception of:

- aggregate information, for instance tabular data published in statistical bulletins;
- information that is anonymised in accordance with the Information Commissioner's Office's anonymisation code of practice (2012);

- information disclosed as per the Statistics and Registration Service Act 2007 (c. 18) as described *supra* (Hunt, 2016).

These objections therefore do not apply to data sharing required for official statistics production. This nevertheless entails important data limitations for academic researchers, who ceased having access to pseudonymised information on patients with objections of type 2.

To date, the only sources of primary care data are the Clinical Practice Research Datalink and the Quality and Outcomes Framework return. The Clinical Practice Research Datalink is an annual extract of records from a sample of approximately 400 General Practices taking part in the research. Data has been collated since 1987 and can be linked by NHS Digital to other clinical and commissioning data sets (Herrett et al., 2015). A commercial equivalent of this data warehouse is known as The Health Improvement Network (Blak et al., 2011). The Quality and Outcomes Framework was introduced in 2003–04 together with the General Medical Services contract as a voluntary framework of incentives to prescribe treatments as well as to identify and record specific risk factors such as hypertension or cigarette smoking (NHS Primary Care Commissioning, 2011). While these were not explicitly public health targets as much as quality of care and accountability targets, the schemes' guiding principles contribute towards the national objective to reduce health inequalities. The Quality and Outcomes Framework comes with a set of national indicators relevant to monitor performance and improvements, which can be complemented by other indicators from a menu set by the National Institute for Health and Care Excellence. These indicators are subject to similar limitations as hospital data in terms of representativeness, quality, and confounding by professional and patient behavioural factors. Further questions have been raised with regard to the sincerity of these data on the basis that they they determine the size of payments benefiting the very organisations that collect them (Martin and Wright, 2009).

Although administrative health data sources are very precious, they require adjustment in that they suffer from three main shortcomings: coverage, quality and confounding.

Coverage is a well-identified limitation to the epidemiological representativeness of healthcare records. Some patients will not make use of NHS services in spite of their condition due to personal reasons (e.g. illegal residence), absence of easy access to healthcare services, premature death, preclinical period, etc. In addition, self-selection to healthcare attendances based on socioeconomic characteristics leads to what has been referred to as an 'iceberg phenomenon' by Last (1963).

Furthermore, data flows are subject to variations in reporting and quality across providers and time; with issues such as breaks in time series, corrupted and lost data, change in clinical definitions, recording practices (e.g. change of clinical records system), perceptions and framing of illnesses and health problems, or even changes in operational objectives. These issues may not be entirely addressed by quality and standards checks

Finally, confounding of the relationship between epidemiological characteristics (or need for healthcare) and rates of attendance in health services is a key limitation to the validity of information from care records. A particular pattern of confounding can arise from the spatial pattern concomitant to healthcare supply and demand factors. A large body of literature has illustrated variations in outpatients admission decisions depending on the number of beds available and the length of waiting lists across the UK (Kirkup and Forster, 1990). More recent literature has tended to look at spatial processes determining the effects of supply and demand on geographical variations in clinical procedures; for instance the association between a woman's probability to have a lumpectomy over a mastectomy depending on the distance to the nearest oncology department (Liang et al., 2008). Overall, the balance between consultations of primary care, outpatient flows and admission decisions vary across areas depending on the ease of access, waiting lists, attitudes, lifestyles and behaviours. This means that patients are likely to be in contact more or less early in the progression of illness, with substantial impacts on health outcomes. Research in health economics (see for instance Morgan et al., 1987) has thus argued against using commissioning datasets to monitor need or allocate resources, conscious of the inverse care law phenomenon. There is scope for a better utilisation of these data with appropriate models that reflect unequal access to health care. An example of such an approach has been given by Congdon and Best (2000).

2.4 Conclusion

This brief overview presented the range of population health indicators in demand to research health inequalities and improve outcomes of the health service. It also shows that the very nature of the information needed has become more complex. The gradual devolution of public health and health care planning missions is increasing intelligence requirements for very small geographies and demographic subgroups—if not individuals. Simultaneously, the nature of these statistics is becoming more complex, with demands for adjusted and highly specific indicators, relating to attributable fractions of certain risk factors, quality-adjusted life years,

amenable deaths and excess mortality, or predicted need/spending. Concepts such as need or risk are latent characteristics that are not directly observed and therefore cannot be easily estimated. Off-the-shelf indicators may not be sufficient to allocate resources efficiently, reduce inequalities, design prevention interventions or specialised commissioning.

While user requirements grow in sophistication, productivity challenges on statistical production are becoming very apparent. Spending reduction across the government statistical service, inflation in costs of fieldwork, and long-standing fall in survey response rates are affecting censuses and surveys in the overall statistical system. New methodology is needed to ensure reliability and accuracy, but also to enhance the range of products and information provided by them. Over the last decade, local health and lifestyle surveys, the General Household Survey and the Integrated Household Survey core have been discontinued. Population statistics will need to be supplemented by new, more efficient modes of collection and estimation. For instance, primary care registrations have become the core source of and improved methodology for MYPEs. Synthetic estimation is now used in England for a range of disease prevalence and behavioural risk factor statistics. The facilitated access to household geographical identifiers opens new possibilities to apply more advanced designs based on hierarchical Bayes (HB) and empirical best (EB) methods and spatial modelling. Decennial censuses involving full enumeration are expected be phased out in the future (ONS, 2016b), as is the case in other developed nations (US Census Bureau, 2008). With them, the set of detailed population attribute statistics providing information on socioeconomic and occupational characteristics is likely to be reduced. As a consequence, population health statistics will need to rely further on data produced by healthcare organisation and government departments, as well as on a re-centred core of surveys with smaller sample sizes and more efficient designs.

The increasing demand for better statistics and productivity is, overall, strengthening the economic case for effective use of data collected by administrations. The health data infrastructure, although extremely rich, is not used to its full potential. The methodology underpinning indicators is set to become more complex. There is scope to combine household surveys, which enjoy good internal validity and coverage, with institutional and clinical register information already collected and available in large quantity. Methodological progress in statistical record matching and new legislation are also opening perspectives to link information at the individual level across NHS records, death registrations and surveys such as the Health Survey for England. This represents a considerable opportunity to improve the design of

health statistics in years to come.

Statistical models and model-based estimation are regularly praised for their potential to enhance the information drawn from various sources of data. Though already well established in disease prevalence synthetic estimation, they are likely to become central in estimating a wider variety of indicators. It is therefore important to review methodological aspects in the inference from surveys and registers at a local level.

A wide range of sources are in use to support public health decisions. Although most indicators now stem from administrative sources, these are not immune to quality issues. Their direct interpretation can be spurious due to the iceberg phenomenon common to all hospital and activity-based statistics. Yet, this review shows that population sources have their own limitations. Household surveys are generally too costly to produce estimates for small population categories and particularly small geographical areas. As for decennial censuses, their cost, questionnaire design constraints, as well as the relatively infrequent rate of update are increasingly compelling reasons to develop new methodology to produce sources of population health intelligence, possibly building on multiple sources of data.

The UK Statistics Authority (2016b, section 6.ii.) has noted that 'the health and care statistics landscape is data rich, but information poor: the importance of analysis has been neglected, as has been support for other analysts and researchers'. To date, methodological work has remained insufficient and to a large extent provided by academia.

The context in which this PhD aims to make a contribution is therefore defined as much by the increasing needs in local intelligence as by the pressure to reduce cost and do without a traditional census. Answers may lie in improvements in the statistical methodology, and particularly in the use of statistical models making use of cheaper and sometimes richer administrative data. In the context of the Census Transformation Programme and growing interest in disease and lifestyle prevalence modelling, there is an opportunity to become more creative with population health statistics. Moving away from a sometimes excessive reliance on the default self-rated health indicators present in the traditional decennial census and social surveys can be the opportunity to make better use of other data sources where quality is improving, as it is currently the case with data collected by the NHS. This demands a better understanding of approaches to modelling health needs and determinants of health. This has been the object of a large amount of literature dedicated to the prediction of epidemiological characteristics and health care needs of populations.

Chapter 3

Model-Based SAE in Public Health: A Systematic Review of Statistical Designs and Validation Studies

Abstract

This paper examines applications of statistical modelling to the estimation of health characteristics for small areas. A total of 160 citations identified from a systematic search of academic citation databases are reviewed. We analyse research studies with respect to the choice of likelihood functions, predictors, random effects, controls for informative survey design. The paper also appraises reporting quality and approaches to validation. The literature is reviewed more generally in terms of the demographic and spatial assumptions implied by the modelling design, recent methodological innovations, substantive and practical uses, as well as constraints and limitations associated with data availability. Although synthetic estimation remains by far the most popular technique, empirical and hierarchical Bayes alternatives make it possible to relax some of the most restrictive modelling assumptions. We note strong developments accommodating for informative sampling designs. Access to georeferenced survey microdata remains a challenge, but availability of auxiliary small area social, economic, environmental and demographic auxiliary datasets is improving. Greater consistency in the way validation studies are conducted could improve transparency in the literature, help researchers learn from previous investigations, and contribute to increasing trust in and acceptability of model-based estimation methods.

Keywords: Public health intelligence, small area estimation, disease mapping, spatial modelling, spatio-temporal modelling, health demography.

3.1 Introduction

Information on the health and social conditions of local populations is valuable to monitor health inequalities as well as to plan local healthcare and prevention strategies. Yet, estimation of local characteristics is often difficult in presence of small samples or unstable rates of events such as deaths or hospitalisations. Furthermore, raw prevalence and incidence rates are misleading measurements of risk which must be adjusted and smoothed in several ways. As a result, the use of regression modelling not only to calibrate/stabilise information from surveys and registers, but also to interpret spatial and temporal trends, has become popular in the last decade.

Notwithstanding rapid developments in SAE and Bayesian disease mapping, adoption in various fields of social statistics is unequal. While atlases of rare disease and epidemiological investigations of hospital statistics are now routinely based on those methodologies, extensions to fields such as transport safety, social research and official statistics production have proven slower. One explanation is that model-based estimation rests on a range of assumptions regarding the demographic and spatial structure of health outcomes which are not always explicit to end users.

Despite the existence of periodic reviews on methodological advancements in the field of SAE, practical applications and uses in public health surveillance and intelligence have been insufficiently examined. The aim of this systematic methodological review is to (a) provide an accessible overview of existing capability in the area of health SAE and some disease mapping studies (b) make modelling assumptions and their implications more understandable to practitioners (c) appraise the design of validation studies and highlight what can be learnt from them.

The paper is organised as follows: a first section provides definitions and background on the research area under scrutiny (see sections 3.2.1–3.2.2), as well as an introduction to the most typical small area models (see section 3.2.3). Section 3.3 presents the search and information extraction strategy, section 3.4 results, and section 3.5 widens the discussion on spatial structures, gaps in SAE design methodology and challenges involved in model-driven inference.

3.2 Background

3.2.1 Definitions

In public health research, a degree of ambiguity subsists between disease mapping and SAE. Some publications define SAE as model-based or model-assisted inference from a survey on small domains or small areas, separate from disease mapping notwithstanding some innovations borrowed from it, for instance spatial priors (Gomez-Rubio et al., 2010). Others treat area-based (lattice) disease mapping as a subdivision of SAE that relies on spatial modelling (Baptista et al., 2015). The present review adopts a wide definition encompassing all studies which involve an explicit model-based prediction approach (therefore excluding basic poststratification) which borrows strength from one or more covariates and/or a complex and explicit covariance structure incorporating dependence across space, time or any other dimension. This, by default, excludes the large number of studies applying crude shrinkage or smoothing estimation, especially with mortality and hospitalisation rates.

Within the broad SAE literature, we recognise three approaches to borrowing strength with models:

- **Survey SAE** is heavily based on survey sampling theory, and therefore models a survey direct estimator usually under a normal likelihood, as this is the natural stochastic process underpinning the error of sample estimators. This area of research focuses on variance smoothing for survey estimates, with strong efforts to model informative sample designs. Models often borrow strength from one or several covariates.
- **Disease mapping and 'small area analysis'** traditionally based on Poisson, binomial and Bernoulli likelihoods. This approach has developed from early research on shrinkage estimation (James-Stein, empirical Bayes) and focuses on shrinking estimates with excessive variance towards a global mean, or even spatially smoothing them (local shrinkage). The use of covariates is less common. This approach has focused on register data for rare events such as death. A recent expansion of this field is spatio-temporal epidemiology, borrowing developments from the age-period-cohort modelling literature (Bernardinelli et al., 1995; Sun et al., 2000).
- **Poststratification synthetic estimation** is typically the most rudimentary as it tends to ignore survey sampling designs. It analyses data from a purely mechanistic viewpoint with unit-level models, 'under the assumption that the

small areas have the same characteristics as the large area' (Gonzalez, 1973, p. 33). Disease prevalence models used in synthetic designs often bear a resemblance to those used in the epidemiological literature to estimate demographic relative risks. More occasional synthetic estimation studies are designed at the area-level, and model between-area heterogeneity using area-level aggregate characteristics. This approach is often not directly concerned with estimating local parameters as much as with producing a summary of expected morbidity at a local level given existing epidemiological knowledge on risk factors. Local sample sizes are typically small if not zero.

The above categories are no longer necessarily mutually exclusive. The present review identifies occasional convergence of methods in recent studies, particularly around very original mixed designs drawing inference from a combination of survey and register data. Two further approaches could be mentioned. First, the 'model-assisted' approach, grounded in traditional survey sampling, shares some developments with the model-dependent SAE approach, but has so far limited application in health research. It is not covered in the present review and we refer to Särndal et al. (1992), Lehtonen and Veijanen (2009) and Särndal (2010a) for an overview. Second, a variety of techniques used in the 'microsimulation' literature (Rahman and Harding, 2016), while sharing some of the design objectives of model-based SAE—and occasionally some of its assumptions—are not usually grounded in a theory of statistical inference. They are therefore not included in the scope of this review either.

3.2.2 Existing reviews

Reference material by Rao and Molina (2015), Lehtonen and Veijanen (2009) and Datta (2009) provides detailed technical descriptions of many techniques developed to date. In addition, a number of periodic reviews have summarised research progress in the field (Rao and Yu, 1994; Rao, 1999; Pfeffermann, 2002; Saei and Chambers, 2003a; Chambers et al., 2006; Whitworth, 2013; Pfeffermann, 2013; Rao, 2014). In contrast, applications in population health research have not been reviewed with a wide scope in relation to modelling and design characteristics, validity, data sources and sample sizes needed. The lack of applied review impedes the scrutiny and development of model-based public health statistics. Comparatively, and in spite of a slower adoption of hierarchical model-based methods, applications to road traffic safety have for instance been much better reviewed (Lord and Mannering, 2010; Dupont et al., 2013; Mannering et al., 2016).

A handful of publications have reviewed narrower aspects, starting with developments of spatial modelling. Pascutto et al. (2000) and Lawson et al. (2000) previously examined implications of spatial priors and hyperpriors in disease mapping, including in the presence of outliers. This is complemented by Wall (2004) who compared the spatial structures implied by simultaneous and conditional autoregressive models. Excellent technical descriptions of space-time designs for SAE have also been produced by Sun et al. (2000), Best et al. (2008) and Baptista et al. (2015) on health topics. Systematic reviews by Paul-Shaheen et al. (1987) and Scarborough et al. (2009) examined respectively the design of small area health analyses and the validity of 16 synthetic SAE health applications, operationalising traditional social science research validity criteria. Srebotnjak et al. (2010) and Zhang et al. (2015) proposed more data-driven approaches to testing the validity of SAE outputs.

Some aspects remain little discussed, especially around the design of data collection strategies . With regard to sample allocation, Longford (2006) and Molefe et al. (2015) gave reviews in a model-assisted framework, but no such solutions are available in a model-based framework. As for inferential aims (detection, ranking, prioritisation of interventions, etc.), only Longford (2015) provides a review, while specific problems are examined by Miaou and Song (2005), Abellan et al. (2008) and Li et al. (2012) in their own domains.

The present review thus examines the range of techniques currently used and their substantive implications. Although the use of hierarchical model-based approaches has been well reviewed in fields such as health care provider profiling (Landrum et al., 2000; Normand et al., 2016), safety monitoring (Racz and Sedransk, 2010), the adjustment of mortality and hospitalisation rates (Lawson et al., 2000), meta-analysis (Congdon, 2014) or school rankings (Goldstein and Rasbash, 1996; Lockwood et al., 2002), the application to health information has been the object of comparatively less review. Yet, these have specific implications in terms when appraising the underlying assumptions of some model-based health statistics, or using the estimates for practical purposes such as ranking areas, monitoring change, or allocating resources. Consequently, some users may encounter difficulties in understanding and using these statistics, and data holders may be reluctant to apply these strategies to their own data. This explains why validation studies have been in high demand.

In this paper, we take the view that much of the validity of model-based SAE for health is dependent on the availability of good data and good models. The former prerequisite is a question of availability of auxiliary data from an exhaustive source at the population level and it is important to understand what these data requirements

are. The latter pertains to the type of model selection and outcomes data available to produce a good working model. This involves understanding what important predictors are, their quality, and common criteria used to appraise 'good models'. Overall, both depend on understanding what the most important predictors are which play an important role in the data generation process of local health outcomes across a population, which is also the object of this review.

3.2.3 Modelling notions

We introduce the basic terminology as follows. We refer as 'domain' to units for which prediction is desired. In most instances domains would be small geographical subdivisions, but it often occurs that the targeted prediction domain is a cross-classification of demographics trait and geographical area of residence; for instance ethnic group within a given district. We therefore use the umbrella term 'domain' to avoid confusion. Model-based SAE for public health is best represented by considering a working model predicting a characteristic of interest in a population U of size N partitioned into a set of small domains $d = \{1, \ldots, D\}$ made up of patients, households or individuals $i = \{1, \ldots, n\}$. These domains have respective population sizes $\{N_1, \ldots, N_d\}$. Data are available from a sample s of size n with small area sizes $\{n_1, \ldots, n_d\}$, made of individual-level values y_{id}. These can be continuous values in the case of a body mass index (BMI), binary in the case of a health status, or counts in the case of hospital admissions. Before treating the mathematical expression of small area models, we describe their most important components.

A first stage called the *sampling model* defines the sampling process with reference to a small area population parameter of interest: mean, proportion or rate. Respectively, sample values y_{id} can be:

- continuous, generated by a normal distribution with area mean μ_d;
- binary values from a Bernoulli distribution with prevalence/probability of success p_d;
- counts from a Poisson variable with rate λ_d.

A second stage often referred to as the *linking model* focuses on reproducing processes driving the variation across small areas. In other words, the linking model is used to summarise all that is known about the spatial structure in this outcome through the model's covariance matrix. The covariance matrix reproduces the data's spatial structure thanks to a combination of a deterministic part (fixed effects) and a stochastic part (random effects). The deterministic part reflects the correlation with characteristics $\mathbf{X}$ of small areas (such as urban/rural class, poverty, deprivation) or

of the individuals making up their populations. The model borrows strength from this information using some generalised linear relationship $\mathbf{X}\boldsymbol{\beta}$. The stochastic part reflects over-dispersion driven by 'missing risk factors' beyond the deterministic process. This residual between-area heterogeneity can be described as an unstructured latent random variable $\boldsymbol{\upsilon} = \{\upsilon_1, \ldots, \upsilon_D\}$, often normal, typically unobserved, but which can be estimated thanks to multilevel or hierarchical modelling. Sometimes, over-dispersion in the parameter of interest exhibits spatial dependence, meaning that area residuals υ_d will tend to be similar in small areas which are close to each other (spatial autocorrelation). In those cases, a further variable $\boldsymbol{\nu} = \{\nu_1, \ldots, \nu_D\}$ can be incorporated which takes into account spatial dependence information from a geographical information system (GIS). The specific definition of this variable can either be another random variable with a predefined underlying spatial covariance function, or a fixed effect covariate produced using a non-parametric technique. Finally, some designs have an additional component ϕ which here denotes more complex variance structures, for instance space-time interactions, survey design effects or exposure to spatially smooth unobserved risk factors, when the estimation of an unconfounded dose-response relationship is of interest.

General linear model The unit-level model by Battese et al. (1988) consists of a basic sampling model with normal errors and a linking model describing the relationship between domains:

$$\begin{aligned}
\text{Sampling} \qquad & y_{id} \mid \mu_d, \sigma_\epsilon \sim \text{Normal}(\mu_d, \sigma_\epsilon) && (3.1)\\
\text{Linking} \qquad & \mu_d \mid \sigma_\upsilon, \boldsymbol{\beta}, \mathbf{X} \sim \text{Normal}(X_{id}\boldsymbol{\beta}, \sigma_\upsilon) && (3.2)\\
\text{Working model} \qquad & y_{id} = X_{id}\boldsymbol{\beta} + \upsilon_d + \epsilon_i && (3.3)\\
\text{EBLUP} \qquad & \hat{\bar{y}}_d^{\text{EBLUP}} = \bar{X}_d\boldsymbol{\beta} + \upsilon_d && (3.4)
\end{aligned}$$

υ, ϵ are identically normally distributed random area and sampling effects, respectively. $\mathbf{X}$ is a matrix of individual- and area-level covariate data X_{id} (auxiliary data) known for the entire population. Prediction can be carried out using expression (3.3) at the unit-level by ignoring sampling error $\boldsymbol{\epsilon}$, before aggregating all units i into domain aggregates. Alternatively, it is possible to use a simplified aggregate-level equivalent (3.4) simply with domain-level covariate means $\bar{X}_d = N_d^{-1} \sum_{i \in d} X_i$. Both lead to the empirical best linear unbiased predictor (EBLUP), a form of EB prediction.

A competing area-level approach examines the sample mean (maximum likeli-

hood estimate) $\hat{\bar{y}}_d = \sum_{i \in d} y_{id}/n_d$ described by Fay and Herriot (1979).

$$\begin{aligned} \text{Sampling} \qquad \hat{\bar{y}}_d \mid \mu_d, \sigma_\epsilon &\sim \text{Normal}(\mu_d, \sigma_\epsilon) && (3.5) \\ \text{Linking} \qquad \mu_d \mid \sigma_\upsilon, \boldsymbol{\beta}, \mathbf{X} &\sim \text{Normal}(\bar{X}_d \boldsymbol{\beta}, \sigma_\upsilon) && (3.6) \\ \text{Working model} \qquad \hat{\bar{y}}_d &= \bar{X}_d \boldsymbol{\beta} + \upsilon_d + \epsilon_d && (3.7) \end{aligned}$$

Here the model is identified by presuming that the sampling error of $\hat{\bar{y}}_d$ is known in each area d, which is the case provided that the sampling process itself is known. The simplest case occurs if variable y_{id} is a 0/1 indicator with success probability p_d. Under simple random sampling, the binomial law implies that the sampling variance of $\hat{\bar{y}}_d$ is $\text{Var}(\hat{\bar{y}}_d) = p_d(1 - p_d)/n_d$, usually estimated by $\widehat{\text{Var}}(\hat{\bar{y}}_d) = \hat{\bar{y}}_d(1 - \hat{\bar{y}}_d)/n_d$. If sample means $\hat{\bar{\mathbf{y}}}$ are computed for very small samples, this estimator is typically unstable and requires smoothing. The Fay-Herriot model has been applied for situations where $\hat{\bar{y}}_d$ is not the maximum likelihood estimator, for instance design-based estimators obtained with complex sampling designs. More complex variance estimation is then required.

Both the unit-level and area-level model relax the regression synthetic estimator based on linear model $y_i = X_i \boldsymbol{\beta} + \epsilon_i$ or $y_d = X_d \boldsymbol{\beta} + \epsilon_d$. Synthetic estimation, used in population health since initial investment by the US National Center for Health Statistics (1968), neglects any between-area variation that is not predicted directly by fixed effects. It is rarely the case that $\mathbf{X}\boldsymbol{\beta}$ can be expected to predict the entirety of the between-area variance.

Generalised linear model While survey sampling theory is built around asymptotic properties of normal distributions, other data likelihoods are commonly used with discrete health outcomes: Bernoulli, binomial, Poisson, Weibull, exponential, mixture distributions. Appropriate likelihoods for health data are reviewed in detail by Flanders and Kleinbaum (1995). With those, models are specified with linking functions which makes their computation more complex, but may better represent the sampling error of the outcome of interest. In this case, the response is not modelled directly: a link with the population parameter of interest is required.

$$\begin{aligned} \text{Sampling} \qquad y_{id} \mid \eta_{id}, \ldots &\sim \pi(\eta_{id}, \ldots) && (3.8) \\ \text{Linking} \qquad g(\eta_{id}) \mid \sigma_\upsilon, \boldsymbol{\beta}, \mathbf{X} &\sim \text{Normal}(X_{id}\beta, \sigma_\upsilon) && (3.9) \\ \text{Working model} \qquad g(\eta_{id}) &= X_{id}\boldsymbol{\beta} + \upsilon_d && (3.10) \end{aligned}$$

where $\pi(\cdot)$ is a known density function (Bernoulli, binomial, etc.) and $g(\cdot)$

a linking function, most commonly the logarithmic or logistic link, occasionally probit, negative binomial or log-log links. In this case, the unit-level synthetic estimator is just:

$$\theta_d^{\text{SYN}u} = \frac{n_d}{N_d}\hat{\bar{y}}_d + \frac{N_d - n_d}{N_d}\sum_{i\in d} g^{-1}(X_{id}\hat{\boldsymbol{\beta}}) \tag{3.11}$$

Omitting i in (3.8-3.10) and substituting $\bar{y}_d$ for y_{id} in (3.8), we obtain an area-level model (often used in disease mapping) and the synthetic estimator of η_d is then:

$$\theta_d^{\text{SYN}a} = \frac{n_d}{N_d}\hat{\bar{y}}_d + \frac{N_d - n_d}{N_d} g^{-1}(X_d\hat{\boldsymbol{\beta}}) \tag{3.12}$$

The **empirical best predictor** rests on the observation that the synthetic predictor θ_d^{SYN} is biased, while the unbiased direct or maximum likelihood estimator $\hat{p}_d$ is too unstable to be used. The EB predictor (and its Bayesian equivalent) combines both estimators as a trade-off between bias and variance by weighting them as a function of the variance of the maximum likelihood estimator (see Efron and Morris, 1973; Jiang and Lahiri, 2006). The more precise the maximum likelihood estimator is, the more weight it is given. The EB predictor does not have a closed-form expression but can easily be computed with Monte Carlo integration. We take the example of an area-level binomial logistic model:

$$\begin{aligned}
\theta_d^{\text{EBP}}(\mathbf{X}, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2) &= \frac{n_d}{N_d}\hat{\bar{y}}_d + \frac{N_d - n_d}{N_d}\text{E}_M(\eta_d|\mathbf{X}, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2) \\
\text{E}_M(\eta_d|\mathbf{X}, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2) &= \\
&\exp(X_d\hat{\boldsymbol{\beta}})\frac{\text{E}_\xi\left[(\sum_{i\in d} y_i + 1)\hat{\sigma}\xi - (n_d + 1)\log(1 + \exp(X_d\hat{\boldsymbol{\beta}} + \hat{\sigma}\xi))\right]}{\text{E}_\xi\left[\sum_{i\in d} y_i\hat{\sigma}\xi - n_d\log(1 + \exp(X_d\hat{\boldsymbol{\beta}} + \hat{\sigma}\xi))\right]}
\end{aligned} \tag{3.13}$$

where $\text{E}_M(\cdot)$ denotes a model-based expectation and $\text{E}_\xi(\cdot)$ is evaluated over $\xi \sim \text{N}(0, 1)$.

A simpler and faster to compute approximation of the above predictor has proved popular: the **naive empirical best predictor** or empirical plug-in predictor (EPP) is computed by introducing component $\boldsymbol{\upsilon}$ from model (3.10):

$$\theta_d^{\text{naive}} = \frac{n_d}{N_d}\hat{\bar{y}}_d + \frac{N_d - n_d}{N_d} g^{-1}(X_d\hat{\boldsymbol{\beta}} + \upsilon_d) \tag{3.14}$$

Studies comparing the performance of the naive and full EB conclude that

the difference is small with large samples (Jiang and Lahiri, 2006) despite the added computational burden. The two basic model structures described above have been extended in many ways, by refining the area effects $\boldsymbol{v}$, relaxing some of the distributional assumptions on both sampling and linking stages, or adapting them to account for complex sampling designs. Once a model is specified and estimated, it must be tested to establish confidence in its reliability. This is the objective of model diagnostics which establish whether the model accurately describes all determinants of the between-area variation in the parameter of interest. This step is important to establish the credibility of the working model and can involve techniques such as hypothesis testing, residual checks or leave-one-out cross-validation.

3.3 Methods

3.3.1 Search strategy

Citations were searched on Scopus, Ovid/EMBASE, PubMed/MEDLINE, Web of Science™ Core Collection and BIOSIS Citation Index. The query was designed to target the SAE literature and take the subset of citations relating to common health-related outcome or determinant. For that reason keywords were kept as general as possible, with a decision to include deprivation, unemployment and road safety. We targeted citations which titles, abstracts, keywords and indexing keywords included at least one of 'small area estim*', 'synthetic estimat*', 'model-based estimat*' or 'model-assisted estimat*' and at least one of 25 words related to public health outcomes and determinants, including deprivation, behavioural and lifestyle risk factors and safety outcomes. Queries and their corresponding numbers of results are presented in appendix (Table A.1).

This search strategy is imperfect due to its restrictive first clause (containing a reference to SAE or model-based estimation). This is due to the difficulty in separating SAE from disease mapping studies as well as the heterogeneity in denominations used in disciplines such as geography, demography and actuarial science. We therefore introduced 41 references identified from forward and backward citations which were relevant despite not being reached in the search. Among those, 8 road traffic safety studies were published at a recent date were not explicitly linkable to the SAE literature based solely on their abstracts and titles, despite adopting now well-established disease mapping modelling designs. While models have been used for some time to measure risk factors associated with traffic accidents, they had until recently not been used for the purpose of identifying hotspots and producing pre-

cise road safety statistics. Their substantial methodological overlap with other SAE studies justifies examining them too.

In contrast, we excluded income and poverty mapping studies from the scope of this review. These studies correspond to somewhat different set of methodologies, mostly developed in the economic literature. This area of research is in particular dominated by the 'World Bank' method (Elbers et al., 2003) and has already been reviewed extensively elsewhere (see for instance Christiaensen et al., 2012; Tarozzi and Deaton, 2009; Haslett et al., 2010a,b; Molina and Rao, 2010).

Citations were screened to meet the following inclusion criteria (see flowchart in Figure 3.1):

- description of a statistical design involving predicting prevalence of a health-related status;
- application with real data or analysis and validation of existing small area estimates;
- SAE as opposed to pure smoothing: use of covariate auxiliary information within the survey model and/or of explicit random effect structures.

The following exclusion criteria were applied:

- studies of income or poverty;
- implicit synthetic estimation without regression modelling (poststratification estimates);
- microsimulation.

3.3.2 Analysis

We counted the number of models per study. In some studies reporting several models for one or more outcomes, we counted one model per health outcome and geographical scale of prediction, even when a wider range of models were considered during model selection. We also recorded the basic likelihood and prior structure of each model and undertook classifying them.

Differences exist between frequentist EB/EBLUP, empirical Bayesian and fully Bayesian approaches to inference, not least in their mathematical presentation and computational implementation. These differences are yet of limited relevance to the scope of the present review. Indeed, all three approaches implement predictive assumptions by using entirely comparable model structures. Notwithstanding, the HB approach is recognisable in that (a) it takes into account uncertainty in random components (the parameters determining the level of spatial heterogeneity); and (b) it produces full posterior distributions for small area indicators of interest, enabling

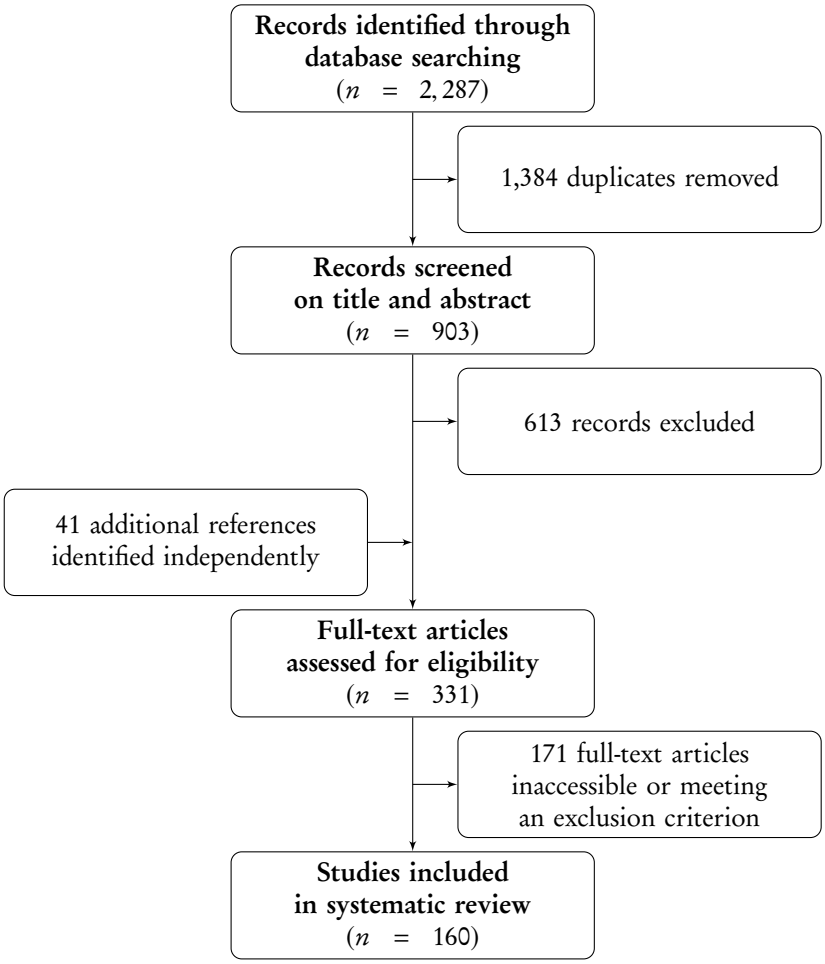


Figure 3.1 | Study identification and inclusion flowchart

Table 3.1 | Breakdown of models analysed by type of health outcome

Category	Number of models	Percentage
Health statuses	50	19%
Behavioural risk factors	42	16%
Disease diagnoses	33	12%
Economic status	30	11%
Psychiatric morbidity	27	10%
BMI and overweight	14	5%
Care utilisation	15	6%
Mortality	11	4%
Immunisation	9	3%
Health insurance	9	3%
Reproductive health	8	3%
Traffic collisions	8	3%
Other	12	4%
Total	268	100%

important post hoc analyses (Richardson et al., 2004). Studies were consequently marked according to three categories: synthetic; EB/EBLUP/empirical Bayes; and HB. In addition, they were classified as area-level or unit-level. For each study, the following characteristics were recorded: year; region; number of models described; type of health outcome; number and description of the set of domains; population of inference; basic specification and covariance structure of the best fitting model; validation method (comparison with direct estimates, comparison with external estimates, simulation); type and source of fitting data; fitting data sample size (in the case of surveys); type and source of auxiliary data; out-of-sample prediction (attempted, not attempted, not reported); whether accessed data geographical identifiers (yes, no, not reported); borrowing strength (space, time, space-time, neither). The percentage of the between-area variance explained by the best fitting model was also calculated. Although it is difficult to report this characteristic in a truly comparable way, we used the model's coefficient of determination (R^2) in the case of area-level linear models used for synthetic prediction. In the case of mixed models used for synthetic, EB or HB prediction, the percentage of level-2 variance explained by covariates is calculated, when possible, by comparing the level-2 variance of a full model with that of a null model without covariate. This is the simplest approach achievable given the information present in published papers (see Xu, 2003 for a discussion).

3.4 Results

3.4.1 Descriptives

3.4.1.1 Study characteristics

This section describes characteristics of 160 citations, most of which are from North America (61%) and Europe (29%). Seven studies were from the Asian subcontinent, 5 from Africa, 4 respectively from Israel, Mexico, Chile, New Zealand and 1 study conducted on an international scale with countries acting as small areas. Over half (53%) were published after 2010. They report on a total of 268 models, most of which concerned health statuses (disability, general health), behavioural risk factors such as smoking, drinking and physical activity, diagnosed or clinically proven disease prevalence, economic status, and psychiatric morbidity (see breakdown in Table 3.1). These correspond to indicators in high demand for public health planning. A smaller number of investigations have looked at indicators of immunisation coverage and health insurance.

Seventy-eight per cent of studies considered in this review analysed survey data (n = 125), 16 per cent analysed data from various types of registers (n = 26), the remaining 6 per cent reporting on models applied to census data (n = 2), purposive research studies (n = 4), or pooling information from a combination of sources (n = 3). Out of 125 designs drawing inference from sample surveys only, 86 relied on health surveys (including health examination surveys), 32 on general social and labour force surveys, 6 relied on a Demographic and Health Survey (DHS) or a health surveillance system. We identified one study modelling school survey results (Leroux et al., 1996). Out of 23 studies drawing inferences from register data, 7 were based on traffic collision reporting systems, 7 on health care records, 5 on death registers, 4 on cancer registers, 2 on birth registers and 1 on an army register.

The domains of inference were extremely varied. Most studies looked at administrative subdivisions such as US counties, UK districts, but also finer statistical output geographies. A small number of studies investigated unusual domains, such as purposive grid squares (Abellan et al., 2008), schools (Li and Zaslavsky, 2010; Li et al., 2010), road segments and intersections (Miranda-Moreno et al., 2007; Mitra and Washington, 2007). Some studies based on synthetic estimation do not refer to specific domains as their models are designed purely at the individual level (see Benzeval and Judge, 1994, 1996).

Sampling fractions and total sample sizes relative to number of domains are scarcely discussed outside of validation studies. Authors have little control over

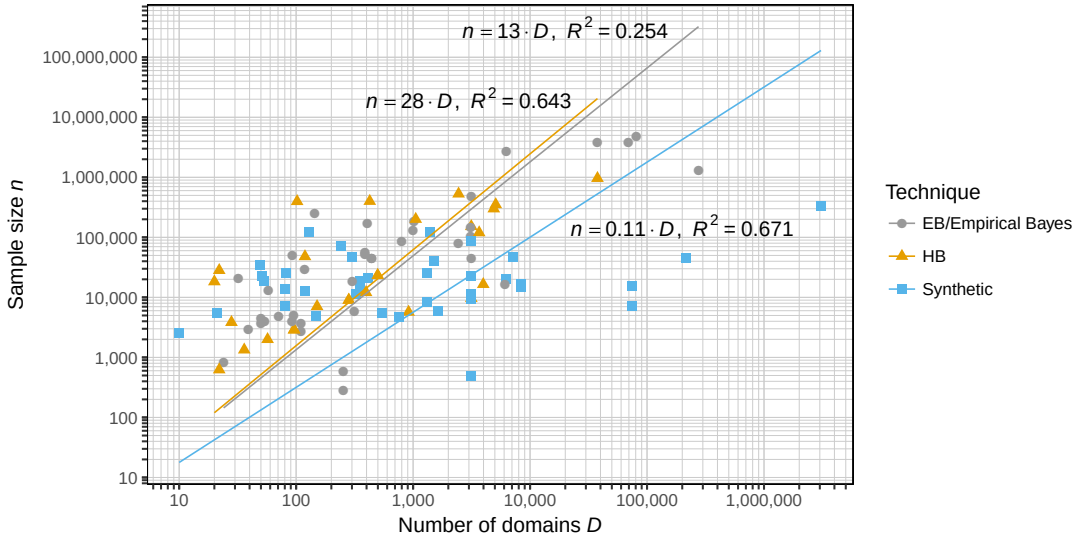


Figure 3.2 | Plot of survey sample size to number of domains by prediction technique with affine lines of best fit ($N = 104$ models)

the sample size of the fitting data they use since almost all studies were based on secondary data sources. They are however generally able to decide on the size of the domains. Despite this, it appears that there is very little relationship between the number of domains and the study sample sizes in the subset of studies relying on survey samples (see Figure 3.2). For some synthetic estimation studies, the (albeit large) sample size is even inferior to the number of domains, for instance Zhang et al. (2013) make inferences at the census block-group level, or Congdon (2009b). Sample sizes are also often abnormally small even for studies based on EB or HB designs, where we find an average sample size per domain of respectively 13 and 28. Yet, only a handful of studies even consider the issue. Cook et al. (2016) do not publish EB predictions for domains with fewer than 20 observations out of concerns for reliability. Lopez-Vizcaino et al. (2013) find that below a minimum area sample size of 90, their HB predictors cannot achieve relative standard errors (RSEs) inferior to 20 per cent. This attests to the relatively crude understanding of design implications, at least in applied research.

3.4.1.2 Prediction methods and model parameterisations

Synthetic, EB and HB prediction are represented evenly in our inclusions and across most categories, aside for a very clear domination of HB methods in register-based studies (81%). Over half (55%) of studies were designed with unit-level models, 47 per cent of which use synthetic estimation. In substantive papers, it is not always obvious whether the EB (3.13), the naive plug-in predictor (3.14) or the syn-

thetic estimator (3.11) are used or, in the case of the first two techniques, what is done about areas with no observations. This is especially obvious when journal space is limited and when is no incentive to provide supplementary material. We did find that almost all EBs are in fact naive 'plug-in' predictors despite the full text making no explicit reference to it.

Synthetic estimation remains popular when the sample microdata are provided with masked primary sampling units (the geographical location of which is not communicated, see for instance Twigg and Moon, 2002; Congdon, 2008a) or when the ratio of the sample size to the number of domains is too small for EB or HB prediction be reliable (Congdon, 2009b; Zhang, 1999). It comes with a disadvantage identified early on by Gonzalez and Hoza (1978, p. 8): 'the use of synthetic estimates is limited by the method's inability to account properly for local factors'. This is commonly referred to as the 'synthetic assumption'. Models cannot predict all local outcomes unless a covariate is so important that it is justified to use it directly as an indicator. One variation of synthetic estimation is known as 'multilevel synthetic estimation' (Heady and Ruddock, 1996; Twigg et al., 2000). While it also assumes that all areas be identical outside of differences captured by fixed effect covariates, authors are nevertheless able to reflect unaccounted residual between-area heterogeneity through wider prediction intervals. Indeed, Farrell et al. (1997a) and Heady et al. (2003) used a naive approximation to the mean squared error (MSE) of the synthetic estimator which explicitly takes into account this residual variance $\widehat{\mathrm{MSE}}(\boldsymbol{\theta}^{\mathrm{SYN}_u}) = \sigma_v + \mathrm{Var}(\boldsymbol{X}\hat{\boldsymbol{\beta}})$. By doing so, even if EB/HB point estimates are not available or that the data set is too sparse for them to be reliable, multilevel synthetic interval estimates at least display the lack of sensitivity of the synthetic point estimate.

EB and HB prediction relies on mixed effect models to produce a hybrid predictor between the biased synthetic predictor and the noisy direct sample estimator. Such models are built on various sampling and linking distributions listed in Table 3.2 which, respectively,

(a) take into account the sampling error processes for the direct estimator and
(b) relax the synthetic assumption of between-area homogeneity with the help of latent random domain effects.

Domain random effects fulfilling (b) are in the vast majority of studies assumed to be independently and normally distributed, as per the original designs proposed by Fay and Herriot (1979) and Battese et al. (1988). These are unstructured and

often referred to as 'normal exchangeable random effects'. Other distributions are suitable when normality is not a reasonable assumption due to a skew (log-normal or arcsine) or outliers (Student's t). In practice, little is known about the underlying distributional characteristics of domain random effects. Nonparametric approaches to EB/HB have been proposed by Opsomer et al. (2008) which evade distributional assumptions, yet they have not been applied to population health outcomes. Semi-parametric methods have also been put forward. Chaudhuri and Ghosh (2011) have developed an empirical likelihood approach to Fay-Herriot and unit-level Battese et al. (1988) models, placing no specific distributional assumption on domain random effects. A competing semi-parametric approach exists in M-quantile models used with both continuous (Tzavidis et al., 2008) and discrete outcomes (Tzavidis et al., 2015; Chambers et al., 2016). The hierarchical domain structure is replaced by M-quantile regression coefficients averaged over units of every domain. This gives rise to domain-level models used for prediction. Although these semiparametric approaches have been applied to health outcomes in statistical case studies, they remain to be tested and adopted by practitioners.

We note that some studies only specify random effects for some rather than all domains. Twigg et al. (2004) specified random regional effects despite carrying out prediction for smaller domains. Similarly, some studies predicting health characteristics for cross-classifications of geographical areas and demographic groups assume homogeneity in the demographic groups' risk ratios across areas; see for instance Olives et al. (2013). This approach is computationally appealing, sometimes even inevitable due to data sparsity, but it implies homogeneity in social groups which is in many studies not tested, when it is at all made explicit.

Recent studies have investigated exploiting the mutual dependence of normal effects with structured priors. These include the multivariate normal prior, the conditional autoregressive (CAR) prior and random walks priors such as first-order random walk (RW1). These are used extensively in applied research so as to exploit dependence structures in terms of spatial proximity, geographical contiguity, time adjacency or some notion of proximity across social groups, for instance closely related age groups. These structures and their use is the object of further discussion in sections 3.4.4 and 3.5.1.

We pay close attention to these structures as they determine the meaning and uses that can be attached to model-based estimates. From a substantive viewpoint, a few studies are transparent with regard to the type of inferences to be drawn, and make the range of acceptable uses explicit. Although many synthetic studies do not discuss implications of failing to account for residual heterogeneity, Walsh and

Table 3.2 | Common parametric specifications for sampling and linking models

Sampling model	Linking model	Observations
Normal	Normal	Transformations are sometimes used to ensure normal sampling errors: logarithmic (Fay and Herriot, 1979), logistic (binomial) (Mercer et al., 2015; Song et al., 2016), arcsine Dempster and Hwang (1994); Jiang and Lahiri (2001); Krapavickaitė and Rudys (2015). Extensions have been proposed with correlated or multivariate sampling error for panel surveys and correlated outcomes (Elston et al., 1991; Paez et al., 2010; Söhl et al., 2016; Li and Zaslavsky, 2010)
Normal	Student's *t*	Outlier robust. Example: Xie et al., 2007
Bernoulli	Normal	With logistic link function.
Binomial	Normal	With logistic link function.
Binomial	Beta	Example: Martuzzi and Elliott, 1996
Poisson	Normal	With logarithmic link function.
Poisson	Student's *t*	With logarithmic link function. Example: Pascutto et al., 2000.
Poisson	Gamma	With logarithmic link function.
Gamma model		Special parameterisation suitable for alcohol consumption: see Rehm et al., 2010; Castillo-Carniglia et al., 2015
Beta model		Jonker et al., 2013a.

Haseeb (2014) for instance state that their synthetic predictions are not a reflection of the local epidemiology but instead can be used to identify areas with a socioeconomic profile which suggests that epidemiological investigations are most likely to be needed. Other reports insist on the need to validate estimates (Twigg et al., 2000; Scarborough et al., 2009), but only few offer explicit and scientific rules of interpretation of ranks, or detection of hotspots (Miaou et al., 2003; Miaou and Song, 2005; Congdon, 2008c) or unusual trends (Abellan et al., 2008; Li et al., 2012). It is likely that the growing sophistication of models will allow incorporation of explicit inferential aims in applied statistical designs.

3.4.2 Auxiliary predictors

Censuses and MYPEs are the main sources of auxiliary data in included studies. By far the most common model predictor is age, followed by sex and ethnicity. Studies also increasingly rely on data from administrative statistics compendiums such as the US Area Health Resources File (US Department of Health and Human Services, Health Resources and Services Administration, Bureau of Health Workforce, 2016), the German INKAR database (Bundesinstitut für Bau-, Stadt- und Raumforschung, 2016) or public health departments' data warehouses (Georgia Department of Public Health, 2016). These constitute a precious infrastructure which

facilitates model building. Seven studies use principal components analysis as a data reduction method to make the most of the wealth of variables available and achieve very high percentages of between-area variance explained by covariates. Area classifications for urban-rural, socioeconomic and deprivation categories are particularly frequently used. A few studies used school and hospital register data, but unlike the point-referenced geostatistical literature, none used GIS or satellite data.

In most cases, auxiliary covariates are themselves subject to sampling error, or other forms of measurement error. Many studies estimate auxiliary covariates from survey samples subject to sampling error; Charlton (1998) used a household survey and a one per cent sample of UK 1991 census records. Twigg et al. (2000) estimated area covariates from the fitting survey data itself as the geographical location of households is provided in a masked form in their dataset. In the US, the American Community Survey has become the standard source of population characteristics data (US Census Bureau, 2008). As for register data and population estimates, they are commonly affected by coverage error, seasonal bias and inaccuracies at a small level (zu Erbach-Schoenberg et al., 2016; Cairns et al., 2016). Error in auxiliary data, whether for administrative or sample survey source, is almost never discussed despite being a common issue. Ybarra and Lohr (2008) produced the only design in included studies which takes into account error in auxiliary data with a measurement error model. This method has not been used in later studies included in this review.

Auxiliary data can be difficult to obtain at the desired level of prediction, but we note three situations where it may not be required at all:

- For various reasons, studies can include important unit-level confounding variables in the linking model, which are then discarded at the prediction phase, therefore absolving the need for exhaustive auxiliary data. This is often the case when the model is applied to estimate an age-adjusted parameter (in which case the model equation is applied to a standard population rather than to MYPEs) or when it is desirable to remove the influence of a confounding factor known to distort the intended interpretation (for instance, adjusting for the excessive influence of a local care home on a small area's morbidity rate, see Jonker et al., 2013b). In other studies, a similar approach is adopted when prediction is required for hypothetical populations or subpopulations of interest. For instance, Paez et al. (2010) estimate the average trip length for specific senior populations (unrepresentative of the actual resident population) to measure accessibility to health care services in specific urban areas. Inequality decomposition studies also proceed this way: Cook

et al. (2016) estimate mental health care need disparities using a 'recycled prediction' approach, in which they calculate the unwarranted shortage/excess in health care need across various ethnic groups.

- Some studies (including survey-based studies), carry out pure smoothing without covariates. This is frequent in space-time filtering studies (for instance Barker and Thompson, 2013; Boulieri et al., 2016a). The absence of covariate can yet reduce the design's efficiency in terms of prediction variance.
- Some studies analyse adequately-powered datasets which do not themselves require filtering or smoothing. They instead use modelling to predict an unobserved latent variable. For instance, Fabrizi et al. (2016) use a small area latent class model to predict the prevalence of a range of disability clusters using data from multiple items and scales in a household survey.

3.4.3 Area- and unit-level modelling

The unit-level model has an older history in the field of population health statistics, where evidence from national trials and investigations has long been 'applied' to local populations through age-standardisation, demographic methods, and regression models (Levy, 1979). This approach has later been developed further to make fuller use of data available for local populations from decennial censuses (Heady and Ruddock, 1996), as well as to implement more efficient EB (Battese et al., 1988) or HB (Malec and Sedransk, 1993) designs. Area-level designs, although present in earlier synthetic designs (Gonzalez and Hoza, 1978) owing to data limitations, have been introduced to decompose sampling and between-area variance by Fay and Herriot (1979) and and have been popular since.

Area- and unit-level have very different implications both in terms of survey sampling and in terms of modelling correlations with social, economic or geographical covariates. Area-level models may not make full use of the information held by the auxiliary information when it is available at the unit level; for instance turning existing individual-level data into aggregates may lose some of the breadth of information. A particular advantage of unit-level general linear models discussed in section 3.2.3 is that they can be developed in circumstances where the fitting data is individual-level, and the auxiliary data for the entire population is only available as aggregates. This is, however, not the case with generalised linear mixed models due to the link function which breaks the direct connection between the maximum likelihood estimator and the model's linking stage. In practice, this means that mod-

els based on a Bernoulli, binomial or Poisson sampling likelihood require unit-level auxiliary data for prediction, because the simplified EBLUP (3.4) does not exist for such models. This is not to say that both specifications are equivalent, even with general linear models. The absence of direct equivalence between area- and unit-level relationships is a well documented ecological fallacy problem (see Tranmer and Steel, 1998, 2001; Tranmer et al., 2003; Namazi-Rad and Steel, 2015).

The traditional Fay-Herriot model assumes normally distributed outcomes. Transformation of the response at the sampling stage, or at the linking stage of the model (for instance log, logit, arcsine-square root) may be necessary, although challenging in many ways. Proportion estimates can be unstable with small or zero denominators since the log link, for instance, is undefined for zero. Proportions equal to 1 can also cause issues with logit transformations. Only the logit transformation guarantees that predictions will be bounded by 0 and 1. Unmatched models under a variety of link functions were studied by You and Rao (2002), and a nonparametric Fay-Herriot model that avoids relying on transformations has been described by Porter et al. (2015b). Area-level models generally requires a very good knowledge of the survey design in order to specify a suitable variance function giving the sampling stochastic effects, as such a function is often in practice not known (Ha et al., 2014).

Area-level models, whether based on general linear or generalised linear mixed models, are in theory well adapted to large datasets. This is because the stable variances of large samples may not require to shrink or smooth them, but also because fewer difficulties are likely to arise with the transformation of responses, especially when the dependent variable is a proportion. In addition, an area-level model is often much easier to manipulate with large datasets since its covariance matrix has smaller dimensions than that of a unit-level model.

3.4.4 Structures for spatial, temporal and demographic dependence

Recent studies feature more sophisticated covariance structures which exploit the dependence between small areas or domains under some form of adjacency, proximity or network arrangement. Such structures are needed to reflect the continuity of health risks across time, space, and closely related social groups. From a statistical viewpoint, they are intended to reduce the prediction MSE, particularly bias. The fundamental covariance structure implementing such dependence structures is the multivariate normal distribution. Dependence is specified in the model variance-covariance matrix with non-zero off-diagonal covariance parameters. The resulting

distribution produces correlated random effects depending on the size of the covariance relative to the diagonal (variance) parameters.

3.4.4.1 Spatial dependence

Typical spatial models rely on an extension of this basic concept. Besag (1974) theorised a dependence structure known as 'intrinsic conditional autoregressive (ICAR)' which relies on a Markov conditional independence assumption under which area random effects are independent realisations of a normal variable, conditionally on proximity. This conditionality materialises into a spatial weights matrix determining the level of covariance between two given areas as a function of their adjacency or spatial proximity. The combination of an ICAR with an unstructured normal exchangeable (NEX) random effect forms the CAR or convolution prior. Instead of shrinking data towards the global mean as with a simple NEX prior (global shrinkage), this approach shrinks data towards local means (local shrinkage). In theory, this should result in smoother predictions, in spatial terms. In appropriate circumstances, it should also equalise prediction MSE across space.

In public health, spatial models have a strong substantive justification. Health outcomes exhibit spatial similarity or clustering (Lorant et al., 2001) and justify exploiting information on distances and spatial adjacency collated using a GIS. A basic illustration in disease mapping is discussed by Marshall (1991a); Mollie and Richardson (1991). The justification for local smoothing is best summarised by Richardson and Monfort (2000, p. 210):

> 'For studies at both the small and large scale geographical scale, it is clear that in most cases there will be unidentified or unmeasured confounders, some of them potentially varying continuously over space. There are many examples of such confounders including genetic characteristics, sun exposure and dietary habits. Spatial structure is therefore introduced to account indirectly for the aggregated effect of those unobserved risk factors'.

Although this justification relates to observational epidemiology models and act as a precaution against spurious inferences, it is also valid for prediction, in order to ensure that models successfully reproduce the spatial pattern of risk.

Out of 36 studies relying on spatial modelling, only 2 studies (Pereira et al., 2013; Drewnowski et al., 2014) implemented a simultaneous autoregressive (SAR) covariance specification for a Fay-Herriot model. Only one (Kokki et al., 2001) used the ICAR structure, thus assuming absolute smoothness in the risk surface. All

others adopt a more complex design coupling the ICAR with basic NEX effects. The resulting CAR structures couple a smooth risk surface with area-specific disruptions accounting for unobserved risk factors exhibiting a different structure. We include in this the Leroux-Lei-Breslow (LLB) structure proposed by Leroux et al. (2000), and applied in three studies (Congdon, 2008c, 2009b; Baptista et al., 2015). Few studies adopt other spatial structures; we only note one exception with Paech et al. (2010) who used fixed effect polynomial expansions of Cartesian coordinates.

3.4.4.2 Demographic dependence

Some designs aim to produce estimates for domains defined as a cross-classification of small areas and demographic subgroups. Some rely on independent random effects, either by modelling social groups separately (see for instance Dwyer-Lindgren et al., 2013), or by using independent covariance structures. Herrador et al. (2011)'s model of unemployment rates employs a single vector of independent exchangeable random effects of women's and men's local area effect with sex-specific variances. Other designs instead assume an interaction between area effects and social groups and pool some of the model effects: we refer to this method as 'shared component prediction', by reference to shared component models (Knorr-Held and Best, 2001). They specify correlated random effects and borrow strength not just from other areas, but also from other social groups within areas. These designs recognise that in spite of inequalities, social groups in a given area are likely to be similar: Malec et al. (1999) specifies random effects of six sex by ethnicity groups to have correlated random effects in each area, and Richardson et al. (2006) specify a correlation between sex-specific risks across areas. These are implemented with a multivariate normal structure. Such models are sometimes known as 'shared components models' due to the fact that they recognise that different social groups share some risk factors, and have become very common over the past decade. Another implementation of risk dependence between social groups within areas has been used by one author (Congdon, 2006a,b, 2009a,b; Congdon and Lloyd, 2010), with models implementing a random walk structure (identical to an ICAR or CAR structure) to reflect the fact that health risks are similar in adjacent age groups. Finally, 13 studies model several health outcomes jointly using multivariate response models, for instance to model morbidity and mortality simultaneously (Congdon, 2008a), or predict rates of road collision of different severities (Miaou and Song, 2005; Boulieri et al., 2016b). The primary goal is to gain efficiency by exploiting the correlation in these health outcomes.

3.4.4.3 Temporal dependence

Time is another source of dependence in data and models are well established which stabilise area unemployment time series independently (see for instance Tiller, 1992). Rao and Yu (1994) first introduced the combining this time filtering with borrowing strength nationally to increase the precision of these rates. Progress have been reviewed in detail elsewhere (Pfeffermann, 2002), but a number of small area studies proposed original solutions. Important contributions have been made which combine (a) seasonality and temporal dependence in time-specific rates (mostly unemployment) and (b) temporal dependence in the sampling error due to the design of longitudinal panel surveys and the rotating group bias (Hwang and Dempster, 1999).

To avoid excessive time dependency Pfeffermann and Tiller (2006); Pfeffermann et al. (2014) propose a two-stage benchmark procedure at the national level and then at the census division level which corrects prediction in case of rapid changes in time series caused by events such as a market crash or terrorist attacks. Similarly, epidemics can cause rapid changes in mortality which autoregressive models may fail to adapt to in the short term.

More sophisticated designs have emerged in the space-time disease mapping literature, when there is interest in measuring local trends rather than point epidemiological prevalence. A simple interactions between space and time by Bernardinelli et al. (1995) introduces a simple interaction between trend (slope of time) and area, while Sun et al. (2000) and Miaou et al. (2003) use an unstructured time effect interacted with areas and age group. Structured space time interactions are later proposed by Knorr-Held and Besag (1998) to exploit the spatial dependence in local trends.

3.4.5 Prediction for out-of-sample domains

It is not unusual that no data are available in some of the target domains. While synthetic predictors can be computed provided that auxiliary data is available, the default frequentist EB predictors (3.13) and (3.14) are undefined in the event of a zero sample size due to the unobserved exchangeable random effect. There is no standard approach for out-of-sample frequentist EB prediction, and studies do not always explain how they compute those. Those that do address this point usually compute the synthetic estimator in lieu of the undefined EB (Pfeffermann and Sverchkov, 2007; Eberth et al., 2013b,a). However, one study (Zhang et al., 2014) imputes missing random effects with the arithmetic mean of neighbouring areas' random effects, in an attempt to recognise the spatial autocorrelation in random effects.

Fully Bayesian studies face no such obstacle since HB posterior distributions are still defined even with missing data. In such situations, posteriors will typically be wide since the prior and particularly the linear predictor $\mathbf{X}\boldsymbol{\beta}$ take precedent, leaving some uncertainty. When explicit structures are defined in the model, for instance conditional autoregressive structures based on time, space or age, posteriors can still be relatively narrow thanks to the ability to borrow strength from adjacent areas or groups. Although this aspect is not discussed in depth in the design justifications of included studies, we note that frequentist work by Saei and Chambers (2005, 2011) has demonstrated possible efficiency gains under a spatially autoregressive model.

3.4.6 Validation methods

Validity has been a strong concern from the earliest attempts to develop inference based on synthetic estimators, due to difficulties in ascertaining the magnitude of their inherent bias. This has remained a concern with HB and EB prediction, the consistency of which is still conditional on the model being true. We examine validation work from the angle of (a) model validity and (b) prediction validity.

3.4.6.1 Model validity

Forty-six per cent of studies report some form of model comparison or a selection procedure. Few describe a systematic covariate selection strategy, possibly owing to the lack of demonstration that automated model selection is reliable. Only Heady et al. (2003) describe in detail a stepwise forward selection process combined with examination of the T^2 statistic. HB designs have the advantage of offering Bayesian covariate and random effect selection, using a simple 0/1 indicator with a Bernoulli prior. This method produces a full posterior probability reflecting uncertainty around whether or not the 'true model' includes a given main effect or interaction effect (Kinney and Dunson, 2007). It can be combined with Bayesian decision rules (Congdon, 2006d, 2009a). A more detailed review of criteria found in the literature is presented point by point below:

- **Goodness-of-fit** The most common model comparison criteria are the coefficient of determination for linear model and the Akaike information criterion (AIC) for generalised linear models, even in the case of mixed models where its use is questionable. With HB designs, the more natural deviance information criterion (DIC) is almost always used. Although an excellent procedure to balance model parsimony and fit, the DIC is not suited to all model structures and can be too lenient with complex models (Spiegelhalter

et al., 2002; Plummer, 2008). Another frequently encountered metric involved in hierarchical model selection is the pseudo-marginal likelihood, the sum of cross predictive ordinates over all areas(Congdon, 2008b).

- **Sensitivity to Bayesian priors** HB studies occasionally examine the sensitivity to the choice of hyperpriors—particularly those placed on the variances of random effects—by plotting outputs from competing models or measuring their correlation (Sun et al., 2000; Pascutto et al., 2000; Ha et al., 2014). Sensitivity to the choice of priors is regarded as a sign of excessively subjective priors, they reflect strong existing knowledge.
- **Percentage of the between-area variation predicted by covariates** It is calculated by reference to a basic random means model without covariates. As pointed out by Heady et al. (2003), small area models of disease may need to be compared with a model with age and sex variables rather than a null model, because prediction is intended to be age-standardised anyway. Brown et al. (2001, p. 1) observed that *'the expected values defined by the model underlying the small area values should be "good". That is, they should explain a significant proportion of the variation in the small area values of interest. Note that for models that include random effects, these do not contribute to the expected value.'* There is no evident threshold to observe in relation to the proportion of between-area variance explained. Some authors have stated or cited an absolute minimum threshold of 40 per cent without further justification (Pickering et al., 2004; Szatkowski et al., 2015; Zhang et al., 2013). In fact, we find that neither synthetic nor EB/HB studies which report on this measure show any evidence of clustering around a particular threshold (Figure 3.3). in addition, the interpretation of this metric is dubious since there is no closed-form function describing its relationship to MSE. While this is undoubtedly a useful measure to check against coefficients of determination quoted in previous research and assess the model performance, it is worth noting that it has not been examined in the statistical literature. Its properties and particularly its stability with sparse datasets remain understudied.
- **Unit-level prediction diagnostics** A small number of studies perform checks usually applied for individual-level prediction, reporting sensitivity, specificity, or area under the curve (AUC) statistics. Although these can provide an idea of model fit, they are of limited value when exploring the reliability of small area predictions across domains.
- **Other methods** Frequentist authors have devised a sampling-based model selection procedure known as the 'fence' method (Jiang et al., 2008, 2009; Jiang

and Nguyen, 2012; Rao, 2011; Pfeffermann, 2013). It is also computationally challenging and is not present in any of the applied literature considered here. Another approach, the Bayes factor, which enjoys renewed popularity in statistical testing, remains little employed in fully Bayesian hierarchical designs. This is partially due to its complexity and sensitivity on priors. A pseudo-Bayes factor approach proposed by Gelfand et al. (1992) is used by Congdon (2009a). This is an underexplored topic, despite arguments that Bayes factors may be the only practical method to assess model uncertainty (Longford, 2012).

3.4.6.2 Prediction validity

Studies included in this review use a combination of comparison with direct estimates (34%, $n = 54$), with an external source of data (16%, $n = 26$), and simulation studies (17%, $n = 27$). While some studies use very reliable external comparators (for instance Jia et al., 2004), other investigations are less conclusive in that they rely on imprecise external survey estimates (Twigg and Moon, 2002), or untested administrative sources such as general practice registers (Public Health England, 2014c) or care activity aggregates without evidence that they are necessarily linearly related to the outcome of interest (Congdon, 2008a; Hudson, 2009; Hudson and Soskolne, 2012). Independently of the type and quality of the comparator, a number of methods are used to validate model-based estimates:

- **MSE/CV** The most common and natural way of measuring error beyond a simple scatterplot is the MSE or its squared root, which is often regarded as equivalent of a standard error. In order to bring it to a standard scale, its square root is divided by the value of the parameter of interest, resulting in the relative root MSE, also referred to as RSE or CV. This measure can present difficulties in terms of the stability and reliability of this ratio since the predictor rather than the true value is used in the denominator. As long as the reliability of MSE linearised approximations, bootstrap estimates or posterior variance is established for all domains and subdomains, traditional official statistics rules on reliability can be applied. Although there is no agreed international standard (Amoako Johnson et al., 2010), thresholds of between 20 per cent (ONS, 2014e) and 30 per cent Klein et al. (2002); US Census Bureau (2013); Wang et al. (2015) are frequently applied.
- **Model-based simulation** With many statistical designs, the lack of guarantees around the accuracy of both the numerator and denominator of the RSE

justifies the use of simulation to provide reassurance. Some studies carry out simulation experiments under an assumed small area working model and are intended to test the influence of a potential distributional violation in the data. In these cases, a particular violation is identified as a risk, and the model is tested on a synthetic dataset so as to determine its robustness in the presence of such a violation. Other model-based simulation studies are conducted using a model fitted on real data. This produces a superpopulation model from which a synthetic study population and survey samples can be simulated. In this case studies usually intend to demonstrate that the estimation strategy is sound as long as the assumed model holds. Many of these simulation studies are meant to verify the statistical reliability of predictions, for instance the coverage property of 95 per cent interval estimates or the sensitivity and specificity of detection of a random effect (Malec et al., 1997; Abellan et al., 2008; Boonstra et al., 2008), the bias of a prediction MSE approximation, the suitability of a prior to reproduce some spatial or temporal processes (Choi et al., 2011; Dwyer-Lindgren et al., 2014a; Lee and Mitchell, 2013), or an approach to account for a given non-ignorable sampling design (Xie, 2004; Raghunathan et al., 2007; Liu et al., 2014; Ha et al., 2014). It is important to understand that they are not suitable to test whether a model is appropriately specified.

- **Design-based simulation** More rarely, studies conduct design-based simulation experiments, when they have access to a large population such as a census (Ugarte et al., 2009) or multiple waves of a large survey (Barker and Thompson, 2013; Krapavickaitė and Rudys, 2015) from which survey sampling can be simulated. Such simulations are extremely powerful to scrutinise the suitability of a model across all areas and subdomains since the simulation is carried out on the actual population with the full complexity of its spatial surface and heteroskedasticity, which cannot be reproduced faithfully under a simple model-based simulation experiment.
- **Average deviation** Deviation metrics averaged across areas are a commonly-used method when area-specific estimates of MSE are not available or reliable. Initially, Gonzalez and Waksberg (1973) showed that the design variance of a synthetic estimate could be estimated provided access to an unbiased estimate

of the truth (maximum likelihood estimator) with known design variance:

$$\mathrm{MSE}(\hat{\bar{y}}_d^{\mathrm{SYN}u}) = \mathrm{E}(\hat{\bar{y}}_d^{\mathrm{SYN}u} - \hat{\bar{y}}_d) + \mathrm{Var}(\hat{\bar{y}}_d) + 2\,\mathrm{Cov}(\hat{\bar{y}}_d^{\mathrm{SYN}u}, \hat{\bar{y}}_d) \tag{3.15}$$

$$\widehat{\mathrm{MSE}}(\hat{\bar{y}}_d^{\mathrm{SYN}u}) = (\hat{\bar{y}}_d^{\mathrm{SYN}u} - \hat{\bar{y}}_d)^2 - \widehat{\mathrm{Var}}(\hat{\bar{y}}_d) \tag{3.16}$$

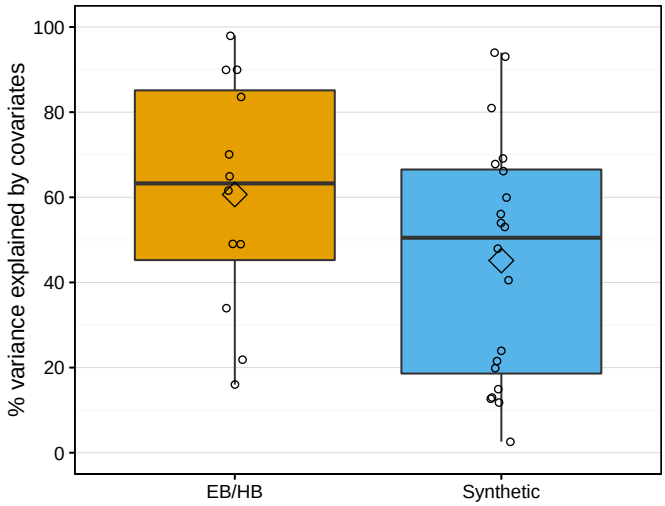
The resulting estimates of area-specific MSEs are unstable and Gonzalez and Waksberg (1973) recommend averaging them across areas to obtain an average estimate. The principle of averaging deviation metrics across areas has remained although some studies have used it for EBs and/or neglecting the second term in (3.16) (Davis et al., 2010; Goodman, 2010; Li et al., 2010; Zhang et al., 2014), where the assumption that the covariance term in (3.15) is negligible is questionable, as EBs with large domain samples will in practice be very close to the direct estimator. As has been noted before in official statistics, this approach is in any case unlikely to satisfy user needs, since 'the official in area i is not interested in an average error over all of the areas. The official likes to assess the error is in his or her domain' (Pfeffermann et al., 2015, p. 459).

Brown et al. (2001) proposed a more rigorous approach by testing the fit of model-based predictions compared to direct estimators using a χ^2 goodness-of-fit test taking into account area-specific variances of the direct estimator. Under central limit assumption, this test ought to detect global bias as well as over- or undershrinkage. While this test is appropriately specified for synthetic estimator, it also fails to take into account the dependence between EB predictors and direct estimators, since the EB predictor is a shrinkage estimator based on the direct estimator.

- **Correlation coefficient/coefficient of determination** Correlation levels are very sensitive to the sampling error of the comparator estimates themselves (see illustration by Zhang et al., 2011, 2013). Although it is possible to use weighted correlation coefficients to adjust any variation in precision across domains (see the concordance correlation coefficient used by Srebotnjak et al., 2010), comparators with large standard errors overall will yield poor correlations even if the model-based estimates are excellent. In those cases, some studies limit correlation analyses to a subset of domains with reliable direct estimates (Cadwell et al., 2010) and/or by aggregating predictions to a higher level at which comparators are more reliable. The limitations of correlation coefficients are yet well identified. Bland and Altman (1986) proposed a graphical method to circumvent them. The Bland-Altman plot of the dif-

ference between model-based and direct estimators to the average of those estimators displays more clearly any potential concentration of bias across the range of estimate values. Szatkowski et al. (2015) apply it and detect higher bias in larger values of smoking rates.

- **Cross-validation** Cross-validation is the natural procedure to evaluate prediction error and in particular MSE. It consists of fitting the model on a proportion of the available data set, predicting the items left aside, and finally comparing predicted and actual items. This can be regarded as a form of design-based simulation using the fitting dataset itself. This procedure is not often used or reported at length in applied citations included in this review (Olives et al., 2013; Twigg et al., 2000). A discussion of cross-validation in epidemiological models can be found in Greenland et al. (2016). In the HB literature, cross-validation is made prohibitive in hierarchical models by the computational burden of Markov chain Monte Carlo. Posterior predictive model checks have been proposed by Marshall and Spiegelhalter (2003) and Stern and Cressie (2000) as an alternative. The posterior predictive *p*-value, which examines the discrepancy between the data and posterior, is also popular (Gelman et al., 1996; You and Rao, 2002; Datta et al., 1999). This approach has already been critically appraised by Gelman et al. (1996); Sinharay and Stern (2003).
- **Calibration** Some studies test predictions for local departures from direct estimates by aggregating them up to higher units, for instance at the level of regions or large units, at which direct estimates are available and reliable (Malec et al., 1999; Twigg et al., 2004; Srebotnjak et al., 2010; Eberth et al., 2013a; Zhang et al., 2013; Eke et al., 2016). This 'calibration diagnostic' measures the 'amount of scaling required to calibrat[e] a set of model-based small area estimates' (Brown et al., 2001). Other studies do not aggregated estimates. They instead conduct comparisons against direct estimate only in large domains where direct estimators are reliable (Olives et al., 2013; Davis et al., 2010; Dwyer-Lindgren et al., 2014b; Srebotnjak et al., 2010). This has been criticised in the peer review for Srebotnjak et al. (2010). It is also apparent that large domains are given a lower amount of shrinkage by design, so that EB by construct closer to the direct estimators. This has been illustrated by Hudson (2009) in a case study.



Note: ⋄ mean; — median; □ quartiles; | 1.5 × inter-quartile range

Figure 3.3 | Box plots of the proportions of between-area variances explained by covariates ($N = 32$ models)

3.5 Discussion

This systematic review of 160 studies shows progress in the adoption of model-based methodology in academic public health and, to a more limited extent, in national statistical offices. Recent years have seen the publication of sophisticated statistical design making very efficient use of multiple sources of data. The structure implied by various models is better understood thanks to efforts to develop more realistic models, yet their implications are not always obvious to all statistics users.

3.5.1 Borrowing strength

SAE borrows strength of prediction from an entire study area to estimate a parameter at a local level. This in part rests on the premise that some health determinants have a uniform effect on health risks across the study area. In particular, age is commonly treated a homogeneous risk factor across space. When standardised health statistics are needed, the overwhelming majority of studies either (a) standardise an area-level health response variable directly or indirectly or (b) model a unit-level response variable as a function of age fixed effects and then compute area predictions by applying the model to a standard population (Olives et al., 2013). This treatment of age is just one illustration of the homogeneity problem; other risk factors such as ethnic group or economic status are frequently modelled similarly. Synthetic estimation borrows strength exclusively from modelling risk factors in a

homogeneous manner and the limitations of this approach lead to well-identified issues in terms of bias, since measuring and accounting for all risk factors through fixed effects is unrealistic.

As a result, a general warning accompanies synthetic and other model-based estimates with regard to the rigidity of their basic assumptions: for instance, a widely-used UK guide to source of public health intelligence reads:

> 'synthetic estimates are based on a model and represent the expected prevalence of lifestyle behaviour for an area, given the demographic and social characteristics of that area. They do not take account of any additional local factors that may have influenced lifestyle in the local population (e.g. local health improvement initiatives). The estimates should therefore not be used to monitor performance or change over time' (Association of Public Health Observatories, 2005, p. 8).

A growing literature has been concerned with producing estimates which are less biased and more sensitive to local departures from global patterns of association between observable determinants and health risks. Since Clayton and Kaldor (1987) presented an early effort to model area-specific relative risks of lip cancer using a shrinkage estimator, the idea of reconciling the (biased) synthetic estimator with the (noisy) maximum likelihood estimator has taken root. From a modelling viewpoint, this reconciliation involves modelling unobserved differences in risk factors across domains or areas as random latent variables, which reproduce residual between-area patterns which are not predicted by fixed effects. This review has identified many approaches to do this, including semi- and nonparametric solutions. But parametric—and in particular normal—random effects remain a dominant tactic since Fay and Herriot (1979). Yet, this review shows that the sole use of an unstructured random effect in EB/HB designs does not automatically address all concerns. We review in turn simple random effects; correlated random effects; mixture structures; and other types of structures in three separate subsections.

3.5.1.1 Simple random effects

The model-based SAE literature overwhelmingly relies on normal assumptions for random effects. It is a very strong assumption; we review what it implies (1) in substantive terms and (2) in statistical terms.

First, an area- or domain-specific normal exchangeable (NEX) random effect does not necessarily relax all the rigidities of the synthetic assumption. Simplifying assumptions are still made which assume homogeneity in relative risks or odds ratios

across areas for some demographic groups. We illustrate this point by returning to the case of age fixed effects in models. Even in case random area effects are included, models tend to ignore variations in age-specific rates of a given health outcome. In the event such variations are very prominent, relying on fixed effects can lead to biased predictions by confounding area relative risks and age by area relative risks: typically, an area where a given age category exhibits very high risks will have an inflated standardised rate. Yet this homogeneity assumption can still be a strong advantage from a practical viewpoint: it can reduces the need for reliable small area MYPEs for survey SAE since the model can just be applied to a standard population and produce small area age-standardised predictions (we note that MYPEs are still required to standardised rates in registers where denominators are unknown). In short, there is always a tension between an assumption and a design constraint set by the data or computational feasibility. The literature contains many examples where such assumptions remain undiscussed, particularly assumed homogeneity of ethnicity risk ratios across space (Pfeffermann and Sverchkov, 2007; Knutson et al., 2008; Dwyer-Lindgren et al., 2013), despite clear evidence that most heterogeneity across areas is taking place in these groups (Malec et al., 1999). Uncertainty around the suitability of these modelling decisions is a key reason why predictive checks remains such an important concern in SAE.

Second, the default NEX specification assumed by all basic parametric model is not an automatic answer to all questions of heterogeneity: it assumes that random effects (a) follow a normal density, (b) are identically distributed (b) are independent and exchangeable, in other terms, unstructured. Research has sought to better understand limitations associated with these.

(a) Fabrizi and Trivisano (2010) demonstrated that predictive posteriors are particularly poor in case of deviations from normality in the distribution of random effects. Outlier-robust models based on Student's t-distribution Xie et al. (2007) and the exponential power distribution (Fabrizi and Trivisano, 2010; Fabrizi et al., 2016) have been used for area effects, while non-parametric alternatives based on p-splines to estimate area effects are discussed by (Salvati et al., 2010). A semi-parametric approach modelling non-normal sampling error based on their empirical distribution is used by Porter et al. (2015b) for unemployment rates with spatial dependency and by (Porter et al., 2015a) with multivariate dependency. These however remain little used outside of statistical journals. We suspect this is in part due to a lack of power in tests detecting violations of normality. It is in practice very difficult to detect these in sparse data, particularly when it affects random effects.

In addition, a lack of software implementations until recently may explain why robust methods have so far not been implemented a lot in the applied literature.

(b) In heteroskedastic populations, the assumption of identically distributed effects can cause issues in failing to recognise that the between-area variance is different across time periods or social groups such as age or ethnic groups. A direct consequence of failure to recognise unequal variance can be suboptimal shrinkage of the various output cross-classifications, which is likely to cause bias. In this case specifying several NEX effects with unequal variances is often necessary. It nevertheless requires very strong prior knowledge. Models have been designed to detect instances of unequal variance across space or time (see section 3.5.1.3). Since they demand a large amount of data, they have so far only been applied to registers. A large amount of studies on unemployment rates and time series survey-based SAE, while not exploring unequal between-area variances, have been concerned with unequal sampling variances (Pfeffermann, 2002).

(c) The assumption of exchangeability (independently distributed effects) reveals to be incorrect in most situations in public health modelling due to the presence of spatial autocorrelation in most health outcomes (Lorant et al., 2001), a sign that risk factors tend to be similar in closely situated areas (). Efforts have been made to adapt covariance structures to reflect new understanding of dependence in health risks across areas, time periods, and social groups.

3.5.1.2 Correlated random effects

The convolution prior (CAR) remains the most common model structure to account for spatial dependence in public health. No studies use ICAR area effects alone as the resulting predictions would be too spatially smooth. The mathematical specification of CAR differs from the SAR covariance structure described by Whittle (1954) and traditionally suited for linear mixed models, although Bhati (2008) developed an implementation for Poisson models. The substantive difference between those two structures remains little understood. In a case study on US state average college entrance exam scores, Wall (2004) found that there were no material differences in outputs obtained with SAR and CAR priors.

We also note a persistent research gap in the definition of spatial weights making up CAR and ICAR structure matrices. These are usually defined as the default binary indicator coding for contiguity with spatial neighbours, or for pres-

ence within a 30 km radius of an area's centroid in the case of Ghosh et al. (1999). Only Miaou et al. (2003) used the inverse distance separating area centroids as spatial weights with a CAR structure. The impact of spatial weight definitions is rarely explored in disease mapping. To our knowledge, few studies investigated this: Best et al. (1999) found a better fit on the Scottish lip cancer dataset with contiguity compared to distance-based. In contrast, distance-based weight provided a better fit in an investigation on risks of birth defects in Australia by Earnest et al. (2007). It is also worth noting that reparameterised variations of the CAR structure have been designed by LLB (Leroux et al., 2000) and MacNab (2003) and can compare favourably as demonstrated in a series of simulation and case studies Lee (2011); Baptista et al. (2015); Rampaso et al. (2016), although this finding is challenged under some indicators of performance. A number of SAE studies included in this review used this specification. Autoregressive models are a rapidly developing field seeing a constant addition of new models (see for instance Lawson et al., 2015).

Comparatively less attention has been given to structures of dependence across groups. Since Malec et al. (1999) first proposed correlated effects in SAE in a study of obesity across sex and ethnic groups, designs based on multivariate normal and multivariate CAR priors have become common in the applied literature to exploit the dependence across sex and ethnic groups mostly (Cadwell et al., 2010; Congdon, 2009a; Congdon and Lloyd, 2010; Richardson et al., 2006). Dependence across age groups is also to be expected. Because ageing is a form of biological and social time, it is intuitive that close age groups have more similar risks than groups that are further apart. This basic concept has been established in dynamic survival analysis studies and age-period-cohorts by placing on age and cohorts the same autoregressive filters previously used for time. This property thus is heavily used in actuarial science to smooth life tables and mortality rates, where the benefits of 'age graduation' are recognised. In disease mapping, we find a small number of implementations in Congdon (2006a); Congdon and Lloyd (2010); Congdon (2009a,b); Davis et al. (2010); Fahrmeir and Lang (2001); Jonker et al. (2013a), using the first- and second-order random walks priors.

3.5.1.3 Mixture structures

While fully structured models can achieve a more efficient smoothing of noisy spatial and demographic variations in disease risk, they risk smoothing out genuine local departures from the global health risk structure. Over the last decade, great progress has been made with spatially adaptive (nonstationary) structures and

models tuned around detecting unusual areas. They make it possible to detect and reproduce discontinuities in complex health risk surfaces (sometimes referred to as step changes or boundaries, see Lee and Mitchell, 2013), for instance urban obstacles (Congdon, 2008a), unusual trends (Li et al., 2012) or mild irregularities compared to the traditional ICAR structure (Green and Richardson, 2002). We distinguish two main categories of statistical designs:

- **Discrete mixtures** Model can be designed which alternate between two or more structures, effectively differencing exceptional areas or outliers from a general pattern. Li et al. (2012) use a space-time interaction design mixing (a) a basic structured CAR area structure and RW1 time structure and (b) an RW1 time structure with an area-specific variance. The mixture parameter is given a very informative Bernoulli prior with a small (5 to 15 per cent) selection rate ('sensitivity') for the area-specific structure. While this prior is informative, Li et al. (2012); Boulieri et al. (2016a) find only limited sensitivity to it in their results. This model has proved to be well specified to detecting unusual trends, while the continuous mixture of Abellan et al. (2008) is more likely to detect areas which rates are unstable across time as observed by Duncan et al. (2016). Lawson et al. (2010) proposed a mixture model allowing for more than two time trends component, using a discrete mixture of multivariate latent first-order autoregressive (AR1) time trends components. The number of components is set by the user and the discrete mixture parameter is specified under a multinomial prior, with the underlying probabilities determined by a multivariate CAR hyperprior. This model in effect classifies areas into a user-defined number of time patterns, although model selection can be used to balance the number of time patterns with model parsimony. These time pattern 'clusters' are by design likely to be contiguous due to the CAR hyperprior. This work was extended in Choi et al. (2011) with spatially correlated mixture weights used as mixture parameters. Their model reflects the fact that clusters of temporal patterns should be similar or contiguous in space. In a case study, they demonstrate that this approach can improve the cluster detection capabilities of the model.
- **Continuous mixtures** In addition to the already discuss convolution mixture (CAR) and its variations (LLB), Lawson and Clark (2002) specified a model with a NEX area effect and a mixture between an ICAR and a 'discrete jump' effect, mixed by a parameter defined on $[0, 1]$ with a Beta prior. The discrete jump effect consists in a normal prior for the area effect in which the mean is a function of the heterogeneity within the clique of neighbours.

Congdon (2007) proposed a 'spatially adaptive convolution' structure based on a continuous mixture between ICAR area effects and unstructured NEX area effects, placing a Beta density on the mixture parameter. In this case, the balance between the structured and the unstructured spatial variance varies across areas. With regard to space-time designs, mixture models also bee proposed. In the work cited above, Lawson et al. (2010) also proposed a model with multiple time trend components, this time with a continuous mixture parameter under a log-normal prior, which mean is spatially correlated with a multivariate CAR hyperprior. Instead of classifying areas into discrete clusters, this model instead mixes different time trends in a spatially smooth way. Choi et al. (2011) also studied the case of spatially correlated mixture weights for these continuous mixtures, using a multivariate CAR prior set on the mean of the mixing weights.

The space-time detection model by Abellan et al. (2008) is specified with specify a mixture of unstructured area effects combining two NEX effects with unequal variance, using a mixture parameter uniformly distributed over $[0, 1]$. We also note that continuous mixtures have been applied on fixed effect spatial structures. With a more traditional mixed model, Congdon (2000) proposed a mixture between fixed effects of three types of spatial interaction: distance, driving time, ambulance travel time, with a Dirichlet prior on the mixture parameter. More recently,

3.5.1.4 Other structures

The other area of future research is the use of more radical approaches to relax assumptions of global smoothness exist. To date, these primarily include nonparametric SAE relying on penalised splines (Opsomer et al., 2008), nonparametric geostatistical methods such as geosplines, and principal component neighbour matrices (Griffith, 2000; Thayn and Simanis, 2013). The semi-parametric empirical likelihood approach has also been adapted to spatial models, introducing spatial covariates under the assumption of orthogonality with other predictors, that is assuming independence between the spatial covariate and the other predictors (Hanks et al., 2015; Chaudhuri and Ghosh, 2011; Porter et al., 2015a). Such a method has been applied to SAE by (Porter et al., 2015c,b). Other approaches known as 'wombling' are based on CAR equivalents where spatial weights are not set to be fixed, but instead estimated with a log-normal prior (Lawson et al., 2010), or a Bernoulli logistic prior dependent on covariates, which can be inverse distances between area centroids (Lu

et al., 2007). Rodrigues and Assunção (2012) propose to use a mixture of neighbourhood matrices. This is a very active research area with as of yet limited applications to SAE. Some implications remain relatively unexplored, particularly with regard to the design requirements of such prior structures in terms of data and spatial patterns.

A final, arguably more radical development around adaptive spatial structures has been brought by the Bayesian wombling literature, in which spatial weights are treated as random rather than fixed. Evident issues of overparamerisation have led to elegant applied developments, some of which have been implemented in the public health literature. Lu et al. (2007) use a CAR prior with the usual contiguity spatial weights matrix, but use a logistic prior on non-zero spatial weights to detect *boundaries*, namely borders where two or more contiguous areas exhibit such contrasted outcomes that CAR does not hold unless the spatial weight is set to zero. This approach reduces the set of random effects to just non-zero weights (contiguous areas).

In light of computational difficulties, solutions have been tested which address the issue of overparameterisation. These approaches strip models from any prior structure on weights and the associated additional hierarchical level. Lee and Mitchell (2013) use a spatial weight matrix selection algorithm which selects a best fitting or most likely set of weights. In contrast, Lee and Mitchell (2012) predict spatial weights of contiguous areas with a fixed effect regression, using the discrepancy in area covariates in two adjacent areas as a predictor of a boundary. The resulting spatial weights matrix is used as fixed in a CAR prior.

3.5.2 Model complexity and realism

The examination of studies published to date shows that advanced model-based designs have made it possible to optimise health risk adjustment and smoothing with the assistance of a wide variety of information, whether from surveys, registers, GIS, population estimates, etc. There are, however, limits to refining model specifications. A stringent challenge in SAE lies in the balance between complexity and efficiency, this balance being largely determined by the amount of information available from the data or, more occasionally, from prior knowledge. In short, simple models may oversmooth variations in risk across domains, while heavily structured models are often difficult to estimate and test and, ultimately, risk overfitting. This problem becomes evident with sparse datasets, complex health risk surfaces with discontinuities and multivariate dependence structures. In practice, small datasets inherently yield simpler models even in situations where more complexity would be required. Furthermore, a lack of good predictors results in loading a large burden of

prediction onto stochastic processes with highly parameterised multivariate random covariance structures. This creates challenges with regard to designing SAE research, selecting models, and validating them.

3.5.2.1 Synthetic assumptions in model structures

Despite the development of very flexible structures, limitations posed by data—sometimes even, as Gołata (2014) note, by statistical training—result in a dominance of the simplest model structures. These default structures can convey an illusion of generality, but simulation work has revealed that they can also prove rigid and equally deserving of cautious examination. The vast majority of studies reviewed here use either synthetic estimation, or rudimentary structures to account for differences between areas and groups. While synthetic estimation should be of no particular concern as long as very good predictors are available, this is rarely the case outside of situations where the predictor is already so informative in its own right that SAE becomes unnecessary. For all other situations, basic modelling designs often lead to assuming that, for instance, groups are identical across areas, or that areas situated close to one another are not more likely to have similar outcomes compared to areas separated by a long distance. This can inflate predictions' MSE or worse, their bias, particularly with sparse data sets.

Despite this, implications of modelling designs reviewed in this paper are not systematically investigated with a simulation study or equivalent powerful method. Above all, they are rarely made evident to end users. First, we note many descriptions of synthetic estimation fail to test whether models really do predict most of the variance in health risks between area, and to take measures to reflect the bias occurring when substantial unaccounted differences remain. The use of fixed covariates and of specific spatial structures is also not systematically justified especially in recent papers. By comparison, the earliest studies contain very thorough explanations, such as the one provided by **?**, p. 6: 'The underlying rationale for synthetic estimation is that the distribution of a health characteristic does not vary among populations of states except to the extent that states vary in demographic composition. In other words, the method assumes that the incidence or prevalence of a health characteristic would be the same for two states if their composition were the same with respect to such demographic variables as age, sex, race, family income, family size, place of residence, and industry of the head of the family.'

Synthetic estimation remains a valid approach in some instances. First, many datasets are still not accessible with geographically identifiers, out of concerns for

privacy: at least 27 studies we reviewed faced this constraint. This makes it difficult to attach area-level predictors or model between-area heterogeneity appropriately. Second, the synthetic assumption is sometimes necessary, either because data are too sparse to allow more advanced EB/HB designs (Congdon, 2008a), or because it is needed so as to operationalise a specific research goal. This is the case for instance with investigations on health inequalities, with some studies using a procedure described as 'recycled prediction' (Paez et al., 2010). This is also the case with studies of unwarranted variations in health care utilisation Benzeval and Judge (1994). Congdon (2006b) explain that 'translation of national prevalence rates for such [psychiatric] disorders into estimates for health areas should at a minimum allow for variations in risk by socioeconomic and ethnic group.' This is thus a pragmatic approach based on a subset of factors known to drive variation in healthcare, which are particularly attached to social equality imperatives set out in the UK health service. The limit with this approach is that it still assumes that ethnic communities are homogeneous across parts of the country, which is difficult to establish. Here the authors make decisive assumption around what is a factor of 'warranted' variation and assume that residual variation is 'unwarranted'.

Concerns exist with more advanced models as well. Gelman and Price (1999) warned against the risk of oversmoothing, suggesting that some model-based estimates can sometimes be as misleading as crude estimates. Both can be unreliable due to unequal sample sizes across areas, but for different reasons. In areas with small sample, crude estimates are unstable, while model-based are likely to be too smooth despite the lack of information. The apparent appeal of parametric CAR structures should not occult the risk of excessive smoothness, particularly when complex multivariate structures are defined—a prime example being Ryan's (2009) discussion of Tassone et al. (2009). It could be argued that smoother estimates may not be worse than highly noisy raw ones. Yet it remains difficult to evaluate their usefulness, or decide whether decidedly excessive smoothness is to put down to data sparsity (i.e. likelihood structure/Bayesian priors taking precedents) or to overstructured models. This situation will probably require a combination of statistical developments in model criticism techniques and consensus building in the research field. Parallel to this, SAE appraisal faces similar issues as design experiment when it comes to the increasingly model-driven inference, and experimental design methodology seems to increasingly take a design and simulation approach to define analysis plans (Adrion and Mansmann, 2012). Addressing concerns against the model-based approach implies thinking of SAE less as post hoc method and more as a research approach outlining a range of inferential aims and proceeding with designing a statistical plan

implementing these aims under cost constraints.

We recommend that future studies observe a simple reporting standard which is to avoid referring to synthetic estimation, regression estimators or hierarchical prediction without a closely associated, explicit and non-technical description of assumptions regarding heterogeneity in health risks across the population. This description should achieve clarity when it comes to substantive interpretation when comparing results over time, place, or demographic subgroups. For instance, it should be apparent that health outcomes are assumed to be identical for all individuals belonging to a given age-ethnicity cross-classification across small areas within a state or region, in the absence of better knowledge suggesting this is not the case. It should be clear in this instance whether this decision is supported by external evidence, or by the lack of residual heterogeneity across areas once all variables are entered into the model; typically, a residual unstructured random effect for the target domain should have very small and negligible variance. A variety of tests have become available to this end (Datta et al., 2011; Gelfand, 1995; Gelman et al., 1996; Albert, 1999; Stern and Cressie, 2000; Vaida and Blanchard, 2005). Given the difficulty to detect model misspecification with sparse datasets or non-flagrant departures (Sinharay and Stern, 2003), further methodological developments are needed to determine the level of confidence that can be placed in a given set of model-based estimates, be it through decision rules or survey and modelling design rules insuring reliable inferences.

3.5.2.2 Acceptability in social and official statistics

Although model selection is an important part of the development of any model-based design, and that the DIC has become an established and accepted model appraisal criterion, prediction validity remains the primary concern in social statistics due to the risks of model misspecification. Compared to academic public health and public health practice, official statistics is bound by stronger impartiality and objectivity design requirements of international codes of conducts (UK Statistics Authority, 2009). Transparency and robustness requirements justify validation procedures reviewed in section 3.4.6. Rigorous model validation does not yet suffice in official statistics with survey samplers challenging the very foundations of model-based inference. There is a lack of consensus over prominent features of model-based inference, namely 'a thorough interplay between a theoretical content and empirical content' as well as assumptions which are 'unverifiable but not entirely out of place, to save survey resources or to bypass other practical difficulties' (Särndal, 2010b,

p. 113).

It is a particular concern that clients 'will always require more than is specified at the design stage' (Fuller, 1999, p. 344) and that SAE as it currently stands in public health sometimes means that models are expected to predict quantities that are either unobserved or poorly observed. Instead, taking the standards expected in official statistics and good clinical research practice, it is not unreasonable to set requirements that the data collection is sufficient to contradict key model assumptions if they are in fact incorrect. As a commonly encountered decision making and hypothesis testing problem, this approach can be developed provided that consensus develops around quality thresholds. In official statistics, such thresholds are usually set with respect to RSE targets which rest on unbiased design-based estimates of variance. In a model-based framework, it is necessary to also set a level of acceptable uncertainty around RSE and detection rate estimates.

We can observe from the published literature that inferences are drawn by assuming that the range of specifications (or Bayesian priors) considered are reasonable, and that the available data is adequate to fit and select models. In particular, the validity of standard errors of predictions depends on this validity, and it is very likely that any deviation from the model assumptions cannot be detected solely from the data unless remarkable and spread across many of the targeted domains of estimation. Ideally, statistical designs should be informed as much as possible by prior knowledge on relationships and patterns in data and an ability to determine what data is required to estimate a given model and carry out prediction in good conditions. The challenge raised by model-based estimation, beyond conveying the estimais to translate to end users how rudimentary the model is thought to be, which should not solely be left to publishing RSEs or confidence intervals. When it comes to the publication of raw statistics, metadata should include a substantial description of modelling assumptions and limitations. Validation studies are an alternative way of achieving this (see *infra*).

3.6 Conclusion

This methodological review has highlighted current developments in model-based population health estimates. It highlights the rise of methodological advancements, mainly random walk priors, spatial modelling, shared component modelling, latent variables models and appropriate treatment of measurement error in covariates. It does also highlight some limitations. A key challenge of these methods is to provide confidence around the specifications of the model. We found a limited

number of validation studies, but even those are carried out along standards that can largely be improved. In particular, we recommend that future studies include further cross-validation, comparison with historical data when available, and above all, well-designed simulation studies. This presupposes to improve the dissemination and training on spatial and mixture modelling developments as well as simulation-based validation designs.

This review concludes that progress remains to be made specifically in the area of public health intelligence, as noted by others (Webber et al., 2014; Briggs et al., 2016). We argue that a design approach is needed to address the key challenges of model-based population health statistics production. The theory of model-based estimation has been developed with permanent warnings against inferential risks; Little (2004, p. 551) noted that 'all models are simplifications and hence subject to some degree of misspecification', while Longford (2015, p. 66) squarely observed that 'a lot of the theory is concerned with deriving estimators that are efficient, or nearly so, sometimes in uncongenial circumstances, using models known not to be valid.' Since model assumptions are difficult to convey to end users, Longford argues that estimating point estimates and standard errors is not enough, and that a design approach is needed which integrates the client's information needs and effectively prevents incorrect post hoc analyses. Historically, disease mapping focused on modelling existing datasets and the question of designing data collection systems was not considered. With the spread of model-based estimation to survey SAE, the need for appropriate sampling design methodologies is becoming clear, especially in comparison of the dramatic developments witness in clinical trial designs in recent decades. This would also benefit disease mapping more widely, as *ad hoc* data collection such as national hospital audits could be designed by making the most of nationwide clinical and health care data infrastructures. There is potential to (a) increase the precision of intelligence from such audits and (b) use audit information to correct for measurement error of clinical datasets, such as diagnostics coding errors, missing data and undercoverage.

If a such a design approach can be developed, SAE offers potential to produce more statistics, with key advantages being (1) safe outputs due to the lower disclosure risk obtained through smoothing, (2) statistical efficiency, (3) adjustment for nonresponse, measurement error in self-reported information (such as BMI), (4) correction for undercoverage and underreporting, for instance by coupling information from databases and surveys on crime, traffic accidents, or disease registers. The entry cost into these levels is not negligible, and requirements in terms of validation and transparency to end users remain a challenge, but efforts to increase familiarity

with model-based methods through reviews such as the present one can help make these methods more accessible. Furthermore, data controllers should be aware of the value of making geographically identified information available safely to researchers requesting it for the purpose of SAE, as opportunities for robust model selection and validation are very limited when respondents' geographical location is not available.

replaced[id=PD]Finally, it is worth emphasising that model-based estimationThis is very encouraging for low and medium income countries, where statistical infrastructure may be insufficient to support public health strategies at the local level. A small number of studies in African and South Asian sites confirms practical feasibility and uses in those countries thanks to the existence of decennial censuses and DHSs. More work could be done to further spread the use of SAE. This presupposes continuous capacity building efforts internationally, as highlighted by Gołata (2014).

Chapter 4

Aims and Methodology

4.1 Introduction

Previous chapters have examined current directions of research in public health applications of model-based SAE and emphasise some of the gaps and obstacles currently faced in widening the use of such techniques to produce morbidity indicators. Methodological developments are needed around the validation and design of SAE in this area, and this thesis brings applied contributions, with the review of the decennial UK census as a starting point. These contributions come in the form of three academic papers included in chapters 5, 6 and 7. The present chapter provides context on some of the research design decisions leading to these contributions. Section 4.2 states the overarching research aim, breaks it down into a series of objectives, and introduces the papers individually with regard to how they operationalise the thesis' range of objectives. Although it does not replace methods sections included in each of the empirical papers, section 4.2 provides additional background on the constraints encountered during this research, and clarifies the research contributions arising from the combination of these papers. Finally, section 4.3 presents some baseline empirical review work that was conducted in preparation of the overall thesis design. Topics of particular importance for chapters 2 and 3 are examined, particularly the nature of the health indicators studied in this thesis as well as survey questionnaire and mode effects.

4.2 Aim and strategy

4.2.1 Research problem

The English and Welsh health data infrastructure is particularly developed, with a range of household surveys, health examination surveys, administrative registers and care records return systems characterised by a high coverage thanks to the organisation of healthcare in those countries. Yet, there is wide recognition that this infrastructure has not been exploited to its full potential (ONS, 2003; UK Administrative Data Taskforce, 2012; UK Statistics Authority, 2016b). Amongst commonly invoked reasons are restrictions on data access, ill-adapted legislation, skill shortage in analytical staff, and gaps in methodology. Such gaps have mostly to do with statistical developments needed in data linkage, anonymisation as well as register- and model-based inference.

This background is nevertheless evolving rapidly: the future of population health statistics for small areas is increasingly seen as dependent on the availability of reliable model-based methodologies. There is a strong interest in establishing these and demonstrating that health and lifestyle atlases, amongst other general population health outputs, can be protected and even enhanced by statistical modelling in the future. The systematic review in chapter 3 provides an up-to-date appraisal of remaining challenges, namely the use of spatial modelling, the planning of survey data collection in accordance with pre-defined SAE goals, and the validation of working models.

Simultaneously, we note the particular goals set by the ONS' Census Transformation Programme to demonstrate the feasibility of using models to estimate population attributes reliably at the level of LADs, MSOAs, and possibly beyond. With this view, the present thesis undertakes applied development work based on current census morbidity indicators, namely LLTI and SRH, with more in-depth work on the former. The importance of these indicators is well established internationally, being at the root of summary population health indicators produced to monitor health inequalities. They are also part of the harmonised European core health indicators and remain very present in empirical research, particularly health economics and social epidemiology. Finally, fieldwork testing and research on LLTI (see sections 4.3.1–4.3.4) shows that it is predicted to a large degree by the presence at the individual level of functional limitations and doctor diagnoses. Therefore, modelling challenges resolved in this research are likely to be relevant to future work on other morbidity and prevalence indicators. This constitutes a further motivation to

dedicating more in-depth attention to LLTI.

4.2.2 Aim and research questions

The overarching aim of this thesis is to further research into applying model-based estimation to public health morbidity indicators, taking into account both

(a) advancements made possible by research on spatial smoothing techniques and
(b) constraints of the UK National Statistics infrastructure, particularly around the Census Transformation Programme.

This research was designed in such a way as to provide an accessible assessment of how much information can really be obtained from small area health predictive models. In spite of numerous published simulation study results regarding the performance of predictors from the EB and HB families, knowledge on the bias and precision of various modes of inference does not easily translate into typical measures of quality end users often rely on. Unlike with traditional design-based methods, no closed-form formula can easily answer this question. It results that trust in any given modelling design is difficult to build and requires comprehensive testing. Other well-known causes of unreliability for small area designs include auxiliary data measurement error, departures from assumptions of normality and spatial smoothness, non-ignorable informative survey design, or household nonresponse. They all contribute to the overall challenge of predictive modelling, which is that any model, as complex as it may be, is a simplified representation of reality, in other words, the wider set of statistical processes at work which determine the spatial distribution of health statuses.

This thesis follows an applied approach, focusing on specific health statuses. It reviews ways of modelling their spatial distribution and assesses their effectiveness in three studies, according to the technical specifications already set by the ONS (2014b): providing health indicators with a fixed precision target for populations as small as 3 per cent of the total headcount of each MSOA in England and Wales. This involves a quadruple challenge:

- specifying simple, yet reliable models
- restricting individual-level auxiliary data to sources which do not depend on decennial enumeration (administrative registers)
- validating these models
- designing a data collection strategy suitable to efficiently estimate these models.

As exposed in chapter 3, recent developments in disease mapping borrowed

in SAE allow one to consider also using more advanced shrinkage estimation techniques to accommodate for this challenge. Given the high level of similarity in health characteristics at small geographical scales, the same spatial smoothing approaches as used in lattice-based disease mapping could be used. In addition, non-spatial smoothing, which borrows strength across closely-related demographic subgroups (for instance across age groups, known to actuaries as graduation) has been envisaged before and should be examined in this context. The difference lies in the fact that in SAE, smoothing is applied to variability arising from sampling error rather than from the rarity of incidence of events such as cancers or deaths.

Although the assumption that epidemiological data fits an underlying smooth mathematical function can be regarded as a strong one, it has survived the test of decades of medical and actuarial research. In addition, statistical modelling rests on established foundations of probabilities and statistics and is a natural framework to reflect uncertainty in parameters one seeks to estimate. However, these assumptions being reasonable at the first glance does not imply reduced need for model validation. Their examination is therefore a major preoccupation across this research.

To summarise, the present thesis examines the following research questions;

1. What existing model-based methodologies can be used to produce estimates of the prevalence of chronic morbidity across LADs and MSOAs in England and Wales?
2. Does the spatial distribution of chronic conditions differ across age and ethnic groups?
3. What contextual factors and auxiliary covariates can be used to strengthen predictions?
4. Can strength be borrowed across space and demographic subgroups and does this bring about material efficiency savings?
5. What precision can be achieved using existing household surveys?
6. How can an efficient data collection strategy be designed to reach a required precision target?
7. How large a sample is needed to produce good predictions according to quality standards expressed by the ONS' Census Transformation Programme?

After considering the context exposed in chapters 2 and 3, we set eight research objectives for this research, as described in Table 4.1 below:

Table 4.1 | Overview of research objectives and corresponding academic papers

Number	Objective	Chapter number
O1	To describe the structure of chronic illness in England and Wales in 2011 with regard to demographic risk factors (age, sex, ethnicity) and other ecological confounders correlating with levels of morbidity, such as mortality, deprivation and health care utilisation.	6
O2	To develop and select models predicting the percentage of usual residents reporting (a) poor/very poor SRH and (b) severe activity limitation in their day-to-day activity for every (a) LAD and (b) MSOA in England and Wales in 2011.	5, 6
O3	To produce point and interval predictions of these target parameters, and compare them with 2011 census estimates in terms of value correlation, rank correlation, absolute and relative mean deviations.	5
O4	To assess the relative efficiency of synthetic estimation, EPP, HB and spatial HB prediction.	5, 7
O5	To design a statistical approach to determining sample size requirements of model-based SAE.	7
O6	To apply this approach to plan a data collection strategy for 2011 as a demonstrating project for 2021 and beyond.	7
O7	To examine the feasibility of joint modelling of demographic subgroups and assess any potential efficiency gains.	7

4.2.3 Strategy and contents of academic papers

This thesis follows an applied approach, focusing on specific self-reported health statuses from the census. This choice of outcomes is justified by the availability of recent and very disaggregated census data on health topics, and the opportunity to access similar indicators in geocoded survey microdata through the Secure Data Service. The comparability is partially ensured thanks to the ONS' cross-survey questionnaire harmonisation programme, as exposed in more detail in section 4.3.4. Census data does not only provide reliable estimates for the validation work; also provide a very detailed population from which simulations studies can readily be designed. It seems difficult to envisage completing such validation work without resorting to population data sets of comparable dimensions. Furthermore, the census does not raise the kind of access and quality challenges that researchers still face with administrative data sources to this day.

This applied work reviews ways of modelling the spatial distribution of these health statuses and assesses the effectiveness of model-based SAE in three studies: an observational study, a comparative study and a simulation study. This comes from an apparent need for translational research examining the applicability of some models and techniques developed in the literature to statistical production.

4.2.3.1 First academic paper

The first academic paper (chapter 5), entitled 'Estimating Small Area Health Status Indicators' is a direct implementation of objectives O2–O3, examining what it currently feasible with existing data and basic model-based designs. This analysis was designed as a real-world attempt to produce model-based small area estimates by (a) making use of the facilitated access to geographically identified survey microdata under a special licence and (b) using the opportunity to compare these estimates against the recently published 2011 census estimates at LAD and MSOA level. A systematic review of household surveys conducted prior to 2011 compared candidates to use in this small area study (see sections 4.3.3 and 4.3.4). This concluded in favour of using the English Housing Survey (EHS) for this first study. This choice was motivated by the survey's efficient sampling design, its large sample size as well as analyses confirming its consistency with census estimates of SRH and LLTI at the national level. In addition, a large proportion of its responses were obtained from a proxy interview with the head of household. This is believed to have a potentially similar interview mode effect to the mode effect of the census, which consisted in single self-administered paper questionnaires sent to every household in 2011. Access to the

EHS datasets with household geographical identifiers was granted by the UK Data Service under the ONS' approved researcher scheme. Microdata were made available on a remote computer session through the Secure Data Service. In order to take into account constraints imposed by the government's plan to abandon traditional full enumeration in censuses after 2021, the set of covariates was voluntarily restricted to those which could realistically be expected to remain available after 2021. This *de facto* excluded detailed socioeconomic population characteristics on employment, social class and material deprivation. Preliminary analysis on census LLTI data had identified a strong covariate in local rates of emergency hospitalisations. These are used to compensate for the lack of individual-level socioeconomic information. The analysis compares synthetic, naive EB and EB prediction as per objectives O3–O4, but not HB prediction. Hierarchical and particularly spatial hierarchical modelling were not an option as they require specialist software which could not be installed on the Secure Access virtual environment. The manuscript of this academic paper is envisaged for submission to *Statistics in Transition* or *Population Health Metrics*.

4.2.3.2 Second academic paper

The second academic paper (chapter 6) is entitled 'The Spatial Structure of Chronic Morbidity' and originates from the need to conduct initial research on the structure of chronic morbidity (objective O1). This required a large source of data to conduct spatial analyses that are traditionally demanding in statistical power. The 2011 census table DC3201 publishing headcounts of LLTI by five ethnic groups by four age groups was selected to that end. This paper investigates the structure in the between-area variance in LLTI prevalence for a range of age, sex and ethnic subgroups, as existing analysis had until then focused on the analysis of spatial autocorrelation of aggregate prevalence. This study produces information which confirms the very high levels of variance in LLTI prevalence across LADs and even more so across MSOAs. It highlights demographic groups where this dispersion is particularly pronounced, paying specific attention to ethnic groups in perspective of objective O7. Global autocorrelation descriptive analysis provides reassurance around the potential to employ methods originally developed in disease mapping to smooth rates of rare events (O4). Based on this, further analysis is required to refine the specification of small area models. First, it is necessary to confirm that this level of autocorrelation subsists when covariates are used in a regression model to predict the prevalence of LLTI. Second, if parametric spatial modelling is to be used, spatial interaction matrices will be needed, but there is no prior knowledge

of an optimal approach to deriving their underlying spatial weights. This paper responds to this gap by deriving three types of spatial weights based on the Queen contiguity method, the *k*-nearest neighbours method, and a novel method based on origin-destination subnational migration data rather than spatial boundary data. This analysis is conducted on all age, sex and ethnic groups separately to detect potential differences in best-fitting spatial weights across groups. Finally, this paper takes into account the unexpected prominence of LLTI prevalence differences across ethnic groups and reports more advanced analyses of differences in their respective spatial structures. The global spatial structure analysis is complemented with local autocorrelation analysis uncovering spatial clusters in prevalence for some minority ethnic groups. Model residual analysis also provides an indication of the level of covariance in the between-area variance in LLTI across ethnic groups, which is important to assess the potential to borrow strength not only *across space* in a given ethnic group, but also *across ethnic groups* in a given area. Results from all analyses strongly influenced subsequent design and focus of this thesis. This academic paper was published in August 2016 in *International Journal of Health Geographics*.

4.2.3.3 Third academic paper

The third academic paper (chapter 7) is entitled 'Planning Hierarchical Bayes Small Area Estimates'. Its initial purpose was to evaluate efficiency gains from a model taking into account both the spatial and the demographic structures of dependence in odds of morbidity (O4, O7). This was envisaged to involve simulating survey sample realisations from 2011 census tables, and applying more sophisticated models than developed in the first two academic papers. These more advanced models would incorporate structures of dependence by borrowing strength across (a) neighbouring areas, (b) age groups and (c) ethnic groups. This initial plan would, in addition, have provided a basic assessment of the kind of sampling fractions that could be required in order to produce small area estimates meeting the quality standards targeted by the ONS in their census review (O6). All analyses were designed to be applied at the MSOA level, which made the computational aspect of this work particularly intensive given the dimensions of census local characteristics tables for these geographical areas.

Preliminary work showed that the range of of computationally feasible models was particularly narrow. The range of realistic sampling fractions (below 5 per cent) investigated produced fitting datasets that were generally too sparse to accommodate complex models. Models predicting prevalence for 18 age by sex or 20 ethnic by sex

groups, in particular, proved too complex. We also found that, at such a fine level, violations of distributional assumptions became a concern: small cell counts present in the census for small cross-classifications of area, age and ethnic groups meant that the UK population was itself an excessively small realisation of the assumed super-population model. This manifested as ill-behaved area-level odds of poor health for ethnic minority groups. Preliminary investigation also revealed the relative paucity of prior knowledge available to incorporate in design priors. This was to the detriment of effective model selection.

This forced to reconsider the design of this paper on the basis of simpler models, with fewer output cross-classifications. We reduced the geographical span of the study to three regions in Northern England and focused on predicting 6 age-specific prevalence proportionsper small area. This allowed us to focus on more theoretical work on experimental and survey design conducted in parallel. The revised paper is organised from a design angle rather than the more applied testing and design-based simulation work initially envisaged. The paper proposes an sample size determination (SSD) approach iterating through model-based simulations, continuing research efforts already noted in Bayesian geographical sampling design and multi-centre clinical trial design.

Chapter 7 thus supports the more fundamental objective O5. It is structured as a methodological paper proposing a Bayesian SSD approach, which is implemented in a case study on 2011 census data. This case study is designed with the context and constraints set by objective O6. This case study looks at statistical outputs of the kind normally produced by the ONS with censuses (objective O7). This implementation is based on the CV quality standards used to design survey collection by the US Census Bureau, the US Center for Disease Control and Prevention, or the UK Government Statistical Service, including the ONS and NHS Digital.

The paper combines two simulation studies. A set of model-based simulations are produced as part of the SSD algorithm. A design-based study using the census as a population then provides triangulation, verifying the assumptions used in the model-based study, particularly model validity. It also provides interesting findings with regard to efficiency savings targeted by objectives O4 and O7. This is again in line with the overall objective of this thesis to provide real-world evaluations of practical statistical designs. This manuscript is intended for submission to *Journal of the Royal Statistical Society – Series A (Statistics in Society)*.

4.3 Data sources

Due to its purpose, this thesis relies mostly on the most recent census data for two main reasons. First, as argued in chapter 6, the near absolute coverage of the census facilitates spatial analyses needed in order to identify the demographic and spatial structure in self-reported health statuses, and develop sound modelling specifications. Second, this doctoral project was funded to conduct translational research, comparing model-based population health estimates directly with reliable design-based estimates, instead of studying the properties of model-based estimation under abstract scenarios. The census is thus a very natural candidate to provide such reliable estimates. Using the census has two consequences on this research design.

First, the set of study outcomes is limited to those collected in recent censuses; SRH and LLTI. In order to better understand the characteristics of those health outcomes in terms of reliability and generalisability to other population health statuses, it is necessary to grasp a few key facts from the history with of their development and validation. Section 4.3.4 gives a brief overview of UK survey research on these health outcomes, highlighting some characteristics on questionnaire effect, reliability and harmonisation. Section 4.3.2 focuses on their implementation in UK censuses since 1991, particularly in relation to LLTI, which is the focus of all three academic papers in this thesis. This provides valuable knowledge to understand the purpose, construct, determinants and implications of these health statuses.

A second consequence is that suitable survey and auxiliary data sources to be identified to predict the prevalence of those two health statuses and compare outputs against the census. Section 4.3.3 provides the background and justification for the household survey selected for chapter 5, together with some comparability analyses reported in section 4.3.4. Section 4.3.5 then explains and justifies the choice of auxiliary data used in this same paper.

4.3.1 Self-reported health outcomes in UK household surveys

The earliest attempts to survey health conditions in the UK date back to the monthly Survey of Sickness, carried out between 1944 and 1952 as an the extension of the 1941 Wartime Social Survey. It implemented some of the first definitions of illness as a subjective measurement, as performing medical examinations at scale seemed unrealistic at the time (Logan and Brooke, 1957). An illness, understood as a condition listed in the ICD 1948, was measured based on the respondent's perception that something was presently disturbing his or her well-being. Initially, the

survey distinguished between 'new illness' (started in the current month), 'residual illness' (recurrent condition declared less than a year before the interview), 'chronic illness' (continuous or recurring ailment declared at least a year before the interview), in order to operationalise the requirement to measure disease prevalence as well as incidence on a monthly basis. Setting the period of reference to one year was not only a convenient span for respondents to remember the chain of events, it also matched some of the criteria underpinning many ICD codes. One version of the questionnaire iterated the following questions;

- 'How was your general health during [3 months period]?'
- 'Did you have any illness, ailment, poisoning or injury, or trouble with long-standing complaints during these months?'
- Additional prompts at discretion (for example, 'How does this affect you? Where do you feel it?')

The questionnaire underwent several changes across the life of the survey, so that questionnaire ordering effects are well documented. For instance, asking respondents about any recent visit to a hospital, doctor or dentist before reading the illness question proved to lead to significantly fewer reports of minor and ill-defined recurrent illnesses, and significantly more minor and undefined illnesses (Logan and Brooke, 1957, p. 24). This was regarded as a consequence of the more intensive interviewing taking place before this question is asked, as well as the ease to remember minor episodes of illness. This experience inspired questions on self-assessed morbidity in the General Household Survey, initiated in 1971. The Department of Health and Social Services required measurements of need and use of social and health services, for the purpose of planning (Whitehead, 1988). The General Household Survey questionnaire covered acute illness, long-term illness, limiting illness, as well as visits to family practitioners, hospitals and social services. The wording of the question proved to be easily answerable and to correlate narrowly to use of primary care and hospitals. The survey was designed with the awareness that measurement error was likely to be high, so that a wide variety of measurements were included. Short-term conditions were expected to be difficult to remember and the open-ended question ('what was the matter with you?') coded according to the ICD classification might not prove to enjoy good reliability. In addition, the absence of precise definition of 'incapacity' ('did you have to cut down at all on the things you usually do because of (this illness/disability or some other) illness or injury?') across those at work and other persons made the statistic unclear. Subsequent surveys deployed more efforts in operationalising specific definitions of incapacity. Overall, chronic illnesses were also recorded using an open question and classified into four levels of

Table 4.2 | Overview of previous long-term illness questions collected as part of the General Household Survey

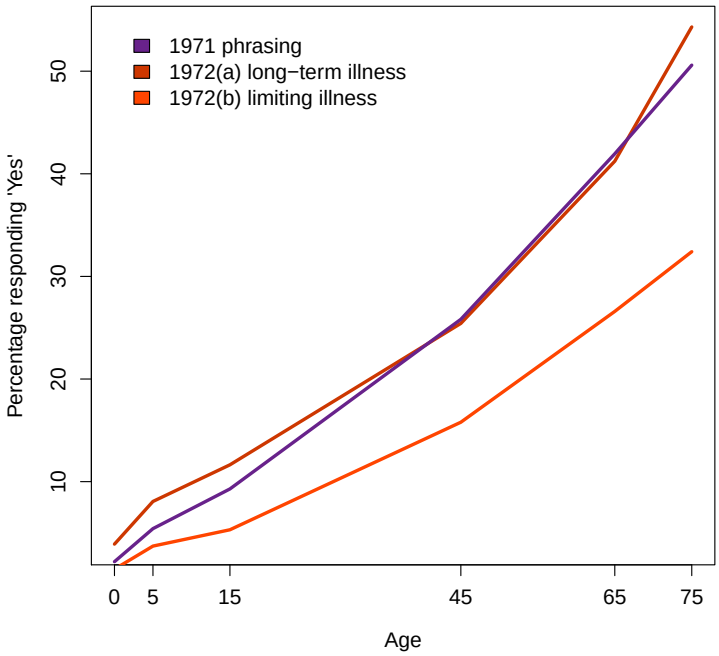
1971	*'Do you (or any of your children under 15) suffer from any long-standing illness, disability or infirmity which limits your activities compared with most people of your own age?'* [Followed by questions on health problems.] *'During the 2 weeks which ended last Sunday, did you (or any of your children under 15) have to cut down at all on the things you usually do because of (this illness/disability or some other) illness or injury?'*
1972	*'Do you suffer from any long-standing illness, disability or infirmity?'* [If yes:] *'Does this limit your activities compared with most people of your own age?'*
1973	*'Do you suffer from any long-standing illness, disability or infirmity?'* [If yes:] *'Does this limit your activities in any way?'*
1976	*'Do you have any long-standing illness, disability or infirmity? (By long-standing illness, I mean anything which has troubled you over a period of time or which is likely to affect you over a period of time)'*

severity using a predetermined ICD ranking and information collected on the length of time during which the respondent was incapacitated. The most important derived rates were;

- sickness rate,
- prevalence rate,
- incapacity rate,
- medical consultation rate.

The questionnaire underwent several modifications in subsequent years. One important change was made to the chronic illness question (see Table 4.2) as the first year of fieldwork revealed that respondents did not always pay attention to the second part of the question was completely pronounced, which relates to activity limitation. An independent pilot survey conducted in September 1971 included a split version of the question (OPCS Social Survey Division, 1973, pp. 271-273). Results showed higher levels of reports of activity limitation (see Figure 4.1). The split question became standard starting from the first quarter of 1972 and has remained common in most UK social surveys until recent harmonisation changes. This was particularly important considering the emphasis on limitations in activities had been initially designed to capture motivations for using the health service. This measure-

Figure 4.1 | Comparison of the questionnaire routine effect in the General Household Survey before and after splitting of the limiting illness question (adapted from OPCS Social Survey Division 1973, pp. 271-273)



ment of sickness proved to indeed correlate very strongly with use of GPs, in-patient and out-patient services. Another substantial change occurred in 1973 when the reference to old age, deemed too restrictive, was dropped.

Another self-reported health outcome was introduced in the General Household Survey in 1977. Given the limited year-to-year variation in many of the measurements included until then (OPCS Social Survey Division, 1977), further research is requested on respondents' perceptions on their own health with the question reported in Table 4.3. It was found to correlate well with service use or to conditions that are usually regarded as justifiable need for health care or public health intervention. Patients tend to approach their GPs when their assessment of their own health deteriorates, and therefore this question responds to a clear operationalised measurement of service use. Until 2005, the phrasing remained mostly unchanged until a new version conform to the Survey on Income and Living Conditions (SILC) harmonisation programme was introduced. The reference to a twelve-month period was dropped and the scale was increased to five categories. This led to a fall by about ten points in the percentage in good health (Smith and White, 2009). The new question was found to correlate more closely to the functional LLTI questions.

The question of comparability across space and demographics was raised for

Table 4.3 | Overview of previous SRH question phrasings in censuses and the General Household Survey

General Household Survey original phrasing	'Over the last twelve months would you say your health has on the whole been good, fairly good, or not good?'	General Household Survey 1976–2007, 2001 census
SILC harmonised question	'How is your health in general? Is it... very good, good, fair, bad, very bad?'	General Household Survey 2005–2011, 2011 census

SRH with even more severity than with LLTI. 'One problem with relying on the patient's own view of matters that are not entirely sensory lies in the fact that the patient's internal assessment may be seriously limited by his or her social experience. To take an extreme case, a person brought up in a community with a great many diseases and few medical facilities may be inclined to take certain symptoms as "normal" when they are clinically preventable.' (Sen, 2002, p. 860). Subgroups of a population of study may turn exhibit substantially different cut-point levels when reporting an illness or rating their own health, regardless of their health state analysed from a biomedical viewpoint. This has been described as 'reporting bias' and controlled for using latent trait analysis against a benchmark such as the SF-36 questionnaire. Common factors recognised to be driving heterogeneous reports are age, sex, education or income (Etilé and Milcent, 2006; Hernández-Quevedo et al., 2005). A substantial body of literature has also looked at cross-national differences in reports and potential for adjustment or recalibration of these measurements (see for instance Hernández-Quevedo et al. 2010; Pfarr et al. 2012).

Beyond health statuses, measurements of disability have been the object of substantial research. Existing ranking of severity of functional disorders came under severe criticism for failing to take into account social determinants to disability (Abberley, 1992). Following a cabinet publication on disability (UK Cabinet Office. Prime Minister's Strategic Unit, 2005), in-depth research was conducted the design of social surveys in Great Britain (see in particular ONS 2007; White et al. 2011; ONS 2011b). It ended in the definition of a disabled person in line with the Disability Discrimination Act 2005 and the Equality Act 2010 as 'someone with a (physical or mental) health condition or illness, lasting or expected to last for 12 months or more, which impairs their functioning and reduces their ability to carry out day-to-day activities' (White et al., 2011, p. 11). At that stage, the survey construct of disability was found to correspond to the original LLTI concept. The word 'disability' itself was effectively removed from all questions, after cognitive inter-

viewing revealed that it did not meet the expected consensus amongst respondents. A variety of questions were tested to reach agreement on four questions measuring, in line with UK legislation and the SILC harmonised framework, respectively long-standing health condition/illness, type of condition/impairment, activity restriction by extent and duration (see Table 4.4).

A pilot run in the 2009 General Lifestyle Survey articulated a short limitation question 'Would you say your activities are limited or strongly limited?' (`LimitL6`), which enjoyed good results thanks to its short and plain wording. It also turned out to yield responses very comparable to other more detailed phrasings used in the Labour Force Survey (LFS), Family Resources Survey (FRS) and Life Opportunities Survey (LOS), based on the detailed Disability Discrimination Act definition of limitations; namely substantial difficulties with any of the following prompted items—moving, lifting, manual dexterity, continence, communication, memory and concentration, sense of danger, physical co-ordination, and other health problem or disability (see `DisDif` in the LOS, section 4.3.3). It has the advantage of being simpler and reflects the broader definition of substantial adverse long-term effect on one's ability to carry out day-to-day activities as given by the Equality Act 2010 (c. 15), which is somewhat more general than adverse effects on the series of capacities listed in the Disability Discrimination Act. For those reasons, it was retained for the 2011 census and deemed broadly comparable to the 2001 question phrasing. All four questions presented in Table 4.4 are asked exclusively to adults over 16; responses concerning children aged under 16 and persons unfit to answer are only elicited from proxies.

4.3.2 Health statuses in UK censuses

In consultations leading to the 1991 census, the Department of Health and Social Security made the case for a morbidity question to be included. The objective was to better understand the expressed and unexpressed need for healthcare services and improve funding allocation formulae. In the context of NHS reforms announced by the 1989 White Paper (UK Department of Health, 1989), the Department was planning the introduction of an allocation model solely based on standardised mortality ratios. Standardised mortality ratios can be seen as a proxy of a population's health and of limited use to develop a prevention-based health service. But primary care and in-patient statistics suffered from other weaknesses, with the usual issues of quality, consistency, and the well-documented 'iceberg problem' of unexpressed needs, especially given the unequal offer and rates of use of health-

Table 4.4 | Overview of harmonised questions for the measurement of long-lasting conditions, activity limitation by extent and by duration, and impairment (White et al., 2011; ONS, 2011b)

Long-lasting health condition	[Introduction:] 'This question asks you about any health conditions, illnesses or impairments you may have.' [Question:] 'Do you have any physical or mental health conditions or illnesses lasting or expected to last for 12 months or more?' 1. Yes 2. No
Activity restriction harmonised question[*†]	[Introduction:] 'This question asks about whether your health condition or illness currently affects your ability to carry-out normal day-to-day activities, either a lot or a little or not at all. In answering this question, you should consider whether you are affected whilst receiving any treatment or medication for your condition or illness and/or using any devices such as a hearing aid, for example.' [Question:] 'Does your condition or illness/do any of your conditions or illnesses reduce your ability to carry-out day-to-day activities?' [Running prompt:] 1. Yes, a lot 2. Yes, a little 3. Not at all
Length of activity limitation[†]	For how long has your ability to carry-out day-to-day activities been reduced? [Running prompt:] 1. Less than six months 2. Between six months and 12 months 3. 12 months or more
Impairment harmonised question	[Introduction:] 'The purpose of this question is to ask you about any impairment(s) you have. In answering this question, you should consider whether you are affected in any of these areas whilst receiving any treatment or medication or using devices to help you such as a hearing aid for example.' [Question:] 'Do you have any health conditions or illnesses which affect you in any of the following areas?' [Show card and code all that apply asking the respondent to read out which apply to him/her:] 1. Vision (for example blindness or partial sight) 2. Hearing (for example deafness or partial hearing) 3. Mobility (for example walking short distances or climbing stairs) 4. Dexterity (for example lifting and carrying objects, using a keyboard) 5. Learning or understanding or concentrating 6. Memory 7. Mental health 8. Stamina or breathing or fatigue 9. Socially or behaviourally (for example associated with autism, attention deficit disorder or Asperger's syndrome) 10. Other (please specify) 11. None of the above

[*] Question included as part of the 2011 UK census.
[†] SILC 'severely hampered/hampered to some extent in daily activities'

Table 4.5 | 1989 census test and post enumeration survey: responses to the LLTI

χ^2 $p < 0.01$		Census test form No	Yes	Total	Census resp. rate
PES	No	85%	1%	86%	93%
interview	Yes	3%	11%	14%	95%
	Total	88%	12%	100%	95%

Note: PES: post-enumeration survey.
Source: Adapted from Pearce and Thomas 1990, p. 28, n = 3,591, excluding 190 missing observations.

care services across the nation. On the other hand, social surveys offered insufficient power to estimate small area statistics with direct estimators within reasonable budgets. The decennial census thus emerged as an opportunity to collect this kind of information at modest marginal costs, as previous research had shown that self-assessed measurements could achieve the basic target of measuring proxies for health care needs.

Preliminary research on questionnaire design was inspired by knowledge accumulated by questions on acute, long-term illness, and visits to family practitioners and hospitals in the General Household Survey since 1971. These questions had proved, overall, to constitute good proxies for measurement of health care need, since they appeared to be good operationalised questions measuring conditions objectively leading populations to using, or at least in need of using health care services. Two initial questions were investigated by the Office of Population Censuses and Surveys (Thompson, 1995). A multiple choice question on disability followed by short-term and long term illnesses yes/no questions were tested in autumn 1987 (see Figure 4.2). This exercise was designed to provide evidence with regard to the acceptability, reliability and understanding of these questions. Fourteen census districts (areas covered by one census officer, responsible of a team of enumerators) were selected in 7 local authorities of England and Scotland in such a way as to cover rural, areas, small towns, inner cities, university towns, and London (Pearce et al., 1988; White, 1990; Pearce and Thomas, 1990); 85,500 households were handed out questionnaires. The 1989 final census conducted in April on 90,000 households from 6 local areas. Answers were linked with data from an additional Post-Enumeration Survey conducted by a separate directorate of the Office of Population Censuses and Surveys on about 2,300 households, of which 1,500 had co-operated in the census test (Table 4.5).

Overall, the disability question proved to be unreliable, for several reasons. Of

all census test respondents, 21 per cent did not provide any answer to it. The Post-Enumeration Survey also suggested that those nonrespondents were more likely to be disabled. On the other hand, 70 per cent of respondents to the Post-Enumeration Survey who were classified by fieldworkers as disabled along the interview had not ticked any disability on the census test form. It became clear that this question did not meet census quality standards, and should thus be dismissed. Interestingly enough, the same decision was made as in 1921, when the infirmity question asked between 1851 and 1911 was discontinued over concern around its quality (Logan and Brooke, 1957; Whitehead, 1988).

The short-term limiting illness question, which does not anyway relate as closely to health care services use, did not yield very strong results either; only 50 per cent of answers provided on the test form agreed with answers collected during the post-enumeration survey, which could also be due to inaccurate recalls as well as to mode effects. In contrast, Table 4.5 shows that responses exhibited higher rates of consistency as far as the LLTI question is concerned (96% valid answers, which is goes down to 91% when including nonresponse to the census test). The pen-and-paper mode of collection seems to lead to morbidity underestimation, suggesting post-enumeration fieldworkers were able to elicit more exhaustive answers, or help respondents recall better. Complementary analysis confirmed that the LLTI question correlates very strongly with use of general practice and hospital services, providing confirmation of findings obtained from ten years of General Household Survey research (see previous section).

Based on such results, the LLTI yes/no phrasing was deemed both more reliable and more meaningful than the short-term illness question and disability multiple choice (Whitehead, 1988). The post-enumeration survey also revealed moderate levels of objections to that question (6%, compared to 7% for the ethnicity question). Since the General Household Survey had shown that this was the most robust predictor of hospital use, the final questionnaire retained an LLTI question phrased as follows: *'Do you have any long-term illness, health problem or handicap which limits your daily activities or the work you can do? Include problems which are due to old age'*.

Maps of standardised morbidity could subsequently be produced, and results were used to develop more advanced funding allocation models based on a wider range of empirical indicators. The reviewed weighted capitation formula was rolled out in 1996–1997. The main difference with General Household Survey estimates, beyond their precision, stemmed from the inclusion of a prompt regarding age-related health problems. The greater granularity meant that disability-free life ex-

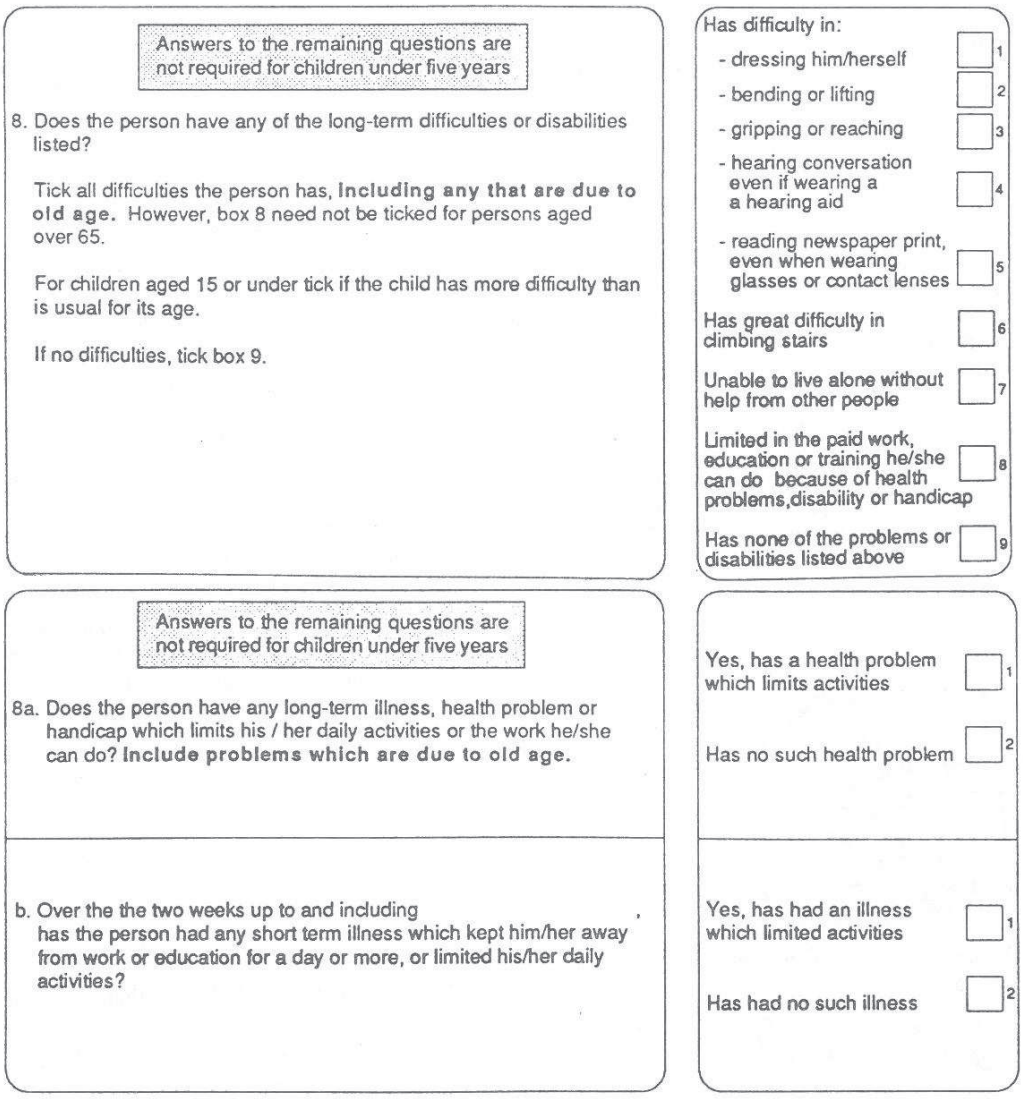


Figure 4.2 | 1987 census wording test: health schedule (reproduced from: Pearce and Thomas 1990, p. 29).

pectancies could now be estimated at a smaller level. This was done by the ONS for the first time at ward level after the 2001 census (ONS, 2001b). This allowed to investigate geographical inequalities further, with tools such as the slope and relative indices of inequality (for a review of indicators, see Mackenbach and Kunst 1997). In addition, unlike household surveys, the census brought coverage of communal establishments, which provided the necessary information to publish health-adjusted life expectancy estimates for these populations as well.

Results also informed academic research to a great extent, especially in the field of health economics with regard to funding allocation and spatial inequalities in need for and access to health care. Substantial work was carried out around the spatial

distribution of health inequalities, neighbourhood effects, ethnicity density effects, etc. Further research tried to establish how consistent self-reported health measurements were across demographics. This is important for the purposes of measuring inequality in health, as there are risks that respondents' assessment is influenced by the health of their relatives, who might share both the same socioeconomic, age or ethnic background and therefore health risks. This task evidently remains a challenge considering how tenuous the boundary between subjectivity, heterogeneity, and under- or over-reporting. Respectively, respondents of different demographic cross-sections of the population may differ in their reports of illnesses for various reasons; differences in ('objective') biological health, differences in the severity with which they reflect upon their own health, differences in the health of people that are close to them, or even differences in economic or social circumstances that soften or harden the burden of disease (e.g. receiving help from a carer, having the financial means of improving lifestyles and reducing the burden of work or caring for others). There have been suggestions that the 1991 output exhibited different cut-off points in reports of LLTI; for instance Senior (1998) failed to explain the entirety of the excess of morbidity measured in Wales and North West England, and argued that this could be a sign of regional differences in propensity to recall or report such an illness all other things being equal. To date, there has nevertheless been little convincing evidence of the existence of substantial subnational measurement error related to potential 'cultural' effects with LLTI. This may yet be more of an issue with other subjective measurements.

Whilst this has not affected results and their spatial distribution since 1991, census questionnaires have undergone some minor changes over time (see Figure 4.3). In 2001, the LLTI question was extended to Northern Ireland and the phrasing was altered to a minor extent, mostly with regard to the explicit use of the word 'disability'. The new phrasing was again well received and the census quality survey carried out in 1999 shortly after the dress rehearsal showed rates of agreements between responses provided on forms and in interviews of about 94 per cent, most of the differential stemming from difficulties arising in proxy responses (ONS, 2005). A question on SRH similar to the one used in the General Household Survey since 1976 was also introduced and met less success with respondents (88% agreement rate). In 2011, new question phrasings were used following testing in the General Lifestyle Survey, so as to harmonise measurements with the European Community Health Indicators list. The LLTI question was replaced by a question on activity limitation retaining the old age prompt, but adding a condition of duration in line with SILC standards. In contrast, the SRH question lost the reference to a period of

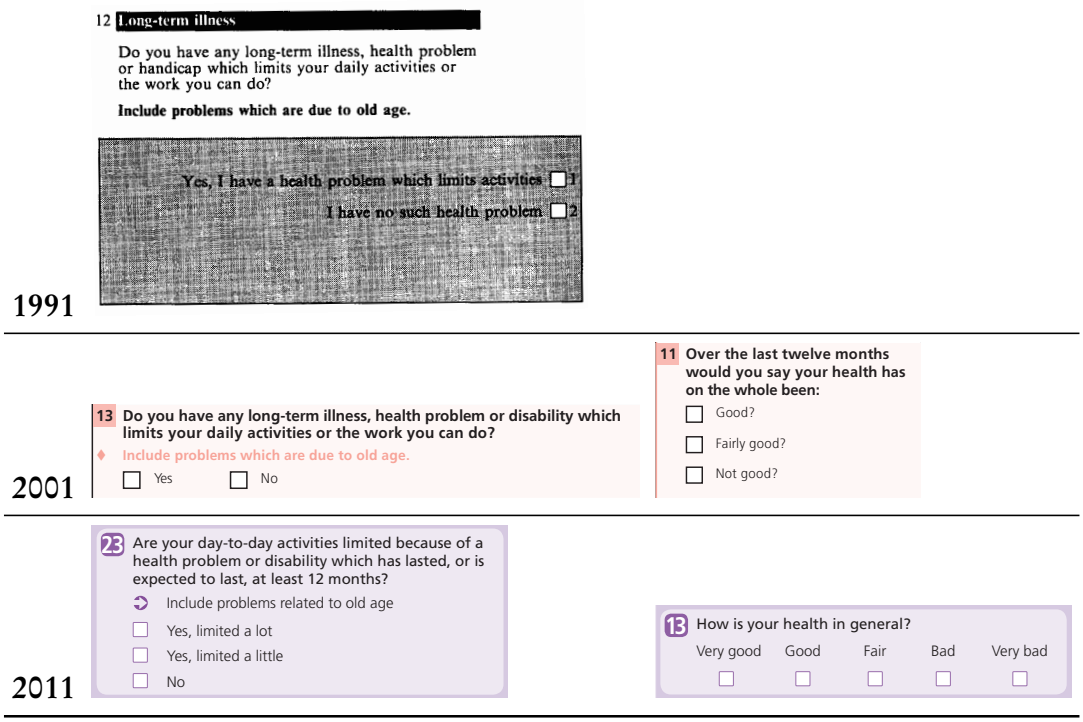


Figure 4.3 | Longstanding illness and SRH questions in successive censuses (reproduced from ONS 2001a, 2002, 2010b)

twelve months and implemented the SILC five-item Likert scale. These wordings are therefore valid to measure healthy life expectancy and disability-free life expectancy according to definitions of the basket of European Community Health Indicators. A last difference resides in the questionnaire ordering; compared to 2001, questions were not longer consecutive but rather set in far-apart sections of the census form.

4.3.3 Choice of survey data

A number of social surveys accessible to the academic community were considered to carry out this work. The choice itself was based on the following criteria:

1. the availability of household geographical identifiers
2. the available sample size and survey design effect
3. the level of comparability with 2011 census health status measures (time period, question phrasing, mode effect, aggregate consistency)
4. the population of inference.

It is possible to distinguish between two groups of surveys; surveys collected using multistage sampling and surveys collected by sampling addresses directly.

Several surveys follow a multistage selection scheme, with postcode sector serving as primary sampling units (PSUs). These include the Health Survey for

England, the General Lifestyle Survey, the Living Costs and Food Survey, the National Statistics Opinions Survey. They feature question phrasings similar to those used in the census, which would have been ideal for this research. However, there are good reasons why clustered designs would ideally be avoided in the context of SAE. They are traditionally less efficient by nature, meaning that for equal sample size, clustered designs yield larger sampling variances. Furthermore, most UK multistage surveys are clustered by postcode sector, a small subdivision of an MSOA. It is unlikely that more than one postcode sector is going to be drawn in most MSOAs. This constitutes an issue for the type of models used in SAE. An identification problem may arise if MSOAs are the chosen prediction output geography. This is because it will be very difficult to ensure sufficient postcode heterogeneity within MSOAs to estimate unbiased variance components. In other terms, since most MSOAs will contain exclusively one PSU, the model will not easily make the distinction between hierarchical levels. Inference thus risks becoming unstable. As a result, in order to simplify the task, it seemed preferable to restrict this research to unclustered surveys as much as possible.

Other surveys do not feature such shortcomings as they do not involve any area sampling at all; the LFS, the EHS, the FRS and the LOS are clustered exclusively by households. They feature very small design effects and are thus good sources of data to compare health questions phrasings (see Table 4.7). The Crime Survey for England and Wales (CSEW) is also included considering its clustering is for the most part at the MSOA level, and that the phrasing of health questions is very close to that of the census. It is important to note that the population of inference varies across these surveys. Most social surveys are restricted to residents of private households aged 16 years and over. The EHS and the census nevertheless collect information on children aged under 16. In the LFS, respondents aged over 75 years were ignored in this analysis. They were asked on first contact whether they wished to answer health questions, which arguably leads to items missing not-at-random (self-selection). Finally, to date, few social surveys have covered the estimated 952,700 population living in communal establishments in 2011, and none of the three makes exception to that rule.

Census day was on 27 March 2011. Data collated to fit models for census days consist of microdata collected on first contact (first wave) of the LFS up to one year before and after this date, in order to reach a critical level to undertake analysis of spatial variance. The revolving design of the LFS around waves JM (January-March), AJ (April-June), JS (July-September) and OD (October-December) implies that ideally datasets retained for pooling would be JM 2010 to JM 2012. It is unlikely that the

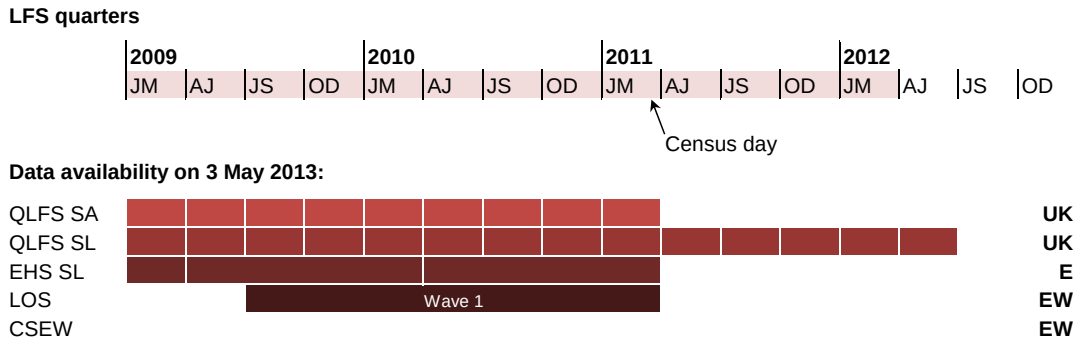
Table 4.6 | Breakdown of the expected sample size for the LFS dataset

Number of cases (valid or imputed) on wave 1 in 2009	England & Wales	Scotland
LFS	83,148	8,528
Sampling fraction (%)	0.148	0.161
Expected average per LAD	238	266
Expected average per MSOA	11	6

Note: Estimated for 348 LAD and 7,201 MSOAs across England and Wales, 32 council areas and 1,235 Scottish 2001 MSOAs (smaller than their E&W counterparts, yet to be redesigned for 2011 output). Based upon estimated usual resident population (2011 census).

population would feature substantial differences one year before or after census day. Based on achieved sample sizes found for 2009 (see Table 4.6, slightly higher than those expected for 2010–2012), it is possible to hope reaching an approximate total size of 160,000 cases for England and Wales, and 30,000 for Scotland. This would be tantamount to a 0.29 per cent sampling fraction, and comes close to what is computationally feasible for a complex hierarchical model. The usable sample size once case-wise missing data are filtered out, and considering response rates have been falling further since 2009, could be as low as 140,000.

These numbers represent an assessment of the total number of cases collected by the ONS. The number of cases distributed, and geo-referenced, is different. **Data geo-referenced at the LAD level** is available from Quarterly Labour Force Survey (QLFS) Special Licence datasets, which are deposited for the entire period. However, **data geo-referenced at the MSOA level** is only available from Secure Access datasets. The third edition (3 May 2013) of the QLFS Secure Access datasets do not cover quarters beyond JM 2011, that is to say beyond census day. Such datasets are deposited at irregular dates (the second edition was deposited in August 2011), so that it does not seem recommended to hope seeing more data being deposited before 2015/16. An estimate of sample sizes attainable for this project for spatial analysis of variance at the MSOA level is difficult to produce. Assuming that each quarter, the LFS provides between 15 and 20 thousand wave 1 cases, we could hopefully reach 90,000 cases. The actual number of valid answers to LLTI and SRH questions is likely to be considerably smaller, due to item non-response, etc.



Note: SA: Secure Access; SL: Special Licence.

Figure 4.4 | Calendar representation of waves of the LFS around Census day

Table 4.7 | Comparison of self-reported health questions in four current household surveys in 2010–11

	LFS	EHS	LOS	CSEW	FRS
Population of inference	Residents of **UK** private households aged 16 years and over. If aged 75 years and over, only if willing to continue with health questions.	Residents of **English** private households aged 16 years and over.	Residents of **GB** private households aged 16 years and over.	Residents of **English and Welsh** private households aged 16 years and over.	Residents of **UK** private households aged 16 years and over. **Children and household members aged under 19 years, unmarried, or in education and training are recorded by proxy.**
Response rate	60 per cent (wave 1)	59 per cent	60 per cent	75 per cent	59 per cent
Licence	Special Licence – Secure Access	Secure Access	Special Licence	Special Licence	Special Licence – safe room
Geographical identifiers	LAD/MSOA	MSOA	LAD	MSOA	LAD
Sample size	10,000 UK households per quarter. This is equivalent to **80,000 English and Welsh individuals per year**.	Circ. 17,000 English households (April to March). Three years of data can provide up to **80,000 English individuals**.	Circ. 15,000 GB households (+4,000 with partial interviews). Information for up to **32,000 E+W individuals**.	Circ. **46,000 English and Welsh households (1 individual per household)**.	Circ. **25,500 UK households** (April to March) each year.
Long-standing illness	*'Do you have any health problems or disabilities that you expect will last for more than a year?'* (`LNGLIM`)	*'Do you have any long-standing illness, disability or infirmity – by long-standing I mean anything that has troubled you over a period of time or that is likely to affect you over a period of time?'* (Integrated Household Survey (IHS) core variable `LSIll`)	*'Do you have any long-standing illness, disability or infirmity – by long-standing I mean anything that has troubled you over a period of time or that is likely to affect you over a period of time?'* (IHS core variable `LSIll`) *'Do you have any long-standing physical or mental impairment, illness or disability? By "long-standing" I mean anything that has affected you over a period of at least 12 months or that is likely to affect you over a period of at least 12 months.'* (`HEALTH`)	*'Do you have any of the following long-standing physical or mental health conditions or disabilities that have lasted or are expected to last 12 months or more?'* [CODE ALL THAT APPLY] 1. Blindness, deafness or other communication impairment 2. Mobility impairment, such as difficulty walking 3. Learning difficulty or disability, such as DownâĂŹs syndrome 4. Mental health condition, such as depression 5. Long-term illness, such as Multiple Sclerosis or cancer 6. Other long-standing health condition or disability 7. None of these	*'Do you have any long-standing illness, disability or infirmity? By "long-standing" I mean anything that has troubled you over a period of at least 12 months or that is likely to affect you over a period of at least 12 months.'* [THIS IS A QUESTION OF OPINION] (`HEALTH`)

	LFS	EHS	LOS	CSEW	FRS
Limiting illness	*'Does this health problem affect the KIND of paid work that you might do?'* (`LIMITK`) *'Does this health problem affect the AMOUNT of paid work that you might do?'* (`LIMITA`) *'Do these health problems or disabilities, when taken singly or together, substantially limit your ability to carry out normal day to day activities? If you are receiving medication or treatment, please consider what the situation would be without the medication or treatment?'* (`HEALIM`)	*'Does this illness or disability (Do any of these illnesses or disabilities) limit X's activities in any way?'* (IHS core variable `IllLim`)	*'Does this illness or disability (Do any of these illnesses or disabilities) limit your activities in any way?'* (IHS core variable `IllLim`)	*'[Does/do] your health condition[s] or [disability/disabilities] mean that your day today activities are limited? Would you say you are…* [READ OUT] 1. *Severely limited* 2. *Limited but not severely* 3. *or not limited at all?'*	'*Does this physical or mental illness or disability (Do any of these physical or mental illnesses or disabilities) limit your activities in any way?'* (`HProb`) *'Have you ever had a long-term illness, disability or infirmity that affected your activities? (By long-term I mean anything lasting for more than a year)'* (`DisDifP`) *'Did this health problem or disability (Did these health problems or disabilities, when taken singly or together) substantially limit your ability to carry out normal day-today activities? If you were receiving medication or treatment, please consider what the situation would have been without medication or treatment.'* (`DDATreP`) *'Can I just check, have you ever been diagnosed with a health condition which could substantially affect your day-to-day activities in the future? This is an opinion question.'* (`DDAProg`, introduced in 2008–09, NO guidance provided on this question since 2009–10)
General health		*'How is your health in general; would you say it was… Very good, good, fair, bad, or very bad?'*	*'How is your health in general; would you say it was… Very good, good, fair, bad, or very bad?'*	*'I would now like to ask you for a few further details about yourself [and your household]. How is your health in general? Would you say it is… Very good, good, fair, bad, or very bad?'*	

LFS	EHS	LOS	CSEW	FRS
Difficulties		'*Does this/Do these health problem(s) or disability(ies) mean that you have substantial difficulties with any of these areas of your life? Please read out the numbers from the card next to the ones which apply to you.* [SHOWCARD] 1. Mobility (moving about) 2. Lifting, carrying or moving objects 3. Manual dexterity (using your hands to carry out everyday tasks) 4. Continence (bladder & bowel control) 5. Communication(speech, hearing or eyesight) 6. Memory or ability to concentrate, learn or understand 7. Recognising when you are in physical danger 8. Your physical co-ordination (e.g. balance) 9. Other health problem or disability 10. None of these' (`DisDif`)		*'Does this/Do these health problem(s) or disability(ies) mean that you have substantial difficulties with any of these areas of your life?'* [SHOWCARD] 1. Mobility (moving about) 2. Lifting, carrying or moving objects 3. Manual dexterity (using your hands to carry out everyday tasks) 4. Continence (bladder and bowel control) 5. Communication (speech, hearing or eyesight) 6. Memory or ability to concentrate, learn or understand 7. Recognising when you are in physical danger 8. Your physical co-ordination (e.g. balance) 9. Other health problem or disability 10. None of these' (`DisDif`) [Probe and code at this question all substantial difficulties the illness causes the respondent. This is important to allow DWP to capture and analyse the extent to which disabilities/health problems affect the areas of respondents' lives.]

	LFS	EHS	LOS	CSEW	FRS
Health conditions	*'Do you have…'* (`HEAL*`) [PROMPTED FOR EACH OF THE FOLLOWING:] *'problems or disabilities (including arthritis or rheumatism) connected with your arms or hands?; … legs or feet?; … back or neck?; difficulty in seeing (while wearing spectacles or contact; lenses)?; difficulty in hearing?; a speech impediment?; severe disfigurements, skin conditions, allergies?; chest or breathing problems, asthma, bronchitis?; heart, blood pressure or blood circulation problems?; stomach, liver, kidney or digestive problems?; Diabetes?; depression, bad nerves or anxiety?; Epilepsy?; severe or specific learning difficulties?; mental illness or suffer from phobias, panics or other nervous; disorders?; progressive illness not included elsewhere (eg cancer not; included elsewhere, multiple sclerosis, symptomatic HIV,; Parkinson's disease, Muscular Dystrophy)?; other health problems or disabilities?'*	*'What type of illness or disability do you/does [name] have?'* [CODE ALL THAT APPLY] 1. Vision 2. Hearing 3. Learning difficulty 4. Mobility 5. Breathing problems 6. Heart disease 7. Mental health problems 8. Other	*'Have you ever been diagnosed with any of the conditions shown on this card?'* [SHOWCARD (HIV, cancer, multiple sclerosis)] (`DDAProg`) *'Now I'd like to talk about chronic health conditions. Do you have any of the following long-term conditions that have lasted or are expected to last 12 months or more and that have been diagnosed by a health professional…'* [SHOWCARD] (`ICond, CondOth`) 'Asthma or severe allergies ; Heart condition or disease ; Kidney condition or disease ; Cancer ; Diabetes ; Epilepsy ; Cerebral Palsy ; Spina Bifida ; Cystic Fibrosis ; Muscular Dystrophy ; Migraines ; Arthritis or rheumatism ; Multiple Sclerosis ; Paralysis of any kind ; Any other long-term condition not already covered (please specify); None'		

All these surveys focus on residents of private households. As exposed before, ascertaining the population of inference can be difficult in the case of individuals who may also be regarded as falling within category of communal establishment residents. This is, in particular, the case of students, student nurses, and people living in retirement homes. In the case of the EHS, the question is asked explicitly and respondents in that situation are coded as living in residential housing. In contrast, the LFS does not include an explicit question but household members are nevertheless coded with regard to their type of residence based on interviewer knowledge.

4.3.4 Comparative analysis of SRH and LLTI in three UK surveys in 2011

Previous sections emphasised concerns around the internal validity of morbidity-related questions, especially in terms of comparability across demographics. The questionnaire effect is another concern which has been well documented along with the experience gained in previous surveys. Phrasings in three of our five candidate surveys are examined with regard to differences in estimates potentially related to variations in question phrasings or orderings. In all three cases, health questions were asked about two-thirds through the interviews. It is therefore difficult to assess exactly how the nature of preceding questions may have affected answers, due to their sheer number. Nevertheless, the structure of the LOS questionnaire has more obvious implications. Interestingly enough, wave one of the LOS contained both the core IHS health question and alternative phrasings. First, core IHS question are asked, including;

- 'How is your health in general; would you say it was... very good, good, fair, bad, or very bad?' (`QHealth1`);
- 'Do you have any long-standing illness, disability or infirmity - by long-standing I mean anything that has troubled you over a period of time or that is likely to affect you over a period of time?' (`LSIll`);
- 'Does this illness or disability (Do any of these illnesses or disabilities) limit your activities in any way?' (`IllLim`).

They are followed by a large section on employment, before carrying on with the LOS questionnaire, around two thirds through. The following questions are then asked:

- 'Do you have any long-standing physical or mental impairment, illness or disability? By "long-standing" I mean anything that has affected you over a period of at least 12 months or that is likely to affect you over a period of at

least 12 months' (`Health`);

- 'Does this/Do these health problem(s) or disability(ies) mean that you have substantial difficulties with any of these areas of your life? Please read out the numbers from the card next to the ones which apply to you.' (`DisDif`);
- 'Can I just check, do you receive medication or treatment without which your health problems (when taken together), would substantially affect your life in the areas we have been discussing?' (`DDATre`);
- 'Now I'd like to talk about chronic health conditions. Do you have any of the following long-term conditions that have lasted or are expected to last 12 months or more and that have been diagnosed by a health professional...' SHOWCARD (`ICond`);
- [followed by questions regarding whether the respondent received a treatment for those, how he or she was operating thanks to this treatment, and how often the condition limited the kind of activities he or she could do].

This is the opportunity to compare the IHS standard questions with more operationalised phrasings. In such situations, two competing effects may often be observed, all other things being equal. More detailed questions or prompts lead some individuals to report difficulties they had not until then necessarily regarded as such, either because they are used to them, or because they are not frequently envisaged as serious medical conditions (for example, poor eyesight partially corrected by glasses). Conversely, a more detailed question may prove to give individual less leeway to include rare or atypical conditions.

A first comparison is drawn between two question phrasings for *longstanding illness*. Although aggregate estimates are identical for either question, the breakdown (see table below) reveals that a substantial 10 percent share of respondents provide a different answer across the two questions. This illustrates how very subjective and unstable such questions can be. There is no obvious indication that the order of questions could be influencing answers (for instance, the second question does not lead to more respondents declaring an illness).

The second comparison is made around questions on *limiting illness*, which follow those on *longstanding illness*. Although the detailed breakdown of answers to `DisDif` is not published, variable `DVDDACur` codes whether the respondents reported experiencing at least one of the difficulties prompted. Results are not entirely neutral considering that the `DVDDACur` question is preceded by the IHS question, as well as by a long series of other questions likely to prepare respondents to make a more in-depth assessment of their heath. By the time they reach this detailed question, they are already more likely to report a wider range of difficulties, let alone

Table 4.8 | Comparison between two LOS longstanding illness phrasings phrasings (source: ONS. Social Survey Division 2012)

χ^2 $p < 0.001$		LOS `lsill` (IHS core) No	Yes	Total
LOS `Health`	No	60%	5%	65%
	Yes	5%	30%	36%
	Total	65%	35%	100%

Table 4.9 | Cross-tabulations of LOS disability questions with routine LOS and IHS questions on longstanding and limiting illness

(a)

χ^2 $p < 0.001$		LOS `illim` No	Yes	Total
LOS `DisDif`	No	70%	3%	73%
	Yes	8%	19%	27%
	Total	78%	22%	100%

(b)

χ^2 $p < 0.001$		LOS `lsill` No	Yes	Total
LOS `DisDif`	No	62%	12%	74%
	Yes	3%	23%	26%
	Total	65%	35%	100%

(c)

χ^2 $p < 0.001$		LOS `Health` No	Yes	Total
LOS `DisDif`	No	64%	10%	74%
	Yes	0%	26%	26%
	Total	64%	36%	100%

Source: Author's own calculations. Adapted from ONS. Social Survey Division 2012.

if they had been asked using a pen-and-paper census form. These caveats in mind, a contingency table (Table 4.9a) indicates that reading out types of impediments potentially encountered by respondents owing to their condition seems to slightly mitigate underreporting. This leads to a higher rate of morbidity reports; 5 per cent additional reports are recording, net of the 3 per cent respondents who were not classified as encountering those difficulties despite reporting a limiting illness at the IHS question.

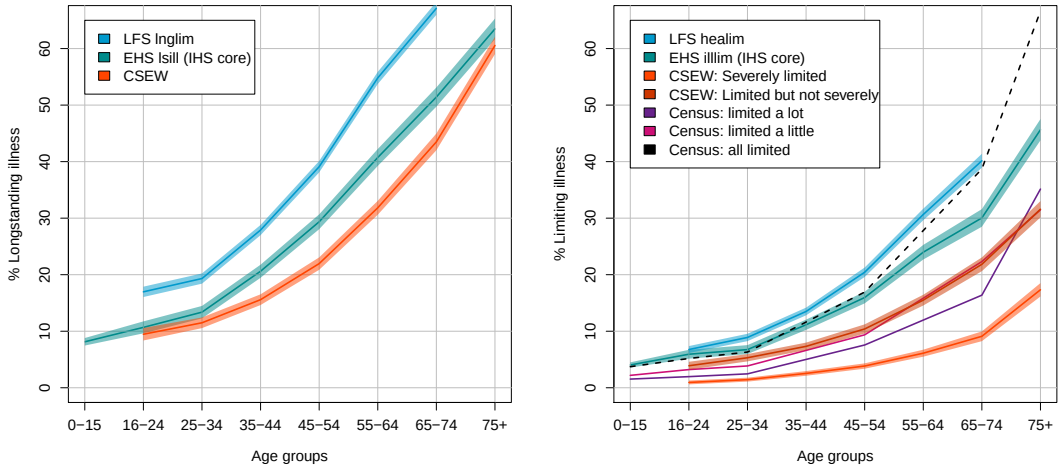
Beyond this aggregate analysis, it is expected that question phrasings are going to yield very different demographic patterns. For instance there is not the same understanding of the question whether people think they should compare to other individuals their age or not. The LFS has a very broad definition of the 'longstanding condition' which means it is expected problem related to old age would tend to be included by respondents. The next section explores these demographic patterns. It

is organised according to two sets of comparison depending upon the geographic scale surveys are available at; first we look at England (LFS, EHS/IHS, CSEW), and then Great Britain (LFS and LOS).

The LFS `lnglim` question phrasing is the most general as it refers very broadly to 'health problems'. Other surveys use more complicated structures and wordings. The LFS is, of our four surveys, the one that yields highest morbidity estimates. Its `healim` question on limiting illness comes after two questions on ability to work, and it is therefore likely that respondents will focus on how their illness limit their ability to do housework, socialise and engage in non-work related activities. By doing so they envisage broader situations than if they were asked a standalone question on day-to-day activities, which on its own could be interpreted as referring essentially to 'work activities'. In addition, the LFS prompt with regard to medical treatments makes the phrasing even more general. It can be thought to remove ambiguity for individuals who function perfectly normally only thanks to medication. Other question phrasings do not make that distinction explicit, and it is possible that the same individuals would in that case not report being limited in their day-to-day activities. Comparisons between the LFS and IHS questions (answers taken from the EHS, see Figure 4.5b) show that the LFS indeed consistently yields higher estimates. This observation does not necessarily imply the somewhat unique LFS question phrasing of `healim` makes it unsuitable to analyse disability. Regional patterns in response to both Annual Population Survey and General Lifestyle Survey were compared by Smith et al. (2011), who found on the contrary lower disability and disability-free life expectancy estimates in the APS, especially after the age of 65. However, this did not affect country or age group gradient in disability-free life expectancy.

The wording of the Crime Survey, which is closest to 2011 census questions, yields more conservative estimates on the other hand. Finally, census estimates from table DC3302EW were confronted to the set of limiting illness estimates (*nota:* across age group 35–64 years, only one average estimate was computed) and seem to be partially similar. Data suggests a very strong resemblance in the profile of responses between the Crime Survey (limited but not severely) and the 2011 census (limited a little). However, taking all limitation categories together (see black dotted line), it seems that elderly respondents report being a lot more seriously limited in their activities. The LFS and EHS estimates are thus only distantly related to census age patterns in answers. Gender differentials were also briefly examined, although they are virtually non-existent in either case.

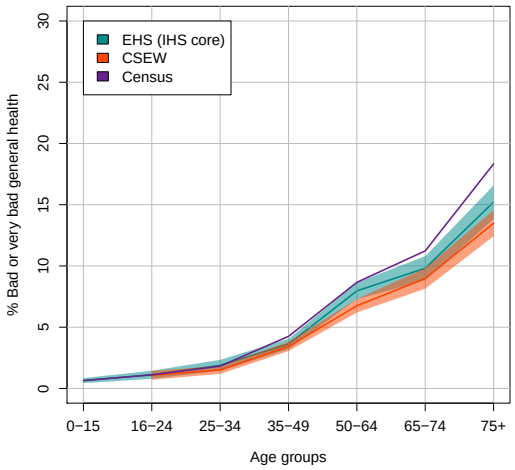
Figure 4.7 reports similar analyses for the poor/very poor SRH responses. The



(a) Longstanding illness **(b)** Limiting illness

Source: Author's own calculations. Adapted from 2011 table DC3302EW; UK Department for Communities and Local Government 2013; ONS. Social Survey Division, Northern Ireland Statistics and Research Agency. Central Survey Unit 2015, 2014a,b,c; Home Office & TNS-BMRB 2012.

Figure 4.5 | Age-specific prevalence of (a) longstanding illness and (b) limiting illness in the LFS, EHS, CSEW and the census in England in 2011 with 95% confidence intervals

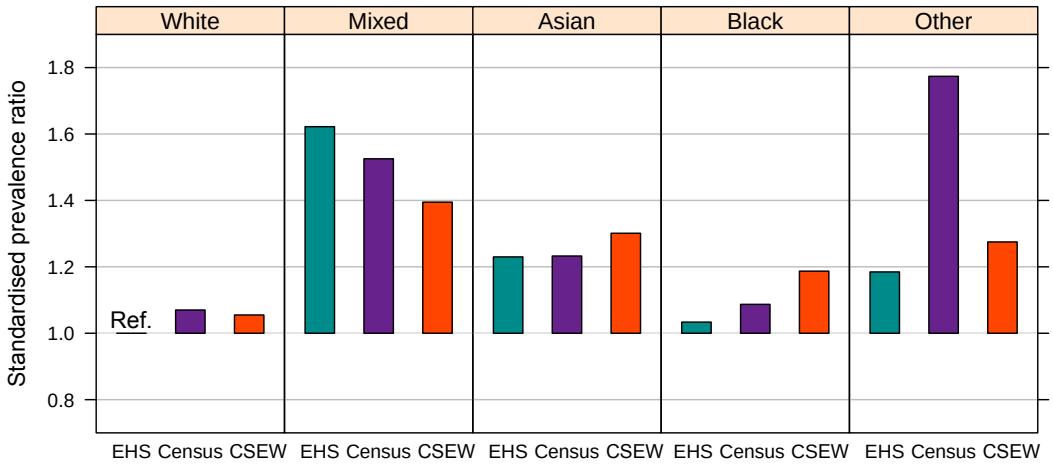


Source: Author's own calculations. Adapted from 2011 tables LC3206EW, DC3205EWr, DC3302EW; UK Department for Communities and Local Government 2013; Home Office & TNS-BMRB 2012.

Figure 4.6 | Bad/very bad SRH in England in 2011 by age and ethnic groups in the EHS, CSEW and the census with 95% confidence intervals

Figure 4.7 | Bad/very bad SRH by ethnic group in the EHS, CSEW and the census in England in 2011: indirectly age-standardised prevalence ratios and 95% confidence intervals

(a) Point estimates



(b) 95% confidence intervals

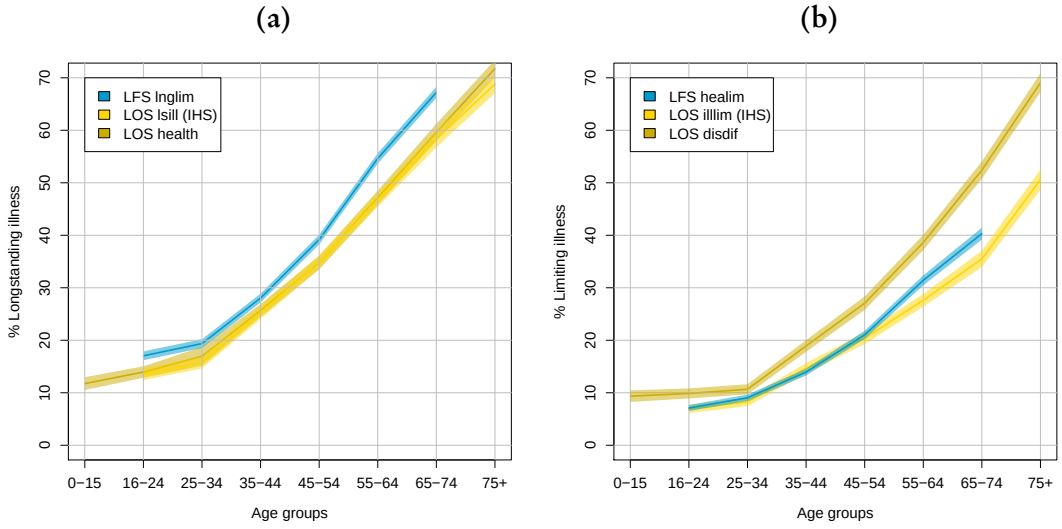
	EHS		Census		CSEW	
	Lo. bound	Up. bound	Lo. bound	Up. bound	Lo. bound	Up. bound
White	1.000	1.000	1.042	1.068	1.070	1.070
Mixed	1.054	2.190	1.006	1.784	1.515	1.536
Asian	1.110	1.350	1.159	1.444	1.229	1.236
Black	0.963	1.104	1.037	1.337	1.084	1.090
Other	0.962	1.407	1.011	1.539	1.758	1.789

Note: 1.000 = EHS prevalence for white respondents. The population of inference for the census consists of all usual residents, including residents of communal establishments. The population of inference for survey estimates consists only of residents of private households.

Source: Author's own calculations. Adapted from 2011 tables LC3206EW, DC3205EWr, DC3302EW; UK Department for Communities and Local Government 2013; Home Office & TNS-BMRB 2012.

phrasing is identical across surveys and the 2011 census, although the census again stands out due to its short, pen-and-paper questionnaire, as well as its particular mode of contact. The housing and crime surveys do not seem to differ significantly from one another. In order to check for any potential questionnaire effects by ethnic group, prevalence levels are indirectly standardised for all three sources across the 16 to 64 years age group (see Figures 4.7 (b) and (c)). Census returns disaggregated by age and ethnic groups cannot be obtained for residents of private households only; unlike with age schedules, with are for residents of privates households only, standardised prevalence ratios reported for the census may not be compared rigorously with surveys. 95 per cent confidence intervals suggest that census estimates are somewhat higher than survey estimates, which is consistent with the pattern seen on the age curves, even in the absence of communal establishments. Significant differences are only found amongst respondents identifying to 'other' ethnicities, a very small demographic making up approximately 1.1 per cent of the English population in 2011. SRH thus seems to constitute an appropriate benchmark at the national level considering the uniform phrasing. This suggests that differences previously observed for longstanding illness and activity limitations are not so much attributable to the sampling design or differences in nonresponse bias. Such differences may thus be attributable to residual confounders and non-sampling error; question phrasing and questionnaire ordering.

Figure 4.8 | Morbidity schedules for (a) longstanding illness and (b) limiting illness in the LFS and the LOS in Great Britain in 2011 with 95% confidence intervals



Source: Author's own calculations. Adapted from ONS. Social Survey Division, Northern Ireland Statistics and Research Agency. Central Survey Unit 2015, 2014a,b,c; ONS. Social Survey Division 2012.

We also compared LOS (LOS), LOS (IHS), and LFS estimates at GB level. Since census data is yet to be released in Scotland, census estimates cannot yet be compared at this geographical coverage. Once again, LFS longstanding illness estimates were found to be higher (see Figure 4.8). LOS and IHS phrasings, on the other hand, returned comparable estimates. As for limiting illness, it appears that they do differ substantially on the other hand. Only the detailed LOS, since it makes very explicit what type of difficulties may be regarded as 'limitations' in day-to-day activities, yields much higher estimates, especially for elderly people. This is probably reinforced by the fact that it is asked at the end of a long survey focussing on health.

In conclusion, this exploratory analysis reveals acute differences in morbidity reports depending on questionnaire routines and wordings. Comparisons with LOS and census estimates make it possible to explore with more certainty differences that can be explained by non-sampling error, mode effect, and detailed prompts. It appears that answers are very unstable, and that a large part of the discrepancy comes from questionnaires themselves, and vary highly across demographic cross-sections. In the context of the census review, these results do not argue much in favour of reiterating the IHS strategy for the most unstable question, although it may prove more suitable to more objective and specific measurements. The high instability in such morbidity proxies, although it remains mild at the aggregate level, should remain a substantial concern when dealing with demographic cross-classifications.

4.3.5 Auxiliary individual-level data

Once a valid model has been estimated using a first source of epidemiological evidence, predictions can be derived by applying it to the population at risk. To this end, auxiliary data on the detailed population headcount broken down by all characteristics included in the model is needed for every geographical area making up the target population. Issues may arise in estimating auxiliary headcounts in the absence of a 'closed' population, especially with high degrees of spatial and demographic disaggregation. This section reviews the main sources of population data in England and Wales, namely decennial censuses and MYPEs in intercensal periods.

Decennial censuses have traditionally referred to the population on 'Census night' by convention (27 March 2011 in the case of the latest census). Since 2001, they have covered all 'usual UK residents', defined as individuals who *'is in the UK and has stayed or intends to stay in the UK for a period of 12 months or more, or had a permanent UK address and was outside the UK and intended to be outside the UK for less than 12 months'* (ONS, 2009, p. 6). Unlike in 2001, special provision was made to

collect information on 'short-term residents' on forms, and ensure they would not count towards usual residents. In addition, enumerators were given explicit instructions with regards to residents with addresses in communal establishments; *'Students, and children at boarding school, should be counted as usually resident at their term-time address. They should also be counted as usually resident at their permanent/family address (if different), but only limited information will be collected'* (ONS, 2009, p. 7), and *'If a person has already spent or expects to spend six months or more in a communal establishment, for example, a care home, hospital or hostel, then their usual residence is that communal establishment. Otherwise usual residence would be at the UK home address and the person should be classified as a visitor at the communal establishment'* (ONS, 2009, p. 8). In practice, the ONS proceeded to removing multiple returns; when students were registered in a student hall, their term-time address was kept as the reference one insofar as possible, thanks to special enumeration procedures in student halls and overcount adjustment (ONS, 2012c).

Aside from decennial UK censuses, MYPEs are currently released every year in June in England and Wales. These are the official population headcounts published yearly by age and quinary sex bands for LADs, CCGs and super output areas, including MSOAs, and they serve as the reference estimates for the exposed-to-risk in more epidemiological work. Although single year of age tables are also published for LADs, they are not yet published for MSOAs owing to concerns regarding their precision. They are nevertheless obtainable on more officious request (Paul Norman, personal communication, 10 September 2013). It is likely that these will become more reliable in the future as secondary administrative data use increases in the future (ONS, 2014f). In addition, the ONS has undertaken work to develop MYPEs for ethnicity groups broken down by age at the district level (ONS, 2011c, 2012d). Experimental statistics were published until 2009 and are currently undergoing quality testing. Other series include population estimates by marital status and very old age population estimates. These do not offer enough disaggregation to be of interest in this research.

MYPEs are relevant to this thesis since they are regarded the auxiliary data usable to carry out SAE, or to use as the denominator in a prevalence proportion. Although census data (relative to census day population) is used to perform quality analyses, in reality MYPEs should be used when possible for all other prediction tasks. This restricts the types of models that will be examined here; in this thesis, we claim that single year of age MYPEs and ethnicity MYPEs have or will in the near future meet sufficient standards to be considered as auxiliary data for small estimation. As a result, models considered here will be restricted to the following

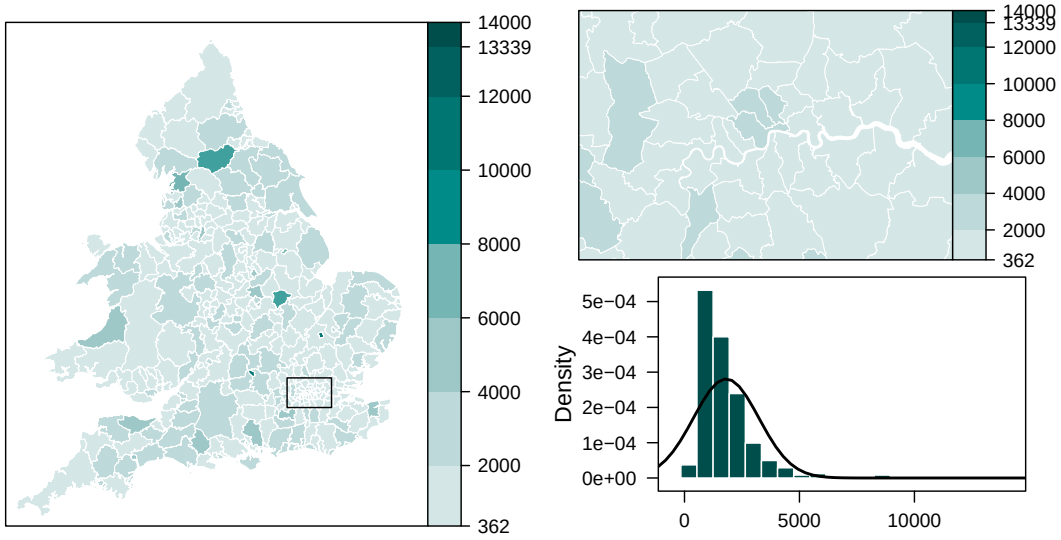


Figure 4.9 | Map of residents of communal establishments (per 100,000 usual residents) by merged LAD in England and Wales in 2011 (source: 2011 census tables DC2101EW, DC2117EWla)

demographic cross-classifications at the individual level—which is not incompatible with the addition of area-level information:

- Sex × single year of age (2 × 86)
- Sex × ethnicity × age bands (2 × 5 × 3)

There is hope that ethnicity population estimates will improve in the future, notably through an increase in the age disaggregation, and publication for super output geographies. Although MYPEs are relevant for SAE, they exhibit a major limitation, in that they do not provide a breakdown of the population residing in private households or in communal establishments.

This can constitute a problem when predictions can only be made for residents of private households, because model parameters can only be estimated for this more limited population of study. This is often the case since residents of communal establishments lie outside of the perimeter of study of almost all social surveys. Whilst residents of communal establishments represented a moderate 1.8 per cent of the total population residing in England and Wales in 2011, this is not a negligible problem if we consider certain areas, such as Oxford, Cambridge, Richmondshire and Rutland, where residents of communal establishments represent close to, and sometimes over, 10 per cent of all usual residents (see Figure 4.9).

The problem of approximating the relevant population size for predictions is not a light one. The decennial census is one of the only sources of information about communal establishments in England and Wales, since MYPEs do not differentiate

between either types of residents. Furthermore, there are complications in ascertaining how certain users should be categorised; this is famously the case with students who have both a term-time address in halls and a home/main address. In practice this means a proportion of census returns refer to an address in halls, whereas in social surveys data would still be obtained about the student either in person or through a proxy.

We further expose the issue of basing predictions on all usual residents with models only valid for residents of private households by examining reports in activity limitation caused by a long-term health issue. It there is no immediate evidence as to whether these categories have very different patterns of morbidity. Previous research (Commander et al., 1997) and data from the 2001 census have tended to suggest that overall, residents of communal establishment exhibited poorer health than residents of private households (Bajekal et al., 2006). Recent evidence by Jonker et al. (2013b) showed that the presence of nurse homes skewed local healthy life expectancy estimates rather substantially. We present indirectly standardised prevalence ratios for strong activity limitation in the 2011 census for each subgroup in Table 4.10. These are to be interpreted by reference to the population of residents of private households living in England and Wales (using age-specific prevalence proportions from table DC3302EW).

Residents are predominantly either young (mostly students) or aged over 75 years (retirees). This creates strong heterogeneity within this group. The age profile of staff and owners is more similar to the profile of private households residents, aside from two categories: there are evidently none below working age, and the population above 85 is rather important, representing 11 per cent of all staff and owners compared to only 2 per cent of residents of private households. As for their relatives, over 56 per cent of them are aged 75 and above.

In addition to differences in the age distributions across all four categories of usual residents, figures suggest departures in morbidity schedules by age. Despite in the absence of detailed age and sex-specific health information regarding staff and their relatives, we can infer that the population living in communal establishments has very different patterns of morbidity compared to residents of private households. This is unlikely to be attributable to differences in the age distribution; the indirect age-standardisation conducted in Table 4.10, based on eight age groups, does suggest a very low prevalence of strong activity limitations amongst all three categories of residents of communal establishments, especially amongst staff and their relatives. It would thus be incorrect to use social surveys to predict the health status of such populations, because it is very different.

Table 4.10 | Residents of private households and communal establishment by age and position in England and Wales in 2011: Headcounts and prevalence of strong activity limitations indirectly standardised by age and sex (source: 2011 census tables DC3302EW, DC3402EWLA, DC4107EWla)

	Private households	**Communal establishments**			
			Other		
	Residents	Residents	Staff and owner	Family or partner	Total other
Observed cases of strong activity limitation	4,444,433	300,205	6,529	18,545	25,074
At risk					
Total	55,071,113	936,994	35,597	32,208	67,805
0 to 15 years	10,537,963	38,365	0	2,804	2,804
16 to 24 years	6,253,980	396,750	5,210	2,696	7,906
25 to 34 years	7,430,851	80,485	7,220	1,968	9,188
35 to 49 years	11,857,419	65,208	7,074	2,075	9,149
50 to 64 years	10,105,262	47,509	7,657	2,343	10,000
65 to 74 years	4,811,995	36,793	1,916	2,129	4,045
75 to 84 years	3,013,263	93,706	2,533	6,050	8,583
85 years and over	1,060,380	178,178	3,987	12,143	16,130
Indirectly standardised morbidity ratio	1.000	2.130	1.445	2.145	1.905
χ^2 statistic		180038.8	896.3	11329.8	10775.6
p-value		<0.001	<0.001	<0.001	<0.001

These difficulties exposed, it is still possible in many cases to approximate the number of residents of private households exposed-to-risk provided a few assumptions are made. In the case of estimating the exposed-to-risk residing in private households by sex and single year of age, it is possible to use the general table DC4107EWla providing usual residents headcounts for all merged LADs, and subtracting residents of communal establishments recoded in table DC4107EWla assuming a uniform distribution of ages within intervals. This is only an approximate method. In the case of the exposed-to-risk by ethnicity (5 groups) and age bands (8 groups), residents for communal establishments can be subtracted from table DC2101EW based on information reported in table DC2117EWla. As for staff and their families, although a breakdown by ethnicity is provided in this table, it is not disaggregated by age group. The only option is to apportion data from table DC4107EWla by assuming that the ethnic profile is identical across age groups. Depending on the situation, it may be preferable to assume the model estimated for private households applies to staff and their relatives in communal establishments.

In contrast, with MSOAs there is currently no data source likely to produce an estimate of the exposed-to-risk for private household residents exclusively, without committing to stronger assumptions. Table DC1104EW provides a breakdown of residents of communal establishments by sex and quaternary age bands. Assuming, once again, a uniform distribution of ages within these age bands, we may subtract these residents from all usual residents recorded in table DC2101EW. As for the second type of cross-classification, with an ethnic breakdown.

Table 4.11 | Summary of options to approximate exposed-to-risk for residents of private households

Structure	Merged LADs	MSOAs
Sex × single year of age	Subtract CE residents appearing in DC1104EW (22 groups) from DC1117EW	Subtract CE residents appearing in DC1104EW (22 groups) from DC1117EW
Sex × age bands × 5 ethnic groups	Subtract CE residents appearing in DC2117EWla (8 age groups, assuming uniform distribution of ages) from DC2101EW (22 age bands)	For each ethnic group, approximate CE residents disaggregated by sex and age appearing in DC1104EW by apportionment assuming the ethnic breakdown measured in the local authority for 8 age groups (DC2117EWla) applies to age groups within the MSOA under consideration. Subtract this approximation from all residents in DC2101EW.

Chapter 5

Estimating Small Area Health Status Indicators Using Administrative Covariates: An Evaluation of Model-Based Predictors Against the 2011 UK Census

Abstract

Background: Model-based estimation is often cited among potential avenues to reduce reliance on traditional censuses as the main mode of production of small area social statistics. This presupposes the existence of reliable working models, which has not yet been demonstrated for health status indicators. *Methods:* We predict the prevalence of self-rated health and long-term limiting illness across English local authority districts (LADs) and middle layer super output areas (MSOAs). We examine the reliability of the synthetic (SYN) and empirical plug-in (EPP) predictors computed using mid-year population estimates, local rates of emergency hospitalisations and 2009–2011 English Housing Survey data (n = 151,000). Internal cross-validation and external comparison against 2011 UK census estimates provide empirical measures of error and bias. *Results:* Simple models only predict 40 to 60 per cent of the between-area variance in odds of poor health. At the LAD level, EPP reduces the average coefficient of variation (CV) by 16 to 38 per cent at the cost of an increase by 16 to 21 per cent in the average relative deviation, an indicator of bias. At the MSOA level, both methods of prediction perform similarly, cutting the average CV by over 70 per cent but with a large increase in the average bias.

Keywords: Small area estimation, Population health indicators, Behavioural risk factor surveillance, Health status, Administrative data, Hospital data, Synthetic estimation, Empirical best prediction.

5.1 Introduction

Subnational information on population health statuses is in high demand in order to help shape public services (Luck et al., 2006; Shah et al., 2014). Although such information can be derived from social surveys, censuses currently provide the necessary statistical power when high level of disaggregation for small demographic and geographical cross-classifications are required. Yet, the rising cost of survey and census fieldwork is leading to spending reviews at national statistical institutes (US Census Bureau, 2015; UK House of Commons. Treasury Committee, 2008) and undermines the case for traditional decennial censuses with long forms and full enumeration. This is forcing statistical institutes to turn to more efficient methodology to produce population health indicators.

Model-based SAE has been proposed as an alternative to traditional censuses. It has the potential to enhance the statistical power of existing household surveys by borrowing strength from auxiliary information available for the entire population. A linking model can be developed with the help of existing household survey data to predict local characteristics of interest, for instance disability or healthy behaviours. Resulting estimates may be more precise than direct survey estimates, provided that the linking model accurately predicts differences across places and demographics. This approach presupposes (a) the availability of relevant auxiliary data and (b) the validity of linking models and resulting small area estimates.

The first condition is now often reasonable thanks to the increasing availability of big data, whether from administrative or private sources (Marchetti et al., 2015; ONS, 2014c). Under suitable legal arrangements, data from records on health insurance, government benefits, primary and secondary care utilisation, death registrations can be shared with statistical offices in a form sufficiently disaggregated in terms of small geographical areas and demographic cross-classifications (gender, age, ethnic group), providing information with coverage across the entire population of study.

The second condition is difficult to ascertain on the sole basis of results from statistical analyses. Previous investigations have emphasised difficulties caused by ab-

sence of an external, reliable benchmark to validate model-based SAE against (Scarborough et al., 2009). Some investigations rely on simulation studies to provide reassurance around the suitability of models (Barker and Thompson, 2013; Raghunathan et al., 2007).

Many studies have already established feasibility of model-based estimation of health status using detailed socioeconomic auxiliary information from decennial censuses. Our aim is to assess the practical capabilities of SAE when such auxiliary information is not available from a high-precision source such as a census. This paper evaluates both of the above conditions in relation to producing univariate health status characteristics which have been available for small areas in England on a decennial basis since 1991 from the UK census: SRH and LLTI. It tests the feasibility of using model-based estimation exclusively with auxiliary information available publicly from hospitals and MYPEs. It also tests the validity of linking models by conducting both internal validation and external validation through a comparison with reliable 2011 UK census estimates of SRH and LLTI across 324 English LADs and 6,676 MSOAs (MSOAs, census geographical units averaging 7,800 residents).

This paper begins with a review of existing small area population health models and methods to scrutinise the quality of resulting estimates. A subsequent section presents our statistical design and procedures for assessing the suitability of predictions using both internal validation (against survey data) and external validation (against census estimates). Two sections then present results of this analysis, and discuss their implications for the prediction of univariate and multivariate population health statistics using simple model-based techniques.

5.2 Model-based estimation of small area population health indicators

5.2.1 Small area models

Model-based small area prevalence estimation is characterised by two main approaches. On the one hand, synthetic estimation relies on simple generalised linear models and does not require one to observe data for all small areas of interest; in many cases, it has been possible to use synthetic estimation without even accessing respondents' geographical location. This may explain why it has remained more popular with public health applications, at least in the UK (Pickett et al., 2000;

Scholes et al., 2008; Twigg et al., 2000). Synthetic estimation may be very efficient provided that an excellent working model can be trusted to hold for the entire population. This presupposes that the model can accurately predict variations in prevalence across all areas. This can be the case when indicators from administrative sources are available which predict a large amount of the between-area variance: for instance, benefit claimant counts have been used to increase the precision of small area unemployment estimates (Ambler et al., 2001).

On the other hand, EB and HB prediction have established themselves as efficient solutions to many SAE problems (Ghosh and Rao, 1994) and relax the assumption that fixed effects can accurately predict all between-area variation. Prevalence proportions can be estimated using a logistic mixed model (Farrell et al., 1997b). This is typically a data intensive approach, requiring detailed and sensitive survey data on all areas of interest. Applications to health statistics have so far remained limited (see chapter 3). Internationally, recent examples include the post-censal estimation of disability rates (Bizier et al., 2009; Fabrizi et al., 2016; You et al., 2014), the estimation of obesity prevalence across US ZIP codes (Drewnowski et al., 2007), or mapping rates of asthma across regions of British Columbia (You and Zhou, 2011). Applications of EB and HB prediction remain more common outside of health statistics; for instance statistics on economic activity (see for instance Curtis 2003; Molina et al. 2007).

5.2.2 Quality standards and validation

Statistical estimators are normally assessed for their efficiency, that is minimising their sampling error. Statistical reliability guidelines are common, often as CV thresholds. The UK's ONS generally regards estimates with a CV above 20 per cent as unreliable. In the context of the Census Transformation Programme, it has set a minimum quality standard of a CV of up to 20 per cent for attributes of populations representing 3 per cent of the MSOA population (ONS, 2013f). This is currently equivalent to an average headcount of 234 residents. The US Census Bureau aims for a median CV of up to 30 per cent (US Census Bureau, 2013) while the National Center for Health Statistics recommend suppression of estimates with a CV higher than 30 per cent (Klein et al., 2002). Yet quality assurance procedures are important to demonstrate the suitability of model-based SAE, in which all inferences depend on the working model's assumptions and specifications being true. In practice, confidence in the working model is difficult to establish and analysts tend to settle for the 'best' model available, using traditional methods of model selection and evaluation

relevant to SAE—mainly deviance measures such as Bayesian and Aikaike information criteria, and their extensions through fence methods (Jiang et al., 2008). There are two main types of additional validation: internal and external.

Internal validation is carried out as part of most investigations in a variety of ways. Model-based estimates can be compared against direct survey estimates in order to detect bias (Brown et al., 2001). When those are not available at a high geographical resolution, for instance to preserve respondent privacy, it is possible to add small area estimates up to a higher level of geography (for instance regions) and conduct the comparison at this higher level, especially when direct survey estimates are deemed reliable (Brown et al., 2001; Curtis, 2003; Twigg et al., 2000). In the case of synthetic multilevel modelling, other practical criteria have been mentioned in the literature. Pickering et al. (2004, p. 56) have suggested that the working model's share of between-area variance explained by the fixed effect predictors should not be below an 'absolute minimum' of 40 per cent. This is on the ground that any remaining between-area variance is typically discarded in synthetic estimation. This results in synthetic estimates being unreasonably close to the national average unless residual between-area variance has been reduced to negligible levels. The justification for recommending the specific cut-off of 40 per cent remains unclear. Scarborough et al. (2009) and Drewnowski et al. (2007) discuss an additional requirement that spatial autocorrelation diagnostics are run on random effect model residuals. Scarborough et al. (2009) regards a Moran's I statistic greater than 0.1 as evidence of model invalidity.

External validation through predictive checks (Gelman et al., 2000; Marshall and Spiegelhalter, 2003), sometimes known as 'leave-one-out' cross-validation, is another common method to verify the specification of models. Splitting a data set into one part to estimate the working model, and another part to validate its predictive power gives information on variable selection and wider model assumptions, for instance parametrisation. When it comes to SAE, feasibility of this method has been investigated by Skinner (2007), showing that it is a valid approach to model selection. However, the practical meaning of cross-validatory estimators is less evident when it comes to validating an entire set of predictions for end users. In contrast, external validation on a full independent dataset and calculation of average deviations is, at least in principle, more transparent. Yet, it is rarely possible, because it requires availability of reliable estimates of the parameters of interest. If such estimates are available then there may not be any need to develop a model-based estimation protocol. This may be a detrimental gap in previous research in the sense that validation against an independent trusted source is conceptually more attractive to practition-

ers.

This paper pursues two main aims: (a) to attempt to estimate standard population health indicators without using detailed socioeconomic information normally available from a census; and (b) to conduct empirical validation of a specific model against a trusted source, in this case census estimates. A recurrent criticism of model-based SAE has been the perceived lack of transparency and accessibility of the assumptions underpinning predictive models, especially to the general public (Rees, 2013). External validation is sometimes regarded as conceptually more appealing to a non-specialist audience. The census being a conventional and trusted source, using it as a benchmark to conduct external validation is not only straightforward; it is also less dependent on somewhat arbitrary testing assumptions and scenarios usually presented in simulation studies. The model-based estimation methods used in this paper are well established, but the emphasis of this work is to provide a practical, practitioner-oriented illustration of the precision compared to basic design-based estimates. External validation against a comprehensive and reliable source also gives the opportunity to look beyond MSE and to appraise the performance of the model in a way that is easily conveyed to non specialists, for instance examining the ability of models to rank LADs correctly by prevalence.

5.3 Methods

5.3.1 Statistical design

We predict the prevalence of LLTI and bad/very bad SRH across LADs and MSOAs in England using the 2009–2011 EHS.

We consider a population U of individuals $i = \{1, \ldots, N\}$ partitioned in demographic subgroups j and areas d with headcounts $\mathbf{N} = \{N_{11}, \ldots, N_{jd}\}$. Data from the EHS is regarded as a random sample s drawn from U achieving sizes $\mathbf{n} = \{n_{11}, \ldots, n_{jd}\}$. It is used to model the prevalence p_d of poor health for every area d making up the population as a function of characteristics of this population. We use the following binomial logistic mixed model:

$$\begin{aligned} y_{jd} &\sim \text{Bin}(n_{jd}, p_{jd}) \\ \eta_{jd} &\equiv \log\left(\frac{y_{jd}}{n_{jd} - y_{jd}}\right) = \mathbf{x}_{jd}\boldsymbol{\beta} + \upsilon_d \end{aligned} \tag{5.1}$$

where y_{jd} is the total numbers of individuals reporting a bad health status (LLTI or bad/very bad SRH); η_{jd} the linear predictor of the log-odds of reporting a bad health status; $\boldsymbol{\beta}$ a q-dimensional column vector of fixed effect coefficients; $\mathbf{x}_{jd}$ the appropriate row of a $JD \times q$ matrix of covariates $\mathbf{X}$ capturing characteristics of demographic subgroup j, and υ_d a realisation of a normal distribution of mean 0 and variance σ^2. All model parameters are estimated using Laplace approximation using software package `lme4` (R Core Team, 2014; Bates et al., 2015). A similar design has previously been described as a 'modified Fay-Herriot method' (Ambler et al., 2001; Silva and Clarke, 2008).

The above model (5.1) comes in two different specifications with regard to covariate matrix $\mathbf{X}$ which are designed to reflect the shape of auxiliary data taken from census tables:

- Model **M1**: p_{jd} is modelled as a function of age (centring required to ease model fitting), its squared transformation, LAD area classifications and area-level indirectly standardised emergency hospital admissions rates (ISARs). In this case, $J = 86$. All persons aged 85 years or more are assumed to be 87.5, as it was found to be the average age for this group.
- Model **M2**: p_{jd} is modelled as a function of age bands (nine binary variables 0–15; 16–24; 25–34; 35–49;50–64; 65–74; 75–84; 85+), ethnicity (five census categories White; Mixed; Black; Asian; Other) and area-level ISARs. In this case, $J = 5 \times 5 = 25$.

Covariate selection was guided by the existing literature and in-depth testing reported in chapter 6. This in particular informed the choice of ISARs as the key predictor over competing options such as age-standardised mortality rates.

We estimate area-level prevalences p_d by aggregation of predictors θ_{jd} thanks to auxiliary population data; $\theta_d = \sum_j N_d^{-1} N_{jd} \theta_{jd}$ of p_{jd} within each area. The synthetic estimator θ_{jd}^{SYN} is computed by applying the reverse logit transformation to model fixed effects:

$$\theta_{jd}^{SYN}(\mathbf{X}, \hat{\boldsymbol{\beta}}) = \text{logit}^{-1}(\mathbf{x}_{jd}\boldsymbol{\beta}) \tag{5.2}$$

In a frequentist framework, the best predictor (BP) is defined as the combination of the direct estimator of $\hat{p}_{jd} = y_{jd}/n_{jd}$ for sampled units and the model-based expectation of p_{jd} for unobserved unit (Jiang and Lahiri, 2001):

$$\theta_{jd}^{BP}(\mathbf{X}, \boldsymbol{\beta}, \sigma^2) = N_{jd}^{-1} n_{jd} \hat{p}_{jd} + N_{jd}^{-1}(N_{jd} - n_{jd}) E(p_{jd} | \mathbf{X}, \boldsymbol{\beta}, \sigma^2) \tag{5.3}$$

where the expectation of $\hat{p}_{jd}$ can be evaluated by Monte Carlo integration:

$$\mathrm{E}_M(p_{jd}|\mathbf{X},\boldsymbol{\beta},\sigma^2) = \exp(\mathbf{x}_{jd}\boldsymbol{\beta})\frac{\mathrm{E}_\xi\left[(y_{jd}+1)\sigma\xi - (n_{jd}+1)\log(1+\exp(\mathbf{x}_{jd}\boldsymbol{\beta}+\sigma\xi))\right]}{\mathrm{E}_\xi\left[y_{jd}\sigma\xi - n_{jd}\log(1+\exp(\mathbf{x}_{jd}\boldsymbol{\beta}+\sigma\xi))\right]} \quad (5.4)$$

where the expected value $E_\xi(\cdot)$ is evaluated by reference to $\xi \sim \mathrm{N}(0,1)$. The BP is usually approximated by the EB predictor, by using parameters $(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ in lieu of $(\boldsymbol{\beta}, \sigma^2)$. Yet, Jiang and Lahiri (2001) have shown that a simpler EPP converges towards the EB predictor at a rate of $O(N^{-1})$.

$$\begin{aligned}\theta_{jd}^{EPP}(\mathbf{X},\hat{\boldsymbol{\beta}},\hat{\sigma}^2) &= N_{jd}^{-1}n_{jd}\hat{p}_{jd} + N_{jd}^{-1}(N_{jd}-n_{jd})\mathrm{logit}^{-1}(\hat{\eta}_{jd}) \\ &\simeq \mathrm{logit}^{-1}(\hat{\eta}_{jd})\end{aligned} \quad (5.5)$$

This predictor assumes that a reliable estimate of $\hat{v}_d$ is available, dependent on the size of sample s in cross-classification jd. We approximate the EPP by bearing in mind that $N_{jd}^{-1}n_{jd}$ is very close to zero and the first term is therefore negligible. Prediction intervals for $\boldsymbol{\theta}^{EPP}$ are constructed by computing upper and lower bounds:

$$\mathrm{logit}^{-1}\left(\hat{\boldsymbol{\eta}} \pm z_{\alpha/2}\sqrt{\widehat{\mathbf{MSE}}}\right) \quad (5.6)$$

where $z_{\alpha/2}$ is the appropriate critical value of the standard Normal distribution for a given probability of type I error α. By convention we set $\alpha = 0.05$.

MSE is estimated using a simple model-based parametric bootstrap, described in Algorithm 1, with $B = 1{,}000$ iterations. This method approximates the prediction error by measuring the empirical mean squared deviation in a population $U^\star$ simulated from the estimated working model. In this approach, random component σ is approximated using estimates $\hat{\sigma}$, which introduces further error, especially as simulation studies have highlighted the poor accuracy of random components estimators for generalised linear mixed models (Carlin and Gelfand, 1990). It has been suggested that naive estimators of MSE have undercoverage of up to order $O^{(n_d-1)}$. Yet this approach is computationally less demanding compared to alternatives such as the double bootstrap (Hall and Maiti, 2006), or the 'small area wild bootstrap', an adaptation of the finite sample bootstrap in a model-based framework proposed by González-Manteiga et al. (2007).

Over B iterations:

1. Simulate random effects $\boldsymbol{\upsilon}^{\star}$ from $\mathrm{N}(0, \hat{\sigma}^2)$
2. Predict random realisations $\boldsymbol{\eta}^{\star}$ using $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\upsilon}}$ (estimated from s) and predictors $\mathbf{X}$
3. Fit a new model on $\boldsymbol{\eta}^{\star}$, to estimate $\boldsymbol{\beta}^{\star\star}$ and $\boldsymbol{\upsilon}^{\star\star}$
4. Predict $\boldsymbol{\eta}^{\star\star}$ using $\boldsymbol{\beta}^{\star\star}$, $\boldsymbol{\upsilon}^{\star\star}$ and predictors $\mathbf{X}$ for the population U.
5. Calculate squared error $(\boldsymbol{\eta}^{\star\star} - \boldsymbol{\eta}^{\star})^2$

Over many iterations, estimate $\widehat{\mathrm{MSE}}_{jd}^{boot} = B^{-1} \sum_{b=1}^{B} \left(\eta_{jd}^{\star\star} - \eta_{jd}^{\star}\right)^2$.

Algorithm 1 | Procedure for the model-based parametric bootstrap

5.3.2 Data sources

The EHS (UK Department for Communities and Local Government, 2015) is a multistage household survey of residents of English private households approached through a systematic random sample of addresses drawn from the Royal Mail's Small User Postal Address File. It is the largest survey available collecting census health indicators. The pooled April 2009/March 2010 and April 2010/March 2011 editions total a sample size of 151,734 valid SRH cases and 151,708 LLTI cases, equivalent to a 0.029 per cent sampling fraction for residents of private households in England. Researchers can access its full version with postcode information under a secure access licence, allowing to link it with area-level characteristics: LAD classifications (ONS, 2010a) and ISARs from Public Health England (2014b).

Census estimates of bad/very bad SRH and LLTI prevalence for residents of households are extracted from table DC3302EW and used as a point of comparison.

Information on the age and ethnicity make-up of all LADs and MSOAs in England would normally be obtained from MYPEs. Because annual population estimates are for all usual residents, including residents of communal establishments (CEs) excluded from the EHS population of study, 2011 census tables (ONS, 2013b) were instead used to reconstruct the 2011 MYPEs of residents of private households disaggregated by age and ethnicity. This required combining information from three census tables (see appendix, p. 277).

5.3.3 Validation

We conduct both internal and external validation of the resulting estimates. We use the EHS under a special licence granting secure access to the geographical location of households, allowing us to compute direct survey estimates for small areas

and to compare them with our model-based estimates. We also validate these estimates against 2011 UK census data at LAD and MSOA level. In 2011, the UK census and the EHS included identically-worded SRH questions, allowing direct comparison. In contrast, the LLTI questions differed somewhat and estimates are likely to have a lower level of agreement (see Taylor et al. 2014).

5.3.3.1 External validation (empirical discrepancy metrics)

Outside a simulation study, and in the absence of straightforward estimators of bias and variance of model-based estimates, comparison with external data is sometimes used. In that case, indicators error and bias have been approximated by averaging performance metrics over all areas rather than over repeated sampling (see for instance Jia et al. 2004; Zhang et al. 2014).

We perform external validation against the 2011 censuson the proportion (relative frequency) scale rather than the odds scale, because it is the most relevant to practitioners when designing surveys and allows an easy comparison with direct survey estimates $\hat{\mathbf{p}}$. External validation is performed by comparing model-based predictors $\theta_d(\mathbf{x}, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ with census 2011 estimates $\hat{\mathbf{p}}^c = \hat{p}^c_{1\leq d\leq D}$ over the D areas. We use the following 'empirical' metrics:

- the empirical Pearson's r coefficient of correlation between predictor $\theta_d(\mathbf{x}, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ and census estimate p^c_d
- the empirical Average Relative Deviation (ARD) is an empirical measure of average bias

$$\text{ARD} = D^{-1} \sum_{d=1}^{D} \frac{\theta_d(\boldsymbol{X}, \hat{\sigma}^2, \hat{\boldsymbol{\beta}}) - \hat{p}^c_d}{\hat{p}^c_d} \tag{5.7}$$

- the empirical Average Squared Deviation (ASD) is an empirical measure of overall error (bias and variance) directly comparable to the average MSE

$$\text{ASD} = D^{-1} \sum_{d=1}^{D} \left[\theta_d(\boldsymbol{X}, \hat{\sigma}^2, \hat{\boldsymbol{\beta}}) - \hat{p}^c_d\right]^2 \tag{5.8}$$

- the empirical Average Coeffient of Variation (eACV) or 'relative standard error' expresses the empirical ASD as a factor of the parameter to be estimated:

$$\text{eACV} = D^{-1} \sum_{d=1}^{D} \left(\sqrt{\left[\theta_d(\boldsymbol{X}, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2) - \hat{p}_d^c \right]^2} / \hat{p}_d^c \right) \tag{5.9}$$

- the proportion of areas allocated to the correct tenth: assuming 2011 census are exact, we count the number of areas for which the predicted rank out of D falls into the same tenth.

We also measure global spatial autocorrelation in the difference between predictions and census estimates, which gives an indication of the extent to which the deviation between predictions and census estimates clusters in space. We use the Moran's I statistic and the random permutation significance test (Cliff and Ord, 1981) on the following fit metric:

$$\frac{\hat{p}_d^c - \theta_d(\mathbf{x}, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2)}{\theta_d(\mathbf{x}, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2)} \tag{5.10}$$

5.3.3.2 Internal validation

It would be unreasonable to regard census estimates as absolute 'true values'. As with any estimate, census estimates are subject to their own sampling error as well as to differential nonresponse bias across areas. More importantly, the EHS's sampling characteristics and interview mode effect are likely to introduce structural differences in estimates, which may not be flagrant when comparing the national prevalence of poor SRH across both sources but may be when the EHS is used as input to SAE. Internal validation against the EHS itself is thus as important as external validation to test whether any bias measured in external validation can be attributed to model misspecification. We conduct two common types of internal validation via three simple tests:

- A χ^2 goodness-of-fit test with $D - 1$ degrees of freedom (discussed by Brown et al. 2001) is performed on the relative deviation between the direct EHS estimate and predicted number of persons with LLTI or poor SRH:

$$\chi^2 = \sum_{d=1}^{D} \frac{\left(\hat{p}_d - \theta_d(\mathbf{x}, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2) \right)^2}{\theta_d(\mathbf{x}, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2)} \tag{5.11}$$

  This goodness-of-fit test statistic is also used to measure the level of agreement between the census and the EHS by plugging census estimates in lieu

of $\theta_d(\mathbf{x}, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ in the above formula.

- We test for global spatial autocorrelation in area residuals $\hat{\boldsymbol{\upsilon}}$.
- We measure the proportion of variance explained by the final model. This is calculated with reference to the random effect variance estimate in a simple random means model (without covariate). The relative difference between the two variances measures the proportion of the between-area variance which is predicted by covariates in the final models.

We also estimate characteristics which can be used for internal validation even in the absence of reliable 2011 census estimates:

- $\widehat{\text{AMSE}}$, the average $\widehat{\text{MSE}}$ over D areas as estimated from the parametric bootstrap results or, in the case of the direct survey estimate, the average of the traditional estimator of standard error of a proportion over D areas.
- $\widehat{\text{ACV}}$, the corresponding average CV:

$$\widehat{\text{CV}} = D^{-1} \sum_{d=1}^{D} \widehat{\text{CV}}_d = D^{-1} \sum_{d=1}^{D} \left(\frac{\sqrt{\widehat{\text{MSE}}^{boot}}}{\theta_d(\mathbf{x}, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2)} \right) \tag{5.12}$$

- We also measure autocorrelation in the difference between predictions and EHS estimates the same way we did against census estimates (see expression 5.10) using the following fit metric:

$$\frac{\hat{p}_d - \theta_d(\mathbf{x}, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2)}{\theta_d(\mathbf{x}, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2)} \tag{5.13}$$

5.4 Results

5.4.1 Agreement between EHS and census estimates

We first examine direct survey estimates of LLTI status and very bad/bad SRH (see Table 5.1). Direct estimates have elevated levels of error as expected from the small sample size. The average coefficient of variation $\widehat{\text{ACV}}$ (estimated using sandwich estimators of standard error) is below the 20 per cent quality threshold across LADs, indicating that many of those areas are below the minimum reliability threshold. The corresponding empirical metric eACV, which is computed by reference to the census across areas, gives similar values. Goodness-of-fit tests provide significant evidence that EHS estimates do not come from the same superpopulation

Table 5.1 | External validation of direct EHS estimates against 2011 census estimates (SRH and LLTI)

	SRH		LLTI	
	LAD	MSOA	LAD	MSOA
D	324	6,766	324	6,766
Pearson's r	0.724	0.312	0.695	0.333
% correct tenth	22.5%	13.8%	21.3%	14.9%
ARD	–8.57E–02	–4.07E–02	–9.98E–02	–6.88E–02
ASD	1.77E–04	3.53E–03	1.03E–03	1.05E–02
eACV	21.2%	87.1%	14.6%	47.7%
$\widehat{\text{AMSE}}$	1.23E–04	2.53E–03	3.68E–04	6.76E–03
$\widehat{\text{ACV}}$	19.6%	68.6%	12.4%	46.2%
Mean estimate (EHS)	4.7%	5.1%	15.4%	16.0%
Mean estimate (census)	5.1%	5.3%	17.2%	17.3%
χ^2 fit with census	738	25,337	1,423	15,268
Degrees of freedom	323	6,765	323	6,765
Significance	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$
Moran's I (census)	0.321	0.259	0.301	0.273
Significance	$p=0.012$	$p=0.087$	$p=0.048$	$p<0.001$

as the census. In spite of the identical wording of the SRH question, test results indicate a statistically significant departure between EHS and census SRH estimates: this could be due to questionnaire ordering and mode effects. As for LLTI estimates, the deviation from census estimates is even more pronounced, as expected from the variation in question wording. In both cases, census estimates exceed direct EHS estimates. EHS bad/very bad SRH and LLTI estimates are, respectively, on average 4 to 9 per cent and 6 to 10 per cent below census estimates. We note some evidence of weak spatial autocorrelation in the relative deviations between EHS estimates and census (see expression 5.13). Moran's I statistics close to 0.3 could suggest that discrepancies between those two sources have a spatial pattern.

We conclude that although census data is a good theoretical comparator, we find evidence that the EHS does not stem from the exact same superpopulation model. This is arguably a common caveat with most forms of external validation and justifies carrying out internal validation in addition to pure cross-validation.

5.4.2 Model characteristics and internal validation

Model coefficients and residual autocorrelation are reported in supplementary material (see appendix, p. 257). For LADs, **M1** predicts a lower proportion of the

Table 5.2 | Internal and external validation of SRH estimates

		LAD		MSOA	
		SYN	EPP	SYN	EPP
	D	324	324	6,766	6,766
M1	Pearson's r	0.830	0.879	0.800	0.767
	ARD	−1.05E−01	−1.04E−01	−1.41E−01	−1.27E−01
	ASD	9.27E−05	7.37E−05	2.22E−04	2.39E−04
	eACV	14.3%	13.2%	20.8%	22.1%
	$\widehat{\text{AMSE}}$	–	4.44E−05	–	1.93E−04
	$\widehat{\text{ACV}}$	–	12.4%	–	25.7%
	% correct tenth	31.2%	33.6%	29.1%	26.7%
	Mean θ_d	4.5%	4.6%	4.5%	4.6%
	Mean $\hat{p}_d^c$	5.1%	5.1%	5.3%	5.3%
	χ^2 fit (internal)	1,701	417	46,886	18,978
	Degrees of freedom	323	323	6,765	6,765
	Significance	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$
	Moran's I (internal)	0.382	0.334	0.278	0.272
	Significance	$p<0.001$	$p=0.004$	$p<0.001$	$p<0.001$
	Moran's I (external)	0.535	0.419	0.755	0.755
	Significance	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$
M2	Pearson's r	0.869	0.895	0.803	0.769
	ARD	−1.59E−01	−1.58E−01	−1.90E−01	−1.78E−01
	ASD	1.14E−04	1.02E−04	2.74E−04	2.77E−04
	eACV	17.1%	16.6%	22.9%	23.7%
	$\widehat{\text{AMSE}}$	–	7.39E−05	–	1.85E−04
	$\widehat{\text{ACV}}$	–	18.7%	–	27.5%
	% correct tenth	32.7%	39.8%	28.7%	26.7%
	Mean θ_d	4.3%	4.3%	4.2%	4.3%
	Mean $\hat{p}_d^c$	5.1%	5.1%	5.3%	5.3%
	χ^2 fit (internal)	1,655	562	48,612	20,521
	Degrees of freedom	323	323	6,765	6,765
	Significance	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$
	Moran's I (internal)	0.308	0.292	0.273	0.269
	Significance	$p=0.027$	$p=0.081$	$p<0.001$	$p=0.003$
	Moran's I (external)	0.515	0.468	0.731	0.552
	Significance	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$

Note: For 25 MSOAs with a null sample size, θ^{SYN} is substituted for θ^{EPP}. Moran's I statistics are calculated with the three-nearest neighbours method.
Source: EHS, 2011 Census, Authors' own calculations.
SYN: synthetic estimator. EPP: empirical plug-in predictor.

Table 5.3 | Internal and external validation of LLTI estimates

		LAD		MSOA	
		SYN	EPP	SYN	EPP
	D	324	324	6,766	6,766
M1	Pearson's r	0.870	0.865	0.834	0.785
	ARD	−1.17E−01	−1.16E−01	−1.22E−01	−1.17E−01
	ASD	6.80E−04	6.62E−04	1.08E−03	1.18E−03
	eACV	12.3%	12.2%	14.4%	15.6%
	$\widehat{\text{AMSE}}$	–	4.11E−04	–	7.52E−04
	$\widehat{\text{ACV}}$	–	12.6%	–	16.8%
	% correct tenth	34.3%	33.6%	31.0%	27.1%
	Mean θ_d	15.1%	15.1%	15.0%	15.1%
	Mean $\hat{p}_d^c$	17.2%	17.2%	17.3%	17.3%
	χ^2 fit (internal)	1,961	389	22,886	12,136
	Degrees of freedom	323	323	6,765	6,765
	Significance	$p<0.001$	$p=0.007$	$p<0.001$	$p<0.001$
	Moran's I (internal)	0.317	0.282	0.286	0.281
	Significance	$p=0.016$	$p=0.145$	$p<0.001$	$p<0.001$
	Moran's I (external)	0.538	0.396	0.736	0.482
	Significance	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$
M2	Pearson's r	0.897	0.880	0.835	0.789
	ARD	−1.60E−01	−1.59E−01	−1.68E−01	−1.62E−01
	ASD	9.79E−04	9.77E−04	1.48E−03	1.55E−03
	eACV	16.0%	15.9%	17.9%	18.4%
	$\widehat{\text{AMSE}}$	–	3.20E−04	–	6.54E−04
	$\widehat{\text{ACV}}$	–	11.9%	–	15.4%
	% correct tenth	34.3%	33.0%	31.1%	27.2%
	Mean θ_d	14.4%	14.4%	14.2%	14.4%
	Mean $\hat{p}_d^c$	17.2%	17.2%	17.3%	17.3%
	χ^2 fit (internal)	2,176	570	26,195	13,877
	Degrees of freedom	323	323	6,765	6,765
	Significance	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$
	Moran's I (internal)	0.321	0.288	0.287	0.282
	Significance	$p=0.012$	$p=0.108$	$p<0.001$	$p<0.001$
	Moran's I (external)	0.518	0.385	0.732	0.481
	Significance	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$

Note: For 25 MSOAs with a null sample size, θ^{SYN} is substituted for θ^{EPP}. Moran's I statistics are calculated with the three-nearest neighbours method.
Source: EHS, 2011 Census, Authors' own calculations.
SYN: synthetic estimator. EPP: empirical plug-in predictor.

between-area variance (53% and 61% for LLTI and SRH respectively) than **M2** (61% and 72% respectively). The fixed part of MSOA models predicts between 37 per cent and 40 per cent of the total between-area variance across all four models. Based on this information only, one would expect **M2** to perform better. Area residuals of both models have statistically significant levels of global spatial autocorrelation (three-nearest neighbours Moran's I statistic of 0.29 and higher)—a sign that important spatial variations remain which are not captured by variations in age, ethnicity and local hospitalisation rates only. In those conditions, the models' stochastic part (random effects) is essential to obtain accurate predictions.

Internal and external prediction validation results are presented in Tables 5.2 and 5.3. The internal χ^2 goodness-of-fit statistic (against direct EHS estimates) can guide model selection by providing a single metric indicating the performance of a given model. This statistic indicates that predictors produced by model **M1** have a better fit, despite this model predicting a lower share of between-area variance. Test statistics indicate a better fit for EPPs compared to synthetic predictors (SYNs), which is to be expected since EPPs are a weighted combination of direct estimates and SYNs. In both cases, goodness-of-fit tests are rejected ($p < 0.001$). At the level of LADs, EPPs computed with model **M1** underestimate prevalence of bad/very bad SRH and LLTI by 2.1 per cent and 1.9 per cent respectively, using per cent direct estimates as a comparator. This is considerably lower than with **M2**. At the MSOA level, underestimation becomes more severe (9.8% and 5.6% respectively).

We compare the $\widehat{\text{AMSE}}$ of EPPs and direct estimates. EPPs of bad/very bad SRH computed with **M1** are again superior, with estimated $\widehat{\text{AMSE}}$ savings of 63.9 per cent and 92.4 per cent for LADs and MSOAs respectively. EPPs of LLTI achieve an estimated reduction in $\widehat{\text{AMSE}}$ by 88.8 and 88.9 per cent for LADs and MSOAs respectively. This brings the $\widehat{\text{ACV}}$ below 20 per cent in both cases.

Global autocorrelation in the deviation between direct estimates and all predictors computed from **M1** are mild (between 0.27 and 0.38) and all except one are statistically significant ($p < 0.050$).

5.4.3 External validation

Point and interval EPPs are plotted against ranks of their corresponding census estimates in Figure 5.1. Discrepancy metrics, averaged over all areas rather than over repeated sampling, provide easy-to-interpret measures similar to estimators of error and bias. These metrics are computed and reported in Tables 5.2 and 5.3.

In addition, the relative efficiency of EPP/SYN to per cent direct estimators

is computed by taking the ratio of their respective empirical discrepancy metrics in Table 5.4. We interpret these metrics in terms of improvements in fit achieved by the various predictors over the direct survey estimator. Although this approach has advantages in allowing us to analyse predictions in greater detail, caution should be exercised in that results in section 5.4.1 have demonstrated that census estimates of both SRH and LLTI are on average higher that per cent estimates. For example, we expect that measurements of ARD taken by comparison with census estimates overestimate prediction bias.

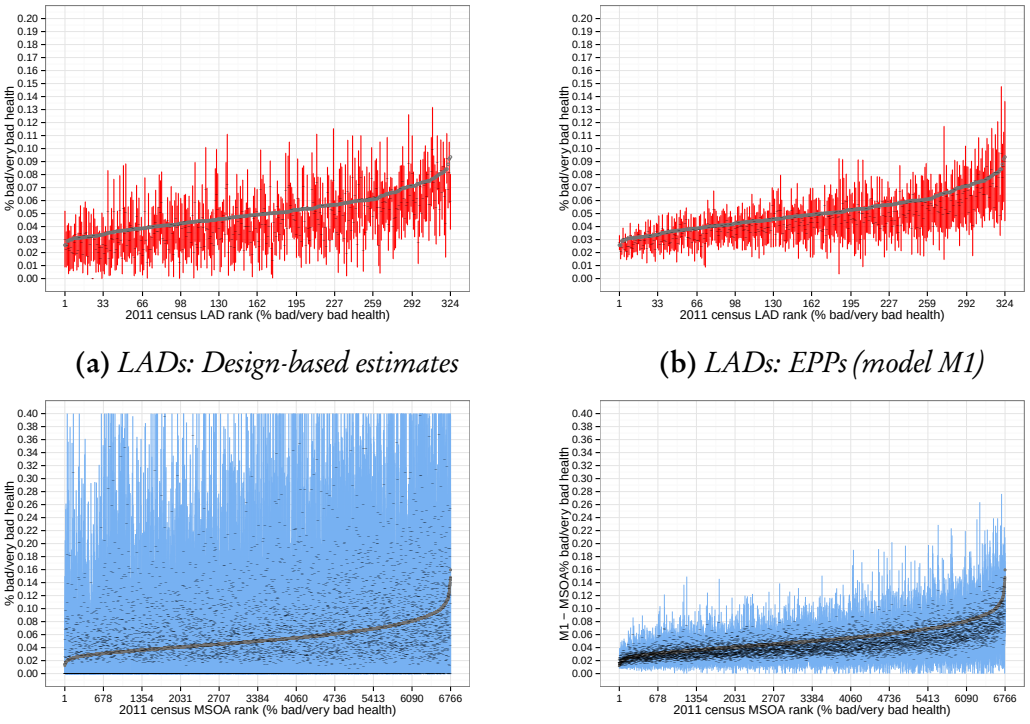
We first examine predictors computed from **M1** (single years of age):

- LAD-level: SRH predictors cut the eASD of direct per cent estimates by 48 per cent and 58 per cent for the SYN and EPP respectively. This comes at the price of an increase in bias, as evidenced by a 21–23 per cent increase in ARD. EPPs prove superior to SYN predictors and they bring LADs estimates to more acceptable levels of precision (eACV of 12% to 13%). We observe similar findings for LLTI predictions although reductions in eACV are more modest.
- MSOA-level: efficiency gains are more substantial (94% and 93% reduction in ASD for SYN and EPP respectively) at a very high cost in terms of bias; the ARD is multiplied by 2.1 (EPP) and 2.5 (SYN). EPPs are manifestly not superior to SYN predictors at the MSOA level, which is likely to be due to the smaller sample size for each area and the fact that the fixed part of MSOA models predicts a higher proportion of the total between-area variance in odds of poor health.

M2 is specified differently, using age bands instead of a polynomial of single

Table 5.4 | Relative efficiency as ratios of empirical discrepancy metrics (model-based predictions to direct estimates)

		SRH				LLTI			
		LAD		MSOA		LAD		MSOA	
		SYN	EPP	SYN	EPP	SYN	EPP	SYN	EPP
M1	ASD	0.52	0.42	0.06	0.07	0.66	0.65	0.10	0.11
	ARD	1.23	1.21	3.46	3.13	1.17	1.16	1.78	1.70
	eACV	0.67	0.62	0.24	0.25	0.84	0.84	0.30	0.33
M2	ASD	0.64	0.58	0.08	0.08	0.95	0.95	0.14	0.15
	ARD	1.85	1.84	4.67	4.37	1.60	1.59	1.78	1.70
	eACV	0.80	0.78	0.26	0.27	1.10	1.09	0.37	0.39



(a) *LADs: Design-based estimates*

(b) *LADs: EPPs (model M1)*

(c) *MSOAs: Design-based estimates*

(d) *MSOAs: EPPs (model M1)*

Figure 5.1 | Plots of bad/very bad SRH estimates against rank of census estimates

years of age, and adding fixed effects for ethnic groups. Despite predicting a higher share of the between-area variance in poor health the accuracy of its predictors is inferior:

- LAD-level: SYN and EPP reduce the empirical error or SRH estimates 36 per cent and 42 per cent respectively, but increase the average relative deviation by a factor of between 2 and 2.5. This again signals that **M2** is not suitable as the scale of underestimation is even greater than with **M1**.
- MSOA-level: gains in precision are of the same order as **M1** but a severe increase in average relative deviation by a factor greater that 3 again suggests severe misspecification in the model.

We note that MSE estimates produced with a simple parametric bootstrap are of the same order of magnitude as the empirical average deviation obtained by comparison with census estimates. Global spatial autocorrelation in the discrepancy between census and model-based estimates is high, with Moran's I statistics of up to 0.7. This strong spatial pattern indicates that the performance of model-based predictions varies considerably across places. Figures 5.2 to 5.5 present maps of the deviation between predictions and census estimates. The colour shade cut-off values highlight areas where predictors are particularly far from census estimates. Spatial

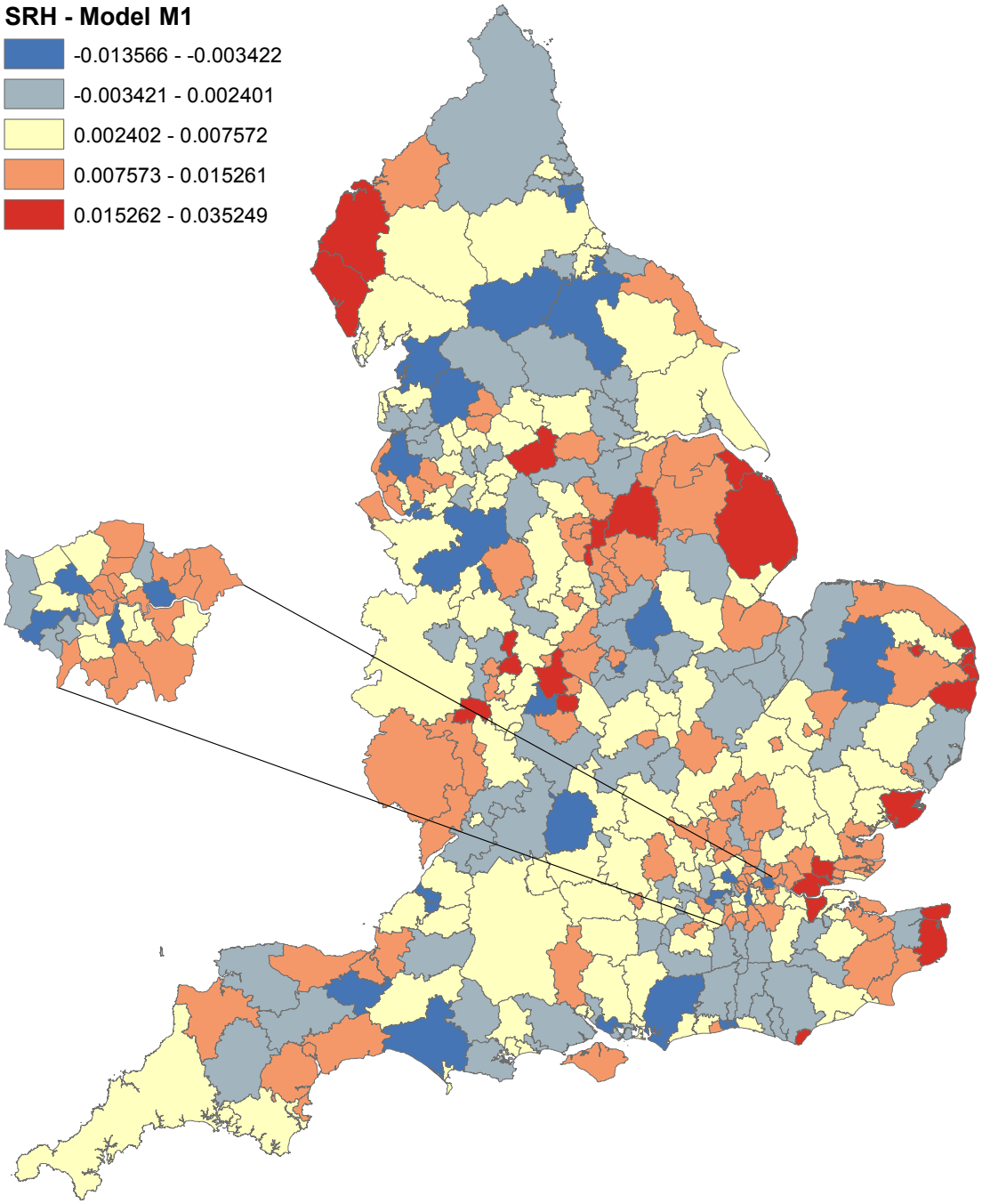
autocorrelation is particularly visible.

5.5 Discussion

The model-based approach presented in this paper remains simple in many respects: model specification, covariates, assumption of homogeneous and independent random effects, method of inference—yet they demonstrate valuable efficiency gains. From a practical viewpoint, our approach borrows strength from both (a) age information and hospital utilisation and (b) other areas (shrinkage estimation). We find it delivers considerable gains in precision, albeit at a certain cost in bias.

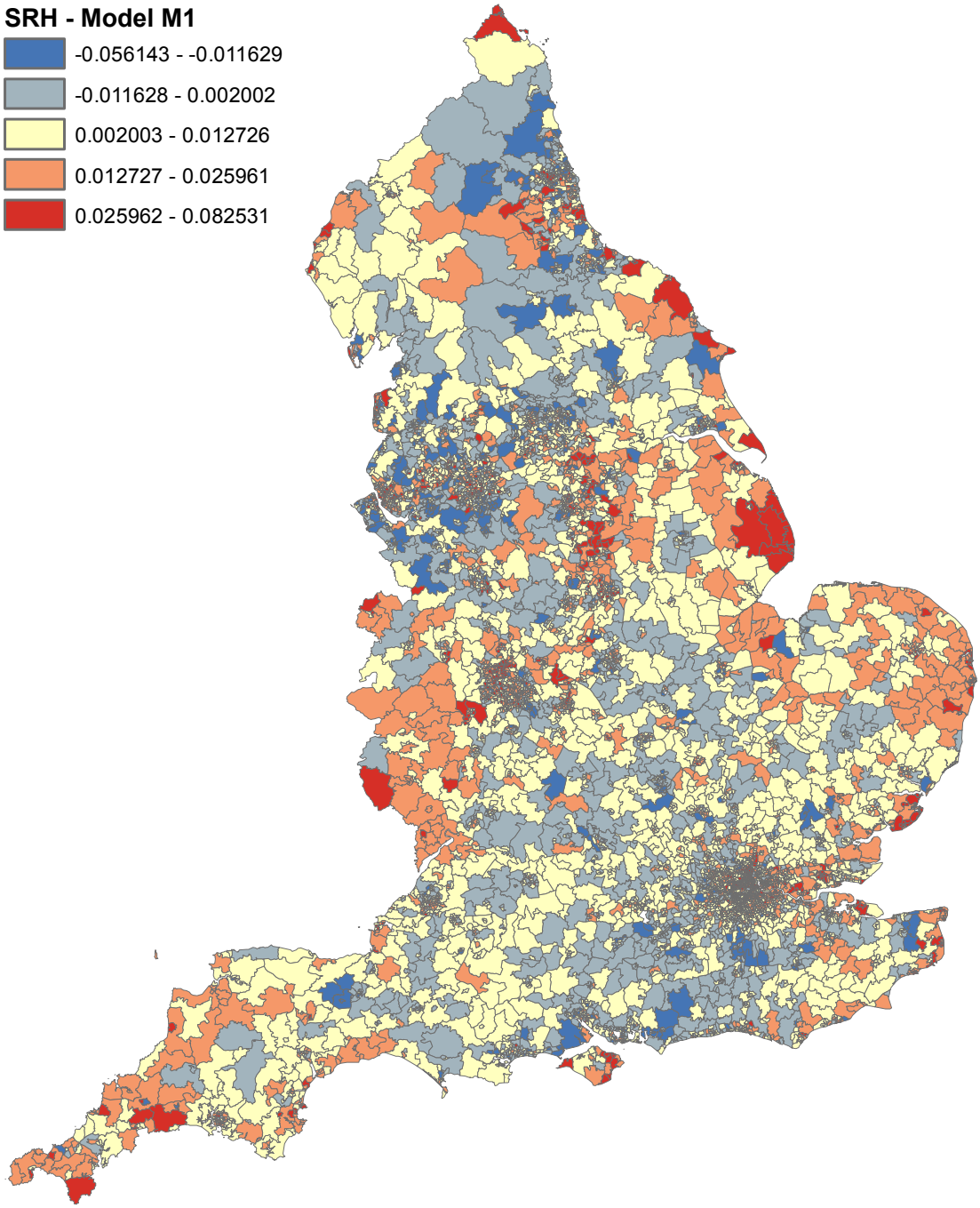
These findings complement existing research by providing strong evidence of empirical gains brought by an empirical best prediction strategy even with simple models. The best models **M1** only predict very limited proportions of the total between-area variance in poor health (circ. 60% at LAD level, 40% at MSOA level). As illustrated in simulations by Rao and Choudhry (2011) both SYN estimates and EPPs are sensitive to the between-area dispersion, which typically increases their MSE and absolute relative bias. In comparison, synthetic disease prevalence models often predict a greater proportion of between-area variance at a level equivalent to MSOAs. Previous work by the ONS (Heady et al., 2003) modelling ward proportions of households with bad/very bad SRH predicted 81 per cent of the between-area variance thanks to the availability of better covariates at the time: area proportion of residents born overseas (from the census) and local rates of social benefits claims (later discontinued due to changes in social benefits). Their attempt to model the proportion of households with adults with an LLTI proved less successful, predicting only 24 per cent of the total between-area variance, but this was attributed to poor reliability of the survey data due to between-interviewer coding variance. A US-based study by Jia et al. (2004) tackled similar research questions on synthetic estimates of the prevalence of severe work disability using a random digit telephone survey. Although this study attempted prediction for fewer domains (circ. 3,000 counties) with a smaller survey, it also found strong predictive power in hospital utilisation rates. In contrast, our study is based on a single-stage sampling design and the availability of data for almost all areas makes it possible to use shrinkage estimation instead of synthetic estimation, which can deliver additional benefits as our results suggest. Yet, the simplicity of our modelling strategy presents limitations of two main kinds.

First, the quality of our predictors remains insufficient compared to user needs expressed for official statistics on population characteristics. This could be addressed



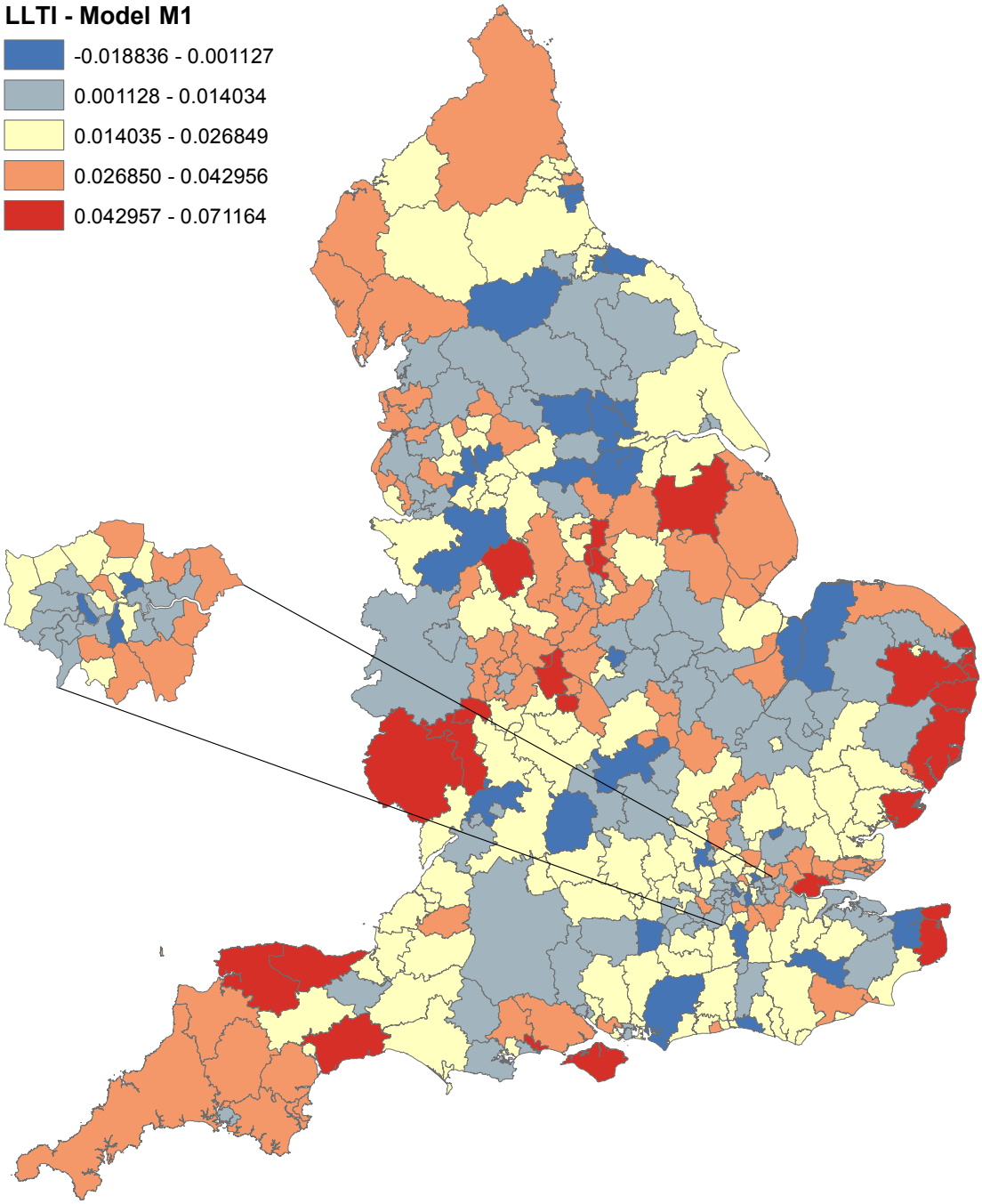
Note: Differences are expressed on the proportion scale (EPP subtracted from census estimate). Negative differences indicate overestimation while positive differences indicate underestimation. The colour shade cut-off values are set using five natural Jenks thresholds.

Figure 5.2 | Model **M1**: Map of deviations between census estimates and EPPs of bad/very bad SRH (LADs)



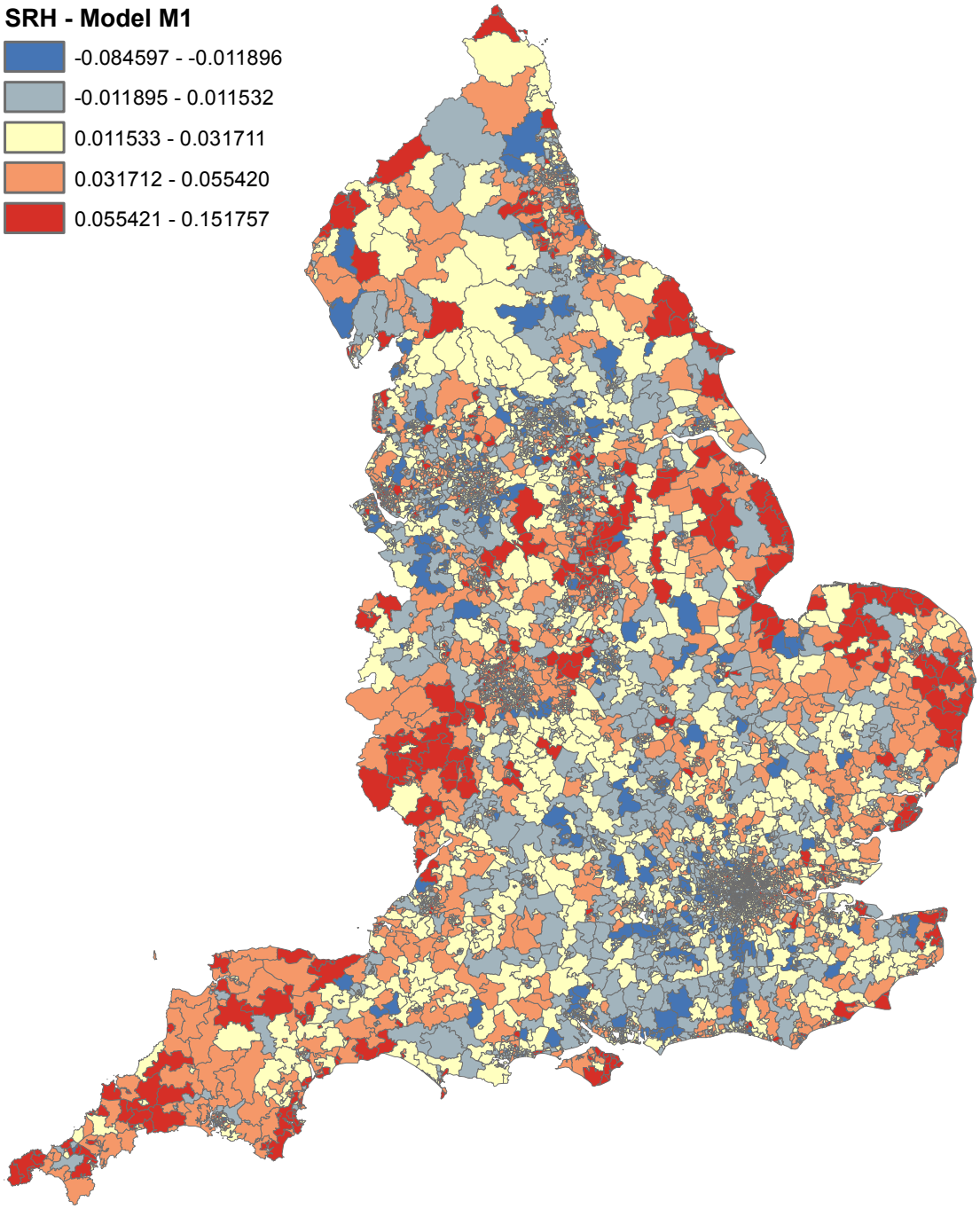
Note: Differences are expressed on the proportion scale (EPP subtracted from census estimate). Negative differences indicate overestimation while positive differences indicate underestimation. The colour shade cut-off values are set using five natural Jenks thresholds.

Figure 5.3 | Model **M1**: Map of deviations between census estimates and EPPs of bad/very bad SRH (MSOAs)



Note: Differences are expressed on the proportion scale (EPP subtracted from census estimate). Negative differences indicate overestimation while positive differences indicate underestimation. The colour shade cut-off values are set using five natural Jenks thresholds.

Figure 5.4 | Model **M1**: Map of deviations between census estimates and EPPs of LLTI (LADs)



Note: Differences are expressed on the proportion scale (EPP subtracted from census estimate). Negative differences indicate overestimation while positive differences indicate underestimation. The colour shade cut-off values are set using five natural Jenks thresholds.

Figure 5.5 | Model **M1**: Map of deviations between census estimates and EPPs of LLTI (MSOAs)

in part by a more sophisticated statistical design, demanding greater software flexibility and computing power than available in many secure data access environments. Given evidence of model misspecification found even in model **M1**, there is scope to reduce bias by benchmarking predictions to make them agree with higher-level direct estimates, for instance regional estimates deemed to be reliable. Benchmarking techniques have been examined for different types of models and predictors (Datta et al., 2011; Pfeffermann and Tiller, 2006; Pfeffermann et al., 2014; Steorts and Ghosh, 2013). This would in particular address the severe underestimation measured in our predictors, thus further attenuating their MSE. In addition, it is also possible to refine the prediction of between-area variation. It is suggested that better covariates (for instance hospitalisation statistics broken down further into age- and ethnicity-specific rates) could improve the predictive power of the model. Independently of this, it is important to recognise that no model is perfect when dealing with health outcomes. Heterogeneity always remains and very often exhibits a spatial structure (Lorant et al., 2001). The strong levels of global spatial autocorrelation identified in our predictions' error supports the use of spatial modelling techniques. Implications in terms of efficiency, parsimony and sample size requirements should be explored further, for instance using a simulation study.

Second, our study raises frequently encountered limitations around the validation of model-based designs. The importance of validation to assess potential misspecification is a recurring theme in SAE due to the mode of inference used with model-based estimation. It has previously been noted that 'in [SAE], as in other survey inference, efficient estimation is generally regarded as superior. A lot of the theory is concerned with deriving estimators that are efficient, or nearly so, sometimes in uncongenial circumstances, using models known not to be valid' (Longford, 2015, p. 66). By model misspecification, we mean an inadequate choice of covariates, parameterisation or covariance structure. This study focuses on the validation process, examining both external and internal validation methods. External validation has advantages, by circumventing risks of overfitting in hypothesis testing. This is especially valuable considering the difficulty to carry out cross-validation and tests of goodness-of-fit internally, since direct estimators and model-based predictors breach the assumption of independence essential in statistical tests (see review by Brown et al. 2001). Nevertheless, good quality external sources of data are not often available, and they may not allow definite comparability, as this study illustrates. These limitations, albeit important, should not overlook other quality assurance requirements implied by SAE. Beyond the reliability of the working model, it is important to assess statistical power in the sense that uncertainty remains around the capacity

to estimate model parameters accurately. From this viewpoint, simulation studies can play a part in designing SAE strategies in addition to conducting internal validation. Recent studies with complex statistical designs now tend to simulate inferences to decide on the reliability of a given set of model estimates or to assess characteristics such as the CV (Barker and Thompson, 2013; Porter et al., 2015c; Raghunathan et al., 2007). This can be a useful way to decide between a complex and a more simple model; while EPPs are more likely to yield a better fit, they may well behave in an unstable way when sample sizes are insufficient, as is the case with our MSOA models. Simulation seems to be a useful option to avoid overfitting unreliable estimates.

The primary implication of this validation study is that further work is needed on planning population health indicators estimation strategies. This paper shows that even very simple methods seem to have the potential to greatly improve quality in univariate statistics and indicate a potential to do much better if more suitable techniques and covariates are used. While we do not attempt to break predictions down by demographic characteristics such as age or ethnicity, this remains a research priority in the UK, where SAE is envisaged to play a role in upgrading national outputs on population characteristics with the help of healthcare utilisation data (ONS, 2016b). Although spatial modelling techniques have the potential to improve the quality of model-based estimators, our results show that SAE in any case demands a large amount of good quality data.

This may need to be addressed with more specific guidance based on simulated statistical power calculations. The ongoing UK Census Transformation Programme is attempting to identify methodological alternatives to full enumeration (long form census), which could involve a 4 to 10 per cent population survey (ONS, 2013d). Similar surveys exist elsewhere. The German census already involves a 10 per cent population survey (Münnich and Gabler, 2012; Ralf et al., 2011). The Netherlands now conduct population censuses using a combination of administrative registers and repeated weighting (Netherlands, 2014) while other countries are embarking on equivalent redesigns of their traditional decennial enumeration (Baffour et al., 2013; Bycroft, 2013; Kukutai et al., 2015). Beyond the census, further work is likely to benefit the wider production of public health statistics.

These results cannot claim strong external generalisability since model-based inferences are inherently dependent on a given working model. Despite this, we expect these to remain relevant for future applications to the same health outcomes in the UK. Results by Lloyd (2015) on univariate census statistics suggests that the spatial structure in LLTI has been relatively stable across a ten-year period across

England and Wales, therefore providing reassurance around the ability to plan statistical designs ahead with the confidence that overall parameters such as between-area heterogeneity and levels of spatial correlation are unlikely to change substantially within a few years. Wider generalisation of our conclusions to the modelling of other health outcomes is possible to the extent that self-reported indicators such as SRH and LLTI have been found to correlate strongly with other health status indicators at the population level (Johnson and Wolinsky, 1993; Johnston et al., 2009; Sutton et al., 1999).

Overall, this study adds to previous research by providing new results on the validation of small area estimates on large datasets: the census and a large survey of 150,000 respondents. Only in the area of economic statistics, Curtis (2003) compared their estimates of unemployment against 2001 UK census estimates. In the case of population health statistics, validation has only been carried out for synthetic estimates. In addition, fewer reliable sources are available that can be used as benchmarks for health topics. Twigg and Moon (2002) compared predictors against three local health surveys while Scarborough et al. (2009) used a range of local surveys and *ad hoc* design-based estimates from a national health survey. In both cases, authors acknowledge limitations in terms of comparability or even reliability of these estimates due to lack of statistical power. Jia et al. (2004) published a similar investigation scrutinising model-based estimates of the prevalence of severe work disability across US counties against the 2000 census, but only examined synthetic estimators. Although our study is not hindered to the same extent by lack of statistical power or unreliable comparator, goodness-of-fit tests between direct survey estimates and the census (Table 5.1) still give evidence of a significant deviation between those two sources. The resulting estimates have a similar bias as synthetic estimates produced by Jia et al. (2004). Overall, despite encouraging efficiency gains, our results show that model-based predictions may not provide the level of accuracy expected by users when examining the low percentage of agreement between the census and predictors in the ranking of areas by prevalence into tenths.

5.6 Conclusion

A recent concern in national statistics research, both in the UK and beyond, has been to understand challenges raised by 'the potential decrease in the amount of detailed, small area data available for use as auxiliary information in [SAE] models' (ONS, 2013e, p. 8) expected from the planned discontinuation of traditional decennial censuses. The ONS' census review has identified two main priorities: the impact

of replacing census auxiliary variables with covariates from administrative sources; and the statistical power of SAE models (*ibid.*).

This paper reports on a quality assessment of simple model-based estimates of health status across small census and administrative geographies of England in 2011. The scenario set for the present research was to only use auxiliary data that could in theory be produced in the absence of a census, in combination with routine household survey data and a common binomial logistic mixed model. This excludes detailed, individual-level socioeconomic auxiliary data. Beyond usual internal validation procedures, this paper cross-validates small area predictions with data from the 2011 census. This enables the production of empirical measures of error in addition to the more traditional prediction intervals, which rely on the very strong assumption that the working model is correctly specified.

It is apparent from our findings that, in the absence of detailed socioeconomic auxiliary data, working models become a lot more basic. Even supplemented with hospitalisation rates, the share of the between-area variance explained by fixed effect predictors remains small. There is a need to take into account between-area heterogeneity using empirical prediction. This comes at a cost: the precision empirical predictors of random effects is dependent on the sample size available in every small area of interest. Despite this, we find that (a) replacing census covariates in disease prevalence models is possible thanks to the existing UK health information infrastructure and (b) that efficiency gains of EPPs are not negligible. We emphasise avenues to improve the statistical design using techniques such as spatial modelling, benchmarking, and more detailed administrative data. Further research is needed to provide a reliable framework to plan small area health estimates which satisfy quality requirements.

Beyond these findings, the present study contributes to making more validation analysis available to researchers and practitioners. Despite considerable interest, adoption of SAE as a standard tool by national statistical institutes remains slow, which is in great part attributed to a reluctance to rely on rigid modelling assumptions for inference (Kordos, 2014). It is in many ways difficult to convey such assumptions to a wide community of health statistics users. It may be even more difficult to detect violations of such assumptions. Our results, examined in conjunction with others, highlight to the need to publish more quality assessments, as well as to address current data and methodology obstacles to SAE. This is likely to help users become more familiar with assumptions underpinning model-based estimation, and have confidence in the large gains in efficiency they bring about.

Chapter 6

The Spatial Structure of Chronic Morbidity: Evidence from UK Census Returns

Abstract

Background Disease prevalence models have been widely used to estimate health, lifestyle and disability characteristics for small geographical units when other data are not available. Yet, knowledge is often lacking about how to make informed decisions around the specification of such models, especially regarding spatial assumptions placed on their covariance structure. This paper is concerned with understanding processes of spatial dependency in unexplained variation in chronic morbidity.

Methods 2011 UK census data on limiting long-term illness (LLTI) is used to look at the spatial structure in chronic morbidity across England and Wales. The variance and spatial clustering of the odds of LLTI across local authority districts (LADs) and middle layer super output areas are measured across 40 demographic cross-classifications. A series of adjacency matrices based on distance, contiguity and migration flows are tested to examine the spatial structure in LLTI. Odds are then modelled using a logistic mixed model to examine the association with district-level covariates and their predictive power.

Results The odds of chronic illness are more dispersed than local age characteristics, mortality, hospitalisation rates and chance alone would suggest. Of all adjacency matrices, the three-nearest neighbour method was identified as the best fitting. Migration flows can also be used to construct spatial weights matrices which uncover non-negligible autocorrelation. Once the most important characteristics observable at the LAD-level are taken into account, substantial spatial autocorrelation remains which can be modelled explicitly to improve disease prevalence predictions.

Conclusions Systematic investigation of spatial structures and dependency is important to develop model-based estimation tools in chronic disease mapping. Spatial structures reflecting migration interactions are easy to develop and capture autocorrelation in LLTI. Patterns of spatial dependency in the geographical distribution of LLTI are not comparable across ethnic groups. Ethnic stratification of local health information is needed and there is potential to further address complexity in prevalence models by improving access to disaggregated data.

Keywords: Spatial autocorrelation, Spatial dependency, Spatial interaction, Spatial weights, Neighbourhood matrices, Disease mapping, Chronic morbidity, Limiting longstanding illness.

6.1 Background

The spatial distribution of chronic morbidity at a subnational level attracts considerable policy interest with relevance for health inequalities, health care planning, and resource allocation. Yet, information on the spatial distribution of morbidity is typically scarce with researchers often reverting to data on mortality or using data on health service use. Intelligence on the small area population prevalence of morbidity has tended to focus on cancer incidence and mortality (Clayton and Kaldor, 1987; Manton et al., 1989), cancer risk factors and screening uptake (Raghunathan et al., 2007), the prevalence of long-term conditions (Congdon, 2008a; Nacul et al., 2007), healthy lifestyles and behaviours (Malec et al., 1999; Kroll and Lampert, 2012). There has also been interest in measuring geographical variations in health needs (Carr-Hill et al., 1994; Gibson et al., 2002) and underdiagnosis of long-term conditions (Soljak et al., 2011).

The challenges involved in developing small area measures of morbidity have led to a range of techniques known as SAE. Model-based approaches to SAE rely on the premise that a chosen statistical model accurately predicts the odds of illness for the entire population. They raise a series of challenges in terms of validity. In the absence of systematic procedures guaranteeing optimal model specification and selection, there is a risk that this modelling process will be ill-informed, introducing bias in the resulting estimates. Reviews have argued that assumptions around the treatment of spatial effects introduces a particular element of subjectivity (Marshall, 1991b; Rao and Molina, 2015, p. 87).

The objective of this paper is to assess spatial dependence between small geographical areas for chronic morbidity. We analyse the geographical distribution of LLTI across England and Wales, focusing on the spatial structure in morbidity both with and without controls for confounders (mortality and hospitalisation rates). We consider global and local autocorrelation statistics for three types of dependence structures: contiguity, nearest k-neighbours and a novel approach building a spatial interaction matrix using origin-destination migration flows. Our analyses are stratified by ethnicity to isolate differences in the spatial structure of morbidity across

different population subgroups. This results from existing interest in monitoring health inequities across ethnic groups. It is currently unclear from the literature how homogeneous the spatial structure of morbidity is across ethnic groups, especially given the complex interaction with existing processes of residential segregation.

The following background section gives a review of existing knowledge on spatial aspects of health determinants, to inform model selection. Aims and methods are then outlined, with a particular emphasis on concepts used to describe spatial structures. A results section then presents both descriptive statistics and model-based analyses of the geographical distribution of LLTI, introducing mortality, hospital admissions and adjacency matrices as predictors of this structure. The paper then concludes by identifying implications for the routine prediction of morbidity prevalence for different geographical units.

6.1.1 Existing knowledge on the spatial structure of chronic morbidity

Much of what is known on the distribution of chronic diseases comes from data on validated self-reported health statuses. LLTI has emerged as a very strong predictor both of chronic morbidity and mortality (Cohen et al., 1995; Jordan et al., 2000; Manor et al., 2001). It has also proved instrumental in measuring health inequalities both across socioeconomic categories and space (Borooah, 1999; Macintyre et al., 2005; Senior, 1998). LLTI has been recorded since 1991 in UK decennial censuses in the form of a question asking whether respondent's day-to-day activities were reduced by a health problem or disability. This information has supported important research into the determinants of health care needs of different populations in different places (Sutton et al., 1999; Congdon, 2002).

The literature provides some information regarding ecological determinants of chronic morbidity and their spatial structure. Analyses have showed that, even once population age and essential demographic confounders are controlled for, adjusted morbidity levels correlate significantly with local socioeconomic characteristics (Borooah, 1999), and the remaining between-area heterogeneity is spatially structured (Shouls et al., 1996b). To examine these 'place effects', Bentham et al. (1995), Martin et al. (1995), Senior (1998), Shouls et al. (1996b,a), Congdon (1995, 2006a, 2008a) and Stafford et al. (2009) have all investigated the association of LLTI prevalence with both individual-level characteristics and area-level contextual variables. Their work has showed that local mortality, unemployment, household overcrowding, ethnic diversity, social renting, proportions of workers employed in mining and other

heavy industries all correlated strongly with standardised ratios of LLTI. These confounders often prove to be similar in places that are near to each other (for instance across urban areas), pointing to distinctive underpinning spatial structures.

A variety of processes have been hypothesised to explain this apparent clustering of long-term conditions across places. On the one hand, it is the case with many health outcomes that a residual spatial pattern can subsist even once observable risk factors or confounders are taken into account (Wakefield et al., 2000). On the other hand, research has argued that population migration not only determines the dispersion of communicable disease, but also provides one of the factors driving the spatial clustering of chronic morbidity. The literature has in particular examined 'health selective' residential migrations as life course processes of selection (Jones and Duncan, 1995). Boyle et al. (2002) have produced evidence that Scottish migrants tend to be healthier than non-migrants, and that healthy migrants are likely to travel longer distances. Further evidence supporting the theory of a 'sorting' effect of migrations on health has been presented by Norman et al. (2005), emphasising the existence of a strong flow of healthy migrants aged 20 to 59 years towards areas with lower levels of material deprivation. A review by Smith and Easterlow (2005) argues that the influence of residential mobility processes on geographical inequalities in morbidity and mortality remains little understood, with mixed results depending on the geographical level of analysis and the health outcomes under consideration. Despite the absence of clear evidence claims for the health sorting effects of migration point to a need to consider how we might use migration data to capture some of the spatial structure in morbidity in a way that proximity may not.

All the above evidence has implications for disease prevalence models. In its most elementary form, model-based SAE fits a model predicting the probability of having a given illness as a function of age, sex, and other individual characteristics. This model is then applied to local population estimates and auxiliary data known for every individual residing in a catchment area in order to produce a local prevalence estimate. This amounts to interpolating prevalence levels known at the national level to local populations using a combination of:

(a) fixed individual-level risk confounders
(b) spatially varying area-level confounders
(c) residual unobserved risk (between-area residual heterogeneity in prevalence)

This last component (c) is essential and explains the popularity of multilevel health models in recent decades (Subramanian et al., 2003), being one of the preconditions to the model's unbiasedness. Residuals capture local departures from the overall average which signals, for instance, excess morbidity. This random compon-

ent avoids assuming for instance that all persons aged 16–24 years have the same prevalence across all areas. This component is difficult to estimate because sample data will typically be small, often well under a few dozen cases. More importantly, the underpinning method assumes that these residuals are independent from one another and often ignores the fact that spatial dependence may persist. Recognising underlying spatial structure makes it possible to borrow information from other areas in order to estimate these components in a more efficient manner (see for instance simulation results by Pratesi and Salvati, 2008).

More research is needed to understand spatial dependence. Spatial structures have previously been described as the result of 'the operation of processes in which spatial relationships enter explicitly into the way the process behaves' (Adams and White, 2006, p. 24). They are often understood as functions of distance or spatial adjacency (neighbours). The science of spatial autocorrelation has largely been dominated by Tobler's First Law of Geography, summarised as 'everything is related to everything else, but near things are more related than distant things' (Tobler, 1970). Contiguity methods, such as Queen, Rook or Bishop, and the k-nearest neighbours method have traditionally been privileged. Although this standard approach is appealing, there are many more ways in which spatial interaction could be defined. In particular, origin/destination migration flow statistics constitute additional evidence of processes of spatial interaction and therefore between-area dependence. Although using such flow metrics to produce spatial weights has been envisaged before Bivand et al. (2013b, p. 271), they have, to the best of our knowledge, not been applied to empirical investigation to date.

Internationally, most research has tended to demonstrate that there is global spatial autocorrelation in many health outcomes even after standardisation (Lorant et al., 2001). This autocorrelation is a sign of spatial similarity in unobserved risk factors (Wakefield et al., 2000). Yet, it remains unclear whether these spatial patterns are homogeneous once we disaggregate by demographic subgroup, and add explicit spatially varying area-level confounders.

This justifies looking further into spatial structures themselves, to inform non-communicable disease mapping methods with a particular focus on the type of constraints placed on the treatment of residual between-place heterogeneity. On the basis of this background we propose to examine the spatial structures of LLTI in a more systematic way, investigating (a) what structures can be uncovered in terms of dispersion, autocorrelation, and contextual effects, (b) whether they are the same across different subgroups (age and ethnicity) and (c) whether they subsist once good area-level covariates are introduced. We aim to address a current gap in know-

ledge regarding the spatial structure of morbidity in England and Wales, but also to reconsider the specification of disease prevalence models.

6.2 Methods

6.2.1 Data sources

We use 2011 census data on LLTI for England and Wales (ONS, 2013a). Although the quality issues concerning self-assessed health information are well documented (ONS, 2014a), a key advantage of using census data lies in the absence of sample size restriction. The 2011 census met a high quality 93 percent person coverage rate for England and Wales (ONS, 2012a), and thus constitutes a unique source of information to establish prior knowledge on the spatial structure of illness. Census data provide sufficient statistical power to examine model-fitting hypotheses which usually cannot be tested with survey data due to lack of power. This is especially true for small population subgroups such as older people and ethnic minorities, whose representation in health surveys is too weak in comparison to the amount of interest they attract. This reduces risks of model overfitting when using a large number of parameters. With the census coverage survey's adjustments for nonresponse (ONS, 2012b), the final sample size used for this analysis is $n = 56{,}075{,}912$.

We examine private households' returns for question no. 23:

> 'Are your day-to-day activities limited because of a health problem or disability which has lasted, or is expected to last, at least 12 months? Include problems related to old age'.

Respondents were able to answer 'Yes, limited a lot', 'Yes, limited a little', or 'No'. Throughout this paper 'LLTI' refers to strong activity limitations ('limited a lot') which has been found to have a better rate of agreement in the post-enumeration Census Quality Survey (ONS, 2014a).

The choice of indicator is justified by two main reasons. First, LLTI has become a central indicator to measure inequalities in health and health needs, to the point of being included in most UK household surveys. It underpins indicators such as the Slope Index of Inequality in health, the disability-free life expectancy, as well as gender and ethnicity gaps in health. These have been for a number of years to inform health service policy aiming to reduce health inequalities (UK Department of Health, 2013). Several of the ONS' products estimate these indicators for local authorities (ONS, 2013g, 2014d), and efforts have been made to publish them for smaller units (Congdon, 2002). Second, although self-reported, the LLTI

health status correlates with important indicators of chronic conditions. In addition to being a good predictor of health service use (Sutton et al., 1999), it is also a strong predictor of diagnoses as defined in the International Classification of Diseases (OPCS Social Survey Division, 1975), although evidence tends to suggest that LLTI tends to underestimate morbidity compared to clinical records or the more demanding SF-36 tool (Jordan et al., 2000).

6.2.2 Statistical methods

This paper aims to address gaps in knowledge regarding the spatial structure of chronic morbidity and provide evidence relevant to build SAE models. We explore spatial heterogeneity in the odds of LLTI at a scale for predictions to be feasible for small ethnic groups: LADs, areas with populations ranging from 34,000 to 1.1 million inhabitants; and MSOAs, census geographical units averaging 7,700 residents. Standard descriptive statistics are used to characterise the spatial structure in odds: variance and autocorrelation. A series of models then analyse this structure conditionally on contextual data (mortality, hospitalisations), using a typical logistic binomial parameterisation:

$$\log\left(\frac{y_{id} + .5}{n_{id} - y_{id} + .5}\right) = \mu_{id} + \upsilon_d = \boldsymbol{x}_{id}\boldsymbol{\beta} + \upsilon_d \tag{6.1}$$

where y_{id} is the number of individuals belonging to a cross-classification i of gender (1, 2), age group ('0–15', '16–49', '50–64', '65+'), and ethnic group ('White', 'Mixed', 'Asian', 'Black', 'Other') reporting an LLTI in a given area d. n_{id} denotes the total number of residents of private households at risk for this same cross-classification, μ_{id} the conditional mean log-odds of having an LLTI (fixed part of the model), $\boldsymbol{\beta}$ a column vector of fixed effect coefficients, and $\boldsymbol{x}_{id}$ a vector of covariates known for all individuals: age, sex and ethnicity dummy variables, as well as area-level characteristics tested in this paper. Random intercepts υ_d are realisations of a random variable $\boldsymbol{\upsilon}$ of mean zero and variance σ^2. We add 0.5 to both the numerator and the denominator of odds to produce 'empirical logits', addressing bias arising from the presence of null denominators (Agresti, 2002; Gart and Zweifel, 1967).

Models are estimated using Laplace approximation with the `R` package `lme4` (R Core Team, 2014; Bates et al., 2015). We use classical model selection techniques; likelihood ratio tests, the AIC and regression coefficient significance. During model selection, attention was also paid to σ^2, the variance of random effects $\boldsymbol{\upsilon}$, which

reflects the between-area dispersion in prevalence that is not attributed to differences in covariates included in the fixed part. The reason why σ^2 is used as a decision factor is that it plays a considerable part in the efficiency of estimation (Longford, 2006). Approximations of the mean squared error of prediction developed by Prasad and Rao (1990) and extended to log-linear models (Jiang and Lahiri, 2001; González-Manteiga et al., 2007) show that the main determinant of prediction error is the size of σ^2 compared to the within-group variance. By attempting to reduce σ^2 as much as possible, we focus on improving the predictive power of the fixed part. This is important when conducting SAE in real world conditions because residuals υ_d will often be estimated with very small sample sizes and therefore subject to substantial error. A strong fixed part μ_{id} is likely to produced better predictions overall.

6.2.3 Defining 'spatial structures'

Global spatial autocorrelation is measured using the Moran's I statistic, with a random permutation test for significance testing (Cliff and Ord, 1981). Local autocorrelation of regression residuals is also examined using a local indicator of autocorrelation (LISA) (Anselin, 1995) and the Moran scatterplot (Anselin, 1996). These are used to detect significant leverage of one set of neighbours on the global (average) level of autocorrelation, thereby signalling a cluster of high or low similarity.

Four types of adjacency matrices were tested (see Table 6.1). **L.A** and **M.A** follow the standard approach and were generated using the `spdep` package (Bivand et al., 2013a; Bivand and Piras, 2015) and boundary shapefiles (ONS, 2011a). They are based on the Queen method: areas were coded as neighbours in the adjacency matrix if their digital boundaries shared at least one point or if two of their respective points were separated by less than 500 metres. This ensures for instance that London boroughs separated by the River Thames are coded as neighbours. The final matrix was edited manually to attach islands to mainland neighbours and verify that no area was left without neighbours. **L.B***k* and **M.B***k* were produced using the *k*-nearest neighbours method for *k* values of 2 to 10, with the view of determining an optimal *k*. All matrices were row-standardised, a procedure that is traditionally used to ensure the positive-definitiveness of correlation matrices in various conditional autoregressive models when spatial weight matrices are not symmetric (Banerjee et al., 2004).

For LADs, additional matrices **L.C** and **L.D** were built using migration flows as a proxy for spatial dependence. There are good reasons why areas further apart could be more closely related to each other given the UK's urban and rural struc-

Table 6.1 | Standardised proximity matrices tested in this paper for between-LAD and between-MSOA autocorrelation

Matrix identifiers *LADs*	*MSOAs*	Method of construction
L.A	**M.A**	Contiguity matrix (with isles attached to the mainland) [spdep + manual adjustments]
L.B*k*	**M.B***k*	*k*-nearest neighbours (based on Euclidian distances between population centroids) [spdep]
L.C*k*	—	Up to *k* LADs from which most new residents originate (ONS, 2013h), binary weights
L.D*k*	—	Up to *k* LADs from which most new residents originate (ONS, 2013h), proximity weights *k*, ..., 3, 2, 1

ture. Proximity is not the only reason why risk factors would be more alike in areas. Intra-national origin-destination migration data published by the ONS (ONS, 2013h) were used to construct spatial weights based on the intensity of flows (see R syntax in Additional file 1). For every LAD, we defined neighbours as the *k* areas from which the most migrants originate, based on the ratio of the total migrants they contributed relative to their respective population sizes. In other words, neighbours are not just those that send most migrants to a given district, they are the ones for which these migrants represent the highest proportion of their respective populations. This is to ensure a fair weighting across all LADs in the process of averaging odds of LLTI, and especially ensure that the resulting neighbours would not systematically be the biggest LADs. If a district *A* sent a large number of migrants to district *B*, but this flow in fact represented a very modest volume relative the entire population of *A*, it would seem excessive to use the odds of poor health of the entirety of district *A* as a smoothing reference for district *B*.

Sensitivity analyses on a subset of LADs suggested that selecting neighbours who send the highest number of migrants or those who send migrants flows which represent the highest proportion of their total population did not alter the eventual list of neighbours substantially. Further analyses (see appendix A) were conducted to establish whether origins and destinations differed substantially depending on the age of migrants. Results showed that excluding younger migrants did not have a strong influence on the resulting matrices. However, we hypothesised that student migrations, which are only temporary, are likely be less determinant of the structure of LLTI than other types of migrations taking place across life. Final spatial weight matrices were therefore generated exclusively based on flows for migrants aged 30 years and over.

Table 6.2 | Between-area variance in odds of LLTI by demographic group for LADs and MSOAs

		MSOA				LAD			
	Age	0-15	16-49	50-64	65+	0-15	16-49	50-64	65+
Females	White	0.000	0.000	0.007	0.032	0.000	0.000	0.003	0.012
	Black	0.145	0.053	0.251	0.585	0.010	0.004	0.013	0.203
	Mixed	0.002	0.009	0.401	1.241	0.000	0.000	0.014	0.065
	Asian	0.028	0.002	0.071	0.699	0.003	0.000	0.010	0.057
	Other	0.181	0.100	0.397	0.882	0.009	0.002	0.033	0.433
Males	White	0.000	0.001	0.008	0.026	0.000	0.000	0.003	0.011
	Black	0.172	0.049	0.231	0.653	0.010	0.001	0.008	0.111
	Mixed	0.004	0.011	0.437	1.165	0.000	0.001	0.015	0.074
	Asian	0.029	0.004	0.082	0.511	0.003	0.003	0.006	0.023
	Other	0.184	0.092	0.355	0.572	0.009	0.002	0.014	0.228

Source: Authors' calculations, 2011 census table DC3201EW (ONS, 2013a).
Note: Cells are shaded according to the decile corresponding to their value.

Table 6.3 | Between-area CVs for odds of LLTI by demographic group for LADs and MSOAs

		MSOA				LAD			
	Age	0-15	16-49	50-64	65+	0-15	16-49	50-64	65+
Females	White	0.535	0.476	0.627	0.399	0.222	0.311	0.440	0.266
	Black	1.359	1.752	1.208	0.901	2.593	1.583	0.895	0.911
	Mixed	1.158	1.187	1.418	1.176	0.882	0.461	0.670	0.540
	Asian	2.002	1.053	1.227	1.092	3.565	0.937	0.812	0.522
	Other	1.076	1.427	0.990	0.898	1.924	0.950	0.971	0.987
Males	White	0.469	0.535	0.640	0.446	0.245	0.337	0.456	0.326
	Black	1.450	1.773	1.243	1.002	2.288	0.714	0.844	0.810
	Mixed	1.304	1.209	1.437	1.198	0.431	0.506	0.673	0.637
	Asian	1.884	1.329	1.392	1.217	2.761	2.261	0.791	0.498
	Other	1.093	1.713	1.161	0.850	1.869	0.967	0.851	1.147

Source: Authors' calculations, 2011 census table DC3201EW, (ONS, 2013a, 2011a).
Note: Cells are shaded according to the decile corresponding to their value.

6.3 Results

6.3.1 Descriptive characteristics

Overall across English and Welsh LADs, the mean odds of LLTI is 9.23×10^{-2} (equivalent to an 8.40 percent mean prevalence) with a variance of 7.01×10^{-4}, equivalent to a 28.7 percent CV. This masks huge differences across subgroups. Examining age, Table 6.2 suggests that the between-area variance in odds of LLTI among older groups is several hundred times that of younger groups. This implies that the level-2 variance is expected to be higher for older age groups. Much of this effect can be attributed to the higher prevalence of LLTIs among older populations; larger odds by definition have larger variances. CVs reported in Table 6.3 confirm this; relative to the average of all odds across England and Wales, the dispersion is of the same order of magnitude across age and gender groups for White populations.

This pattern differs substantially across minority ethnic groups. In the case of ethnic minorities in general, it seems that between-area differences in prevalence are strong for younger groups; even age groups 0–15 exhibit high dispersion in the case of categories 'Black' and 'Other'. We also find higher between-area variance estimates at the MSOA level for these groups: while for the White group, the between-MSOA variance in odds of LLTI is on average two to three times the between-LAD variance, for most other cross-classifications the variance is multiplied by a factor of five to ten.

For both LADs and MSOAs, highest levels of autocorrelation are measured using the three-nearest matrix **•.B3** (see Table 6.4). Similar measurements taken for higher values of k (up to 10 neighbours), not reported in the table, confirmed that increasing the number of neighbours only reduces Moran's I estimates. Estimates for White populations show that odds for older age groups exhibit higher levels of spatial autocorrelation than younger groups. In other words, the spatial clustering of poor health is higher for older age groups. Around retirement age a final wave of intra-national migrations emphasises the clustering of people by health.

Interestingly, there is no evidence of the same pattern occurring for Other ethnic groups. On the contrary, the older the individuals reporting an LLTI, the less they are found to cluster in areas. This implies that odds of poor health for ethnic minorities are not only more dispersed than those of White people; they are also less predictable or, in spatial terms, more random. None of the matrices tested in this investigation uncovered substantial spatial structure in the patterns of illness experienced by ethnic minorities, and these structures are very different from

Table 6.4 | Moran's *I* statistics of spatial autocorrelation in odds of LLTI by adjacency matrix and demographic group

		M.A				M.B3				M.B5			
		0-15	16-49	50-64	65+	0-15	16-49	50-64	65+	0-15	16-49	50-64	65+
Females	White	0.164	0.485	0.556	0.538	0.381	0.644	0.676	0.653	0.310	0.579	0.621	0.604
	Black	0.426	0.228	0.198	0.066	0.569	0.415	0.396	0.307	0.505	0.342	0.331	0.225
	Mixed	0.233	0.133	0.082	0.032	0.403	0.352	0.322	0.282	0.340	0.281	0.244	0.204
	Asian	0.226	0.164	0.109	0.060	0.391	0.347	0.327	0.297	0.317	0.282	0.257	0.218
	Other	0.364	0.245	0.102	0.015	0.526	0.436	0.341	0.257	0.465	0.364	0.264	0.179
Males	White	0.241	0.464	0.564	0.605	0.440	0.626	0.686	0.713	0.364	0.563	0.630	0.664
	Black	0.354	0.221	0.204	0.083	0.511	0.401	0.409	0.318	0.446	0.339	0.342	0.241
	Mixed	0.125	0.111	0.082	0.044	0.344	0.333	0.315	0.281	0.266	0.266	0.237	0.196
	Asian	0.256	0.162	0.119	0.052	0.409	0.357	0.345	0.301	0.332	0.281	0.258	0.218
	Other	0.358	0.164	0.088	0.044	0.517	0.377	0.319	0.284	0.454	0.299	0.241	0.204

		L.A				L.B3				L.D3			
		0-15	16-49	50-64	65+	0-15	16-49	50-64	65+	0-15	16-49	50-64	65+
Females	White	0.348	0.556	0.539	0.566	0.540	0.731	0.694	0.717	0.298	0.471	0.459	0.468
	Black	0.075	0.025	0.169	0.096	0.326	0.265	0.332	0.259	0.094	-0.008	0.137	0.088
	Mixed	0.012	0.345	0.287	0.205	0.300	0.573	0.539	0.398	0.028	0.250	0.263	0.176
	Asian	-0.003	0.081	0.384	0.358	0.252	0.317	0.547	0.559	-0.001	0.048	0.256	0.265
	Other	0.114	0.057	0.093	0.114	0.355	0.311	0.316	0.345	0.101	0.039	0.063	0.064
Males	White	0.353	0.548	0.556	0.596	0.529	0.723	0.712	0.751	0.292	0.455	0.457	0.478
	Black	0.042	0.150	0.140	-0.004	0.313	0.342	0.366	0.279	0.066	0.099	0.098	0.063
	Mixed	0.048	0.298	0.206	0.205	0.315	0.517	0.447	0.396	0.044	0.210	0.186	0.198
	Asian	-0.010	0.007	0.194	0.266	0.250	0.259	0.389	0.445	-0.005	0.005	0.145	0.186
	Other	0.118	0.088	0.059	-0.044	0.328	0.292	0.291	0.232	0.110	0.065	0.086	0.046

Source: Authors' calculations, 2011 census table DC3201EW (ONS, 2013a), ONS migration and digital boundary data (ONS, 2013h, 2011a).
Note: Cells are shaded according to the decile corresponding to their value.

those of White populations. We hypothesise that such heterogeneity relates to the presence of stronger socioeconomic differences across space for ethnic minorities. In these circumstances, it is unlikely that borrowing strength from the structure exhibited by White populations would help make precise inferences about the health of other populations. There is more potential in using other information such as ethnic density data to reduce the variability in the model, as we show in the next section.

These descriptive estimates also provide indications regarding best fitting adjacency matrices. In the case of LADs, levels of autocorrelation measured using the 'migration neighbourhoods' **L.C·** and **L.D·** are lower than with more traditional matrices. Table 6.4 only reports results for row-standardised, ranked neighbours matrix **L.D3**, because the specification of **L.C***k* (binary weights) did not perform as well. In addition, sensitivity analyses found that the age categories included to generate those migration neighbourhood matrices did not have a strong influence on measures of spatial autocorrelation. More research on age-specific adjacency matrices could refine this observation.

We conclude from this exploratory work that levels of dispersion in odds of LLTIs, although comparable between sexes, are very dissimilar depending on age and ethnic groups. They may require separate treatment when it comes to their modelling and prediction. Descriptive estimates of autocorrelation provide a strong suggestion that the three-nearest neighbours method **·.B·** is likely to be the most efficient since it captures highest levels of homogeneity in odds of LLTI. This finding is consistent across all demographic cross-classifications.

6.3.2 Modelling with covariates: Area classifications and data on ethnicity

We now examine the residual geographical variance in odds of LLTI once contextual information (area classification, ethnic density, mortality rates, and health service data) is introduced in a multivariate framework. We seek to establish whether this contextual information predicts the spatial structure in residuals υ, that is to say, shrinks their variance σ^2. In this section, we build a series of models predicting LLTI prevalence for LADs exclusively, since they are the level at which contextual data is most commonly available. We begin by introducing some disaggregation using the 2001 National Statistics area classification of English local authorities produced by cluster analysis (ONS, 2010a). This allows us to treat LADs differently according to the following typology;

— Cities and Services; London Suburbs; London Cosmopolitan (reference category)
— London Centre
— Prospering UK
— Coastal and Countryside
— Mining and Manufacturing.

Welsh contextual data being unavailable for LADs, a coarser specification involving a single dummy variable reflecting higher odds of poor health in Wales is retained. To reflect hypotheses of an ethnic density effect in the literature (Pickett and Wilkinson, 2008), the census estimate of the proportion of the district population identifying as the same ethnicity is incorporated as a covariate for those ethnic groups where such an addition improves the model fit. This is the case for all groups but Black and White populations. With other groups, this covariate improves fit as measured by the AIC and substantially reduces the between-area variance (−12 % for Asian and Other, −7 % for Mixed). This forms specifications for a baseline model *(M0)* fitted separately on data for each of the five ethnic groups (see appendix, Table A.8). Area-level residuals exhibit mild autocorrelation (Moran's I comprised between 0.3 and 0.6).

6.3.3 Local mortality and hospitalisation data

We continue with contextual information on mortality and hospitalisations which are expected to be associated with some of the unobserved risk factors modelled through random effects until now. We aim to test whether this information absorbs either between-area heterogeneity or its spatial structure.

Existing evidence (Jordan et al., 2000, 2003; Sutton et al., 1999) demonstrates that individuals are very likely to report an LLTI if they have had or are about to seek a medical diagnosis. In addition, there is a well-known association at the population level between self-reported poor health and local mortality rates (Bentham et al., 1995). Though non-linear, this association has been exploited for SAE using bivariate life table models (Congdon, 2002) and relational logistic models (Marshall, 2009). The bivariate response model, relevant for the data at hand, gave a particularly poor fit and was immediately discarded. Instead directly age-standardised mortality rates (SMRs) from death registrations (ONS, 2013i) were transformed through Z-standardisation and used as a straightforward covariate. Models *(M1)* (see Table A.9) result from best model selection among a range of specifications for each ethnic group separately. We compared sex-specific SMRs, overall SMRs, and interaction

with gender dummies. In the case of Black populations, no association with mortality was found. Gains in terms of reduction of between-area variance in random intercepts are important, especially in the case of Mixed ethnic groups, where σ^2 is almost halved.

While mortality data does help predict local prevalence of LLTI, it arguably remains distantly related to chronic morbidity amongst the living. We compare its predictive power with that of ISARs for 2008–2013 (Public Health England, 2014b) on the one hand, and elective admissions (Public Health England, 2014a) on the other hand. This is with the hypothesis that prevalence of LLTI and rates of hospitalisation share common determinants (socioeconomic characteristics, lifelong exposure to health determinants). For all ethnicities, rates of emergency admissions are found to be associated with larger regression coefficients and improvements in fit. They are thus selected as the preferred covariate. We then test interaction effects between (a) sex and age variables, (b) mortality and (c) emergency admissions proceeding by backward elimination based on the best sets of covariates, leading us to the set of final models *(M2)* reported in Table A.10.

Overall, for White populations, the new specifications reduce the AIC by over 9,800, and the between-area residual variance by 45 percent. The effect is less marked for ethnic minorities. The between-area residual variance remains stable for Black populations and is cut by about 20 percent for other minorities. Emergency hospitalisations exhibit a strong association with morbidity rates and make the biggest improvement to the models. English areas with observed emergency admissions in excess of 10 percent relative to the expected number of admissions (based on age-specific rates of admission for England overall) exhibit odds of LLTI for White persons aged 50 to 64 years on average 16 percent higher compared to areas in line with England's overall admissions rate. Models remain very different across ethnic groups and the association with admissions rates is weaker for non-White population, but disaggregation of admissions statistics by ethnic group could yield stronger associations in future investigations.

Residuals of each of the final models were examined in detail. Table 6.5 shows that residuals correlate only very weakly across ethnic categories. This constitutes further evidence that the spatial structure is specific to each of those population groups. Autocorrelation statistics confirm that accounting for differences in mortality and hospitalisation rates does not reduce the spatial autocorrelation in residuals. It reduces the random variability across LADs substantially without offsetting the extent to which deviations of a district's odds of LLTI from the mean correlate with the deviation measured in neighbouring LADs. From the viewpoint of pre-

Table 6.5 | Matrix of pairwise correlation in random intercepts between models *(M2)*

	White	Black	Asian	Mixed	Other
White	-				
Black	0.181	-			
Asian	0.107	0.375	-		
Mixed	0.405	0.432	0.311	-	
Other	0.212	0.360	0.338	0.364	-

Note: Cells are shaded according to the decile corresponding to their value.

dictive modelling, it constitutes an advantage; introducing area-level covariates does not reduce the potential information can be borrowed from neighbouring areas using relevant autoregressive model specifications. In addition, area-level predictors did not lead to important outliers emerging which could signify local departures from the global association with mortality and hospitalisation rates. Aside for individuals from Other ethnic minorities, there is strong evidence both from normal quantile-quantile (Q–Q) plots and Shapiro-Wilks tests that residuals follow a normal distribution.

Figure 6.1 presents a series of maps of final model residuals (on the odds ratio scale), which illustrate by how much the fixed part of the model should be multiplied in order to reach the census estimate of odds of LLTI. Unshaded areas indicate predictions falling within ±10 percent of the census estimate. Red shades signal LADs where the fixed part of the model underestimates odds by more than 10 percent while blue shades LADs where it overestimates odds by over 10 percent. With Moran's I estimates close to 0.5, we conclude that half of the deviation between odds for a given district and the national mean is on average shared by its three nearest neighbours. Figure 6.2 examines the local contribution of each clique of LADs towards the global measure of spatial autocorrelation. LISAs are calculated, regressed against the model residuals and plotted for each ethnic group separately. Each of the bottom left quadrants signals statistically significant outliers in red, which can be regarded as area residuals which exhibit significant higher or lower similarity with their three nearest neighbours than average, and therefore have particular leverage of the global level of autocorrelation. Together with the maps, it becomes apparent that the chosen modelling and spatial specifications leave important clusters of unexplained risk factors, which are dissimilar across ethnic groups. The Asian model in particular exhibits a lot of heterogeneity in the strength of spatial dependence between LADs, with very strong clusters emerging for instance in parts of Lancashire, Merseyside and Yorkshire, Nottingham and Leicester, as well as North East

and South West London boroughs.

6.4 Discussion

Previous work on the 2011 census has highlighted the presence of a strong spatial structure in univariate morbidity statistics (Lloyd, 2015). Analysis reported in this paper presents a deeper examination of multivariate aspects of this spatial dependence. Descriptive estimates suggest that the variability in odds of poor health across groups and places is larger than can be expected from just looking at crude prevalence estimates. For instance, area effects are often thought to correlate strongly across age groups, as reflected in random walk priors proposed by Congdon (1995). Our analysis looking at ethnicity provides strong evidence that patterns of spatial dependency in the odds of LLTI differ substantially across ethnic groups. The covariate-adjusted spatial structure of LLTI in White people only moderately correlates with that for Mixed ethnic groups. Structures of LLTI of all other groups correlate very weakly with each other. Descriptive estimates for ethnic minorities also reveal that levels of spatial autocorrelation are higher for young people, in constrast with the increased autocorrelation measured among older age groups in White people. Reasons for this difference are unclear and call for further research. One can hypothesise that cohort exposure is different for ethnic minorities and that older people reporting a minority ethnic identity have more diverse histories of exposure to risk factors, or are not affected by the same health-selecting processes of residential segregation.

The rationale for stratifying our analysis by ethnic group resides in the substantial interest in understanding variations in health care need across different population groups. Since the LLTI indicator has been used as a proxy for health care need (Vallejo-Torres et al., 2009), it is interesting to understand whether care needs of different ethnic groups are stable across different places. Our findings are in line with work by Finney et al. (2014) and confirm that knowledge on patterns and determinants of local ethnic health gaps remains insufficient. Overall, disaggregation of ethnicities reveals more variation than would arise purely out of the combination of local age characteristics and chance. Our finding is also consistent with previous investigations by Shouls et al. (1996b,a) relying on factor analysis to classify LADs with respect to known area-level aggregate health estimates. Our analysis of spatial autocorrelation patterns confirms that even when accounting for other common population health measurements such as rates of hospitalisation and mortality, which we assume capture important unobserved risk factors, the significant

remaining between-area heterogeneity still exhibits strong, almost unaffected spatial patterns in a way that is specific to each ethnic group. This is a sign of very different health needs and has been identified as an important area of current research (Nazroo, 2014).

We can draw implications for predictive modelling. In addition to measuring disparities in health needs which are not already contained in mortality and hospitalisation statistics, the distinct spatial pattern of overdispersion in the final model confirms the importance of reviewing assumptions on random effects in multilevel health models. While the assumption of spatially independent residuals may be sufficient in many descriptive epidemiology studies, it introduces risks of substantial variation and clustering in the quality of small area prediction across space, especially in the presence of underpowered sample data. This has seldom been raised as a validity issue with disease prevalence prediction models (Scarborough et al., 2009) though the importance of testing for the existence of significant between-area heterogeneity was noted by Datta et al. (2011).

This paper gives a practical illustration of the implications of assuming independence of random effects across areas. In our results, the degree to which the fixed part of the model underestimates or overestimates odds of LLTI is highly dependent on error in neighbouring areas. It implies that, in the absence of sufficient individual-level auxiliary data (e.g. from a census) or area-level predictors (e.g. statistics on health utilisation, social or occupational characteristics), there is a greater need to explicitly model these spatial structures not just using covariate adjustments, but also incorporating spatial information explicitly into regression models. The literature has identified several routes for doing so (Dale and Fortin, 2014). A common stochastic approach is the use of spatial or conditional autocorrelation functions, by introducing spatial matrices into the model's covariance structure (Besag, 1974). A competing approach is the use of spatial trend surfaces (polynomial functions of the geographic coordinates, Adams and White, 2006, or Euclidean distance matrix eigenvalues (Borcard and Legendre, 2002; Legendre and Fortin, 2010; Voutilainen and Sherwood, 2015) as regressors in a standard generalised linear model.

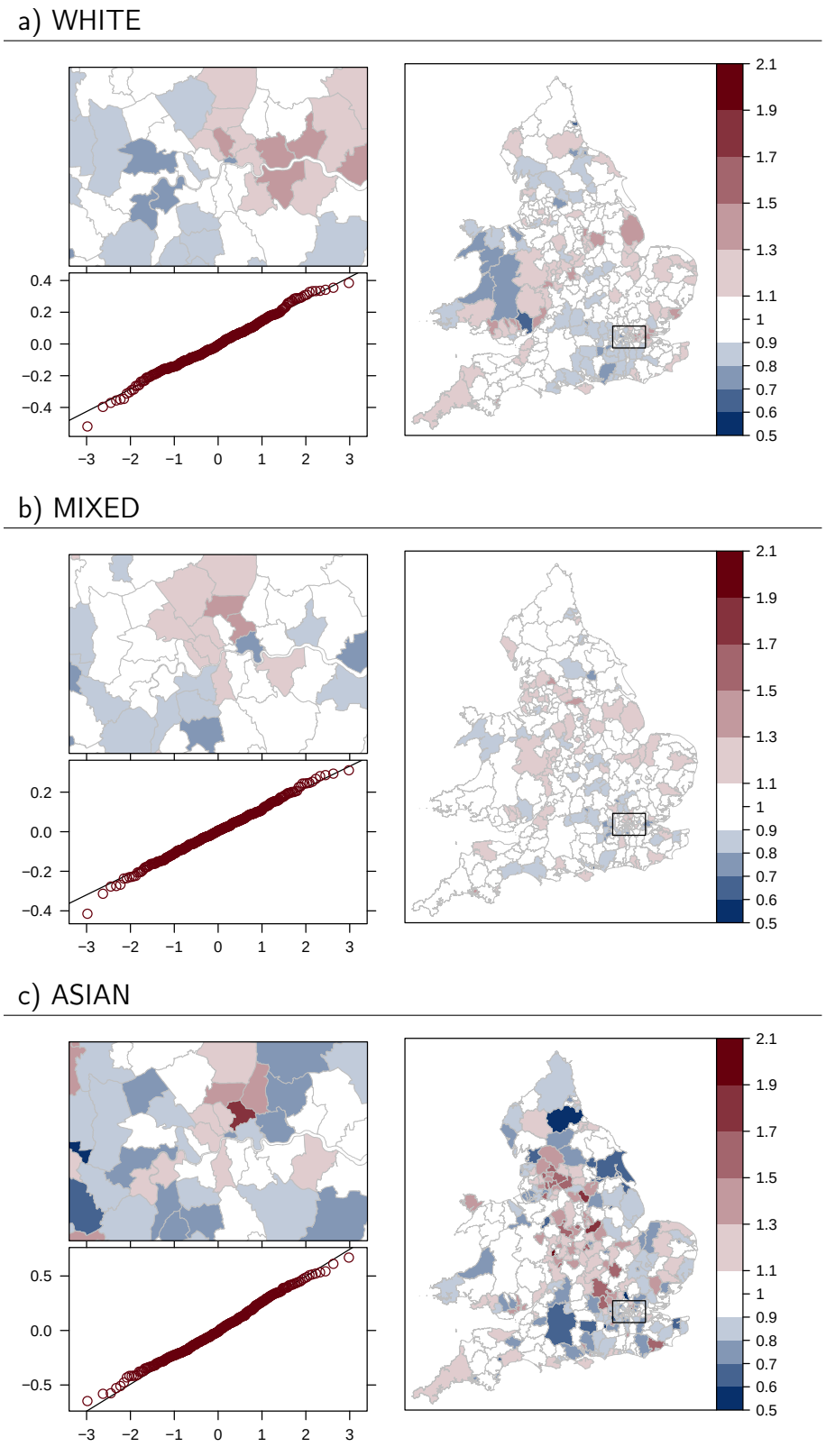
All these techniques presuppose that the structure underpinning spatial processes in the data is well understood. In addition to reviewing existing structures defined around both adjacency and proximity (k-nearest method), the purpose of this paper was to test the relevance of an alternative definition of spatial structure based on residential migration. Levels of autocorrelation reached with this method indicate that while it is not the best fitting method for LLTI, it does capture a non-negligible spatial interaction.

Incorporating spatial information explicitly into regression models requires good prior knowledge regarding the study outcome. The main benefit of using a large population source, such as the census in this paper, is to be able to conduct additional tests on local levels of autocorrelation. In our case, substantial local clusters were apparent in the residuals for the Asian model, suggesting that the range of covariates used was particularly inappropriate to predict local odds, even once global autocorrelation was taken into account. This constitutes further evidence of the need to better understand the spatial structure of chronic conditions.

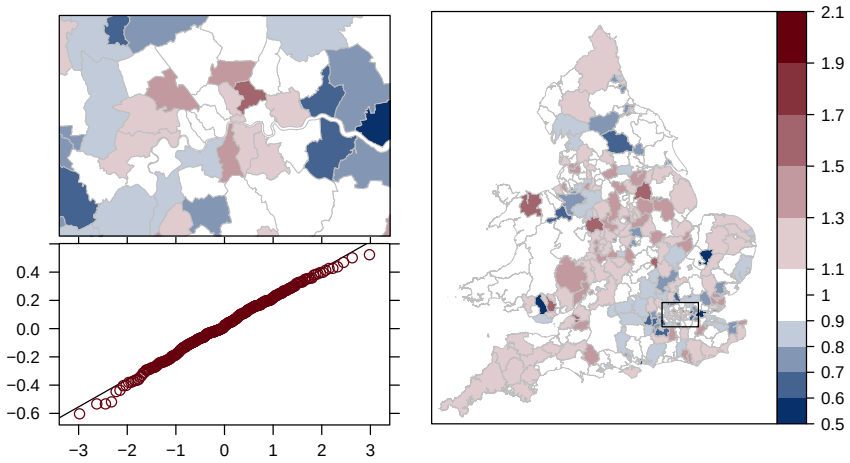
Our study indicates that SAE remains a data intensive task. It remains difficult to predict LLTI with simple models without introducing socioeconomic information on local populations from a source such as the census (Darlington et al., 2015). Looking at between-area heterogeneity (Figure 6.1), it is apparent that geographical inequalities remain which prove difficult to predict. With these models, about 46.8 percent of LADs require an adjustment of the fixed part of the model by at least ±10 percent for White populations while 15.8 percent require an adjustment of at least ±20 percent in order to reach the actual odds of LLTI. The latter figure is of 9.5 percent for Mixed ethnic groups, 15.8 percent for Other ethnic groups, 33.6 percent for Black populations and as high as 40.2 percent for Asian populations. A reasonably large sample of data is required for every area of interest in order to reach a precise predictor of random parameters $\boldsymbol{v}$. This has implications for power calculations to obtain good quality empirical best predictors in presence of such residual variability. Moreover there are also questions around the properties of synthetic estimators, which only make use of the fixed part of the prevalence model, in situations where the between area residual variance σ^2 is not negligible. Such estimators currently underpin the majority of UK disease prevalence models. These issues point to the importance of tests for heterogeneity recently examined by Datta et al. (2011) and Molina et al. (2015).

While models can help produce estimates for small populations, hypothesis testing can prove limiting. The range of possible small area model specifications is virtually limitless. Shortcomings are likely to arise especially in cases where models demonstrate similar levels of unexplained variance and spatial clustering in their random part. This highlights the importance of large-scales studies such as censuses in providing reliable auxiliary information for small groups.

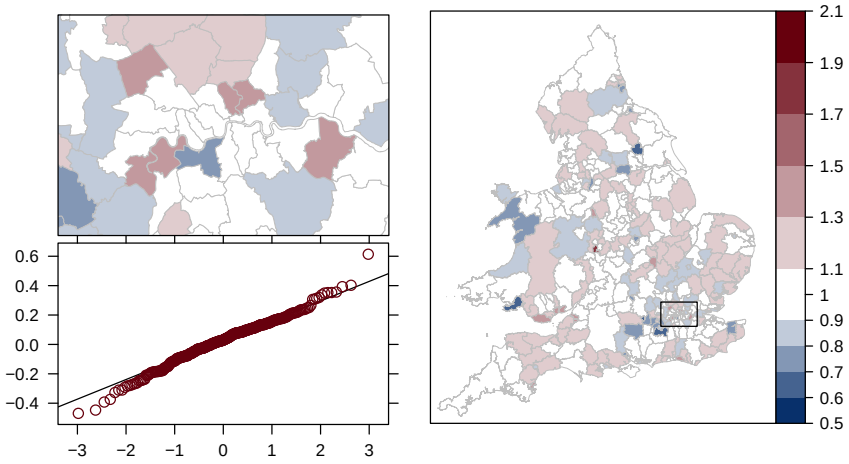
Figure 6.1 | *Model (M2):* Q–Q plots of area residuals against a normal distribution and maps of transformed residuals e^{v_d} (odds ratio scale)



d) BLACK

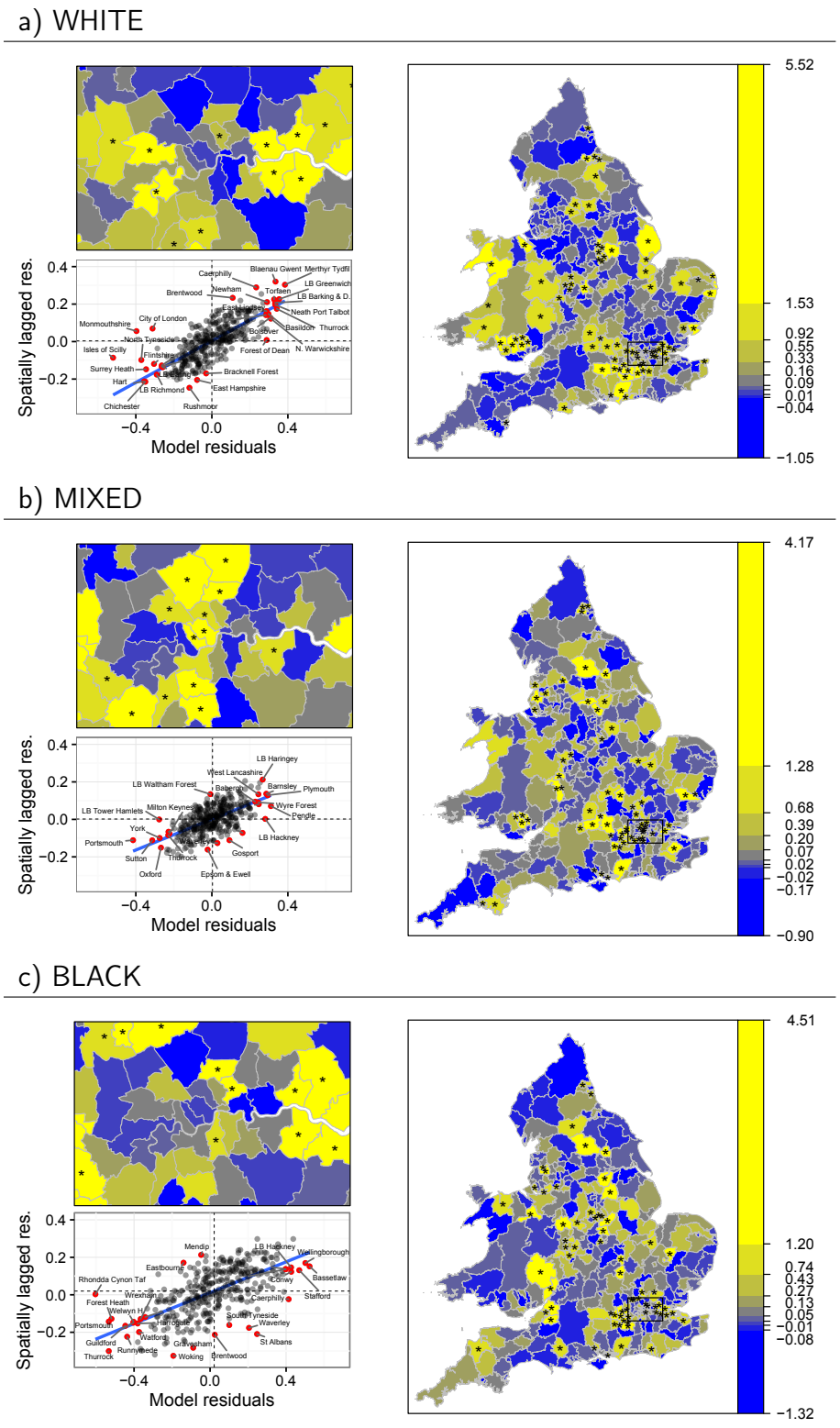


e) OTHER

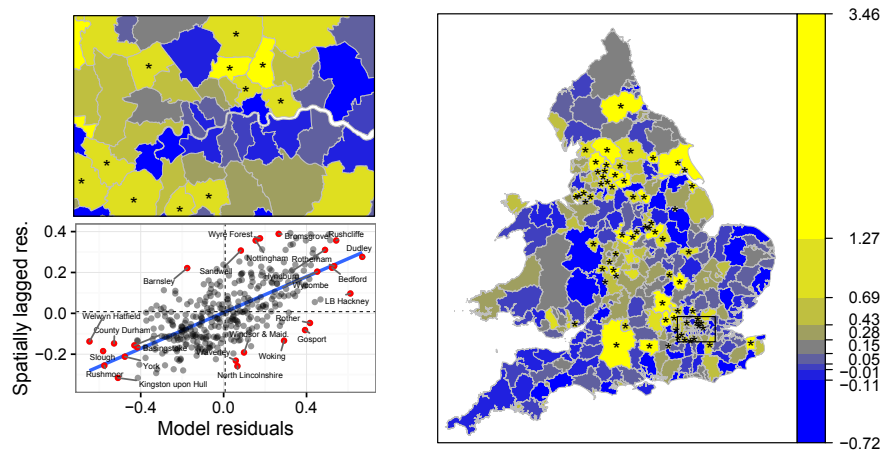


Plots: Residuals of model *(M2)* are compared to a theoretical normal distribution with the same mean and standard deviation to assess normality.
Choropleths: Model residuals are converted on the odds ratio scale using the exponential function to map heterogeneity in odds of LLTI across areas once differences in covariates are taken into account. Shades of *red* (*blue*) signal areas where the prevalence of LLTI is *higher* (*lower*) than expected given local age characteristics, area classification and rates of emergency hospitalisations.

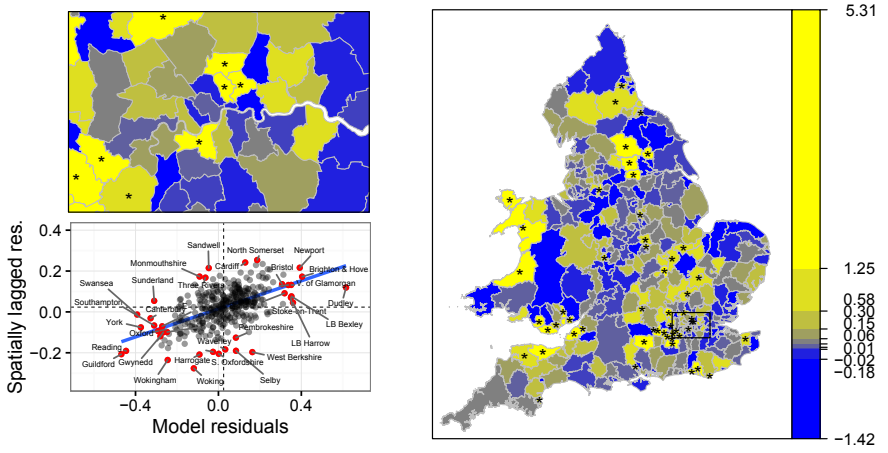
Figure 6.2 | *Model (M2):* Maps of LISA with significant clusters (asterisks) and Moran scatterplots of area residuals



d) ASIAN



e) OTHER



Moran scatterplots Global spatial clustering of LLTI is represented graphically as the relationship between area residuals (on the logit scale) and the spatially lagged area residuals. Some neighbourhoods exhibit *higher*-than-average clustering and appear above the line of best fit. *Significant clusters* are marked with a *red dot*.
Choropleths: Shades of *yellow* indicate areas with a *high* LISA, while shades of *blue* indicate areas with a *low* LISA. Statistically significantly *higher*-than-average LISAs are marked with an *asterisk* (*) and indicate presence of a statistically significant spatial cluster at the 95 per cent confidence level

6.5 Conclusion

The key contributions of this paper relate to (i) new descriptions of the spatial structure in LLTI both in terms of dispersion and autocorrelation and (ii) implications for predictive modelling and SAE.

With regard to the first point, we present greater disaggregation than previous investigations in this area (Lloyd, 2015; Shouls et al., 1996b,a) and emphasise the importance of ethnicity and alternative conceptualisations of 'spatial structure'. We provide a systematic analysis of best-fitting spatial structures and give an applied example of a new method to build adjacency matrices using migration data. Further research could examine the predictive power of disaggregating migration interaction according to demographic characteristics (age and ethnicity being strong determinants in spatial terms). It would also be worth considering spatial interaction beyond the notion of symmetry, by examining hypotheses where *A* being a 'neighbour' of *B* does not imply the reciprocal. Alternative approaches have proposed treating spatial weights as random parameters to be estimated rather than as fixed data (Lee and Mitchell, 2013). This may reduce subjectivity in model specification, arguably at a certain computational and precision cost.

Our second contribution concerns the applied relevance of the paper to concerns related to predictive modelling, particularly around planning efficient SAE strategies. In the UK, there is sustained interest in information for small geographical areas (ONS, 2015a), in a context where local population health surveys have almost entirely disappeared due to rising fieldwork costs and falling nonresponse (McCluskey and Topping, 2009). Persistent, and often widening health inequalities are a concern internationally (Mackenbach, 2003; France. Haut Conseil de la Santé Publique, 2009; Bleich et al., 2012), with improvements in small area public health monitoring among major policy recommendations to tackle such problems (on Social Determinants of Health, 2008).

Subnational monitoring of morbidity levels raises particular statistical challenges. Our results show that geographical variability in the odds of LLTI are greater than expected not only from sampling error and differences in local populations' age distributions, but also in relation to levels of mortality and healthcare utilisation. Odds of LLTI also exhibit a larger between-area variance for ethnic minorities compared to White populations.

From a methodological viewpoint, we acknowledge limitations commonly encountered in disease mapping. In addition to data limitations themselves (insufficient disaggregation of age bands, quality issues usually expected from hospital data), this

research relies on complex models. Only taking into account main fixed effects of model *(M2)*, the number of candidate models is 2^{12}. When taking into account the different types of hospital and mortality covariates, the number of possibilities rises to millions. Specifications with complex random effects and spatial autocorrelation structures could in addition be considered, raising this number even higher. Overall, this concern, well-identified in predictive modelling (Austin and Tu, 2004; Greenland, 1989; Twigg et al., 2000), represents a challenge in transparency and reproducibility of public health information.

Overall, our results emphasise the importance of detailed contextual information on population characteristics and spatial structures in the production of working models that can be trusted to hold for the whole population. Modelling techniques can be applied which make use of the spatial clustering illustrated in this paper to improve prediction. Yet, like empirical prediction, these require access to good-quality survey data with individual geographical identifiers (for instance postcode sectors) for all of the targeted small areas. In the future, geographic masking (Zandbergen, 2014) may offer safer alternatives in situations where geographical identifiers are too disclosive to be released. This study also highlights the importance of local health care statistics to improve the predictive capability of models. Further disaggregation of these data sources by ethnic group and groups of medical conditions at the local level is likely to help improve future disease prevalence models.

Chapter 7

Bayesian Sample Size Determination for Planning Hierarchical Bayes Small Area Estimates

Abstract

This paper devises a fully Bayesian sample size determination method for hierarchical model-based small area estimation with a decision risk approach. A new loss function specified around a desired maximum posterior variance target implements conventional official statistics criteria of estimator reliability (coefficient of variation of up to 20 per cent). This approach comes with an efficient binary search algorithm identifying the minimum effective sample size needed to produce small area estimates under this threshold constraint. Traditional survey sampling design tools can then be used to plan appropriate data collection using the resulting effective sample size target. This approach is illustrated in a case study on small area prevalence of life limiting health problems for 6 age groups across 1,956 small areas in Northern England, using the recently developed Integrated Nested Laplace Approximation method for spatial generalised linear mixed hierarchical models.

Keywords: Sample size determination, small area estimation, disease prevalence models, hierarchical Bayes, Integrated Nested Laplace Approximation, official statistics, national statistics.

7.1 Introduction

When small area estimates from survey samples are imprecise, model-based estimation is often used to borrow strength and produce more efficient estimators. HB prediction, in particular, has been shown to yield good results and make efficient use of available survey samples (Ghosh and Rao, 1994). When it comes to applying this method to real world problems, an important question faced by practitioners is how to determine the sample size needed to produce estimates achieving a desired precision, generally expressed as a target standard error or RSE. Although some analytical solutions have been proposed in the model-assisted literature, this problem has remained relatively unexamined in model-dependent SAE.

This paper is organised as follows: in section 7.2, we briefly review existing work on Bayesian SSD for clinical trials and other investigations, and draw implications for planning model-based SAE studies. Section 7.3 describes a model-based simulation procedure to determine effective sample size (ESS) requirements under a maximum relative posterior variance constraint. This approach is applied in section 7.4 using 2011 UK census data on chronic health and a spatial hierarchical generalised linear mixed model. A design-based simulation study confirms both (a) the validity of solution produced by the SSD and (b) efficiency savings achieved over traditional survey sampling direct estimation. Results are presented in section 7.5 and inform a discussion of the influence of Bayesian priors and applicability to more complex sampling designs in section 7.6.

7.2 Review of SSD for hierarchical models

In the frequentist approach, target sample sizes are usually determined by reference to the sampling distribution of the target parameter under a given survey sampling plan. With complex model-assisted or model-dependent statistical designs, this sampling distribution is typically unknown. This is particularly the case when a study examines a model parameter under a multivariate statistical model, for instance relative risks in a survival model with covariate confounding.

Amongst those complex statistical designs are two-stage hierarchical models commonly needed to estimate some kind of effect in multicentre randomised controlled trials or in multilevel observational studies (e.g. studies of educational attainment across classes and schools). This effect can be a treatment effect, a regression coefficient or some other cluster-specific characteristic. We distinguish between two categories of studies based on their inferential motivation:

1. studies aiming to detect a global effect with a predefined statistical power;
2. studies aiming to produce an estimate of the *effect size* within each centre/cluster with a predefined precision.

7.2.1 SSD for detection

In category (1), success is defined as collecting a sufficient number of observations across a sufficient number of centres or clusters so that the target statistical power is achieved in the model of interest. Achieving the target statistical power allows study investigators to reject a null hypothesis if the size of the treatment effect exceeds a clinically meaningful threshold set at the design stage. In observational study designs, examples are varied which optimise, for instance, the number of clusters and the number of units within clusters. Several software applications now exist across the educational, behavioural and wider multilevel literature (Snijders, 2005; Cools et al., 2008; Browne et al., 2009; Zhang and Wang, 2009). In interventional research, including clinical trials, some design problems have closed-form solutions (Raudenbush, 1997), while others are more complex and require computer-intensive Monte Carlo experiments to reach a solution. Joseph et al. (1995) proposed algorithms relying on a binary sample size search between 0 and the frequentist binomial sample size requirement to optimise the SSD solution at a reasonable computational cost. The adequate sample size is determined as the smallest sample size meeting the chosen constraints.

7.2.2 SSD for prediction

The present paper is concerned with studies belonging to category (2), in which success is defined as collecting a sufficient number of observations to estimate the treatment effect with a predefined precision within each centre or cluster. Unlike the first category of designs, research on SSD for a vector parameter of interest resulting from a working model (rather that a simple scalar) is not as developed. Design methodology has mostly been considered from the angle of observational studies, particularly clinical quality audits. Normand and Zou (2002) explored in

a series of simulation studies the effect of the number of clusters as well as of the sample size on the efficiency of audits under beta-binomial hierarchical models borrowing strength and adjusting for confounders. Zou and Normand (2001) and Zou et al. (2003) proposed an SSD algorithm applicable to three-stage models in order to identify optimal designs for hospital benchmarking clinical audits. Their design criterion is an upper bound for the average width of posterior intervals over all study centres. Their method involves a grid search of the minimum target sample size (a sample size sweep in a predetermined sequence of candidate values). It is implemented with Monte Carlo simulation using the prespecified model, in turn: sampling from the model priors; simulating a realisation from the model; predicting the corresponding parameter; and computing the level of compliance with the target posterior interval width. Bayesian decision theory and Monte Carlo solutions of this kind are attracting growing interest for complex SSD problems in hierarchical designs for medical studies. Another area of application of Bayesian SSD techniques is the design of spatial point patterns data collection. Diggle and Lophaven (2006) have examined the design of spatial sampling strategies with simple models, and conclude that sampling simulations are inevitable, albeit computationally burdensome. The authors nevertheless recognise that new model estimation methods are likely to reduce this burden.

7.2.3 SSD for small area studies

To the best of our knowledge, SSD techniques have seldom been applied to plan model-based SAE studies, that is when a single working model is used to compute one or more predictors within each small area in a given study population. Empirical or hierarchical Bayes prediction is required for each of those small areas. Although simulation studies are ubiquitous in the SAE literature, they have been used either to illustrate efficiency gains obtained from more complex modelling designs (Jonker et al., 2013a; Porter et al., 2015c; Ross and Wakefield, 2015), or to validate a small area model against historical data (Barker and Thompson, 2013). Closed-form SSD solutions are only available for the simplest models: see Molefe (2011) in the model-assisted literature, and Raudenbush (1997) who looks at implications on SSD of introducing a model covariate. The availability of SSD methods for real-world statistical needs is yet particularly essential with model-based estimation since precision depends not just on the type of predictor selected, but on covariates and random components included in the model (Rao and Choudhry, 2011; Rotondi and Donner, 2009). This is in part because the MSE of predictors depends not solely

on a known sampling distribution, but also on a working model incorporating predictive structure. No such methods are available for emerging developments such as spatial smoothing models (Pratesi and Salvati, 2008; Gomez-Rubio et al., 2010), or small area models borrowing strength from time (You and Zhou, 2011), age or cohort effects (Congdon, 2006c).

7.2.4 Prerequisites of SSD

SSD presupposes (1) design constraints in the form of a set of criteria against which to optimise sample design; and (2) prior information on the population of study.

The implementation of design constraints is straightforward in closed-form frequentist solutions which focus on estimator variance targets. In fully Bayesian SSD, design criteria are often treated as a decision rule determined by a loss function (Adcock, 1988, 1997). A typical decision rule for studies preoccupied with hypothesis testing is based on a function of statistical power (opposite of the false negative rate) and significance level (opposite of the false positive error rate), as illustrated by Sahu and Smith (2006). Yet other rules have been proposed for detection studies; Joseph et al. (1995) considered three such design objectives: achieving a desired average coverage probability for highest posterior density intervals; satisfying a maximum average length for those intervals; and a combination of the two constraints. As for studies interested in prediction, decision rules have in majority been specified in relation to the width of interval estimates; see for instance Joseph et al. (1995) and Zou and Normand (2001). These overlap conventional government survey design criteria defined in terms of precision, either based on an estimator's margin of error or RSE. For instance the 2000/01 English Local boost of the LFS was designed with a frequentist approach to ensure an acceptable precision, defined as a maximum RSE of 20 per cent for design-based estimators of economic activity headcounts of 6,000 districts (Hastings, 2002, p. 40). A model-based equivalent is the relative root MSE, while a fully Bayesian equivalent would be to consider a function of the relative posterior variance, for instance the number of districts which fail to meet a maximum relative posterior variance threshold.

The second prerequisite of SSD, namely prior information around the parameter of interest, is treated very differently in the frequentist and the Bayesian apparach (Adcock, 1997). In the frequentist paradigm, SSD is entirely determined by the sampling distribution of the study target parameter, which itself depends on its population variance, which is typically unknown. In the absence of knowledge

regarding the population variance of the study parameter, the standard frequentist approach to SSD involves plugging an assumed value for this variance in a closed formula. The outcome is therefore entirely dependent on how conservative this assumption is and it is often necessary to overestimate the population variance. In contrast, fully Bayesian SSD does not handle unknown parameters using a plug-in method but instead using explicit priors and hyperpriors (Adcock, 1997). The literature contains a variety of examples. When determining a sample size for a binomial proportion, Joseph et al. (1995) and Zou and Normand (2001) set the scale and rate (hyperparameters) of the beta distribution (prior) believed to determine an overdispersed binomial distribution of interest. In biomedical research, such hyperpriors are generally elicited from pilot data, previous studies and subjective expert opinion (Spiegelhalter and Freedman, 1986).

At the first glance, particularly when relying on pilot or historical data to form a prior, it seems intuitive to use a single prior for both the design and the analysis. In other words, the set of priors used to simulate prediction under various sample size scenarios is also incorporated in the working model. Yet Spiegelhalter and Freedman (1986) and Sahu and Smith (2006) have argued in favour of separating design and fitting priors on scientific grounds. Regulations on biomedical trials sometimes impose that the data are analysed under a state of pre-experimental knowledge, that is without incorporating knowledge from data produced in previous studies. Although such historical data can be valuable in optimising SSD, it is not necessarily desirable to introduce them in the analysis itself as this can be left to subsequent meta-analytic studies. Similar constraints may exist in official statistics, where the reliance on informative priors is sometimes subject to objections (Fienberg, 2011). When the only available source of prior knowledge is historical data, there may be reasons to restrict its use to SSD. This provides further assurance in the solution of SSD while determining sample requirements to produce a sufficiently precise estimate without having to pool data from previous statistical bulletins. This is by no mean the only way to proceed, but it is sometimes desirable to treat the elicitation of design priors and fitting priors separately.

On the one hand, design priors retain a strong influence on the outcome of any SSD procedure and its success. Particular attention has been given to robust prior elicitation—that is priors that do not convey excessive confidence compared to the existing knowledge, and which are flexible enough to offer protection against misspecified models. With regard to SAE, priors can be elicited from previous survey waves or pilots: routine government survey data are typically abundant. In principle, the most simple type of design prior can consist of hyperparameters of

random components since they determine the level of shrinkage in HB prediction and, by way of consequence, posterior variance. Such hyperpriors can be elicited from appropriate marginal posteriors obtained from the combination of pilot data with a vague uninformative prior (which can be the fitting prior). Due consideration must be given to how much belief can be placed into the stability of these parameters across years or across surveys, and it may be necessary to apply a small discount to their influence (see De Santis 2007).

On the other hand, fitting priors can be everything between vague and informative. The use of uninformative hyperpriors for random components is not always an option as it often leads to difficulties in estimating models. Both Markov chains Monte Carlo and integrated nested Laplace approximation (INLA) encounter numerical difficulties with complex models, especially with spatial models. Fong et al. (2010) suggest specifying weakly informative priors for Gaussian random effects by using the log Student t distribution (with one or two degrees of freedom) and predefined lower and upper bounds for the range of 95 per cent of realisations. Hyperpriors can then be deduced in the form of the scale and rate of a Gamma distribution. More recently, Simpson et al. (2015) have addressed more complex models with a combination of random effects and proposed a weakly informative 'penalised complexity prior' based on some belief of the scale of random effect (standard deviation).

7.3 Model-based SSD

We consider the estimation of a small area characteristic (e.g. economic status, illness, income, marital status) under a given working model $\mathcal{M}$. Let the population U be partitioned into small areas $d = \{1, \ldots, D\}$. $\mathbf{N} = \{N_1, \ldots, N_D\}$ and $\mathbf{Y} = \{Y_1, \ldots, Y_D\}$ respectively denote the population size in area d and the characteristic total in area d: this can be the area headcount of individuals with the given characteristic, or the area total (such as total income). We are interested in estimating population means $\overline{Y}_d = Y_d/N_d$ using

- hierarchical working model $\mathcal{M}$, to produce an HB predictor of $\overline{Y}_d$ notated θ_d;
- data from auxiliary covariates $\mathbf{X}$ available for the entire population;
- a sample survey s to be designed.

We seek to determine f, the effective sampling fraction for s, using Bayesian SSD under some design constraints. We remark that area-specific ESSs n_d are such that $n_d \sim \text{Binomial}(N_d, f)$.

7.3.1 Sample size criteria and Bayes decision rule

A conventional frequentist criterion of statistical reliability in official statistics is the RSE or CV. A possible Bayesian equivalent is the relative posterior variance of the HB predictor θ_d. Tabular cells of estimates with a relative posterior variance in excess of 20 per cent are to be suppressed from statistical publications. It is desirable that the overall rate of cell suppression remains low, for example below a threshold $\mathfrak{z} = 0.01$. We implement this requirement through a simple design loss function: the proportion of cell suppression in the dataset weighted by the population headcount of those cells. The total loss $\ell(f \mid \ldots)$ can be thought of as the total headcount of populations eligible to a reliable estimate, whose estimates have to suppressed due to reliability concerns. The weighting introduces a form of trade-off, which prioritises reliable estimates for large populations while being more tolerant of the risk of cell suppression for the smallest cross-classifications, which are inevitably the most demanding in terms of data collection.

$$\ell(f \mid \mathcal{M}, \pi(\tau_\gamma), \pi(\tau_\upsilon), \pi(\tau_\nu), \ldots) = N^{-1} \sum_d [N_d \, I(\mathrm{RSE}(\theta_d) > 0.2)] \tag{7.1}$$

In this expression, $I(\cdot)$ is the indicator function. The overall loss is an intractable function of the sampling fraction f and is conditional on both the working model $\mathcal{M}$ and the set of design and fitting priors $\pi(\cdot)$. Though ℓ has no obvious closed-form expression, it is reasonable to take the premise that it is a monotonically decreasing function of n (for a discussion of the design consistency of HB prediction, see Lahiri and Mukherjee 2007). This facilitates the evaluation of integral $\int \ell(f \mid \ldots) \; df$ over a reasonable interval of sampling fractions $[a, b]$.

Many more specifications can be considered for the loss. ℓ can be defined with respect to the width of prediction intervals such as highest posterior density intervals rather than RSE. It can also be restricted to a subset of domains d in the event that quality standards set by statutory or funding requirements do not apply to all domains d.

Depending on design constraints and costs, it is possible to envisage more sophisticated loss functions inspired from optimal sample design or the 'value of information' approach. In particular, it is conceivable to attribute a price to cell suppressions or to penalised ℓ by a marginal cost function of increasing the sampling fraction. These constitute options for more holistic decision rule and lead to an optimum between cost saving and exhaustive publication. Provided that such refinements strictly depend on f and posterior means or variances, they add no further

dimension to the SSD equation, and the approach described in this paper should remain entirely applicable.

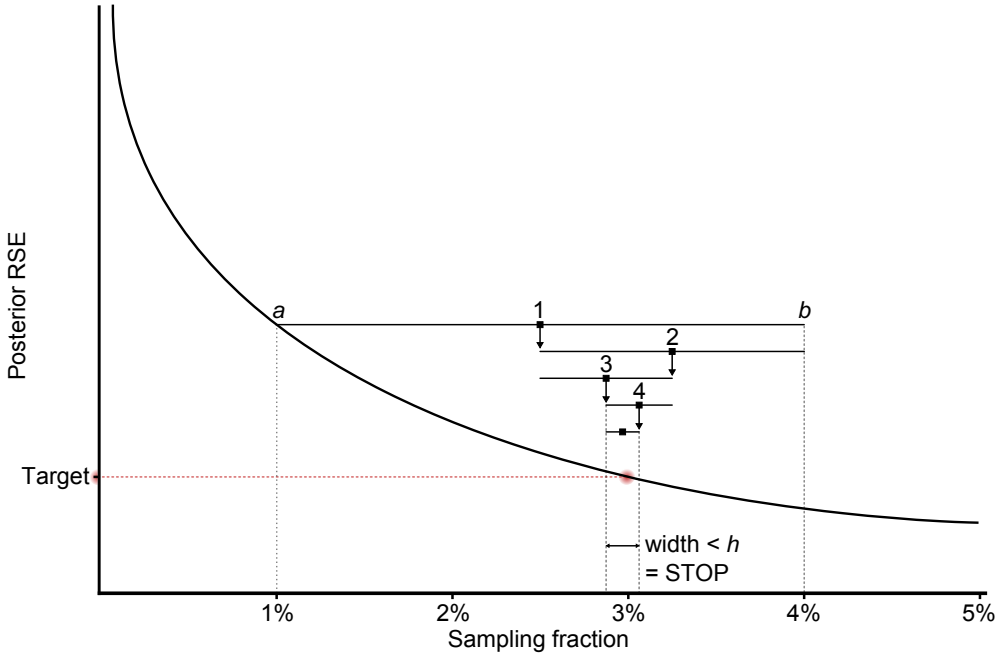


Figure 7.1 | Graphical illustration of the binary algorithm

7.3.2 Sample size minimisation

Simulating the relative posterior variance for all possible sample sizes until the minimum quality standard is attained following a sweep approach is computationally cumbersome. We instead opt for a binary search algorithm (see Figure 7.1 and Algorithm 2) to minimise $\ell(f)$. By progressing iteratively towards the solution with steps of decreasing sizes, we considerably reduce the number of simulations. The algorithm starts with f_1, the midpoint of the interval $[a, b]$, and evaluates $\ell(f_1)$ over many replications of the sampling and model estimation process. This process is then replicated on interval $[a, f_1]$ if $\ell(f_1)$ can be trusted to fall below the maximum tolerated suppression setting $\wp$ or $[f_1, b]$ in the opposite case—that is, if more than $\wp$ per cent of sampling simulations return a strictly positive value for $\ell(f_1)$. The interval of possible solutions f_k is therefore halved repeatedly until the interval width is small enough to be used; this is determined by a tolerance setting h. The upper bound of the final interval is the target ESS solution. The algorithm can be restarted on the final interval to reach an even narrower h. By construct, the number of iterations required to complete the algorithm is the smallest integer $k_{\max}$ such that $k_{\max} > (2h)^{-1}(b - a)$.

It is worth emphasising that this procedure relies on ℓ being a monotonous (not strictly) decreasing function of sampling fraction f. The algorithm progression is mainly determined by a risk setting γ, the probability that cell suppression will exceed $\wp$ for any given f_k. The number of simulations L ordered for each sampling fraction f_k should be set in relation to γ; we suggest that the order of magnitude of L should always be at least equal to the order of magnitude of γ^{-1}.

Input: s_0, $\pi(\theta_d \mid s_0)$, f_a, f_b, $\ell(f_a)$, $\ell(f_b)$, h, $\wp$, L, γ
$k \leftarrow 0; l \leftarrow 0;$
repeat
  $k \leftarrow k + 1;$
  $f_k \leftarrow (f_a + f_b)/2;$
  repeat
    $l \leftarrow l + 1;$
    sample $\theta_d^\star$ from prior $\pi(\theta_d \mid s_0)$;
    simulate $n_{dl}^\star \sim \text{Binomial}(N_d, f_k)$;
    simulate $n_{dl}^\star$ realisations $\bar{y}_{dl}^\star$ from the likelihood of $\theta_{dl}^\star$;
    fit model $\mathcal{M}_l^\star$ on data $\bar{y}_{dl}^\star$;
    estimate posterior density $\pi(\theta_{dl}^\star | y_{dl}^\star, \mathcal{M}_l^\star)$;
    compute relative posterior variances $\widehat{\text{RSE}}(\theta_{dl}^\star)$;
  until $l = L$;
  $\ell(f_k)_l \leftarrow N^{-1} \sum_{d=1}^{D} \left(N_d \, I \left[\widehat{\text{RSE}}(\theta_{dl}^\star) > 0.2 \right] \right)$;
  if $\Pr(\ell(f_k) \leq \wp) < \gamma$ **then**
    $f_b \leftarrow f_k$
  else
    $f_a \leftarrow f_k$
  end
until $f_b - f_a < h$;
return $f_a, f_b, \ell(a), \ell(b)$

Algorithm 2 | Binary SSD algorithm

7.4 Case study

7.4.1 Context

The approach described in section 7.3 is illustrated on a typical tabular ONS census output. This case study examines sample size requirements and efficiency gains under small area models predicting the age-specific prevalence of LLTI across small areas in Northern England. The UK government having announced its goal to discontinue full population enumeration through traditional decennial censuses

Table 7.1 | Example of tabular census output for health characteristics

	LLTI status									
	LLTI			No LLTI			All			
	1	⋯	J	1	⋯	J	1	⋯	J	Total
1	Y_{11}	⋯	Y_{J1}	$N_{11} - Y_{11}$	⋯	$N_{J1} - Y_{J1}$	N_{11}	⋯	N_{J1}	N_1
⋮	⋮	⋱	⋮	⋮	⋱	⋮	⋮	⋱	⋮	⋮
D	Y_{1D}	…	Y_{JD}	$N_{1D} - Y_{1D}$	…	$N_{JD} - Y_{JD}$	N_{1D}	…	N_{JD}	N_D
Total	$Y_{1\cdot}$	⋯	$Y_{J\cdot}$	$N_{1\cdot} - Y_{J\cdot}$	⋯	$N_{1\cdot} - Y_{J\cdot}$	$N_{1\cdot}$	⋯	$N_{J\cdot}$	N

Legend
assumed to be known in case study
quality standard enforced in case study

(Maude, 2014), it is expected that small area population health characteristics will in the future have to be estimated using sample surveys and administrative sources. The aim of the present case study is twofold: (i) to apply the SSD procedure set out in section 7.3.2 to reproduce 2011 census health outputs under conventional UK standards of statistical reliability; (ii) to assess efficiency gained from incorporating more advanced hierarchical prior structures into small area models (borrowing strength from neighbouring areas and age groups).

Future census quality standards published by the ONS (2014b, p. 44) demand a maximum RSE of 20 per cent for population headcounts representing at least 3 per cent of an MSOA's population. Assuming an average resident population of 8,000 across MSOAs, this criterion is equivalent to a margin of error of up to ±94 for a headcount of 240 persons. MSOA zones were originally designed to ensure lower and upper population limits of 5,700 to 11,100 usual residents. By 2014, 10 per cent of these zones no longer met such limits (ONS, 2015b). As a consequence, the future ONS census quality standards in effect require a much smaller margin of error for the smallest MSOAs.

7.4.2 Design constraints on a tabular output

The case study population U is partitioned into MSOAs $d = \{1, \ldots, D\}$ and age groups $j = \{1, \ldots, J\}$. $\mathbf{N} = \{N_{11}, \ldots, N_{JD}\}$ and $\mathbf{Y} = \{Y_{11}, \ldots, Y_{JD}\}$ respectively denote headcounts of members of private households of age group i residing in area d and the corresponding number of individuals reporting an LLTI. We are interested in the LLTI prevalence proportions $p_{jd} = Y_{jd}/N_{jd}$. Table 7.1 illustrates the typical tabular census output corresponding to this data.

In the present case study, the ONS's census quality threshold is applied to tabular cells Y_{jd} (headcounts of individuals reporting an LLTI) but not to cells $N_{jd} - Y_{jd}$ (headcounts of individuals reporting no LLTI). Although the ONS quality standard would normally apply to these cells as well, we consider that most public health statistics producers would only publish prevalence or disease headcounts estimates, not estimates of disease-free headcounts. For simplicity, we neither predict nor test the precision of LLTI-free cells. Furthermore, we set the tolerable cell suppression rate to $\mathfrak{z} = 0$. Because we aim for no suppression at all, there are no longer any benefits in weighting the loss function (7.1) by the underlying population sizes, as discussed in section 7.3.1. In those conditions, we substitute the loss function used at line 12 of Algorithm 2 with the simpler crude proportion of suppression below:

$$\ell(f_k)_l \leftarrow \sum_{j=1}^{J} \sum_{d=1}^{D} \left(I\left[\widehat{\mathrm{RSE}}_{jdl} > 0.2\right]\right) \tag{7.2}$$

Since the quality threshold does not apply to all cross-classifications, the above loss ℓ is only computed over cross-classifications jd such that $Y_{jd}/N_d \geq 0.03$. Because Y_{jd} is in reality unknown, we compute the estimated loss $\hat{\ell}$ over cross-classifications jd such that $\theta_{jd}/N_d \geq 0.03$. We nevertheless report both the true and the estimated loss in results in order to assess any potential divergence in the algorithm's progression.

Using data from the 2011 UK census table LC3101EWLS (ONS, 2013c), 18 age groups were collapsed in order to bring Y_{jd}/N_d to an average level close to 3 per cent or more. With $D = 1,956$ and setting $J = 6$, we have a total $JD = 11,736$ cells, 49 per cent of which are eligible for the quality standard of a 20 per cent RSE. Distributional characteristics of prevalence proportions are plotted in Figure 7.2. Other settings are configured at $f_a = 0.01$, $f_b = 0.04$, $h = 0.01$ and $L = 100$. The acceptable risk of exceeding the tolerable rate of suppression $\mathfrak{z}$ is set to $\gamma = 0.01$.

7.4.3 Model specification

We consider the below working model (7.3). Sample counts y_{jd} are treated as the realisation of a binomial distribution with sample size n_{jd} and success parameter $p_{jd} = \mathrm{logit}^{-1}(\theta_{jd})$.

$$\begin{aligned} y_{jd} &\sim \text{Binomial}\left(p_{jd}, n_{jd}\right) \\ \text{logit}\left(p_{jd}\right) \equiv \theta_{jd} &= \boldsymbol{\beta}_j^{(1)} + X_d \boldsymbol{\beta}_j^{(2)} + \upsilon_d + \nu_{jd} \end{aligned} \tag{7.3}$$

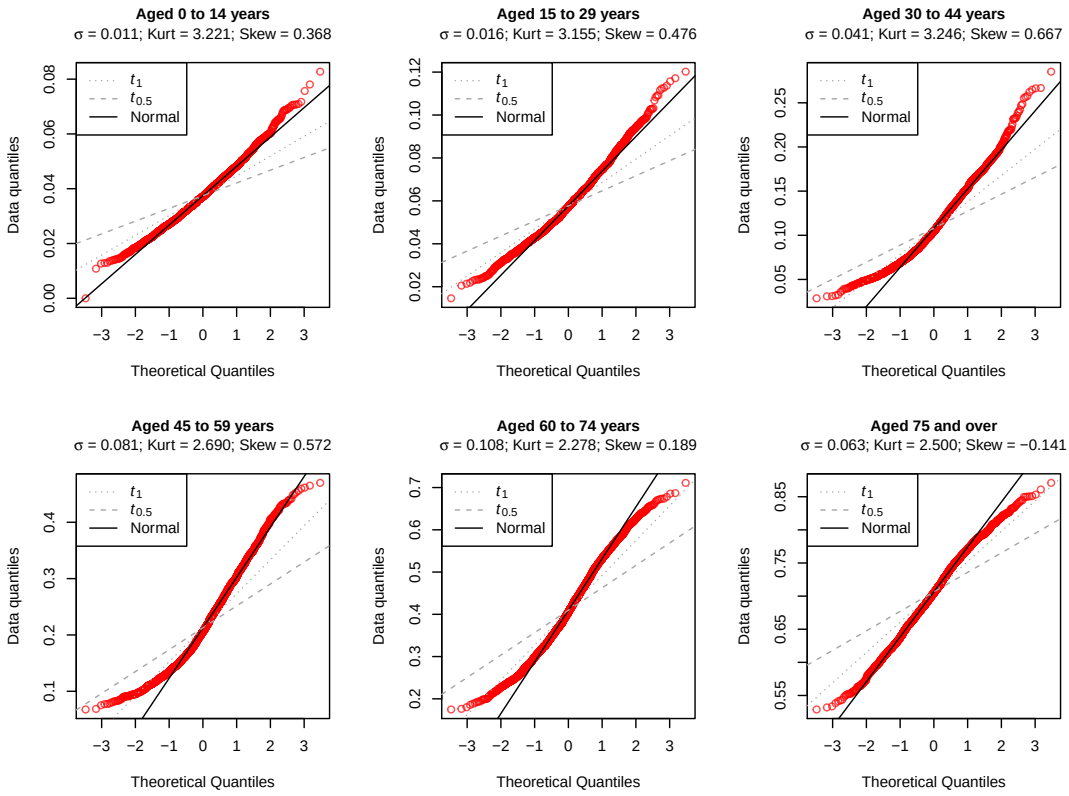


Figure 7.2 | Q–Q plots of age-specific prevalences of LLTI against normal and Student's t distributions (MSOA-level)

where $\beta^{(1)}$ is an i-dimensional vector of age contrasts (with $\beta_1^{(1)} = 0$ for identifiability); $\mathbf{X}$ a matrix of area-level scaled covariates; $\boldsymbol{\beta}^{(2)}$ is a matrix of coefficients controlling the effect of one standard deviation in the covariates on area-level log-odds of LLTI; a structured ICAR area random effect υ_d; an unstructured (exchangeable) area by age random effect ν_{jd}.

Covariates in $\mathbf{X}$ are informed by previous model building work (see chapters 5 and 6) and extracted from public sources; namely: the MSOA- and district-level ISAR (Public Health England, 2014b); the 2015 Income Deprivation Affecting Children Index (proportion of all children aged 0–15 years living in income-deprived families based on tax and benefit departmental database, Public Health England, 2016); the MSOA mean price of residential property sales in 2011 (ONS, 2016c); 2011 mortality ratios indirectly standardised by sex and age (ONS, 2013i); and the 2011 MSOA Rural Urban Classification consisting of five contrasts (ONS, 2016e). No age-specific area covariates were available, which reduces the predictive capabilities of the model. Figure A.1 (see appendix, p. 266) shows the association and coefficient of determination of the MSOA-level prevalence of LLTI with ISARs for

different demographic subgroups. These plots show that despite a strong association with crude LLTI prevalence, association with age-specific prevalence is weaker.

The structured ICAR area effect υ_d is included to model the shared spatial surface in disease prevalence as first introduced by Besag (1974) under a sum-to-zero constraint for identifiability: $\sum_{d=1}^{D} \upsilon_d = 0$. Every structured area effect υ_d is dependent on other area structured effects $\boldsymbol{\upsilon}_{[d]} = \{\upsilon_j, j \neq d\}$ under the following conditional distribution:

$$\upsilon_d \mid \boldsymbol{\upsilon}_{[d]} \sim \text{Normal}\left(\sum_{d \neq j} \frac{w_{dj}\upsilon_j}{w_{dj}}, \frac{1}{\tau_\upsilon \sum_{[d]} w_{dj}}\right) \tag{7.4}$$

where spatial weights w are taken from the spatial dependence matrix $\mathbf{R}_\upsilon^{-1}$ defined as a $D \times D$-dimensional contiguity matrix (Queen's method) as below:

$$w_{ij} = \begin{cases} 1 & i = j \\ 1 & i, j \text{ are neighbours} \\ 0 & \text{otherwise} \end{cases} \tag{7.5}$$

corresponding with the below joint prior density:

$$\pi(\boldsymbol{\upsilon} \mid \tau_\upsilon) \propto \exp\left(-\frac{\tau_\upsilon}{2} \boldsymbol{\upsilon}' \mathbf{R}_\upsilon \boldsymbol{\upsilon}\right) \tag{7.6}$$

$\boldsymbol{\nu}$ has a basic normal exchangeable prior centred around zero with a unique precision τ_ν. The sum of $\boldsymbol{\upsilon}$ and $\boldsymbol{\nu}$ forms the widely used convolution prior (Besag et al., 1991). A discrete mixture of normal exchangeable effects with unequal variances was considered to take into account evidence of heteroskedasticity across age groups. This has not led to substantial improvement and has thus been abandoned.

Models are estimated using INLA (Rue et al., 2009), given the substantial computational gains over Markov Chains Monte Carlo. Simulation results for generalised linear mixed models (Carroll et al., 2015; Grilli et al., 2014) have shown that posterior distributions obtained are virtually perfectly aligned to posteriors obtained with Markov Chains Monte Carlo sampling. Estimation is implemented using software packages `INLA` v 1.698, `R` 3.2.1, and package `R-INLA` (Martins et al., 2013; R Core Team, 2014) on high performance computer IRIDIS 4, using 16-core nodes, each equipped with 2.6 GHz CPUs and 4 GB RAM.

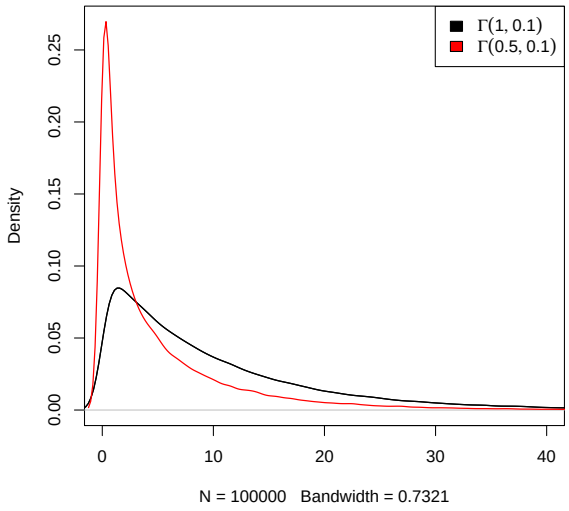


Figure 7.3 | Probability density function of area effects precision hyperpriors

7.4.4 Study and fitting priors

The fitting prior can be formed based on the accumulation of modelling experience on the outcome of interest, particularly around the scale of random effects. It can be difficult to anticipate the magnitude of residual between-area heterogeneity once covariates are used, but the accumulation of evidence over time can guide priors around the magnitude of area random effects expressed as log-odds. We specify inverse Gamma priors for these effects on the basis of previous evidence that once covariates are introduced, their variance is likely to be near 0.1, and that half of this between-area heterogeneity is likely to be spatially structured. Following this, we place equal weight on structured and unstructured effect. To achieve a sum variance of 0.1, we need a mean hyperprior precision of 5 for each area effect. Based on plots we also decide that the variance of this hyperprior should be ten times the mean to avoid dominating the data. This leads to hyperpriors $\Gamma(0.5, 0.1)$ being placed on each of the inverse variance of area effects or $\Gamma(1, 0.1)$ when only unstructured effects are used (see Figure 7.3).

In contrast, design priors are obtained from the marginal posterior density of $\boldsymbol{\theta}$ obtained from fitting $\mathcal{M}$ on a simulated pilot sample s_0 with sampling fraction $f = 0.01$. Where prior knowledge expressed in fitting priors is very informative, the sampling fraction can be relatively small and still produce relatively narrow design priors.

Table 7.2 | Description of design-based simulation scenarios

Scenario	Estimation method	Model structure for predictor θ_{jd}
S1	Direct estimation	–
S2	HB – no covariate	$\boldsymbol{\beta}_j^{(1)} + \phi_{jd}$
S3	HB – covariate	$\boldsymbol{\beta}_j^{(1)} + X_d\boldsymbol{\beta}_d^{(2)} + \phi_{jd}$
S4	HB – covariate + spatial structure	$\beta_j^{(1)} + X_d\boldsymbol{\beta}_j^{(2)} + \upsilon_d + \phi_{jd}$

Note: HB: Hierarchical Bayes; MSOA: Middle Layer Super Output Area.

7.4.5 Design-based simulation

A second series of simulations is carried out with a design-based procedure: samples are taken directly from the census tables using a multinomial law. Unlike with the model-based simulation, results no longer assume that the models are true. Doing so makes it possible to verify the validity of the sample size obtained under the previously described model-based SSD procedure. We also take the opportunity to estimate the relative efficiency in a series of scenarios of ascending complexity, from the simple design-based estimator to an HB predictor borrowing strength from a covariate and an explicit spatial covariance structure. These scenarios notated S1 to S4 are summarised in Table 7.2. To this end, we simulate sampling and estimating each one and compare it with the true population value.

One hundred simple random samples s_b are drawn from U with simulated sizes $n_{jdb} \sim \text{Binomial}(N_{jdb}, f)$ where f is set by the experiment, with no guarantee that all $n_{jdb} > 0$. We simulate sample headcounts such that:

$$y_{jdb} \sim \text{Binomial}(n_{jdb}, N_{jdb}^{-1}\, Y_{jdb}) \tag{7.7}$$

Samples s_b are used to produce model-based estimates $\widehat{\overline{Y}}_{jdb}$ of prevalence $\overline{Y}_{jdb} = Y_{jdb}/N_{jdb}$ of LLTI every area making up the population as a function of characteristics of this population.

To reduce the computational burden of this simulation study, we only estimate relative efficiency under two sampling fractions: 2 and 4 per cent. The number of iterations is set to $B = 400$ for each combination of a scenario and sampling fraction, totalling 2,400 procedures of model estimation and prediction. For scenarios S2–S5, and for each of the chosen sampling fractions f, we compute measures of accuracy:

- the root mean squared error (RMSE)

$$\text{RMSE}_{jd} = \sqrt{B^{-1} \sum_{b=1}^{B} \left(\widehat{\overline{Y}}_{jdb} - \overline{Y}_{jdb} \right)^2} \tag{7.8}$$

- the bias

$$\text{Bias}_{jd} = B^{-1} \sum_{b=1}^{B} \left(\widehat{\overline{Y}}_{jdb} - \overline{Y}_{jdb} \right) \tag{7.9}$$

- the absolute relative bias (absolute relative bias (ARB))

$$\text{ARB}_{jd} = B^{-1} \sum_{b=1}^{B} \frac{|\widehat{\overline{Y}}_{jdb} - \overline{Y}_{jdb}|}{\overline{Y}_{jdb}} \tag{7.10}$$

- the relative RMSE or relative standard error (RSE)

$$\text{RSE}_{jd} = \frac{\text{RMSE}_{jd}}{\overline{Y}_{jd}} \tag{7.11}$$

- the RSE's relative bias (RSEB) (where $\widehat{\text{postMSE}}$ is the posterior variance)

$$\begin{aligned} \text{RSEB}_{jd} &= B^{-1} \sum_{b=1}^{B} \frac{\widehat{\text{RSE}}_{jdb} - \text{RSE}_{jdb}}{\text{RSE}_{jdb}} \\ \widehat{\text{RSE}}_{jd} &= \frac{\sqrt{\widehat{\text{postMSE}}_{jd}}}{\hat{\overline{Y}}_{jd}} \end{aligned} \tag{7.12}$$

These measures are mapped and averaged across small areas d to be reported in tables. For scenario S1, the bias of sample means is by definition zero and we calculate the MSE and RSE using the variance formula of the sample proportion:

$$\text{Var}(\bar{y}_{jd}) = \frac{\overline{Y}(1-\overline{Y})}{n_{jd}} = \frac{\overline{Y}(1-\overline{Y})}{f\ N_{jd}} \tag{7.13}$$

Table 7.3 | Case study results of the SSD procedure after four steps

Step k	Sampling rate f_k	Mean loss $\bar{\ell}(f_k)$	Mean loss $\hat{\bar{\ell}}(f_k)$	Risk $\Pr(\ell(f_k) > 0)$	Risk $\Pr(\hat{\ell}(f_k) > 0)$	Search interval
1	0.010000	0.004000	0.008400	0.063500	0.095200	
1	0.040000	0.000000	0.000000	0.000000	0.000000	
1	0.025000	0.000006	0.000005	0.025000	0.010000	[0.010000,0.040000]
2	0.032500	0.000000	0.000000	0.000000	0.000000	[0.025000,0.040000]
3	0.028750	0.000003	0.000000	0.010050	0.000000	[0.025000,0.032500]
4	0.026875	0.000001	0.000000	0.005076	0.000000	[0.025000,0.028750]

Note: Authors' own calculations. Loss and risk estimated over 200 simulations per step.

7.5 Results

7.5.1 SSD procedure

The basic model (7.3) is fitted to pilot sample s_0 of ESS $n = 146,574$, with 1,952 cells jd containing no observations. This means no data is available for 1.3 per cent of the total number of cells, or 8.8 per cent of cells where the minimum quality standard applies. The fitted model achieves an acceptable DIC of 33,787. The SSD procedure summarised in Table 7.3 produces an effective sample fraction requirement of 2.6875 per cent, equivalent to an ESS of 394,103. Assuming the procedure is correct, this solution is considered to fall at the most 0.1875 percentage points from the true effective sampling fraction achieving a negligible risk of cell suppression. Further iterations would reduce this distance. The algorithm was guided by an estimator of the loss function ℓ which depends on predictors θ to determine which cells should be checked for the quality standard. Any bias on θ can potentially increase or decrease the suppression risk to a substantial degree. In this case study, we note that the estimator $\hat{\ell}$ fails to detect a risk of suppression at step $k = 3$, which leads the algorithm to select a lower interval of sample fractions that it would have done if it had been driven by actual loss ℓ calculated using the census figures Y_{jd}.

Traditional survey sampling methods can determine the most cost-effective strategy to achieve this target ESS. A basic household survey like the EHS is based on a multistage design sampling one household per dwelling per address selected in the address sampling frame, and interviewing all eligible members (by proxy if necessary). Using the 2010/11 microdata sample for England of 11,188 effective responses, we estimated the design effect (DEFF) to 1.16 for the LLTI question. This implies that survey planners should aim for a final sample size at least 16 per cent higher than the ESS found above, to account for the design effect. Yet many surveys have additional sampling stages which can increase the DEFF, for instance selecting post-

code sectors as primary sampling units and then addresses/dwellings/households as secondary sampling units. This is for instance the case of the Health Survey for England for which we estimated a DEFF of 1.44 using a microdata sample from 2011 (NatCen Social Research and University College London. Department of Epidemiology and Public Health, 2013). The DEFF can easily be used in area-level SAE models in a variance function, or in unit-level generalised linear mixed models with the ESS method described by Chen et al. (2014).

7.5.2 Design-based simulation

We verify the validity of the small area model used in this case study as well as the reliability of the loss function estimator across iterations of the algorithm. Design-based simulation results are presented in Tables 7.4 to 7.7. Prevalence proportions observed from the census and predicted in S3 are mapped in Figures 7.4 and 7.5, respectively.

Results provide evidence that the relative posterior variances obtained from the spatial model are inflated by 20 to 40 per cent in basic scenario S2 (Table 7.5) and by as much as 40 to 50 per cent in scenario S3 (Table 7.6). This is likely to make the SSD procedure very conservative as many cells will be unnecessarily suppressed. This is a likely sign of poor model specification. Simulations for other scenarios are reported here to quantify the relative efficiency of the different model- and design-based methods. Efficiency savings from using a particularly model or from an increase in sample size can be derived as the ratio of RMSE of two scenarios. Comparing the scenarios presented in Table 7.2, we find that even a basic model without area covariates (see S1–S2, Tables 7.4 and 7.5) achieves a reduction in root mean squared error (RMSE) by about 49 per cent overall, and up to 74 per cent for youngest age groups ($f = 0.02$). This comes at a very reasonable computational cost and with a bias of between 0.1 and 0.7 percentage point. With a much larger sample ($f = 0.04$), the efficiency gain comes down to approximately 35 per cent.

The addition of covariates delivers strong efficiency gains (see S3, Tables 7.6) with an reduction in RMSE by 72 ($f = 0.02$) and 62 per cent ($f = .04$). This includes a reduction in bias from 0.02 to below 0.01 percentage point.

Comparing results for $f = 0.02$ and $f = 0.04$, we note that doubling the sample size has diminishing returns as the methods becomes more complicated. Whilst it reduces the standard deviation of design-based estimates S1 by 29 per cent, the RMSE of estimates S2 is only reduced by 12 per cent. As for estimates S3 and S4 the reduction by less than 3 and 4 per cent respectively is not material for an

Table 7.4 | Simulation S1: mean accuracy metrics by age group

Age (years)	**S1** *('design-based')*				
	$\overline{\text{RMSE}}$	$\overline{\text{Bias}}$	$\overline{\text{RSE}}$	$\overline{\text{RSEB}}$	$\ell(f)$
Sampling fraction $f = 0.02$					
0–14	0.0379	—	1.0600	—	0.0000
15–29	0.0449	—	0.8048	—	0.0005
30–44	0.0576	—	0.5663	—	0.1922
45–59	0.0752	—	0.3722	—	0.8226
60–74	0.1050	—	0.2660	—	0.8502
75+	0.1436	—	0.2054	—	0.4054
All ages	0.0774	—	0.5458	—	0.3785
Sampling fraction $f = 0.04$					
0–14	0.0268	—	0.7495	—	0.0000
15–29	0.0318	—	0.5691	—	0.0005
30–44	0.0407	—	0.4004	—	0.1907
45–59	0.0532	—	0.2632	—	0.6687
60–74	0.0742	—	0.1881	—	0.3129
75+	0.1015	—	0.1452	—	0.0133
All ages	0.0547	—	0.3859	—	0.1977

Note: Accuracy metrics are aggregated into a mean across all 1,956 areas.

augmentation of this magnitude. This shows that RMSE as a function of sample size is already relatively horizontal in the regions of sample sizes we are investigating, which is not the case for design-based estimators. Regardless of the sample size, bias remains high. Although the average bias across areas remains of the order of 0.1 to 0.3 percentage points, the more meaningful ARB metric reveals an average absolute bias of 5 to 20 per cent of the target parameter. This represents almost all the total RSE. This bias is not very sensitive to the sampling fraction or the type of between-area variance structure used.

The working model was designed to borrow strength by assuming that areas situated near each other were more similar. We thus examined simulation results to verify whether this assumption holds everywhere equally. Figures 7.6 and 7.7 present the RMSE and bias for age-specific MSOA prevalence proportion computed from S3 with $f = 0.020$. Hotpots of high RMSE and bias of 1 to 2 percentage points are visible which exhibit a spatial pattern. This could signal outlier areas, or areas in which that covariates are poorly measured and making the linear predictor particularly biased. The convolution prior used here may not fully and authentically reproduce the spatial pattern of the health status under consideration.

Table 7.5 | Simulation S2: mean accuracy metrics by age group across $B = 400$ replications

Age (years)	**S2** *('HB - no covariates')*								
	$\overline{\text{RMSE}}$	$\overline{\text{Bias}}$	$\overline{\text{ARB}}$	$\overline{\text{RSE}}$	$\overline{\text{RSEB}}$	$\ell(f)$	$\overline{\Pr(\ell > 0)}$	$\hat{\ell}(f)$	$\overline{\Pr(\hat{\ell} > 0)}$
Sampling fraction $f = 0.02$									
0–14	0.0100	0.0010	0.2648	0.3004	0.6734	0.0000	0.000	0.0000	0.000
15–29	0.0149	0.0010	0.2340	0.2750	0.5234	0.0005	1.000	0.0057	1.000
30–44	0.0316	0.0021	0.2876	0.3298	0.1488	0.1921	1.000	0.1027	1.000
45–59	0.0527	-0.0017	0.2317	0.2710	0.0418	0.7449	1.000	0.8589	1.000
60–74	0.0722	-0.0071	0.1576	0.1859	0.0780	0.1670	1.000	0.1649	1.000
75+	0.0568	-0.0069	0.0691	0.0814	0.4325	0.0000	0.000	0.0000	0.000
All ages	0.0397	-0.0019	0.2075	0.2406	0.3163	0.1841	1.000	0.1887	1.000
Sampling fraction $f = 0.04$									
0–14	0.0100	0.0010	0.2547	0.2984	0.4418	0.0000	0.000	0.0000	0.000
15–29	0.0146	0.0012	0.2211	0.2681	0.3228	0.0005	1.000	0.0005	1.000
30–44	0.0280	0.0021	0.2442	0.2893	0.0609	0.1727	1.000	0.1727	1.000
45–59	0.0437	-0.0006	0.1849	0.2229	-0.0075	0.3322	1.000	0.3322	1.000
60–74	0.0600	-0.0047	0.1268	0.1539	0.0220	0.0107	1.000	0.0107	1.000
75+	0.0539	-0.0062	0.0638	0.0773	0.2697	0.0000	0.000	0.0000	0.000
All ages	0.0351	-0.0012	0.1826	0.2183	0.1849	0.0860	1.000	0.0879	1.000

Note: Accuracy metrics are aggregated into a mean across all 1,956 areas.

Table 7.6 | Simulation S3: mean accuracy metrics by age group across $B = 400$ replications

Age (years)	**S3** *('HB – with covariates')*								
	$\overline{\text{RMSE}}$	$\overline{\text{Bias}}$	$\overline{\text{ARB}}$	$\overline{\text{RSE}}$	$\overline{\text{RSEB}}$	$\ell(f)$	$\overline{\Pr(\ell > 0)}$	$\hat{\ell}(f)$	$\overline{\Pr(\hat{\ell} > 0)}$
Sampling fraction $f = 0.02$									
0–14	0.0078	-0.0009	0.2073	0.2134	0.4473	0.0000	0.000	0.0000	0.000
15–29	0.0108	-0.0026	0.1857	0.1929	0.4871	0.0000	0.000	0.0000	0.008
30–44	0.0143	-0.0006	0.1269	0.1371	0.6470	0.0000	0.000	0.0000	0.000
45–59	0.0255	-0.0016	0.1107	0.1222	0.4717	0.0000	0.003	0.0000	0.003
60–74	0.0360	-0.0020	0.0814	0.0903	0.4439	0.0000	0.000	0.0000	0.000
75+	0.0346	0.0024	0.0467	0.0496	0.5439	0.0000	0.000	0.0000	0.000
All ages	0.0215	-0.0009	0.1264	0.1343	0.5068	0.0000	0.003	0.0000	0.008
Sampling fraction $f = 0.04$									
0–14	0.0076	-0.0008	0.2004	0.2080	0.3958	0.0000	0.000	0.0000	0.000
15–29	0.0104	-0.0024	0.1769	0.1865	0.3872	0.0000	0.000	0.0000	0.000
30–44	0.0141	-0.0003	0.1223	0.1359	0.4695	0.0000	0.000	0.0000	0.000
45–59	0.0248	-0.0014	0.1044	0.1190	0.3044	0.0000	0.000	0.0000	0.000
60–74	0.0348	-0.0016	0.0765	0.0876	0.2849	0.0000	0.000	0.0000	0.000
75+	0.0338	0.0024	0.0446	0.0485	0.4106	0.0000	0.000	0.0000	0.000
All ages	0.0209	-0.0007	0.1209	0.1309	0.3754	0.0000	0.000	0.0000	0.000

Note: Accuracy metrics are aggregated into a mean across all 1,956 areas.

Table 7.7 | Simulation S4: mean accuracy metrics by age group across $B = 400$ replications

Age (years)	S4 *('HB – covariates and spatial structure')*								
	$\overline{\text{RMSE}}$	$\overline{\text{Bias}}$	$\overline{\text{ARB}}$	$\overline{\text{RSE}}$	$\overline{\text{RSEB}}$	$\ell(f)$	$\overline{\Pr(\ell > 0)}$	$\hat{\ell}(f)$	$\overline{\Pr(\hat{\ell} > 0)}$
Sampling fraction $f = 0.02$									
0–14	0.0079	-0.0008	0.2072	0.2155	0.2797	0.0000	0.000	0.0000	0.000
15–29	0.0107	-0.0026	0.1845	0.1934	0.3490	0.0000	0.000	0.0000	0.000
30–44	0.0136	-0.0007	0.1187	0.1305	0.5762	0.0000	0.000	0.0000	0.000
45–59	0.0244	-0.0019	0.1039	0.1162	0.4433	0.0000	0.007	0.0000	0.000
60–74	0.0337	-0.0026	0.0750	0.0844	0.4341	0.0000	0.000	0.0000	0.000
75+	0.0345	0.0023	0.0461	0.0495	0.4146	0.0000	0.000	0.0000	0.000
All ages	0.0208	-0.0010	0.1226	0.1316	0.4162	0.0000	0.005	0.0000	0.000
Sampling fraction $f = 0.04$									
0–14	0.0077	-0.0008	0.2014	0.2097	0.2464	0.0000	0.000	0.0000	0.000
15–29	0.0104	-0.0025	0.1768	0.1866	0.2833	0.0000	0.000	0.0000	0.000
30–44	0.0132	-0.0005	0.1133	0.1265	0.4617	0.0000	0.000	0.0000	0.000
45–59	0.0234	-0.0019	0.0974	0.1113	0.3174	0.0000	0.000	0.0000	0.000
60–74	0.0321	-0.0023	0.0702	0.0807	0.3096	0.0000	0.000	0.0000	0.000
75+	0.0332	0.0021	0.0439	0.0477	0.3318	0.0000	0.000	0.0000	0.000
All ages	0.0200	-0.0010	0.1172	0.1271	0.3250	0.0000	0.000	0.0000	0.000

Note: Accuracy metrics are aggregated into a mean across all 1,956 areas.

7.6 Final remarks

This paper has examined the feasibility of a design approach to SAE model-based estimation, by proposing a fully Bayesian design algorithm taking a working model and a pilot sample as inputs. The decision rule is based on a new loss function of the sampling fraction: the weighted cell suppression rate. This approach is intended for survey planners and public health professionals to consider an SAE problem at the survey design phase, rather than using it strictly for post hoc analyses. The availability of historical survey data is likely to allow survey planners to implement similar approaches at a small cost. The Bayesian decision-making approach allows them to make use of whatever information is available—in this case, a pilot sample—but it is perfectly possible to form a comparable design prior on the basis of good information on model random component (random effect variances, spatial fraction). The SSD algorithm then operates on a simulated population that is only as realistic and informative as the design prior is.

Our approach is a direct continuation of previous model-based design decisions considered by Joseph et al. (1995), Zou and Normand (2001) and Sahu and Smith (2006). It establishes a new link to official statistics requirements with a new loss

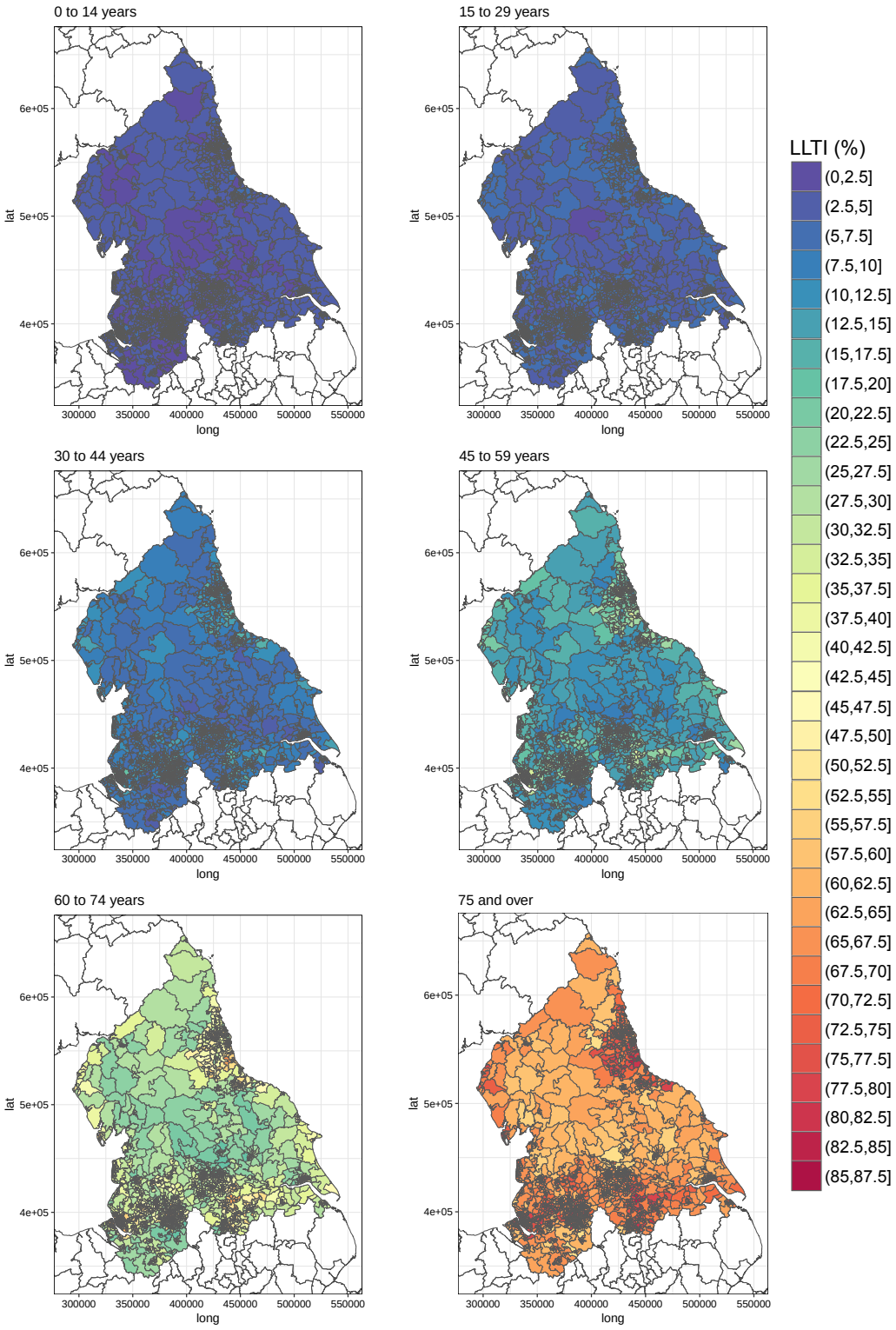


Figure 7.4 | Map of age-specific prevalence proportions of LLTI (in %) in the 2011 census

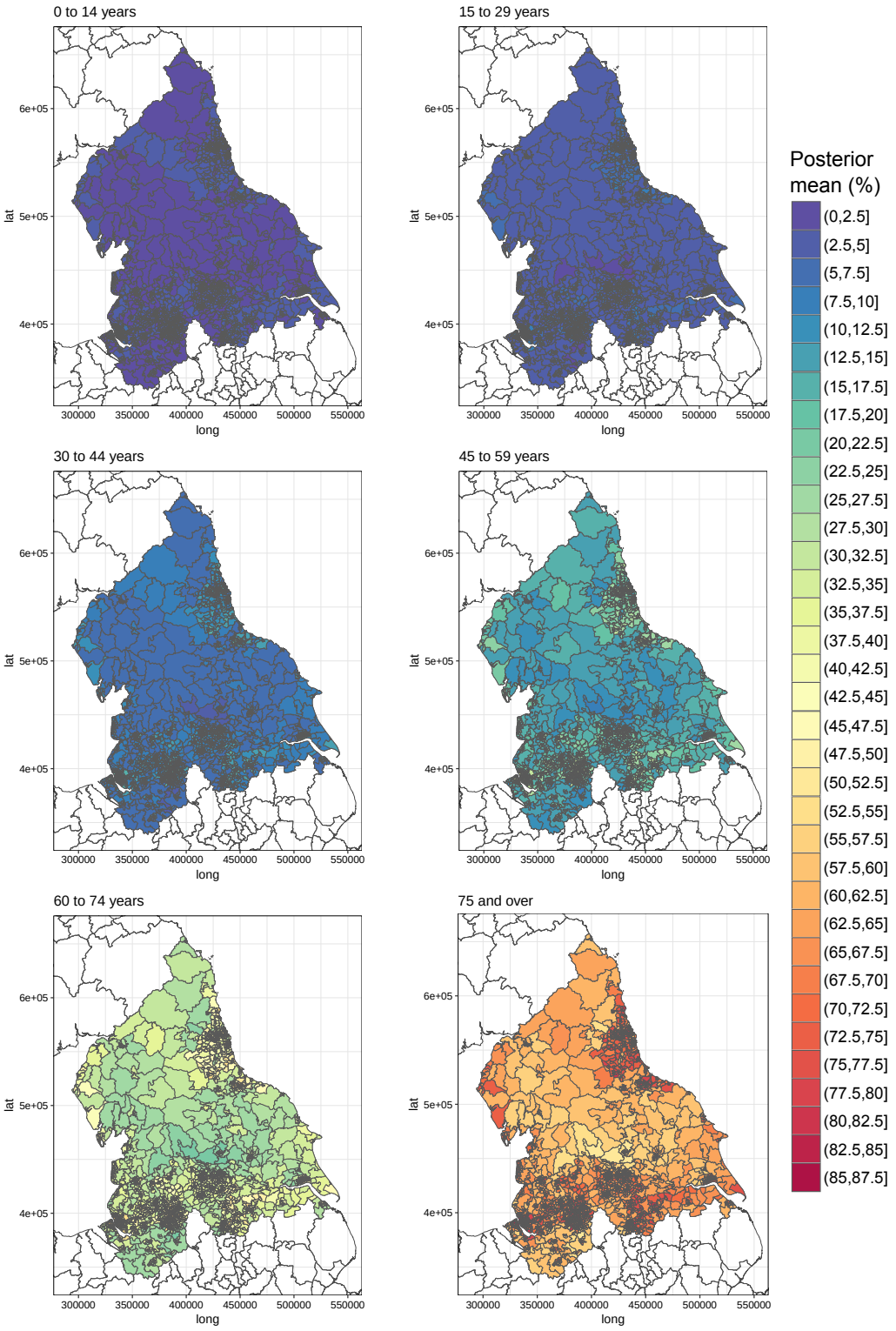


Figure 7.5 | Map of posterior mean age-specific proportions of LLTI (in %) averaged across 400 iterations of scenario S3 with sampling fraction $f = 0.02$

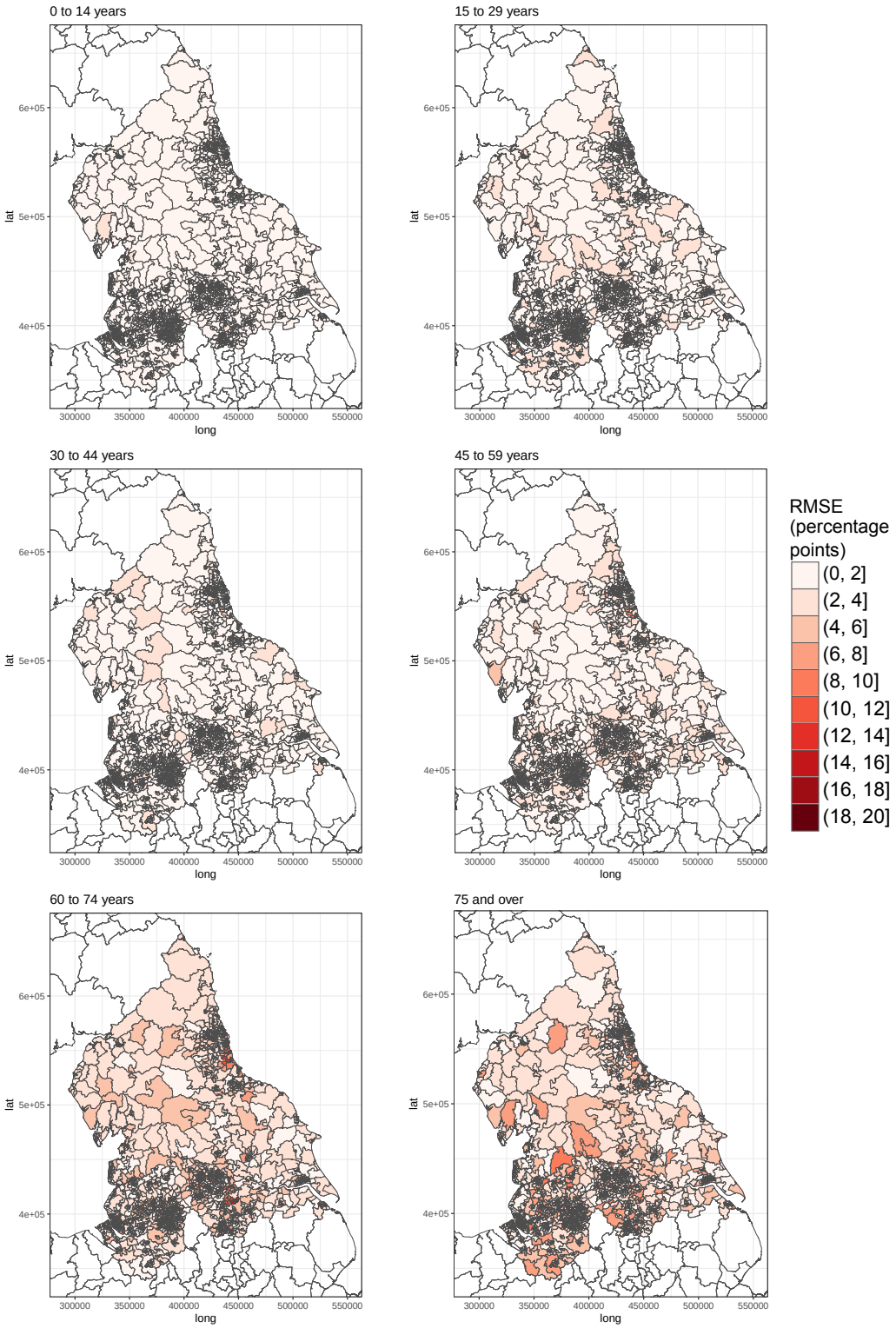


Figure 7.6 | Map of age-specific root MSE in prediction on the proportion scale across 400 iterations of scenario S3 with sampling fraction $f = 0.02$

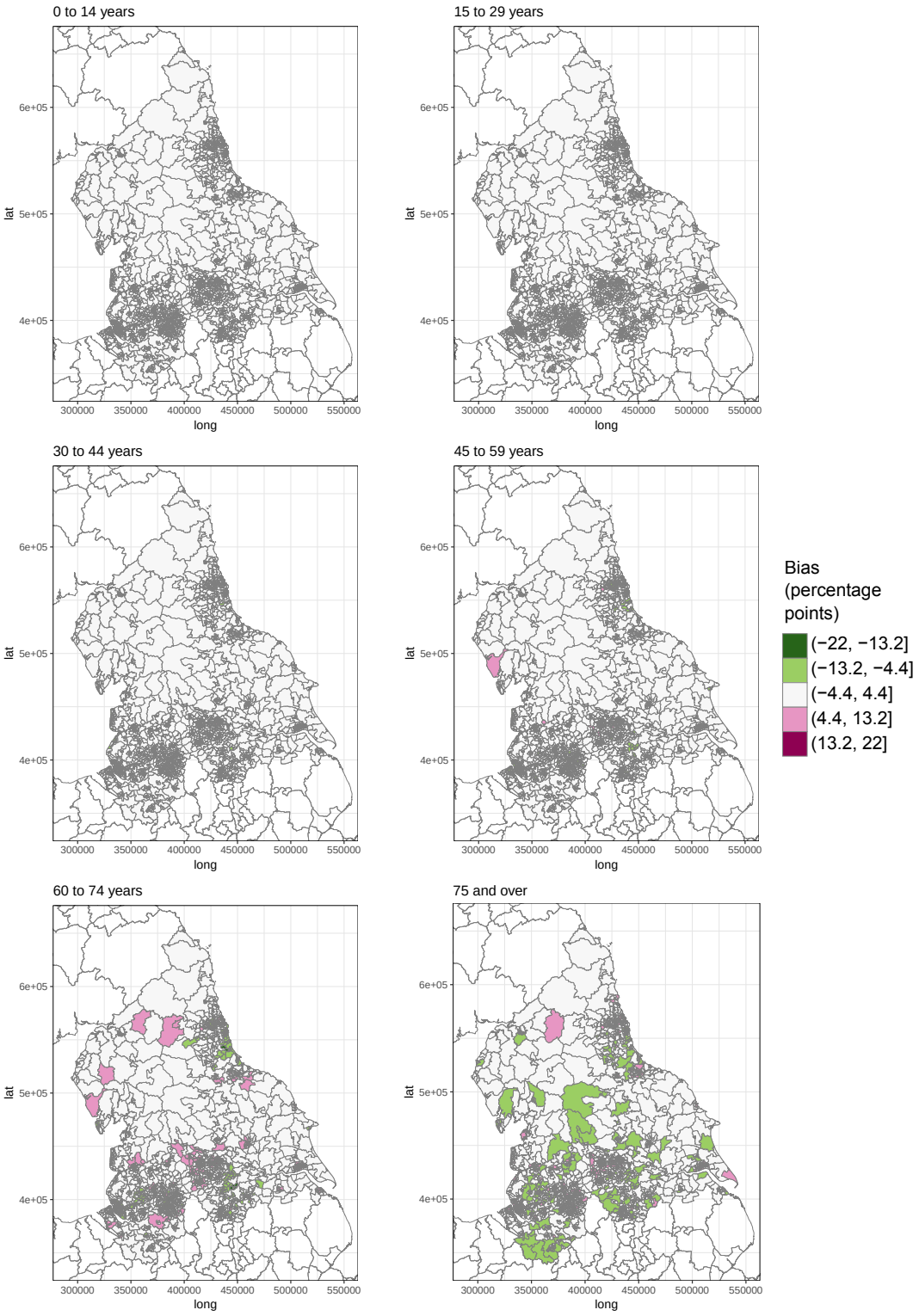


Figure 7.7 | Map of age-specific prediction bias in percentage points across 400 iterations of scenario S3 with sampling fraction $f = 0.02$

function directly relevant to quality criteria present in most statistical codes of practice internationally. This loss function can be calculated for different social subgroups in situations where estimates of equal precision are needed for minority populations, particularly when equality impact assessments demand equal statistical precision for official statistics across different communities. The implementation illustrated in our case study greatly simplifies the design work by expressing the output in terms of effective sampling fraction or ESS, which integrates well with existing survey design tools, knowledge around the DEFF of various sampling designs, and work to address sample attrition caused by nonresponse. In this sense this procedure is not one that is likely to determine optimal sampling schemes; traditional survey sampling being more suitable to this end. But it is a flexible approach which could easily be extended to simulate unequal probability of sampling method (UPSEM) designs incorporating on known population characteristics across areas (for instance ethnic diversity or age characteristics).

While the computational burden of this method can remain high in studies involving a large number of cells, the combination of INLA and the binary search logic brings these costs to levels that are no longer absurd with modern equipment and high performance cloud computing. Most importantly, the binary search allows planners to anticipate the exact number of iterations needed to reach a prespecified solution precision. They can therefore budget the algorithm's computational requirements.

There are limitations to this approach which make it highly dependent on model misspecification and modelling priors which, together with model selection, arguably constitute the three main challenges to model-based estimation. Given known difficulties in identifying spatial models, it is clear that such a procedure is not likely to be reliable outside of models well established across years of data collection and model testing. While this is not unrealistic with government surveys, this does represent a non-negligible design cost which is difficult to incorporate explicitly in planning.

Our case study conducted on a real population also illustrates that the computation of the loss function being dependent on the reliability of the model, model diagnostics and validation are essential to the reliability and stability of the procedure. Results from the design-based simulation studies show the influence of typical model misspecification on estimation bias and how it clusters in space or social groups. Prior formation also remains, as in many other areas of statistics, a challenge. Recent work in this area by Fong et al. (2010) and Simpson et al. (2015) expresses some of the important obstacles faced by practitioners in this area.

This work tends to support other research showing the feasibility of reconciling Bayesian inference and survey sampling, the intersection of which, in this case, is expressed in terms of ESS. This also opens perspectives to combine survey sampling with some very strong developments witnessed in recent years in the design of experiments literature, particularly around adaptive trial designs.

Chapter 8

Discussion

After setting out key population health analytical needs in chapter 2, we systematically reviewed existing research on model-based SAE in chapter 3. This review highlighted limitations of model-based SAE in terms of validity and identified key prerequisites of future applications, namely:

- the adoption of a more homogeneous terminology and set of reporting standards for SAE designs;
- methodological improvements in model selection and diagnostics;
- the development of simulation-based validation work;
- the development of approaches to plan and design model-based SAE outputs.

The collection of papers featured in this thesis makes contributions towards each of the above, based on applied work on health outcomes commonly used in large-scale population investigations. This discussion chapter cross-examines these contributions, highlighting (a) what can be learnt from results of this empirical work; (b) more substantive research contributions achieved across those three papers; (c) key limitations in their methods and generalisability; and (d) implications for future research and development.

8.1 Summary of empirical results

In chapter 5, we examined the predictive capability of four unit-level models designed to estimate the crude prevalence of two harmonised health indicators: SRH and LLTI. The unit-level structure is justified by the large variance of sample proportion estimators with low or null sample sizes, where a normal approximation through the central limit theorem is expected to be unreliable. Rates of emergency admissions explain high proportions of the between-area variance in the prevalence

of health statuses. Despite this, comparisons with census estimates illustrate the lack of fit of synthetic prediction. With regard to objective O5, benefits of using EPP over synthetic prediction are apparent at the LAD level, but the small sample sizes at the MSOA level mean that this technique cannot use enough local information to increase efficiency at this finer level. Comparison of naive EPP and EB prediction showed no flagrant difference.

Both EPP and synthetic methods nevertheless yield negatively biased results, particularly for areas with the highest prevalence of poor health. It is only possible to conclude that the model is misspecified, either because the assumed structure in residual between-area variance does not hold (heteroskedasticity, violations of normality), because the model lacks a few covariates, or because it fails to account for some nonignorable survey design feature. Further tests and diagnostics could not identify the origin of this misspecification. This work is a reminder of the likely shortcomings of model-based SAE.

The validation and diagnostics design of this chapter is strong, thanks to the large sample sizes provided by our fitting dataset and the availability of geographical identifiers locating respondents' area of residence. This means that internal validation against direct survey estimates was possible. We note that these conditions are not met in many of the studies reviewed in chapter 3. This implies that many of these studies report predictions which in practice cannot be scrutinised to the same extent we did in chapter 5. Combining internal and external validation in this investigation provides additional insight.

Though the inadequacy of predictions is apparent from the verification against census estimates, more easily carried-out diagnostics relying on direct survey estimates also evidence it. We conclude not only that (a) the discrepancy is not simply due to the effect of the census self-administration mode but also that (b) it could have been detected even without access to census estimates. Despite this, a key benefit we draw from a comparison with the census is the ability to map discrepancies in estimates area by area, rather than averaging them across the whole population as is usually the case in validation studies. Though these results are not directly generalisable to other health outcomes, they may be helpful to practitioners who seek to interpret the effect of similar levels of unexplained between-area variation on the quality of the resulting estimates across areas. It is manifest from maps 5.2 to 5.5 (pp. 143–146) that, in addition to the global underestimation, the 'missing covariate' effect generates pockets of over- or underestimation in places. The strong spatial autocorrelation in this bias could argue in favour of using spatial modelling to compensate for the model's inability to predict the similarity in prevalence across

geographically adjacent areas (objective O4). This is further examined in chapter 7 in a range of simulations.

With regard to objectives O2–O3, chapter 5 finds that models under consideration have limited predictive power and show evidence of misspecification leading to global underestimation. As a result, only basic univariate model-based small area estimates of bad/very bad SRH and LLTI were produced, leaving the question of disaggregating them by age and ethnic group to simulation work in a third academic paper (objective O8). Overall, the paper nevertheless provides a real-world evaluation of an SAE design, comparing model-based estimates with a widely trusted and recognised source of data. This design is intended to be easy to appraise even by non-specialist readers, as well as to convey a practical illustration as to how much confidence can be placed in small area estimates for chronic morbidity. Estimates are rated in terms of their overall error as well as the correctness of the ranking of areas that can be inferred from them, attracting attention to other properties of interest to end users. The results highlight the difficulty of predicting health characteristics at such a small geographical scale, with limited predictive power. The study concludes that more work is needed to improve the predictive power of models through better covariate auxiliary data, as well as to understand statistical power and sample size requirements implied by what remain highly demanding quality standards.

Chapter 6 examines the level of similarity in odds of LLTI (a) in neighbouring areas and (b) in different ethnic groups within areas in an observational study of 2011 census LLTI estimates. This follows thesis objectives O1 and O7 and findings from chapter 3 that a wide variety of spatial structures are now used in disease mapping and could potentially be useful in survey-based SAE. With this view, several conceptualisations of spatial interaction are examined and although all produce rather high levels of autocorrelation, we note higher levels of autocorrelation for neighbourhood structures allocating few neighbours to each small area. We also measure unequal levels of spatial autocorrelation across ethnic groups. Unequal spatial autocorrelation is also manifest across age groups but only on a univariate basis: we do not investigate residual autocorrelation for these groups using models with covariates. In addition to unequal autocorrelation across groups, we find that spatial autocorrelation in area random effects is nonstationary across space, even once covariates are introduced into the model. For instance, we notice a number of discontinuities in spatial surfaces which form clusters particularly in Wales, around Nottingham, Middlesborough, and generally around Greater London. These are more pronounced for some ethnic minority groups.

The key learning point from this investigation is the complexity in the spatial

pattern of illness across different groups, with unequal variances and levels of spatial autocorrelation. This suggests that borrowing strength across these various dimensions cannot rely on basic models. This chapter also questions the approach taken in many studies reviewed in chapter 3—as well as in our own chapter 5—to model the effect of age or ethnicity as homogeneous fixed effects. Together with the strong evidence of nonstationarity in the spatial autocorrelation of area random effects, this investigation demonstrates the difficulty of specifying appropriate spatial priors (e.g. adaptive spatial structures) as well as models suitable to produce health status predictions for multiple demographic subgroups within each area (shared components prediction). Although better covariates may help, this suggests that the synthetic assumption is not a reasonable approach to adopt in chronic morbidity modelling. Reliable models may prove to be very complex and to require large sample sizes in order to be identified and produce reliable predictions.

Design-based simulations conducted in chapter 7 address questions arising out of the first two empirical papers, by jointly examining the effect of sample size and specification of spatial priors on the bias and error of prediction. Results are broadly encouraging when it comes to the objectives of the Census Transformation Programme: they demonstrate that a sampling fraction in the region of 3 per cent (equivalent to an average ESS of 75 per MSOA or 12 per age by MSOA cross-classification) should be sufficient to deliver reliable estimates down to the level of MSOAs, even with large DEFFs. More importantly, this study quantifies the respective efficiency gains achieved by hierarchical modelling, the addition of covariates, and the addition of a spatial prior. In the case study, it is manifest that the rudimentary spatial structure used across several age groups delivers no material benefits in terms of precision.

It remains that prediction bias remains high, with only moderate reductions achieved from increasing the sample size. Chapters 5 and 7 together illustrate the behaviour of model-based SAE as a trade-off between bias and design variance and the difficulties raised by biased inferences.

To conclude on practical implications of the empirical modelling work, we summarise key findings for census health indicators:

- models can be specified to predict health outcomes without detailed tabular data on the socioeconomic attributes of small areas—emergency health care utilisation data seems to have a high predictive power for the outcomes under consideration;
- caution should be exercised when considering other outcomes, as health care utilisation is not enough to predict prevalence in itself: phenomena such as the inverse care law mean that other variables are at stake;

- rudimentary contiguity-based spatial modelling does not achieve reductions in bias for LLTI despite evidence of a spatial structure;
- more work is needed in order to identify more suitable and parsimonious structures for area random effects in LLTI and SRH models, possibly building on mixture priors;
- further work is needed to understand how similar models could be developed for chronic disease atlases for specific conditions such as diabetes, respiratory diseases or behavioural risk factors.

8.2 Methodological contributions

This thesis makes contributions beyond model building and validation for census health indicators. These are primarily in relation to (a) approaches to validation, (b) a description of useful model structures and (c) a design approach to SAE.

First, this thesis contains both a systematic review and a range of applications of approaches to the validation of SAE: internal validation against the fitting dataset itself, external validation against a comparator, and simulation-based validation. Chapter 3 clarifies the specific value brought by each of those. The simulation-based approach is a gold standard to be developed as much as possible. When data exhibit evident distributional characteristics which cannot be handled easily by the model, whether in terms of normality, design effect or dependence in space, time or some other dimension, it is advisable to simulate a synthetic population from this potential true model and examine implications of using a misspecified model in terms of prediction quality. It remains that planning such simulation studies presupposes a strong knowledge of what the true working model may resemble. This is why internal and external validation remain important. Chapter 5 gives a better understanding of how each of these approaches can be interpreted.

This thesis produces a second contribution with regard to complex model structures beyond normal exchangeable area effects, particularly around developments from the disease mapping literature and other model-based administrative data research. Chapter 3 identifies very strong development in mixture priors for random effects which are particularly relevant to many of the inferential aims that pertain to small area statistics: hot-spot identification, ranking, monitoring across time. Few of these models have been applied to survey-based SAE, potentially due to the difficulty to identify such models on sparse sample data. Yet, progress in computational efficiency are increasingly allowing application of SAE to very large datasets,

which may help identify optimal covariance structures in the future for situations when covariate fixed effects are not enough to incorporate knowledge into working models.

The key contribution of the present thesis is arguably its attempt to develop a design approach to SAE in order to help produce better strategies and remove some of the considerable unknowns currently faced in survey planning. Chapter 7 develops an approach focusing on SSD, which is suitable to examine sample design implications on model-based estimation. It makes it possible to anticipate relative root MSE which we argue is comparable to the more traditional CV criterion used in national statistical institutes. While this development is specified in the very natural fully Bayesian approach, it is not impossible to envisage similar frequentist implementations. In fact, most of the research on designing data collection for hierarchical models use simulation in a frequentist approach (see in particular Browne et al., 2009).

8.3 Limitations

Limitations pertaining to the specific design of empirical chapters can be noted. Chapter 5 acknowledges signs of model misspecification which have potentially serious implications on the reliability of predictions. This paper also does not take into consideration the potentially informative multi-stage sampling design effect of the fitting data set. More recent work by Chen et al. (2014) has since investigated procedures to account for nonignorable sampling designs in generalised linear mixed models. Simulation results demonstrate in particular that binomial logistic mixed models can be used reliably by plugging in an estimate of the DEFF in the numerator and denominator of the logit response (Vandendijck et al., 2016). This presupposes a reliable estimator of DEFFs across small areas which we do not have in the case of the EHS, although this is an area for future research.

Chapter 6 presents rudimentary descriptive analyses of the structure of morbidity. These are constrained by the amount of models we could reasonably investigate for each demographic subgroup under consideration. A key limitation resulting from this is the reliance on Moran's I statistics which do have a direct and straightforward translation into modelling decisions or prediction reliability. While some of the work examining the use of spatial modelling is carried out in chapter 7, there remain gaps in understanding how efficiency gains can be achieved from introducing spatial information into models. This includes applying some of the more sophisticated approaches reviewed in section 3.5.1.3 as well as researching innovative ways of

building and selecting appropriate spatial weights.

More generally, the work featured in empirical chapters leaves gaps in the execution of thesis objectives. Our work does not entirely answer questions related to objective O7, regarding the feasibility of shared component prediction for different demographic subgroups. This is part due to a data limitation. Due to the unavailability of auxiliary data for each demographic subgroups, we could only use strictly area-level covariates, which reduce model fit. It is conceivable that the release of group-specific emergency admission rates (e.g. age or ethnic group-specific) could be of considerable help in predicting morbidity in age, sex and ethnicity cross-classification. Such data exists and could be collated by NHS Digital to improve the fit of shared prediction models.

Shared component prediction involves further modelling challenges. Preliminary analysis for chapter 7 revealed that using a first- or second-order random walk prior for age by area random effects yielded a poorer fit than using simple NEX effects. This is in spite of strong levels of correlation between age-specific odds of LLTI within areas. Instead, model components shared across age groups were restricted to covariate fixed effects and a spatial ICAR area random effect. Further research is needed to determine whether multivariate normal or another form of prior can be used to borrow strength for age by area random effects. As for ethnicity, chapter 6 demonstrates its importance to understand the spatial and population structure of health inequalities. Yet, model selection conducted in preparation this chapter has identified some important obstacles to the shared component prediction of morbidity across ethnic groups. This was mainly due to the absence of straightforward implementation of multivariate normal and multivariate CAR priors in INLA. Although shared component SAE for age by ethnicity cross-classifications is left for a future investigation, it is apparent from early testing that these populations are extremely small in areas such as MSOAs. In practice, it may be difficult to have assurance that basic distributional assumptions hold at such small levels, particularly normality.

Overall, the generalisability of these findings to other health outcomes and other regions is limited. We also note that this thesis does not address the more fundamental question of the acceptability of model-based estimation in official statistics. This more theoretical problem remains a collective concern in official methodology steering groups. While this does not belong to the set of research questions targeted by the present thesis, it remains an important background concern in interpreting implications of this work.

8.4 Implications for future research

Results presented in this thesis are directly relevant to ongoing work conducted by the Census Transformation Programme. Results are encouraging in terms of potential efficiency gains from the application of EB and HB prediction, the ability to achieve output quality targets set by the ONS (2014b), and the fact that a four per cent annual survey envisaged by the ONS (2016b) should provide for survey requirements of model-based SAE, at least for the two health indicators we investigated.

Having demonstrated the basic feasibility and benefits of applying model-based SAE to the Census Transformation work, we can formulate several recommendations for future research. First and foremost, more work should be done on the basis of simulations presented in chapter 7 to investigate predictions for cross-classifications based on sex and ethnic groups in the first instance, as models are likely to be less efficient for minority ethnic groups. Model-based small area estimates of health characteristics across socioeconomic cross-classifications (housing tenure, socioeconomic classification, educational attainment) will also be much sought by users who monitor health inequalities. Beyond health outcomes, similar simulation work can easily be envisaged for outputs on economic activity topics in relation to the literature on state-space models. Further, and across all topics, these aspects should be explored while paying attention to appropriate modelling designs in order to combine data from several surveys into a single SAE model, as pioneered by Xie (2004); Schenker and Raghunathan (2007); Raghunathan et al. (2007); Davis et al. (2010). It is worth noting that accommodating the complex design of many different surveys and modes of administration (personal interviews, internet, paper-based) is likely to raise additional challenges necessitating close examination. Despite this, combining sources of data collections would make maximum use of the existing population statistics infrastructure, particularly labour force and housing surveys, and thus improve the efficiency of the methodology. These developments can also be investigated by building on the design approach developed in chapter 7.

Widening the range of small area population attribute statistics will undoubtedly require to develop new purposive covariates from administrative databases. The present thesis shows the very strong explanatory power of variations in health care utilisation for small area models. This supports the ONS' working hypothesis that health indexes can be developed from linked care records and used as covariate in predictive models. This would rest on new legal bases to disclose NHS commissioning data to the UK Statistics Authority under statutory provisions of the Statistics

and Registration Service Act 2007 (c. 18) and the Digital Economy Act 2017 (c. 30) (ONS, 2017). The access to a much wider range of auxiliary health data sources than was possible in this doctoral research is likely to deliver better-quality SAE models which can fulfil their purpose without the need for computationally intense spatial modelling. We do note, however, that the reliance on external data sources managed by a large number of independent organisations and information systems will deprive government statisticians of much of their faculty to scrutinise and audit all data sources. Furthermore, the attenuation, in the long-term, of the relationship between the target outcome and the available auxiliary data series (in this thesis, emergency hospitalisation rates) cannot be ruled out and further justifies good validation methodology and the involvement of scientists with subject area expertise.

Beyond furthering methodological research into the technical feasibility of model-based SAE, official statisticians must review the public acceptability (particularly with regard to non-specialist end users) of SAE's key principle: that of trading variance against bias. This would demand from national statistical institutes that they carry out in-depth qualitative research on user needs and statistical literacy. If precision is to be gained at the expense of bias, it is arguably a very different kind of error that is introduced to statistical bulletins. It is likely that end users are willing to accept this type of error provided that they are given the assurance that solid quality assurance work is carried out to detect areas with persistent and large bias, and address it when possible.

Addressing the issue of persistent bias in any given domain (a) for one-off or (b) across recurrent statistical publications, it is possible to envisage that statistical methodology focuses on two main developments:

- diagnostics tools for bias; in a similar way disease mapping has developed approaches for hot-spot detection with mixture distributions (see section 3.5.1.3, p. 68), it is possible to envisage quality assurance procedures to detect areas which do not fit the global model. This process can be incremental and affect the specification of future models.
- benchmarking can be used to calibrate predictors with strong persistent bias for recurrent publications, where it is possible to collate sample data in large amounts across several years to correct this persistent bias.

Satisfaction in any quality assurance framework building on such methodological developments can only be attained if:

- these tools are understandable and credible
- they are accompanied by power calculations which can inform end users of the scale of biases that can be detected with a given level of confidence given

the data available.

As for non-recurrent statistical outputs, the application of benchmarking or bias detection techniques is likely to be difficult in the absence of dense or frequent data collection.

In order for model-based SAE to become accepted as standard official statistics methodology, consensus and systematic criteria on what constitutes a good model should be reached. Building such consensus requires to engage both the academic and the non-academic communities with the assistance of structured approaches for consensus-building (Moher et al., 2010). This could take the form of working groups appraising existing practice on the basis of reviews comparable to the one presented in chapter 3. For academics, key areas to review include model selection and diagnostics, parametric assumptions and validation diagnostics. For government statisticians, work on guidance and minimum standards on quality metrics, approaches to validation and quality and methodology reporting standards is still needed. Government statisticians would also be required to consider how to collaborate with academics bringing substantive expertise on discrete topics in the form of model-building and validation. Model-based estimation is currently more common in environmental sciences, public health and developmental economics, where theories and the knowledge of determinants of health are exploited openly in order to build effective models. The use of theory or expert opinion on causal links is traditionally not part of official statistics practices. Model-based estimation presupposes certain assumptions, the objectivity of which may be disputed. In this regard, the conditions in and extent to which substantive expertise is used to guide official statistics methodology have not been greatly explored by the profession in accordance with requirements of impartiality and professional independence set by the statistics codes of conduct (UK Statistics Authority, 2009; European Statistical System, 2011).

As for the non-academic statistics users and the wider public, substantial mixed methods engagement would allow to better understand the acceptability of model-based estimation. There remain important gaps in understanding how end users make sense of such statistics once they understand their methodological assumptions and implications. Better evidence on this would shape the way model-based statistics can be communicated to the public and methodological safeguards ensuring their robustness and acceptability.

8.4.1 Conclusion

Assuming the challenges raised in methodology and public acceptability can be overcome, SAE could deliver substantial benefits in years to come. The production of census-like outputs on a more frequent basis is the most obvious. Beyond, the development of SAE to produce census outputs should open up a range of opportunities for the wider health statistics production in the UK. A range of topics which cannot be included in decennial censuses could in the future become available, for instance more sophisticated measurements of disability, doctor-diagnosed conditions and behavioural risk factors. In this regard, SAE design tools in the spirit of the approach elaborated in chapter 7 should allow individual departments and agencies to examine existing survey data and assess the value of collecting more data to produce small area estimates on health topics of interest. While SAE implies a significant investment in expertise, research and user engagement, efficiency gains illustrated in chapters 5 and 7 should convince government statisticians and the wider public that it will in many cases be worthwhile. The introduction of modelling in statistics production allows to incorporate information from data already available from a growing range of sources, guided by scientific consensus on methods and social theories. In comparison, simpler statistical outputs generated from strictly design-based estimators are very transparent, but fail to deliver the information that statistics users now require. Chapter 2 describes a range of sophisticated analytical needs in terms of ranking, measuring health inequalities, identifying hotspots and temporal trends which increasingly exceed what is provided by traditional statistical outputs.

In high-income countries, some national statistics institutes recognised for their methodological rigour now recognise that meeting growing user needs is becoming almost as important a priority as furthering trust in the impartiality and methodological rigour of statistical outputs. Model-based estimation is often a key instrument in order to operationalise such analytical needs. In low-income countries that do not enjoy such well-developed national statistics infrastructure, model-based estimation is a unique opportunity to derive much-needed statistics to target aid, protect the health of population, and inform investment in community services. Convincing illustrations are provided by Haslett et al. (2013) in the case of malnutrition and by Amoako Johnson et al. (2012) in the case of unmet need for contraception. Examples of applications supporting health protection include the estimation of infant mortality (Asiimwe et al., 2011) and arsenic contamination of water wells (Pal and Lahiri, 2017). Model-based estimation has the potential to make better use of

readily available data from DHSs matched with information available from existing statistics infrastructure (such as a census) or even without any such infrastructure (for instance, satellite imaging and land use classification). In middle-income countries, model-based estimation has the potential to enhance the quality and breadth of official statistics in a similar way as in high-income countries, with the difference that the costs of data acquisition can be considerably lower, for instance when it comes to survey data.

This thesis makes a pragmatic contribution to the model-based estimation debate, by acknowledging that health statistics users express rising needs. Model-based estimates are now filling a gap 'in which users previously had nothing' (Holt, 2007, p. 3). Many end users—and, indeed, statistics producers—will rightly be suspicious of the development of model-based estimation. As note by Holt (2007), 'the statistical judgments of the analyst are more central to modelling, and different statistical models will produce different estimates. Hence, the more comfortable stance of the national statistical office, "the estimates are driven by the design, hence unbiased, and therefore may be trusted," must be replaced with a stronger appeal to the professional competence and integrity of the staff undertaking the analysis' (*ibid.*, p. 4). It may be argued that public health intelligence is one of the areas where this transition may be less controversial thanks to the already well-established collaboration of substantive scientists with statisticians and accepted use of statistical models. We argue that improvements in validation methodology and the development of a design approach to model-based estimates are key requirements to achieve this.

Bibliography

Abberley, P. (1992). Counting Us Out: A Discussion of the OPCS Disability Surveys. *Disability, Handicap & Society*, **7**(2), 139–155. DOI: 10.1080/02674649266780171.

Abellan, J. J., Richardson, S., & Best, N. (2008). Use of Space-Time Models to Investigate the Stability of Patterns of Disease. *Environmental Health Perspectives*, **116**(8), 1111–1119. DOI: 10.1289/ehp.10814.

Acheson, D. (1998). Independent Inquiry into Inequalities in Health Report.

Adams, J. & White, M. (2006). Removing the Health Domain from the Index of Multiple Deprivation 2004 – Effect on Measured Inequalities in Census Measure of Health. *Journal of Public Health (Oxford, England)*, **28**(4), 379–83. DOI: 10.1093/pubmed/fdl061.

Adcock, C. J. (1988). A Bayesian Approach to Calculating Sample Sizes. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **37**(4/5), 433–439.

Adcock, C. T. (1997). Sample Size Determination: A Review. *The Statistician*, **46**(2), 261–283. DOI: 10.1111/1467-9884.00082.

Adrion, C. & Mansmann, U. (2012). Bayesian Model Selection Techniques As Decision Support for Shaping a Statistical Analysis Plan of a Clinical Trial: An Example from a Vertigo Phase III Study with Longitudinal Count Data As Primary Endpoint. *BMC medical research methodology*, **12**, 137. DOI: 10.1186/1471-2288-12-137.

Agresti, A. (2002). *Categorical Data Analysis*. Wiley-Interscience, New York.

Albert, J. H. (1999). Criticism of a Hierarchical Model Using Bayes Factors. *Statistics in Medicine*, **18**(3), 287–305. DOI: 10.1002/(SICI)1097-0258(19990215)18:3<287::AID-SIM18>3.0.CO;2-3.

Ambler, R., Caplan, D., Chambers, R., Kovacevic, M., & Wang, S. (2001). Combining Unemployment Benefits Data and LFS Data to Estimate ILO Unemployment for Small Areas: An Application of a Modified Fay-Herriot Method. In *Proceedings of the 53rd Conference of the International Statistical Institute*, Seoul. International Statistical Institute. URL: http://isi.cbs.nl/iamamember/CD2/pdf/725.PDF [Accessed: 2-2-2017].

AMELI (2011). AMELI – Advanced Methodology for European Laeken Indicators. URL: http://ameli.surveystatistics.net/ [Accessed: 2-2-2017].

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders (5th ed.)*. American Psychiatric Publishing.

Amoako Johnson, F., Chandra, H., Brown, J. J., & Padmadas, S. S. (2010). District-Level Estimates of Institutional Births in Ghana: Application of Small Area Estimation Technique Using Census and DHS Data. *Journal of Official Statistics*, **26**(2), 341–359.

Amoako Johnson, F., Padmadas, S. S., Chandra, H., Matthews, Z., & Madise, N. J. (2012). Estimating Unmet Need for Contraception by District Within Ghana: An Application of Small-Area Estimation Techniques. *Population Studies*, **66**(2), 105–122. DOI: 10.1080/00324728.2012.678585.

Anselin, L. (1995). Local Indicators of Spatial Association – LISA. *Geographical Analysis*, **27**(2), 93–115. DOI: 10.1111/j.1538-4632.1995.tb00338.x.

Anselin, L. (1996). The Moran Scatterplot As an ESDA Tool to Assess Instability in Spatial Association. In Fischer, M., Scholten, H., & Unwin, D., editors, *Spatial Analytical Perspectives on GIS*, pages 111–125. Taylor & Francis, London.

Asiimwe, J. B., Jehopio, P., Atuhaire, L. K., & Mbonye, A. K. (2011). Examining Small Area Estima-

tion Techniques for Public Health Intervention: Lessons From Application to Under-5 Mortality Data in Uganda. *Journal of Public Health Policy*, **32**(1), 1–15. DOI: 10.1057/jphp.2010.46.

Association of Public Health Observatories (2005). Sources of Data on Lifestyle Risk Factors in Local Populations. Technical Briefing No.1. URL: http://www.apho.org.uk/resource/item.aspx?RID=39442 [Accessed: 2-2-2017].

Asthana, S., Gibson, A., Moon, G., Dicker, J., & Brigham, P. (2004). The Pursuit of Equity in NHS Resource Allocation: Should Morbidity Replace Utilisation As the Basis for Setting Health Care Capitations? *Social Science & Medicine*, **58**(3), 539–551. DOI: 10.1016/S0277-9536(03)00217-X.

Austin, P. C. & Tu, J. V. (2004). Automated Variable Selection Methods for Logistic Regression Produced Unstable Models for Predicting Acute Myocardial Infarction Mortality. *Journal of Clinical Epidemiology*, **57**(11), 1138–1146. DOI: 10.1016/j.jclinepi.2004.04.003.

Baffour, B., King, T., & Valente, P. (2013). The Modern Census: Evolution, Examples and Evaluation. *International Statistical Review*, **81**(3), 407–425. DOI: 10.1111/insr.12036.

Bajekal, M., Wheller, L., & Dix, D. (2006). Estimating Residents and Staff in Communal Establishments from the 2001 Census. *Health Statistics Quarterly*, **31**, 42–50. URL: http://www.ons.gov.uk/ons/rel/hsq/health-statistics-quarterly/no--31--autumn-2006/estimating-residents-and-staff-in-communal-establishments-from-the-2001-census.pdf [Accessed: 2-2-2017].

Banerjee, S., Gelfand, A. E., & Carlin, B. P. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, Boca Raton.

Baptista, H., Mendes, J. M., De Almeida, J. C., & Xavier, M. B. (2015). Alcohol Abuse Disorder Prevalence and Its Distribution across Portugal. A Disease Mapping Approach. *REVSTAT – Statistical Journal*, **13**(1), 79–101.

Barker, L. & Thompson, T. (2013). Bayesian Small Area Estimates of Diabetes Incidence by United States County, 2009. *Journal of Data Science*, **11**(1), 249–267. URL: http://www.jds-online.com/files/JDS-1127.pdf [Accessed: 2-2-2017].

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, **67**(1). DOI: 10.18637/jss.v067.i01.

Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. *Journal of the American Statistical Association*, **83**(401), 28–36. DOI: 10.2307/2288915.

Bentham, G., Eimermann, J., Haynes, R., Lovett, A., & Brainard, J. (1995). Limiting Long-Term Illness and Its Associations with Mortality and Indicators of Social Deprivation. *Journal of Epidemiology & Community Health*, **49**(Suppl 2), S57–S64. DOI: 10.1136/jech.49.Suppl_2.S57.

Benzeval, M. & Judge, K. (1994). The Determinants of Hospital Utilization: Implications for Resource Allocation in England. *Health Economics*, **3**(2), 105–116. DOI: 10.1002/hec.4730030207.

Benzeval, M. & Judge, K. (1996). Access to Health Care in England : Continuing Inequalities in the Distribution of GPs. *Journal of Public Health Medicine*, **18**(1), 33–40.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., & Songini, M. (1995). Bayesian Analysis of Space-Time Variation in Disease Risk. *Statistics in Medicine*, **14**(21-22), 2433–2443. DOI: 10.1002/sim.4780142112.

Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **36**(2), 192–236.

Besag, J., York, J., & Mollié, A. (1991). Bayesian Image Restoration, with Two Applications in Spatial Statistics. *Annals of the Institute of Statistical Mathematics*, **43**(1), 1–20. DOI: 10.1007/BF00116466.

Best, N., Richardson, S., & Clarke, P. (2008). A Comparison of Model-Based Methods for Small Area Estimation. *ESRC National Centre for Research Methods*, pages 1–26. URL: http://www.bias-project.org.uk/papers/ComparisonSAE.pdf [Accessed: 2-2-2017].

Best, N., Waller, L., Thomas, A., Conlon, E., & Arnold, R. (1999). Bayesian Models for Spatially Correlated Disease and Exposure Data (with Discussion). In Bernardo, J., Berger, J., Dawid, A., & Smith, A., editors, *Bayesian Statistics 6*, pages 131–156. Oxford University Press, Oxford.

Bhati, A. S. (2008). A Generalized Cross-Entropy Approach for Modeling Spatially Correlated Counts. *Econometric Reviews*, **27**(4-6), 574–595. DOI: 10.1080/07474930801960451.

Bivand, R., Hauke, J., & Kossowski, T. (2013a). Computing the Jacobian in Gaussian Spatial Autoregressive Models: An Illustrated Comparison of Available Methods. *Geographical Analysis*, **45**(2), 150–179. DOI: 10.1111/gean.12008.

Bivand, R. & Piras, G. (2015). Comparing Implementations of Estimation Methods for Spatial Econometrics. *Journal of Statistical Software*, **63**(18), 1–36. URL: http://www.jstatsoft.org/v63/i18/. [Accessed: 2-2-2017].

Bivand, R. S., Pebesma, E., & Gómez-Rubio, V. (2013b). Modelling Areal Data. In *Applied Spatial Data Analysis with R*, pages 263–318. Springer New York, New York, NY. DOI: 10.1007/978-1-4614-7618-4_9.

Bizier, V., You, Y., Veilleux, L., & Grondin, C. (2009). Model-Based Approach to Small Area Estimation of Disability Counts and Rates Using Data from the 2006 Participation and Activity Limitation Survey. In *Proceedings of the 2009 Joint Statistical Meetings: Survey Research Methods Section*, pages 1232–1247, Washington, DC.

Blak, B., Thompson, M., Dattani, H., & Bourke, A. (2011). Generalisability of the Health Improvement Network (THIN) Database: Demographics, Chronic Disease Prevalence and Mortality Rates. *Journal of Innovation in Health Informatics*, **19**(4), 251–255. DOI: 10.14236/jhi.v19i4.820.

Bland, J. & Altman, D. (1986). Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. *Lancet*, **327**, 307–310. DOI: 10.1016/S0140-6736(86)90837-8.

Bleich, S. N., Jarlenski, M. P., Bell, C. N., & LaVeist, T. A. (2012). Health Inequalities: Trends, Progress, and Policy. *Annual Review of Public Health*, **33**(1), 7–40. DOI: 10.1146/annurev-publhealth-031811-124658.

Boonstra, H. J., Van Den Brakel, J. A., Buelens, B., Krieg, S., & Smeets, M. (2008). Towards Small Area Estimation at Statistics Netherlands. *Metron*, **66**(1), 21–49. URL: ftp://metron.sta.uniroma1.it/RePEc/articoli/2008-1-2.pdf [Accessed: 2-2-2017].

Borcard, D. & Legendre, P. (2002). All-Scale Spatial Analysis of Ecological Data by Means of Principal Coordinates of Neighbour Matrices. *Ecological Modelling*, **153**(1-2), 51–68. DOI: 10.1016/S0304-3800(01)00501-4.

Borooah, V. K. (1999). Occupational Class and the Probability of Long-Term Limiting Illness. In *Social Science & Medicine*, volume 49, pages 253–266. DOI: 10.1016/S0277-9536(99)00101-X.

Boulieri, A., Hansell, A., & Blangiardo, M. (2016a). Investigating Trends in Asthma and COPD through Multiple Data Sources: A Small Area Study. *Spatial and Spatio-Temporal Epidemiology*, 0, 1–9. DOI: 10.1016/j.sste.2016.05.004.

Boulieri, A., Liverani, S., de Hoogh, K., & Blangiardo, M. (2016b). A Space-Time Multivariate Bayesian Model to Analyse Road Traffic Accidents by Severity. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **Early View**(January). DOI: 10.1111/rssa.12178.

Boyle, P., Norman, P., & Rees, P. (2002). Does Migration Exaggerate the Relationship between Deprivation and Limiting Long-Term Illness? A Scottish Analysis. *Social Science & Medicine*, **55**(1), 21–31. DOI: 10.1016/S0277-9536(01)00217-9.

Briggs, A. D. M., Wolstenholme, J., Blakely, T., & Scarborough, P. (2016). Choosing an Epidemiological Model Structure for the Economic Evaluation of Non-Communicable Disease Public Health Interventions. *Population Health Metrics*, **14**(1), 17. DOI: 10.1186/s12963-016-0085-1.

Brown, G., Chambers, R., Heady, P., & Heasman, D. (2001). Evaluation of Small Area Estimation Methods – An Application to Unemployment Estimates from the UK LFS. In *Proceedings of Statistics Canada Symposium 2001 Achieving Data Quality in a Statistical Agency: A Methodological Perspective*. DOI: 10.1201/9780203166314.ch1.

Browne, W. J., Lahi, M. G., & Parker, R. M. (2009). *A Guide to Sample Size Calculations for Random Effect Models Via Simulation and the MLPowSim Software Package*. University of Bristol, Bristol. URL: http://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/mlpowsim-manual.pdf [Accessed: 2-2-2017].

Bundesinstitut für Bau-, Stadt- und Raumforschung (2016). INKAR. – Indikatoren und Karten zur Raum- und Stadtentwicklung. Bonn. URL: http://inkar.de/ [Accessed: 2-2-2017].

Bycroft, C. (2013). Options for Future New Zealand Censuses: Census Transformation Programme.

URL: http://www.statistics.govt.nz/~/media/Statistics/surveys-and-methods/methods/research-papers/topss/options-future-nz-censuses.pdf [Accessed: 2-2-2017].

Cadwell, B. L., Thompson, T. J., Boyle, J. P., & Barker, L. E. (2010). Bayesian Small Area Estimates of Diabetes Prevalence by US County, 2005. *Journal of Data Science*, **8**(1), 173–188.

Cairns, A. J. G., Blake, D., Dowd, K., & Kessler, A. R. (2016). Phantoms Never Die: Living with Unreliable Population Data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **179**(4), 975–1005. DOI: 10.1111/rssa.12159.

Carlin, B. P. & Gelfand, A. E. (1990). Approaches for Empirical Bayes Confidence Intervals. *Journal of the American Statistical Association*, **85**(409), 105–114. DOI: 10.1080/01621459.1990.10475312.

Carr-Hill, R., Hardman, G., Martin, S., Peacock, S., Sheldon, T. A., & Smith, P. C. (1994). A Formula for Distributing NHS Revenues Based on Small Area Use of Hospital Beds. *Centre for Health Economics Occasional Paper*, **22**. URL: https://www.york.ac.uk/che/pdf/op22.pdf [Accessed: 2-2-2017].

Carr-Hill, R. A., Hardman, G., Martin, S., Peacock, S., Sheldon, T. A., & Smith, P. C. (1997). A New Formula for Distributing Hospital Funds in England. *Interfaces*, **27**(1), 53–70. DOI: 10.1287/inte.27.1.53.

Carr-Hill, R. A., Sheldon, T. A., Smith, P., Martin, S., Peacock, S., & Hardman, G. (1994). Allocating Resources to Health Authorities: Development of Method for Small Area Analysis of Use of Inpatient Services. *British Medical Journal*, **309**(6961), 1046–1049. DOI: 10.1136/bmj.309.6961.1046.

Carroll, R., Lawson, A. B., Faes, C., Kirby, R. S., Aregay, M., & Watjou, K. (2015). Comparing INLA and OpenBUGS for Hierarchical Poisson Modeling in Disease Mapping. *Spatial and Spatio-Temporal Epidemiology*, **14-15**, 45–54. DOI: 10.1016/j.sste.2015.08.001.

Castillo-Carniglia, A., Kaufman, J. S., & Pino, P. (2015). Geographical Distribution of Alcohol-Attributable Mortality in Chile: A Bayesian Spatial Analysis. *Addictive Behaviors*, **42**, 207–215. DOI: 10.1016/j.addbeh.2014.11.025.

Chambers, R., Salvati, N., & Tzavidis, N. (2016). Semiparametric Small Area Estimation for Binary Outcomes with Application to Unemployment Estimation for Local Authorities in the UK. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **179**(2), 453–479. DOI: 10.1111/rssa.12123.

Chambers, R., van den Brakel, J., Hedlin, D., Lehtonen, R., & Zhang, L.-C. (2006). Future Challenges of Small Area Estimation. *Statistics in Transition*, **7**(4), 759–769.

Charlton, J. (1998). Use of the Census Samples of Anonymised Records (SARs) and Survey Data in Combination to Obtain Estimates at Local Authority Level. *Environment and Planning A*, **30**(5), 775–784. DOI: 10.1068/a300775.

Chaudhuri, S. & Ghosh, M. (2011). Empirical Likelihood for Small Area Estimation. *Biometrika*, **98**(2), 473–480. DOI: 10.1093/biomet/asr004.

Chen, C. X., Wakefield, J., Lumely, T., Lumley, T., & Wakefield, J. (2014). The Use of Sampling Weights in Bayesian Hierarchical Models for Small Area Estimation. DOI: 10.1016/j.sste.2014.07.002.

Choi, J., Lawson, A. B., Cai, B., & Hossain, M. M. (2011). Evaluation of Bayesian Spatiotemporal Latent Models in Small Area Health Data. *Environmetrics*, **22**(8), 1008–1022. DOI: 10.1002/env.1127.

Christiaensen, L., Lanjouw, P., Luoto, J., & Stifel, D. (2012). Small Area Estimation-Based Prediction Methods to Track Poverty: Validation and Applications. *The Journal of Economic Inequality*, **10**(2), 267–297. DOI: 10.1007/s10888-011-9209-9.

Clarke, R., Shipley, M., Lewington, S., Youngman, L., Collins, R., Marmot, M., & Peto, R. (1999). Underestimation of Risk Associations Due to Regression Dilution in Long-Term Follow-Up of Prospective Studies. *American Journal of Epidemiology*, **150**(4), 341–53. DOI: 10.1093/oxfordjournals.aje.a010013.

Clayton, D. & Kaldor, J. (1987). Empirical Bayes Estimates of Age-Standardized Relative Risks for

Use in Disease Mapping. *Biometrics*, **43**(3), 671–81. DOI: 10.2307/2532003.

Cliff, A. D. & Ord, J. K. (1981). *Spatial Processes: Models and Applications*. Pion, London.

Cohen, G., Forbes, J., & Garraway, M. (1995). Interpreting Self Reported Limiting Long Term Illness. *British Medical Journal*, **311**(7007), 722–724. DOI: 10.1136/bmj.311.7007.722.

Commander, M. J., Odell, S., Sashidharan, S. P., & Surtees, P. G. (1997). A Comparison of the Socio-Demographic and Clinical Characteristics of Private Household and Communal Establishment Residents in a Multi-Ethnic Inner-City Area. *Social Psychiatry and Psychiatric Epidemiology*, **32**(7), 421–7. DOI: 10.1007/BF00788183.

Congdon, P. (1995). The Impact of Area Context on Long Term Illness and Premature Mortality: An Illustration of Multi-Level Analysis. *Regional Studies*, **29**(February 2015), 327–344. DOI: 10.1080/00343409512331349003.

Congdon, P. (2000). A Bayesian Approach to Prediction Using the Gravity Model, with an Application to Patient Flow Modeling. *Geographical Analysis*, **32**(3), 205–224. DOI: 10.1111/j.1538-4632.2000.tb00425.x.

Congdon, P. (2002). A Life Table Approach to Small Area Health Need Profiling. *Statistical Modelling*, **2**(1), 63–88. DOI: 10.1191/1471082x02st026oa.

Congdon, P. (2006a). Estimating Diabetes Prevalence by Small Area in England. *Journal of Public Health (Oxford, England)*, **28**(1), 71–81. DOI: 10.1093/pubmed/fdi068.

Congdon, P. (2006b). Estimating Population Prevalence of Psychiatric Conditions by Small Area with Applications to Analysing Outcome and Referral Variations. *Health & Place*, **12**(4), 465–78. DOI: 10.1016/j.healthplace.2005.05.001.

Congdon, P. (2006c). A Model Framework for Mortality and Health Data Classified by Age, Area, and Time. *Biometrics*, **62**(1), 269–278. DOI: 10.1111/j.1541-0420.2005.00419.x.

Congdon, P. (2006d). Modelling Multiple Hospital Outcomes: The Impact of Small Area and Primary Care Practice Variation. *International Journal of Health Geographics*, **5**(1), 50. DOI: 10.1186/1476-072X-5-50.

Congdon, P. (2007). Mixtures of Spatial and Unstructured Effects for Spatially Discontinuous Health Outcomes. *Computational Statistics & Data Analysis*, **51**(6), 3197–3212. DOI: 10.1016/j.csda.2006.11.028.

Congdon, P. (2008a). Estimating CHD Prevalence by Small Area: Integrating Information from Health Surveys and Area Mortality. *Health & Place*, **14**(1), 59–75. DOI: 10.1016/j.healthplace.2007.04.003.

Congdon, P. (2008b). The Need for Psychiatric Care in England: A Spatial Factor Methodology. *Journal of Geographical Systems*, **10**(3), 217–239. DOI: 10.1007/s10109-008-0064-6.

Congdon, P. (2008c). A Spatially Adaptive Conditional Autoregressive Prior for Area Health Data. *Statistical Methodology*, **5**(6), 552–563. DOI: 10.1016/j.stamet.2008.02.005.

Congdon, P. (2009a). Life Expectancies for Small Areas: A Bayesian Random Effects Methodology. *International Statistical Review*, **77**(2), 222–240. DOI: 10.1111/j.1751-5823.2009.00080.x.

Congdon, P. (2009b). A Multilevel Model for Cardiovascular Disease Prevalence in the US and Its Application to Micro Area Prevalence Estimates. *International Journal of Health Geographics*, **8**, 6. DOI: 10.1186/1476-072X-8-6.

Congdon, P. (2014). Meta-Analysis and Multilevel Models. In *Applied Bayesian Modelling*, chapter 5, pages 183–214. John Wiley & Sons, Chichester, UK, 2nd edition.

Congdon, P. & Best, N. (2000). Small Area Variation in Hospital Admission Rates: Bayesian Adjustment for Primary Care and Hospital Factors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **49**(2), 207–226. DOI: 10.1111/1467-9876.00188.

Congdon, P. & Lloyd, P. (2010). Estimating Small Area Diabetes Prevalence in the US Using the Behavioral Risk Factor Surveillance System. *Journal of Data Science*, **8**(2), 235–252.

Cook, B. L., Kim, G., Morgan, K. L., Chen, C.-N., Nillni, A., & Alegría, M. (2016). Measuring Geographic "Hot Spot" of Racial/Ethnic Disparities: An Application to Mental Health Care. *Journal of Health Care for the Poor and Underserved*, **27**(2), 663–684. DOI: 10.1353/hpu.2016.0091.

Cools, W., Van den Noortgate, W., & Onghena, P. (2008). ML-DEs: A Program for Designing

Efficient Multilevel Studies. *Behavior Research Methods*, **40**(1), 236–249. DOI: 10.3758/BRM.40.1.236.

Curtis, D. (2003). Area Random Effects Model-Based Estimation of Unemployment at Local Authority Level. In *7th Meeting of the National Statistics Methodology Advisory Committee Area*. URL: http://www.ons.gov.uk/ons/guide-method/method-quality/advisory-committee/2001-2004/seventh-meeting/area-random-effects-model-based-estimation-of-unemployment-at-local-authority-level.pdf [Accessed: 2-2-2017].

Dale, M. R. T. & Fortin, M.-J. (2014). Spatial Regressions. In *Spatial Analysis: A Guide For Ecologists*, pages 227–239. Cambridge University Press, Cambridge, 2nd edition.

Darlington, F., Norman, P., Ballas, D., & Exeter, D. J. (2015). Exploring Ethnic Inequalities in Health: Evidence from the Health Survey for England, 1998–2011. *Diversity and Equality in Health and Care*, **12**(2), 54–65. URL: http://eprints.whiterose.ac.uk/89993/ [Accessed: 2-2-2017].

Datta, G. S. (2009). Model-Based Approach to Small Area Estimation. In Pfeffermann, D. & Rao, J., editors, *Sample Surveys: Inference and Analysis, Handbook of Statistics Vol. 29B*, chapter 32, pages 251–288. Elsevier Science B.V., Amsterdam, The Netherlands. DOI: 10.1016/S0169-7161(09)00232-6.

Datta, G. S., Hall, P., & Mandal, A. (2011). Model Selection by Testing for the Presence of Small-Area Effects, and Application to Area-Level Data. *Journal of the American Statistical Association*, **106**(493), 362–374. DOI: 10.1198/jasa.2011.tm10036.

Datta, G. S., Lahiri, P., Maiti, T., & Lu, K. L. (1999). Hierarchical Bayes Estimation of Unemployment Rates for the States of the U.S. *Journal of the American Statistical Association*, **94**(448), 1074–1082. DOI: 10.1080/01621459.1999.10473860.

Davis, W. W., Parsons, V. L., Xie, D., Schenker, N., Town, M., Raghunathan, T. E., & Feuer, E. J. (2010). State-Based Estimates of Mammography Screening Rates Based on Information from Two Health Surveys. *Public Health Reports (Washington, D.C. : 1974)*, **125**(4), 567–78. DOI: 10.2307/41435282.

De Santis, F. (2007). Using Historical Data for Bayesian Sample Size Determination. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **170**(1), 95–113. DOI: 10.1111/j.1467-985X.2006.00438.x.

Dempster, A. P. & Hwang, J. S. (1994). Bayesian Implementation of a Complex Hierarchical Model. In Anderson, T. W., Fang, K. T., & Olkin, I., editors, *Multivariate Analysis and its Applications. IMS Lecture Notes Monograph Series. Volume 24*, pages 221–236. Institute of Mathematical Statistics, Hayward, CA. DOI: 10.1214/lnms/1215463798.

Diez Roux, A. V. (2016). On the Distinction—Or Lack of Distinction—Between Population Health and Public Health. *American Journal of Public Health*, **106**(4), 619–620. DOI: 10.2105/AJPH.2016.303097.

Diggle, P. & Lophaven, S. (2006). Bayesian Geostatistical Design. *Scandinavian Journal of Statistics*, **33**(1), 53–64. DOI: 10.1111/j.1467-9469.2005.00469.x.

Dixon, J., Smith, P., Gravelle, H., Martin, S., Bardsley, M., Rice, N., Georghiou, T., Dusheiko, M., Billings, J., Lorenzo, M. D., & Sanderson, C. (2011). A Person Based Formula for Allocating Commissioning Funds to General Practices in England: Development of a Statistical Model. *British Medical Journal*, **343**, d6608–d6608. DOI: 10.1136/bmj.d6608.

Drewnowski, a., Rehm, C. D., & Arterburn, D. (2014). The Geographic Distribution of Obesity by Census Tract among 59,767 Insured Adults in King County, WA. *International Journal of Obesity*, **38**(6), 833–9. DOI: 10.1038/ijo.2013.179.

Drewnowski, A., Rehm, C. D., & Solet, D. (2007). Disparities in Obesity Rates: Analysis by Zip Code Area. *Social Science & Medicine*, **65**(12), 2458–2463. DOI: 10.1016/j.socscimed.2007.07.001.

Duncan, E. W., White, N. M., & Mengersen, K. (2016). Bayesian Spatiotemporal Modelling for Identifying Unusual and Unstable Trends in Mammography Utilisation. *BMJ Open*, **6**(5), e010253. DOI: 10.1136/bmjopen-2015-010253.

Dupont, E., Papadimitriou, E., Martensen, H., & Yannis, G. (2013). Multilevel Analysis in Road Safety Research. *Accident Analysis & Prevention*, **60**, 402–411. DOI: 10.1016/j.aap.2013.04.035.

Dwyer-Lindgren, L., Freedman, G., Engell, R. E., Fleming, T. D., Lim, S. S., Murray, C. J., & Mokdad, A. H. (2013). Prevalence of Physical Activity and Obesity in US Counties, 2001–2011: A Road Map for Action. *Population Health Metrics*, **11**(1), 7. DOI: 10.1186/1478-7954-11-7.

Dwyer-Lindgren, L., Kakungu, F., Hangoma, P., Ng, M., Wang, H., Flaxman, A. D., Masiye, F., & Gakidou, E. (2014a). Estimation of District-Level Under-5 Mortality in Zambia Using Birth History Data, 1980–2010. *Spatial and Spatio-Temporal Epidemiology*, **11**, 89–107. DOI: 10.1016/j.sste.2014.09.002.

Dwyer-Lindgren, L., Mokdad, A. H., Srebotnjak, T., Flaxman, A. D., Hansen, G. M., & Murray, C. J. (2014b). Cigarette Smoking Prevalence in US Counties: 1996–2012. *Population Health Metrics*, **12**(1), 5. DOI: 10.1186/1478-7954-12-5.

Earnest, A., Morgan, G., Mengersen, K., Ryan, L., Summerhayes, R., & Beard, J. (2007). Evaluating the Effect of Neighbourhood Weight Matrices on Smoothing Properties of Conditional Autoregressive (CAR) Models. *International Journal of Health Geographics*, **6**(1), 54. DOI: 10.1186/1476-072X-6-54.

Eayres, D. (2004). Evaluation of Methodologies for Small Area Life Expectancy Estimation. *Journal of Epidemiology & Community Health*, **58**(3), 243–249. DOI: 10.1136/jech.2003.009654.

Eberth, J. M., Hossain, M. M., Tiro, J. A., Zhang, X., Holt, J. B., & Vernon, S. W. (2013a). Human Papillomavirus Vaccine Coverage among Females Aged 11 to 17 in Texas Counties: An Application of Multilevel, Small Area Estimation. *Women's Health Issues*, **23**(2), e131–e141. DOI: 10.1016/j.whi.2012.12.005.

Eberth, J. M., Zhang, X., Hossain, M., Tiro, J. a., Holt, J. B., & Vernon, S. W. (2013b). County-Level Estimates of Human Papillomavirus Vaccine Coverage among Young Adult Women in Texas, 2008. *Texas Public Health Journal*, **65**(1), 37–40. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3901306&tool=pmcentrez&rendertype=abstract [Accessed: 2-2-2017].

Efron, B. & Morris, C. (1973). Stein's Estimation Rule and Its Competitors. An Empirical Bayes Approach. *Journal of the American Statistical Association*, **68**(341), 117–130. DOI: 10.2307/2284155.

Eke, P. I., Zhang, X., Lu, H., Wei, L., Thornton-Evans, G., Greenlund, K. J., Holt, J. B., & Croft, J. B. (2016). Predicting Periodontitis at State and Local Levels in the United States. *Journal of Dental Research*, **95**(5), 515–522. DOI: 10.1177/0022034516629112.

Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003). Micro-Level Estimation of Poverty and Inequality. *Econometrica*, **71**(1), 355–364. DOI: 10.1111/1468-0262.00399.

Elston, J. M., Koch, G. G., & Weissert, W. G. (1991). Regression-Adjusted Small Area Estimates of Functional Dependency in the Noninstitutionalized American Population Age 65 and Over. *American Journal of Public Health*, **81**(3), 335–343. DOI: 10.2105/AJPH.81.3.335.

Etilé, F. & Milcent, C. (2006). Income-Related Reporting Heterogeneity in Self-Assessed Health: Evidence From France. *Health Economics*, **15**(9), 965–81. DOI: 10.1002/hec.1164.

European Statistical System (2011). European Statistics Code of Practice for the National and Community Statistical Authorities. Adopted by the European Statistical System Committee 28th September 2011. URL: http://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-32-11-955 [Accessed: 2-2-2017].

EuroQol Group (2009). EQ-5D-5L Health Questionnaire. English Version for the UK. URL: http://www.euroqol.org/fileadmin/user_upload/Documenten/PDF/Products/Sample_UK__English__EQ-5D-5L_Paper_Self_complete_v1.0__ID_24700_.pdf [Accessed: 2-2-2017].

Evans, R. G. & Stoddart, G. L. (1990). Producing Health, Consuming Health Care. *Social Science & Medicine*, **31**(12), 1347–1363. DOI: 10.1016/0277-9536(90)90074-3.

Fabrizi, E., Montanari, G. E., & Giovanna Ranalli, M. (2016). A Hierarchical Latent Class Model for Predicting Disability Small Area Counts from Survey Data. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **179**(1), 103–131. DOI: 10.1111/rssa.12112.

Fabrizi, E. & Trivisano, C. (2010). Robust Linear Mixed Models for Small Area Estimation. *Journal of Statistical Planning and Inference*, **140**(2), 433–443. DOI: 10.1016/j.jspi.2009.07.022.

Faculty of Public Health (2014). Functions of the Local Public Health System. URL: http://www.fph.org.uk/uploads/FunctionsofthelocalPHsystemFINAL200514.pdf [Accessed: 2-2-2017].

Faculty of Public Health (2016). Good Public Health Practice Framework. April 2016. URL: http://www.fph.org.uk/uploads/GoodPublicHealthPracticeFramework_2016_Final.pdf [Accessed: 2-2-2017].

Fahrmeir, L. & Lang, S. (2001). Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **50**(2), 201–220. DOI: 10.1111/1467-9876.00229.

Farrell, P. J., Macgibbon, B., & Tomberlin, T. J. (1997a). Empirical Bayes Estimators of Small-Area Proportions in Multistage Designs. *Statistica Sinica*, **7**, 1065–1083. URL: http://www3.stat.sinica.edu.tw/statistica/oldpdf/A7n417.pdf [Accessed: 2-2-2017].

Farrell, P. J., MacGibbon, B., & Tomberlin, T. J. (1997b). Empirical Bayes Small-Area Estimation Using Logistic Regression Models and Summary Statistics. *Journal of Business & Economic Statistics*, **15**(1), 101–108. DOI: 10.1080/07350015.1997.10524692.

Fay, R. E. & Herriot, R. A. (1979). Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, **74**(366), 269–277. URL: http://www.jstor.org/stable/2286322 [Accessed: 2-2-2017].

Fienberg, S. E. (2011). Bayesian Models and Methods in Public Policy and Government Settings. *Statistical Science*, **26**(2), 212–226. DOI: 10.1214/10-STS331.

Fillingham, D. (2012). Letter to Rt Hon Jeremy Hunt MP. 17 October 2012. URL: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/213327/DF-letter-to-SoS.pdf [Accessed: 2-2-2017].

Finney, N., Lymperopoulou, K., Kapoor, N., Marshall, A., Sabater, A., & Simpson, L. (2014). Local Ethnic Inequalities: Ethnic Differences in Education, Employment, Health and Housing in Districts of England and Wales, 2001–2011. URL: http://www.runnymedetrust.org/uploads/Inequalitiesreport-finalv2.pdf [Accessed: 2-2-2017].

Fitzpatrick, J. & Jacobson, B. (2003). Local Basket of Inequalities Indicators: Full Report.

Flanders, W. D. & Kleinbaum, D. G. (1995). Basic Models for Disease Occurrence in Epidemiology. *International Journal of Epidemiology*, **24**(1), 1–7. DOI: 10.1093/ije/24.1.1.

Fong, Y., Rue, H., & Wakefield, J. (2010). Bayesian Inference for Generalized Linear Mixed Models. *Biostatistics*, **11**(3), 397–412. DOI: 10.1093/biostatistics/kxp053.

France. Haut Conseil de la Santé Publique (2009). Les inégalités sociales de santé : Sortir de la fatalité. URL: http://www.hcsp.fr/explore.cgi/avisrapportsdomaine?clefr=113 [Accessed: 2-2-2017].

Freeman, G. (2016). The Care Quality Commission and National Data Guardian for Health and Care's Independent Reviews into Data Security, Consent and Opt-Outs: Written Statement (HCWS62). 6 July 2016. URL: http://www.parliament.uk/business/publications/written-questions-answers-statements/written-statement/Commons/2016-07-06/HCWS62 [Accessed: 2-2-2017].

Fries, J. F., Koop, E. C., Sokolov, J., Beadle, C. E., & Wright, D. (1998). Beyond Health Promotion: Reducing Need and Demand for Medical Care. *Health Affairs*, **17**(2), 70–84. DOI: 10.1377/hlthaff.17.2.70.

Fuller, W. A. (1999). Environmental Surveys Over Time. *Journal of Agricultural, Biological, and Environmental Statistics*, **4**(4), 331–345. DOI: 10.2307/1400493.

Gart, J. J. & Zweifel, J. R. (1967). On the Bias of Various Estimators of the Logit and Its Variance with Application to Quantal Bioassay. *Biometrika*, **4**(1/2), 181–187. DOI: 10.2307/2333861.

Gelfand, A., Dey, D., & Chang, H. (1992). Model Determination Using Predictive Distributions with Implementation Via Sampling Based Methods (with Discussion). In Bernardo, J. M., Berger, J. O., & Dawid, A. P., editors, *Bayesian Statistics 4: Proceedings of the 4th Valencia International Meeting*, pages 147–167. Oxford University Press.

Gelfand, A. E. (1995). Model Determination Using Sampling-Based Methods. In W. Gilks, Richardson, S., & Spiegelhalter, D. J., editors, *Markov Chain Monte Carlo In Practice*, pages 145–161. Chapman Hall, London.

Gelman, A., Goegebeur, Y., Tuerlinckx, F., & Van Mechelen, I. (2000). Diagnostic Checks for Discrete Data Regression Models Using Posterior Predictive Simulations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **49**(2), 247–268. DOI: 10.1111/1467-9876.00190.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies. *Statistica Sinica*, **6**(4), 733–760. URL: http://www3.stat.sinica.edu.tw/statistica/oldpdf/A6n41.pdf [Accessed: 2-2-2017].

Gelman, A. & Price, P. N. (1999). All Maps of Parameter Estimates Are Misleading. *Statistics in Medicine*, **18**(23), 3221–3234. DOI: 10.1002/(SICI)1097-0258(19991215)18:23<3221::AID-SIM312>3.0.CO;2-M.

Georgia Department of Public Health (2016). OASIS. – Online Analytical Statistical Information System. URL: https://oasis.state.ga.us/ [Accessed: 2-2-2017].

Ghosh, M., Natarajan, K., Waller, L. A., & Kim, D. (1999). Hierarchical Bayes GLMs for the Analysis of Spatial Data: An Application to Disease Mapping. *Journal of Statistical Planning and Inference*, **75**(2), 305–318. DOI: 10.1016/S0378-3758(98)00150-5.

Ghosh, M. & Rao, J. N. K. (1994). Small Area Estimation: An Appraisal. *Statistical Science*, **9**(1), 55–76. URL: https://projecteuclid.org/euclid.ss/1177010647 [Accessed: 2-2-2017].

Gibson, A., Asthana, S., Brigham, P., Moon, G., & Dicker, J. (2002). Geographies of Need and the New NHS: Methodological Issues in the Definition and Measurement of the Health Needs of Local Populations. *Health & Place*, **8**(1), 47–60. DOI: 10.1016/S1353-8292(01)00035-1.

Gołata, E. (2014). SAE Education Challenges to Academics and NSI. *Statistics in Transition new series and Survey Methodology*, **16**(4 Joint Issue: Small Area Estimation 2014), 611–630.

Gold, M. R., Stevenson, D., & Fryback, D. G. (2002). HALYs and QALYs and DALYs, Oh My: Similarities and Differences in Summary Measures of Population Health. *Annual Review of Public Health*, **23**(1), 115–134. DOI: 10.1146/annurev.publhealth.23.100901.140513.

Goldstein, H. & Rasbash, J. (1996). Improved Approximations for Multilevel Models with Binary Responses. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **159**(3), 505. DOI: 10.2307/2983328.

Gomez-Rubio, V., Best, N., Richardson, S., Li, G., & Clarke, P. (2010). Bayesian Statistics for Small Area Estimation. URL: http://eprints.ncrm.ac.uk/1686/ [Accessed: 2-2-2017].

Gonzalez, M. E. (1973). Use and Evaluation of Synthetic Estimates. In *Proceedings of the Social Statistics Section*, pages 33–36. American Statistical Association.

Gonzalez, M. E. & Hoza, C. (1978). Small Area Estimation with Application to Unemployment and Housing Estimates. *Journal of the American Statistical Association*, **73**(361), 7–15. DOI: 10.1080/01621459.1978.10479991.

Gonzalez, M. E. & Waksberg, J. E. (1973). Estimation of the Error of Synthetic Estimates. In *Paper Presented at the first meeting of the International Association of Survey Statisticians, 18-25 August 1973*, Vienna, Austria.

González-Manteiga, W., Lombardía, M., Molina, I., Morales, D., & Santamaría, L. (2007). Estimation of the Mean Squared Error of Predictors of Small Area Linear Parameters under a Logistic Mixed Model. *Computational Statistics & Data Analysis*, **51**(5), 2720–2733. DOI: 10.1016/j.csda.2006.01.012.

Goodman, M. S. (2010). Comparison of Small-Area Analysis Techniques for Estimating Prevalence by Race. *Preventing Chronic Disease*, **7**(2), A33. URL: http://www.cdc.gov/pcd/issues/2010/mar/09_0026.htm [Accessed: 2-2-2017].

Gravelle, H., Sutton, M., Morris, S., Windmeijer, F., Leyland, A., Dibben, C., & Muirhead, M. (2003). Modelling Supply and Demand Influences on the Use of Health Care: Implications for Deriving a Needs-Based Capitation Formula. *Health Economics*, **12**(12), 985–1004. DOI: 10.1002/hec.830.

Green, P. J. & Richardson, S. (2002). Hidden Markov Models and Disease Mapping. *Journal of the American Statistical Association*, **97**(460), 1055–1070. DOI: 10.1198/016214502388618870.

Greenland, S. (1989). Modeling and Variable Selection in Epidemiologic Analysis. *American Journal of Public Health*, **79**(3), 340–349. DOI: 10.2105/AJPH.79.3.340.

Greenland, S., Daniel, R., & Pearce, N. (2016). Outcome Modelling Strategies in Epidemiology: Traditional Methods and Basic Alternatives. *International Journal of Epidemiology*, **45**(2), 565–575. DOI: 10.1093/ije/dyw040.

Griffith, D. A. (2000). A Linear Regression Solution to the Spatial Autocorrelation Problem. *Journal of Geographical Systems*, **2**(2), 141–156. DOI: 10.1007/PL00011451.

Griffiths, C. & Fitzpatrick, J. (2001). *Geographic Variations in Health. Decennial Supplement No 16*. HMSO, London. URL: http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/rel/subnational-health3/geographic-variations-in-health--ds-no-16-/2001/index.html [Accessed: 2-2-2017].

Grilli, L., Metelli, S., & Rampichini, C. (2014). Bayesian Estimation with Integrated Nested Laplace Approximation for Binary Logit Mixed Models. *Journal of Statistical Computation and Simulation*, (August 2015), 1–9. DOI: 10.1080/00949655.2014.935377.

Ha, N. S., Lahiri, P., & Parsons, V. (2014). Methods and Results for Small Area Estimation Using Smoking Data from the 2008 National Health Interview Survey. *Statistics in Medicine*, **33**(22), 3932–3945. DOI: 10.1002/sim.6219.

Hall, P. & Maiti, T. (2006). On Parametric Bootstrap Methods for Small Area Prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**(2), 221–238. DOI: 10.1111/j.1467-9868.2006.00541.x.

Halonen, J. I., Hansell, A. L., Gulliver, J., Morley, D., Blangiardo, M., Fecht, D., Toledano, M. B., Beevers, S. D., Anderson, H. R., Kelly, F. J., & Tonne, C. (2015). Road Traffic Noise Is Associated with Increased Cardiovascular Morbidity and Mortality and All-Cause Mortality in London. *European Heart Journal*, **36**(39), 2653–2661. DOI: 10.1093/eurheartj/ehv216.

Hanks, E. M., Schliep, E. M., Hooten, M. B., & Hoeting, J. A. (2015). Restricted Spatial Regression in Practice: Geostatistical Models, Confounding, and Robustness under Model Misspecification. *Environmetrics*, **26**(4), 243–254. DOI: 10.1002/env.2331.

Haslett, S., Jones, G., & A., S. (2013). *Small-Area Estimation of Poverty and Malnutrition in Cambodia, National Institute of Statistics, Ministry of Planning, Royal Government of Cambodia and the United Nations World Food Programme, Cambodia*.

Haslett, S., Jones, G., Noble, A., & Ballas, D. (2010a). More for Less? Comparing Small Area Estimation, Spatial Microsimulation, and Mass Imputation. *Proceedings of the 2010 Joint Statistical Meetings: Survey Research Methods Section*, pages 1584–1598.

Haslett, S. J., Isidro, M. C., Geoffrey, & Jones (2010b). Comparison of Survey Regression Techniques in the Context of Small Area Estimation of Poverty. *Survey Methodology*, **36**(2), 157–170.

Hastings, D. (2002). Annual Local Area Labour Force Survey Data for 2000/2001. *Labour Market Trends*, **110**(1), 33–41. URL: http://www.ons.gov.uk/ons/rel/lms/labour-market-trends--discontinued-/volume-110--no--1/labour-market-trends.pdf [Accessed: 2-2-2017].

Heady, P. (2003). *Model-Based Small Area Estimation Series. Small Area Estimation Project Report*. Office for National Statistics, London.

Heady, P., Clarke, P., Brown, G., Ellis, K., Heasman, D., Hennell, S., Longhurst, J., & Mitchell, B. (2003). Model-Based Small Area Estimation Series No. 2. Small Area Estimation Project Report.

Heady, P. & Ralphs, M. (2005). EURAREA: An Overview of the Project and its Findings. *Statistics in Transition*, **7**(3), 557–570.

Heady, P. & Ruddock, V. (1996). Report on a Project to Estimate the Incidence of Psychiatric Morbidity in Small Areas.

Hennessy, D. A., Flanagan, W. M., Tanuseputro, P., Bennett, C., Tuna, M., Kopec, J., Wolfson, M. C., & Manuel, D. G. (2015). The Population Health Model (POHEM): An Overview of Rationale, Methods and Applications. *Population Health Metrics*, **13**(1), 24. DOI: 10.1186/s12963-015-0057-x.

Hernández-Quevedo, C., Jones, A. M., & Rice, N. (2005). Reporting Bias and Heterogeneity in Self-Assessed Health. Evidence from the British Household Panel Survey. *HEDG Working Paper*,

05/04.

Hernández-Quevedo, C., Masseria, C., & Mossialos, E. (2010). Methodological Issues in the Analysis of the Socioeconomic Determinants of Health Using EU-SILC Data. DOI: 10.2785/55316.

Herrador, M., Esteban, M. D., Hobza, T., & Morales, D. (2011). A Fay-Herriot Model with Different Random Effect Variances. *Communications in Statistics - Theory and Methods*, **40**(5), 785–797. DOI: 10.1080/03610920903480858.

Herrett, E., Gallagher, A. M., Bhaskaran, K., Forbes, H., Mathur, R., van Staa, T., & Smeeth, L. (2015). Data Resource Profile: Clinical Practice Research Datalink (CPRD). *International Journal of Epidemiology*, **44**(3), 827–836. DOI: 10.1093/ije/dyv098.

Holt, D. T. (2007). The Official Statistics Olympic Challenge. *The American Statistician*, **61**(1), 1–8. DOI: 10.1198/000313007X168173.

Home Office & TNS-BMRB (2012). British Crime Survey, 2010–2011. [data collection] SN: 6937. January 2012.. DOI: 10.5255/UKDA-SN-6937-1.

HSCIC (2016). Data Quality Maturity Index: Introduction and Methodology. Published 24 May 2016. URL: http://content.digital.nhs.uk/media/21885/Methodology/pdf/DQMI_methodology.pdf [Accessed: 2-2-2017].

Hudson, C. G. (2009). Validation of a Model for Estimating State and Local Prevalence of Serious Mental Illness. *International Journal of Methods in Psychiatric Research*, **18**(4), 251–264. DOI: 10.1002/mpr.294.

Hudson, C. G. & Soskolne, V. (2012). Disparities in the Geography of Serious Mental Illness in Israel. *Health & Place*, **18**(4), 898–910. DOI: 10.1016/j.healthplace.2012.02.008.

Hunt, J. (2016). Direction to the Health and Social Care Information Centre to Process Type 2 Objections. URL: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/517522/type2objections.pdf [Accessed: 2-2-2017].

Hwang, J.-S. & Dempster, A. P. (1999). A Stochastic System for Modeling Labor Force Series of Small Areas. *Statistica Sinica*, **9**, 297–324.

Information Standards Board for Health and Social Care (2012). Commissioning Data Sets (CDS) V6.2 Standard Specification (ISB 0092 Amd 16/2010 Change). URL: http://content.digital.nhs.uk/media/17281/0092162010sspec/pdf/0092162010sspec.pdf [Accessed: 2-2-2017].

Ipsos Mori (2007). Active People Survey Small Area Estimates. Research Study Conducted for Sport England. Final Report. URL: https://www.ipsos-mori.com/Assets/Docs/Publications/Ops_RMC_APS_Document_08_03_10.pdf [Accessed: 2-2-2017].

Jefferies, J. & Fulton, R. (2005). Making a Population Estimate in England and Wales, *National Statistics Methodological Series No. 34*. URL: http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/guide-method/method-quality/specific/gss-methodology-series/gss-methodology-series--34--making-a-population-estimate-in-england-and-wales.pdf [Accessed: 2-2-2017].

Jia, H., Muennig, P., & Borawski, E. (2004). Comparison of Small-Area Analysis Techniques for Estimating County-Level Outcomes. *American Journal of Preventative Medicine*, **26**(5), 453–460. DOI: 10.1016/j.amepre.2004.02.004.

Jiang, J. & Lahiri, P. (2001). Empirical Best Prediction for Small Area Inference with Binary Data. *Annals of the Institute of Statistical Mathematics*, **53**(2), 217–243. DOI: 10.1023/A:1012410420337.

Jiang, J. & Lahiri, P. (2006). Mixed Model Prediction and Small Area Estimation. *Test*, **15**(1), 1–96. DOI: 10.1007/BF02595419.

Jiang, J. & Nguyen, T. (2012). Small Area Estimation Via Heteroscedastic Nested-Error Regression. *Canadian Journal of Statistics*, **40**(3), 588–603. DOI: 10.1002/cjs.11148.

Jiang, J., Nguyen, T., & Rao, J. S. (2009). A Simplified Adaptive Fence Procedure. *Statistics & Probability Letters*, **79**(5), 625–629. DOI: 10.1016/j.spl.2008.10.014.

Jiang, J., Rao, J. S., Gu, Z., & Nguyen, T. (2008). Fence Methods for Mixed Model Selection. *The Annals of Statistics*, **36**(4), 1669–1692. DOI: 10.1214/07-AOS517.

Johnson, R. J. & Wolinsky, F. D. (1993). The Structure of Health Status among Older Adults: Disease, Disability, Functional Limitation, and Perceived Health. *Journal of Health and Social*

Behavior, **34**(2), 105–121. DOI: 10.2307/2137238.

Johnston, D. W., Propper, C., & Shields, M. A. (2009). Comparing Subjective and Objective Measures of Health: Evidence from Hypertension for the Income/Health Gradient. *Journal of Health Economics*, **28**(3), 540–552. DOI: 10.1016/j.jhealeco.2009.02.010.

Jones, K. & Duncan, C. (1995). Individuals and Their Ecologies: Analysing the Geography of Chronic Illness within a Multilevel Modelling Framework. *Health & Place*, **1**(1), 27–40. DOI: 10.1016/1353-8292(95)00004-6.

Jonker, M. F., Congdon, P. D., van Lenthe, F. J., Donkers, B., Burdorf, A., & Mackenbach, J. P. (2013a). Small-Area Health Comparisons Using Health-Adjusted Life Expectancies: A Bayesian Random-Effects Approach. *Health & Place*, **23**, 70–78. DOI: 10.1016/j.healthplace.2013.04.003.

Jonker, M. F., Van Lenthe, F. J., Congdon, P. D., Donkers, B., Burdorf, A., & MacKenbach, J. P. (2012). Comparison of Bayesian Random-Effects and Traditional Life Expectancy Estimations in Small-Area Applications. *American Journal of Epidemiology*, **176**(10), 929–937. DOI: 10.1093/aje/kws152.

Jonker, M. F., van Lenthe, F. J., Donkers, B., Congdon, P. D., Burdorf, A., & Mackenbach, J. P. (2013b). The Impact of Nursing Homes on Small-Area Life Expectancies. *Health & Place*, **19**, 25–32. DOI: 10.1016/j.healthplace.2012.09.012.

Jordan, K., Ong, B. N., & Croft, P. (2000). Researching Limiting Long-Term Illness. *Social Science & Medicine*, **50**(3), 397–405. DOI: 10.1016/S0277-9536(99)00297-X.

Jordan, K., Ong, B. N., & Croft, P. (2003). Previous Consultation and Self Reported Health Status as Predictors of Future Demand for Primary Care. *Journal of Epidemiology & Community Health*, **57**, 109–113. DOI: 10.1136/jech.57.2.109.

Jordan, R. E., Lam, K.-b. H., Cheng, K. K., Miller, M. R., Marsh, J. L., Ayres, J. G., Fitzmaurice, D., & Adab, P. (2010). Case Finding for Chronic Obstructive Pulmonary Disease: A Model for Optimising a Targeted Approach. *Thorax*, **65**(6), 492–498. DOI: 10.1136/thx.2009.129395.

Joseph, L., Wolfson, D. B., & Du Berger, R. (1995). Sample Size Calculations for Binomial Proportions Via Highest Posterior Density Intervals. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **44**(2), 143–154.

Kilpeläinen, K., Aromaa, A., & ECHIM Core Group (2008). European Health Indicators: Development and Initial Implementation. Final Report of the ECHIM Project. URL: http://www.uzis.cz/cz/echim/final-report.pdf [Accessed: 2-2-2017].

Kinney, S. K. & Dunson, D. B. (2007). Fixed and Random Effects Selection in Linear and Logistic Models. *Biometrics*, **63**(3), 690–698. DOI: 10.1111/j.1541-0420.2007.00771.x.

Kirkup, B. & Forster, D. (1990). How Will Health Needs Be Measured in Districts? Implications of Variations in Hospital Use. *Journal of Public Health (Oxford, England)*, **12**(1), 45–50. DOI: 10.1093/oxfordjournals.pubmed.a042505.

Klein, R. J., Proctor, S. E., Boudreault, M. a., & Turczyn, K. M. (2002). Healthy People 2010 Criteria for Data Suppression. URL: http://www.cdc.gov/nchs/data/statnt/statnt24.pdf [Accessed: 2-2-2017].

Knorr-Held, L. & Besag, J. (1998). Modelling Risk from a Disease in Time and Space. *Statistics in Medicine*, **17**(18), 2045–2060. DOI: 10.1002/(SICI)1097-0258(19980930)17:18<2045::AID-SIM943>3.0.CO;2-P.

Knorr-Held, L. & Best, N. G. (2001). A Shared Component Model for Detecting Joint and Selective Clustering of Two Diseases. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **164**(1), 73–85. DOI: 10.1111/1467-985X.00187.

Knutson, K., Zhang, W., & Tabnak, F. (2008). Applying the Small-Area Estimation Method to Estimate a Population Eligible for Breast Cancer Detection Services. *Preventing Chronic Disease*, **5**(1), A10.

Kokki, E., Ranta, J., Penttinen, a., Pukkala, E., & Pekkanen, J. (2001). Small Area Estimation of Incidence of Cancer around a Known Source of Exposure with Fine Resolution Data. *Occupational and Environmental Medicine*, **58**(5), 315–320. DOI: 10.1136/oem.58.5.315.

Kordos, J. (2014). Development of Small Area Estimation in Official Statistics. *Statistics in Transition*

new series and Survey Methodology, **17**(1 Joint Issue: Small Area Estimation 2014), 105–132.

Krapavickaitė, D. & Rudys, T. (2015). Small Area Estimates for the Fraction of the Unemployed. *Lithuanian Mathematical Journal*, **55**(2), 243–254. DOI: 10.1007/s10986-015-9277-9.

Kroll, L. E. & Lampert, T. (2012). Regionalisierung von Gesundheitsindikatoren: Ergebnisse aus der Geda-Studie 2009. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz*, **55**(1), 129–140. DOI: 10.1007/s00103-011-1403-1.

Kukutai, T., Thompson, V., & McMillan, R. (2015). Whither the Census? Continuity and Change in Census Methodologies Worldwide, 1985–2014. *Journal of Population Research*, **32**(1), 3–22. DOI: 10.1007/s12546-014-9139-z.

Lahiri, P. & Mukherjee, K. (2007). On the Design-Consistency Property of Hierarchical Bayes Estimators in Finite Population Sampling. *The Annals of Statistics*, **35**(2), 724–737. DOI: 10.1214/009053606000001262.

Landrum, M. B., Bronskill, S. E., & Normand, S.-L. T. (2000). Analytic Methods for Constructing Cross-Sectional Profiles of Health Care Providers. *Health Services and Outcomes Research Methodology*, **1**(1), 23–47. DOI: 10.1023/A:1010093701870.

Last, J. (1963). The Iceberg: "Completing the Clinical Picture" in General Practice. *The Lancet*, **282**(7297), 28–31. DOI: 10.1016/S0140-6736(63)92546-7.

Lawson, A., Rotejanaprasert, C., Moraga, P., & Choi, J. (2015). A Shared Neighbor Conditional Autoregressive Model for Small Area Spatial Data. *Environmetrics*, **26**(6), 383–392. DOI: 10.1002/env.2346.

Lawson, A. B., Biggeri, A. B., Boehning, D., Lesaffre, E., Viel, J.-F., Clark, A., Schlattmann, P., & Divino, F. (2000). Disease Mapping Models : An Empirical Evaluation. *Statistics in Medicine*, **19**, 2217–2241.

Lawson, A. B. & Clark, A. (2002). Spatial Mixture Relative Risk Models Applied to Disease Mapping. *Statistics in Medicine*, **21**(3), 359–370. DOI: 10.1002/sim.1022.

Lawson, A. B., Song, H. R., Cai, B., Hossainc, M. M., & Huang, K. (2010). Space-Time Latent Component Modeling of Geo-Referenced Health Data. *Statistics in Medicine*, **29**(19), 2012–2027. DOI: 10.1002/sim.3917.

Lee, D. (2011). A Comparison of Conditional Autoregressive Models Used in Bayesian Disease Mapping. *Spatial and Spatio-Temporal Epidemiology*, **2**(2), 79–89. DOI: 10.1016/j.sste.2011.03.001.

Lee, D. & Mitchell, R. (2012). Boundary Detection in Disease Mapping Studies. *Biostatistics*, **13**(3), 415–426. DOI: 10.1093/biostatistics/kxr036.

Lee, D. & Mitchell, R. (2013). Locally Adaptive Spatial Smoothing Using Conditional Auto-Regressive Models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **62**(4), 593–608. DOI: 10.1111/rssc.12009.

Legendre, P. & Fortin, M.-J. (2010). Comparison of the Mantel Test and Alternative Approaches for Detecting Complex Multivariate Relationships in the Spatial Analysis of Genetic Data. *Molecular Ecology Resources*, **10**(5), 831–844. DOI: 10.1111/j.1755-0998.2010.02866.x.

Lehtonen, R. & Veijanen, A. (2009). Design-Based Methods of Estimation for Domains and Small Areas. In Pfeffermann, D. & Rao, J., editors, *Sample Surveys: Inference and Analysis, Handbook of Statistics Vol. 29B*, chapter 31, pages 219–249. Elsevier Science B.V.

Lem, S. (1999). *His Master's Voice*. Northwestern University Press, Evanston, Illinois.

Leroux, B., Maynard, R., Domoto, P., Zhu, C., & Milgrom, P. (1996). The Estimation of Caries Prevalence in Small Areas. *Journal of Dental Research*, **75**(12), 1947–1956. DOI: 10.1177/00220345960750120601.

Leroux, B. G., Lei, X., & Breslow, N. (2000). Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence. In Halloran, M. E. & Berry, D., editors, *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, volume 116 of *The IMA Volumes in Mathematics and its Applications*, pages 179–191. Springer New York, New York, NY. DOI: 10.1007/978-1-4612-1284-3_4.

Levy, P. S. (1979). Small Area Estimation—Synthetic and Other Procedures, 1968–1978. In Steinberg, J., editor, *Synthetic Estimates for Small Areas: Statistical Workshop Papers and Discussion.*

NIDA Research Monograph No 24, pages 4–19. UK Department of Health, Education, and Welfare; National Institute on Drug Abuse. URL: http://archives.drugabuse.gov/pdf/monographs/download24.html [Accessed: 2-2-2017].

Levy, P. S. & French, D. K. (1977). Synthetic Estimation of State Health Characteristics Based on the Health Interview Survey. In *Vital and Health Statistics. Series 2, Data Evaluation and Methods Research. DHEW publication; (PHS) 78-1349*, number 75, pages 1–22. URL: http://www.ncbi.nlm.nih.gov/pubmed/610050 [Accessed: 2-2-2017].

Li, F., Green, J. G., Kessler, R. C., & Zaslavsky, A. M. (2010). Estimating Prevalence of Serious Emotional Disturbance in Schools Using a Brief Screening Scale. *International Journal of Methods in Psychiatric Research*, **19**(S1), 88–98. DOI: 10.1002/mpr.315.

Li, F. & Zaslavsky, A. M. (2010). Using a Short Screening Scale for Small-Area Estimation of Mental Illness Prevalence for Schools. *Journal of the American Statistical Association*, **105**(492), 1323–1332. DOI: 10.1198/jasa.2010.ap09185.

Li, G., Best, N., Hansell, A. L., Ahmed, I., & Richardson, S. (2012). BaySTDetect: Detecting Unusual Temporal Patterns in Small Area Data Via Bayesian Model Choice. *Biostatistics*, **13**(4), 695–710. DOI: 10.1093/biostatistics/kxs005.

Liang, S., Banerjee, S., Bushhouse, S., Finley, A. O., & Carlin, B. P. (2008). Hierarchical Multiresolution Approaches for Dense Point-Level Breast Cancer Treatment Data. *Computational Statistics & Data Analysis*, **52**(5), 2650–2668. DOI: 10.1016/j.csda.2007.09.011.

Little, R. J. (2004). To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling. *Journal of the American Statistical Association*, **99**(466), 546–556. DOI: 10.1198/016214504000000467.

Liu, B., Lahiri, P., & Kalton, G. (2014). Hierarchical Bayes Modeling of Survey-Weighted Small Area Proportions. *Survey Methodology*, **40**(1), 1–13. URL: http://www.statcan.gc.ca/pub/12-001-x/2014001/article/14030-eng.pdf [Accessed: 2-2-2017].

Lloyd, C. D. (2015). Assessing the Spatial Structure of Population Variables in England and Wales. *Transactions of the Institute of British Geographers*, **40**(1), 28–43. DOI: 10.1111/tran.12061.

Lockwood, J., Louis, T. A., & McCaffrey, D. F. (2002). Uncertainty in Rank Estimation: Implications for Value-Added Modeling Accountability Systems. *Journal of Educational and Behavioral Statistics*, **27**(3), 255–270. DOI: 10.3102/10769986027003255.

Logan, W. P. D. & Brooke, E. M. (1957). *The Survey of Sickness, 1943 to 1952*. HMSO, London.

Longford, N. T. (2006). Sample Size Calculation for Small Area Estimation. *Survey Methodology*, **32**(1), 89–96. URL: http://www5.statcan.gc.ca/olc-cel/olc.action?ObjId=12-001-X20060019259&ObjType=47&lang=en [Accessed: 2-2-2017].

Longford, N. T. (2012). "Which Model?" Is the Wrong Question. *Statistica Neerlandica*, **66**(3), 237–252. DOI: 10.1111/j.1467-9574.2011.00517.x.

Longford, N. T. (2015). Policy-Oriented Inference and the Analyst-Client Cooperation. An Example from Small-Area Statistics. *Statistics in Transition new series*, **16**(1), 65–82.

Lopez-Vizcaino, E., Lombardia, M. J., & Morales, D. (2013). Multinomial-Based Small Area Estimation of Labour Force Indicators. *Statistical Modelling*, **13**(2), 153–178. DOI: 10.1177/1471082X13478873.

Lorant, V., Thomas, I., Deliège, D., & Tonglet, R. (2001). Deprivation and Mortality: The Implications of Spatial Autocorrelation for Health Resources Allocation. *Social Science & Medicine*, **53**(12), 1711–1719. DOI: 10.1016/S0277-9536(00)00456-1.

Lord, D. & Mannering, F. (2010). The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives. *Transportation Research Part A: Policy and Practice*, **44**(5), 291–305. DOI: 10.1016/j.tra.2010.02.001.

Lu, H., Reilly, C. S., Banerjee, S., & Carlin, B. P. (2007). Bayesian Areal Wombling Via Adjacency Modeling. *Environmental and Ecological Statistics*, **14**(4), 433–452. DOI: 10.1007/s10651-007-0029-9.

Luck, J., Chang, C., Brown, E. R., & Lumpkin, J. (2006). Using Local Health Information to Promote Public Health. *Health Affairs*, **25**(4), 979–991. DOI: 10.1377/hlthaff.25.4.979.

Macintyre, S., Der, G., & Norrie, J. (2005). Are There Socioeconomic Differences in Responses to a

Commonly Used Self Report Measure of Chronic Illness? *International Journal of Epidemiology*, **34**(6), 1284–1290. DOI: 10.1093/ije/dyi200.

Mackenbach, J. P. (2003). Widening Socioeconomic Inequalities in Mortality in Six Western European Countries. *International Journal of Epidemiology*, **32**(5), 830–837. DOI: 10.1093/ije/dyg209.

Mackenbach, J. P. & Kunst, A. E. (1997). Measuring the Magnitude of Socio-Economic Inequalities in Health: An Overview of Available Measures Illustrated with Two Examples from Europe. *Social Science & Medicine*, **44**(6), 757–771. DOI: 10.1016/S0277-9536(96)00073-1.

MacNab, Y. C. (2003). Hierarchical Bayesian Spatial Modelling of Small-Area Rates of Non-Rare Disease. *Statistics in Medicine*, **22**(10), 1761–73. DOI: 10.1002/sim.1463.

Malec, D., Davis, W. W., & Cao, X. (1999). Model-Based Small Area Estimates of Overweight Prevalence Using Sample Selection Adjustment. *Statistics in Medicine*, **18**(23), 3189–3200. DOI: 10.1002/(SICI)1097-0258(19991215)18:23<3189::AID-SIM309>3.0.CO;2-C.

Malec, D. & Sedransk, J. (1993). Bayesian Predictive Inference for Units with Small Sample Sizes. *Medical Care*, **31**(supplement), YS66–YS70. DOI: 10.1097/00005650-199305001-00010.

Malec, D., Sedransk, J., Moriarity, C. L., & Leclere, F. B. (1997). Small Area Inference for Binary Variables in the National Health Interview Survey. *Journal of the American Statistical Association*, **92**(439), 815–826. DOI: 10.1080/01621459.1997.10474037.

Mannering, F. L., Shankar, V., & Bhat, C. R. (2016). Unobserved Heterogeneity and the Statistical Analysis of Highway Accident Data. *Analytic Methods in Accident Research*, **11**, 1–16. DOI: 10.1016/j.amar.2016.04.001.

Manor, O., Matthews, S., & Power, C. (2001). Self-Rated Health and Limiting Longstanding Illness: Inter-Relationships with Morbidity in Early Adulthood. *International Journal of Epidemiology*, **30**(3), 600–607. DOI: 10.1093/ije/30.3.600.

Manton, K. G., Woodbury, M. A., Stallard, E., Riggan, W. B., Creason, J. P., & Pellom, A. C. (1989). Empirical Bayes Procedures for Stabilizing Maps of U.S. Cancer Mortality Rates. *Journal of the American Statistical Association*, **84**(407), 637–650. DOI: 10.1080/01621459.1989.10478816.

Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Pedreschi, D., Rinzivillo, S., Pappalardo, L., & Gabrielli, L. (2015). Small Area Model-Based Estimators Using Big Data Sources. *Journal of Official Statistics*, **31**(2). DOI: 10.1515/jos-2015-0017.

Marmot, M. (2010). Fair Society, Healthy Lives: Strategic Review of Health Inequalities in England post-2010. URL: http://www.instituteofhealthequity.org/projects/fair-society-healthy-lives-the-marmot-review [Accessed: 2-2-2017].

Marshall, A. (2009). *Developing a Methodology for the Local Estimation and Projection of Limiting Long Term Illness and Disability*. PhD thesis, University of Manchester. URL: http://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.511749 [Accessed: 2-2-2017].

Marshall, E. C. & Spiegelhalter, D. J. (2003). Approximate Cross-Validatory Predictive Checks in Disease Mapping Models. *Statistics in Medicine*, **22**(10), 1649–1660. DOI: 10.1002/sim.1403.

Marshall, R. J. (1991a). Mapping Disease and Mortality Rates Using Empirical Bayes Estimators. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **40**(2), 283–94. URL: http://www.ncbi.nlm.nih.gov/pubmed/12157989 [Accessed: 2-2-2017].

Marshall, R. J. (1991b). A Review of Methods for the Statistical Analysis of Spatial Patterns of Disease. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **154**(3), 421–441.

Martin, D. & Wright, J. A. (2009). Disease Prevalence in the English Population: A Comparison of Primary Care Registers and Prevalence Models. *Social Science & Medicine*, **68**(2), 266–274. DOI: 10.1016/j.socscimed.2008.10.021.

Martin, S., Rice, N., & Smith, P. C. (1998). Risk and the General Practitioner Budget Holder. *Social Science & Medicine*, **47**(10), 1547–1554. DOI: 10.1016/S0277-9536(98)00231-7.

Martin, S., Sheldon, T. A., & Smith, P. (1995). Interpreting the New Illness Question in the UK Census for Health Research on Small Areas. *Journal of Epidemiology & Community Health*, **49**(6), 634–641. DOI: 10.1136/jech.49.6.634.

Martins, T. G., Simpson, D., Lindgren, F., & Rue, H. (2013). Bayesian Computing with INLA: New Features. *Computational Statistics & Data Analysis*, **67**, 68–83. DOI: 10.1016/j.csda.2013.04.

014.

Martuzzi, M. & Elliott, P. (1996). Empirical Bayes Estimation of Small Area Prevalence of Non-Rare Conditions. *Statistics in Medicine*, **15**(17), 1867–1873. DOI: 10.1002/(SICI)1097-0258(19960915)15:17<1867::AID-SIM398>3.0.CO;2-2.

Mathur, R., Bhaskaran, K., Chaturvedi, N., Leon, D. A., VanStaa, T., Grundy, E., & Smeeth, L. (2014). Completeness and Usability of Ethnicity Data in UK-Based Primary Care and Hospital Databases. *Journal of Public Health (Oxford, England)*, **36**(4), 684–692. DOI: 10.1093/pubmed/fdt116.

Maude, F. (2014). Government's Response to the National Statistician's Recommendation. Letter to Sir Andrew Dilnot. 18 July 2014. URL: https://www.statisticsauthority.gov.uk/archive/reports---correspondence/correspondence/letter-from-rt-hon-francis-maude-mp-to-sir-andrew-dilnot---180714.pdf [Accessed: 2-2-2017].

Mays, N. (1995). Geographical Resource Allocation in the English National Health Service, 1971–1994: The Tension between Normative and Empirical Approaches. **24**, S96–S102. DOI: 10.1093/ije/24.Supplement_1.S96.

McCluskey, S. & Topping, A. (2009). Increasing Response Rates to Lifestyle Surveys: A Review of Methodology and 'Good Practice'. URL: http://eprints.hud.ac.uk/6878/ [Accessed: 2-2-2017].

Mercer, L. D., Wakefield, J., Pantazis, A., Lutambi, A. M., Masanja, H., & Clark, S. (2015). Space-Time Smoothing of Complex Survey Data: Small Area Estimation for Child Mortality. *The Annals of Applied Statistics*, **9**(4), 1889–1905. DOI: 10.1214/15-AOAS872.

Meriwether, R. A. (1996). Blueprint for a National Public Health Surveillance System for the 21st Century. *Journal of Public Health Management and Practice*, **2**(4), 16–23. DOI: 10.1097/00124784-199623000-00006.

Miaou, S.-P. & Song, J. J. (2005). Bayesian Ranking of Sites for Engineering Safety Improvements: Decision Parameter, Treatability Concept, Statistical Criterion, and Spatial Dependence. *Accident Analysis & Prevention*, **37**(4), 699–720. DOI: 10.1016/j.aap.2005.03.012.

Miaou, S.-P., Song, J. J., & Mallick, B. K. (2003). Roadway Traffic Crash Mapping: A Space-Time Modeling Approach. *Journal of Transportation and Statistics*, **6**(1), 33–57. URL: https://apps.bts.gov/publications/journal_of_transportation_and_statistics/volume_06_number_01/html/paper_03/index.html [Accessed: 2-2-2017].

Miranda-Moreno, L. F., Labbe, A., & Fu, L. (2007). Bayesian Multiple Testing Procedures for Hotspot Identification. *Accident Analysis and Prevention*, **39**(6), 1192–1201. DOI: 10.1016/j.aap.2007.03.008.

Mitra, S. & Washington, S. (2007). On the Nature of Over-Dispersion in Motor Vehicle Crash Prediction Models. *Accident Analysis and Prevention*, **39**(3), 459–468. DOI: 10.1016/j.aap.2006.08.002.

Moher, D., Schulz, K. F., Simera, I., & Altman, D. G. (2010). Guidance for Developers of Health Research Reporting Guidelines. *PLoS Medicine*, **7**(2), e1000217. DOI: 10.1371/journal.pmed.1000217.

Molefe, W. B. (2011). *Sample Design for Small Area Estimation*. A thesis submitted in fulfilment of the requirements for the award of the degree of doctor of philosophy, University Of Wollongong. URL: http://ro.uow.edu.au/theses/3495 [Accessed: 2-2-2017].

Molefe, W. B., Shangodoyin, D. K., & Clark, R. G. (2015). An Approximation to the Optimal Subsample Allocation for Small Areas. *Statistics in Transition new series*, **16**(2), 163–182.

Molina, I. & Rao, J. N. K. (2010). Small Area Estimation of Poverty Indicators. *Canadian Journal of Statistics*, **38**(3), 369–385. DOI: 10.1002/cjs.10051.

Molina, I., Rao, J. N. K., & Datta, G. S. (2015). Small Area Estimation under a Fay-Herriot Model with Preliminary Testing for the Presence of Random Area Effects. *Survey Methodology*, **41**(1), 1–19. URL: http://www.statcan.gc.ca/pub/12-001-x/2015001/article/14161-eng.pdf [Accessed: 2-2-2017].

Molina, I., Saei, A., & Lombardía, M. (2007). Small Area Estimates of Labour Force Participation

under a Multinomial Logit Mixed Model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **170**(4), 975–1000. DOI: 10.1111/j.1467-985X.2007.00493.x.

Mollie, A. & Richardson, S. (1991). Empirical Bayes Estimates of Cancer Mortality Rates Using Spatial Models. *Statistics in Medicine*, **10**(1), 95–112. DOI: 10.1002/sim.4780100114.

Morgan, M., Mays, N., & Holland, W. W. (1987). Can Hospital Use Be a Measure of Need for Health Care? *Journal of Epidemiology & Community Health*, **41**(4), 269–274. DOI: 10.1136/jech.41.4.269.

Münnich, R. & Gabler, S. (2012). Stichprobenoptimierung und Schätzung im Zensus 2011. *Statistik und Wissenschaft*, **21**. URL: https://www.destatis.de/DE/Publikationen/StatistikWissenschaft/Band21_StichprobenoptimierungZensus1030821129004.pdf;jsessionid=CFA11D1AD702A550F9EFF32AE4DA0532.cae2?__blob=publicationFile [Accessed: 2-2-2017].

Murray, C. J., Ezzati, M., Lopez, A. D., Rodgers, A., & Vander Hoorn, S. (2003). Comparative Quantification of Health Risks: Conceptual Framework and Methodological Issues. *Population Health Metrics*, **1**(1), 1. DOI: 10.1186/1478-7954-1-1.

Murray, C. J. L., Salomon, J., & Mathers, C. (2000). A Critical Examination of Summary Measures of Population Health. *Bulletin of the World Health Organisation*, **78**(8), 981–994.

Nacul, L., Soljak, M., Samarasundera, E., Hopkinson, N. S., Lacerda, E., Indulkar, T., Flowers, J., Walford, H., & Majeed, A. (2011). COPD in England: A Comparison of Expected, Model-Based Prevalence and Observed Prevalence from General Practice Data. *Journal of Public Health (Oxford, England)*, **33**(1), 108–16. DOI: 10.1093/pubmed/fdq031.

Nacul, L. C., Soljak, M., & Meade, T. (2007). Model for Estimating the Population Prevalence of Chronic Obstructive Pulmonary Disease: Cross Sectional Data from the Health Survey for England. *Population Health Metrics*, **5**(1), 8. DOI: 10.1186/1478-7954-5-8.

Namazi-Rad, M.-R. & Steel, D. (2015). What Level of Statistical Model Should We Use in Small Area Estimation? *Australian & New Zealand Journal of Statistics*, **57**(2), 275–298. DOI: 10.1111/anzs.12115.

NatCen Social Research & University College London. Department of Epidemiology and Public Health (2013). Health Survey for England, 2011 [data collection]. SN: 7260. 1st edition, April 2013. Colchester, Essex: UK Data Archive [distributor]. DOI: 10.5255/UKDA-SN-7260-1.

Nazroo, J. (2014). Ethnic Inequalities in Health: Addressing a Significant Gap in Current Evidence and Policy. In *"If You Could Change One Thing...": Nine Local Actions to Reduce Health Inequalities*, pages 92–101. British Academy, London.

Netherlands, S. (2014). Dutch Census 2011 – Analysis and Methodology. URL: https://www.cbs.nl/NR/rdonlyres/5FDCE1B4-0654-45DA-8D7E-807A0213DE66/0/2014b57pub.pdf [Accessed: 2-2-2017].

Newton, J. N., Briggs, A. D. M., Murray, C. J. L., Dicker, D., Foreman, K. J., Wang, H., Naghavi, M., Forouzanfar, M. H., Ohno, S. L., Barber, R. M., Vos, T., Stanaway, J. D., Schmidt, J. C., Hughes, A. J., Fay, D. F. J., Ecob, R., Gresser, C., McKee, M., Rutter, H., Abubakar, I., Ali, R., Anderson, H. R., Banerjee, A., Bennett, D. A., Bernabé, E., Bhui, K. S., Biryukov, S. M., Bourne, R. R., Brayne, C. E. G., Bruce, N. G., Brugha, T. S., Burch, M., Capewell, S., Casey, D., Chowdhury, R., Coates, M. M., Cooper, C., Critchley, J. A., Dargan, P. I., Dherani, M. K., Elliott, P., Ezzati, M., Fenton, K. A., Fraser, M. S., Fürst, T., Greaves, F., Green, M. A., Gunnell, D. J., Hannigan, B. M., Hay, R. J., Hay, S. I., Hemingway, H., Larson, H. J., Looker, K. J., Lunevicius, R., Lyons, R. A., Marcenes, W., Mason-Jones, A. J., Matthews, F. E., Moller, H., Murdoch, M. E., Newton, C. R., Pearce, N., Piel, F. B., Pope, D., Rahimi, K., Rodriguez, A., Scarborough, P., Schumacher, A. E., Shiue, I., Smeeth, L., Tedstone, A., Valabhji, J., Williams, H. C., Wolfe, C. D. A., Woolf, A. D., & Davis, A. C. J. (2015). Changes in Health in England, with Analysis by English Regions and Areas of Deprivation, 1990–2013: A Systematic Analysis for the Global Burden of Disease Study 2013. *The Lancet*, **386**(10010), 2257–2274. DOI: 10.1016/S0140-6736(15)00195-6.

NHS Digital (2016a). Care Information Choices, December 2016. URL: http://www.content.digital.nhs.uk/catalogue/PUB22677 [Accessed: 2-2-2017].

NHS Digital (2016b). Quality and Outcomes Framework. – Prevalence, Achievements and Exceptions Report. – England, 2015–16. – Published 27 October 2016. URL: http://www.content.

digital.nhs.uk/catalogue/PUB22266/qof-1516-rep-v2.pdf [Accessed: 2-2-2017].

NHS England (2012). CCG Outcomes Indicator Set 2013/14: Technical Guidance. URL: http://webarchive.nationalarchives.gov.uk/20160830094443/www.england.nhs.uk/wp-content/uploads/2012/12/ccg-ois-tech-guide.pdf [Accessed: 2-2-2017].

NHS England (2013). Fundamental Review of Allocations Policy – Annex C: Technical Guidance to Weighted Capitation Formula for Clinical Commissioning Groups, 2013. NHS England, Analytical Services. URL: http://webarchive.nationalarchives.gov.uk/20160830094443/www.england.nhs.uk/wp-content/uploads/2013/08/ann-c-tech-guid.pdf [Accessed: 2-2-2017].

NHS Information Centre (2005). Summary of Synthetic Estimates Project. URL: http://content.digital.nhs.uk/catalogue/PUB03630/synt-esti-heal-life-beha-rep-summ.pdf [Accessed: 2-2-2017].

NHS Primary Care Commissioning (2011). Quality & Outcomes Framework Management Guide. – Volume 1. Review Date: March 2011.. URL: https://www.pcc-cic.org.uk/sites/default/files/articles/attachments/qof_volume_1_updated.pdf [Accessed: 2-2-2017].

Norman, P., Boyle, P., & Rees, P. (2005). Selective Migration, Health and Deprivation: A Longitudinal Analysis. *Social Science & Medicine*, **60**(12), 2755–71. DOI: 10.1016/j.socscimed.2004.11.008.

Normand, S.-L. T., Ash, A. S., Fienberg, S. E., Stukel, T. A., Utts, J., & Louis, T. A. (2016). League Tables for Hospital Comparisons. *Annual Review of Statistics and Its Application*, **3**(1), 21–50. DOI: 10.1146/annurev-statistics-022513-115617.

Normand, S.-L. T. & Zou, K. H. (2002). Sample Size Considerations in Observational Health Care Quality Studies. *Statistics in Medicine*, **21**(3), 331–345. DOI: 10.1002/sim.1020.

of Health, U. D. (2000). *The NHS Plan: A Plan for Investment, A Plan for Reform (Cm. 4818-I)*. The Stationary Office.

Olives, C., Myerson, R., Mokdad, A. H., Murray, C. J. L., & Lim, S. S. (2013). Prevalence, Awareness, Treatment, and Control of Hypertension in United States Counties, 2001–2009. *PLoS ONE*, **8**(4). DOI: 10.1371/journal.pone.0060308.

on Social Determinants of Health, C. (2008). Closing the Gap in a Generation: Health Equity through Action on the Social Determinants of Health. Final Report of the Commission on Social Determinants of Health. URL: http://www.who.int/social_determinants/thecommission/finalreport/en/ [Accessed: 2-2-2017].

ONS (2001a). 2001 Census Household Questionnaire for England. URL: http://www.ons.gov.uk/ons/guide-method/census/2011/how-our-census-works/about-censuses/census-history/200-years-of-the-census/census-1911-2001/index.html [Accessed: 2-2-2017].

ONS (2001b). Health Expectancies at Birth for the 2001 Census Wards in England and Wales, 1999–2003. URL: http://www.ons.gov.uk/ons/rel/disability-and-health-measurement/sub-national-health-expectancies/2001/index.html [Accessed: 2-2-2017].

ONS (2002). 1991 Census Household Questionnaire for England. URL: http://www.ons.gov.uk/ons/guide-method/census/2011/how-our-census-works/about-censuses/census-history/200-years-of-the-census/census-1911-2001/index.html [Accessed: 2-2-2017].

ONS (2003). *Proposals for an Integrated Population Statistics System. Discussion Paper Series*. Office for National Statistics, London.

ONS (2005). Census 2001 Quality Survey Rehearsal Report. URL: http://www.ons.gov.uk/ons/guide-method/census/census-2001/data-and-products/quality-of-the-census-data/data-quality-notes-and-reports/census-quality-survey--rehearsal-/index.html [Accessed: 2-2-2017].

ONS (2007). Report from the Review of Equality Data. – October 2007. URL: http://www.ons.gov.uk/ons/guide-method/measuring-equality/equality/equality-data-review/review-of-equality-data--final-report.pdf [Accessed: 2-2-2017].

ONS (2009). Final Population Definitions for the 2011 Census. URL: www.ons.gov.uk/ons/

guide-method/census/2011/the-2011-census/2011-census-questionnaire-content/final-population-definitions-for-the-2011-census.pdf [Accessed: 2-2-2017].

ONS (2010a). 2001 National Statistics Area Classification for Local Authority Districts in England. URL: http://www.ons.gov.uk/ons/guide-method/geography/products/area-classifications/ns-area-classifications/index/datasets/local-authorities/la-data-file.xls [Accessed: 2-2-2017].

ONS (2010b). Census Household Questionnaire for England 2011. URL: http://www.ons.gov.uk/ons/guide-method/census/2011/how-our-census-works/about-censuses/census-history/200-years-of-the-census/census-1911-2001/index.html [Accessed: 2-2-2017].

ONS (2011a). 2011 Census: Digitised Boundary Data (England and Wales). URL: http://census.ukdataservice.ac.uk/get-data/boundary-data.aspx [Accessed: 2-2-2017].

ONS (2011b). Harmonised Concepts and Questions for Social Data Sources. Primary Standards. Long-Lasting Health Conditions and Illnesses; Impairments and Disability. Version 1.0. URL: http://www.ons.gov.uk/ons/guide-method/harmonisation/primary-set-of-harmonised-concepts-and-questions/long-lasting-health-conditions-and-illnesses--impairments-and-disability.pdf [Accessed: 2-2-2017].

ONS (2011c). Population Estimates by Ethnic Group: Methodology Paper. URL: http://www.ons.gov.uk/ons/guide-method/method-quality/specific/population-and-migration/pop-ests/population-estimates-by-ethnic-group/population-estimates-by-ethnic-group-methodology.pdf [Accessed: 2-2-2017].

ONS (2011d). 'Small Area Population Estimates', *Quality and Methodology Information Paper*, March 2011. URL: http://www.ons.gov.uk/ons/guide-method/method-quality/quality/quality-information/social-statistics/summary-quality-report-for-small-area-population-estimates.pdf [Accessed: 2-2-2017].

ONS (2012a). 2011 Census: Local Authority Response, Return and Coverage Rates. Published 11 December 2012. URL: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release--quality-assurance-and-methodology-papers/census-response-rates.xls [Accessed: 2-2-2017].

ONS (2012b). 2011 Census: Methods and Quality Report: The 2011 Census Coverage Assessment and Adjustment Process. July 2012. URL: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release--quality-assurance-and-methodology-papers/coverage-assessment-and-adjustment-process.pdf [Accessed: 2-2-2017].

ONS (2012c). Overcount Estimation and Adjustment. 2011 Census: Methods and Quality Report. July 2012. URL: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release--quality-assurance-and-methodology-papers/overcount-estimation-and-adjustment.pdf [Accessed: 2-2-2017].

ONS (2012d). Population Estimates by Ethnic Group. *Quality and Methodology Information Paper*. URL: http://www.ons.gov.uk/ons/guide-method/method-quality/quality/quality-information/social-statistics/summary-quality-report-for-population-estmates-by-ethnic-group.pdf [Accessed: 2-2-2017].

ONS (2013a). 2011 Census: aggregate (England and Wales) bulk data download (release 3).

ONS (2013b). 2011 Census: Detailed Characteristics (england and Wales) Bulk Data Downloads (release 3) [computer File]. URL: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/bulk-data/bulk-data-downloads/index.html [Accessed: 2-2-2017].

ONS (2013c). 2011 Census Local Characteristics Table LC3101EWLS (Long Term Health Problem or Disability by Sex by Age). URL: https://www.nomisweb.co.uk/census/2011/lc3101ewls [Accessed: 2-2-2017].

ONS (2013d). Beyond 2011: Options Report O2. July 2013. URL: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/beyond-2011-options-report-2--o2-.pdf [Accessed: 2-2-2017].

ONS (2013e). Beyond 2011: Producing Socio-Demographic Statistics. *Beyond 2011: Methods & Policies Report M4*. URL: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/methods-and-policies-reports/beyond-2011--producing-socio-demographic-statistics.pdf [Accessed: 2-2-2017].

ONS (2013f). Beyond 2011: Producing Socio-Demographic Statistics 2. URL: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/beyond-2011-producing-socio-demographic-statistics-2.pdf [Accessed: 2-2-2017].

ONS (2013g). Inequality in Disability-Free Life Expectancy by Area Deprivation: England, 2003–06 and 2007–10. URL: http://www.ons.gov.uk/ons/rel/disability-and-health-measurement/sub-national-health-expectancies/inequality-in-disability-free-life-expectancy-by-area-deprivation--england--2003-06-and-2007-10/index.html [Accessed: 2-2-2017], Released 25 July 2013.

ONS (2013h). Internal Migration by Local Authorities in England and Wales. Research Series, Years Ending June 2009 to June 2011. Reference Tables. 14 November 2013. URL: http://www.ons.gov.uk/ons/rel/migration1/internal-migration-by-local-authorities-in-england-and-wales/research-series--years-ending-june-2009-to-june-2011/index.html [Accessed: 2-2-2017].

ONS (2013i). Mortality Statistics: Deaths Registered in England and Wales by Area of Usual Residence, 2011. URL: http://www.ons.gov.uk/ons/rel/vsob1/deaths-registered-area-usual-residence/2011/index.html [Accessed: 2-2-2017].

ONS (2014a). 2011 Census: Methods and Quality Report: 2011 Census Quality Survey. February 2014. URL: www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/quality/quality-measures/assessing-accuracy-of-answers/2011-census-quality-survey-report.pdf [Accessed: 2-2-2017].

ONS (2014b). Beyond 2011: Final Options Report O4. April 2014. URL: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/methods-and-policies-reports/beyond-2011--final-options-report.pdf [Accessed: 2-2-2017].

ONS (2014c). Beyond 2011: Statistical Research Update. *Methods & Policies Reports*, **M13**. URL: http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/methods-and-policies-reports/beyond-2011--statistical-research-update.pdf [Accessed: 2-2-2017].

ONS (2014d). Disability-Free Life Expectancy by Upper Tier Local Authority: England 2008–10. URL: http://www.ons.gov.uk/ons/rel/disability-and-health-measurement/sub-national-health-expectancies/disability-free-life-expectancy---subnational-estimates-for-england--2008-10/index.html [Accessed: 2-2-2017], Released 1 May 2014.

ONS (2014e). How Do People Rate Their General Health? an Analysis of General Health by Disability and Deprivation.

ONS (2014f). Population Statistics Strategy 2014–2019. URL: http://www.ons.gov.uk/ons/guide-method/method-quality/specific/population-and-migration/population-statistics-research-unit--psru-/ons-population-statistics-strategy-2014-2019.doc [Accessed: 2-2-2017].

ONS (2015a). The 2021 Census Initial View on Content for England and Wales. You Said: A Summary of the Results. November 2015. URL: https://consultations.ons.gov.uk/

census/2021-census-topics-consultation/user_uploads/yousaid_final_report_19.11.15.pdf [Accessed: 2-2-2017].

ONS (2015b). Mid-2014 Population Estimates for Middle Layer Super Output Areas in England and Wales by Single Year of Age and Sex. URL: https://www.ons.gov.uk/file?uri=/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/middlesuperoutputareamidyearpopulationestimates/mid2014/rft-msoa-unformatted-table-2014.zip [Accessed: 2-2-2017].

ONS (2016a). 2021 Census Design Document. URL: http://www.ons.gov.uk/ons/guide-method/census/2021-census/progress-and-development/research-projects/2021-census-design-document.pdf [Accessed: 2-2-2017].

ONS (2016b). Census Transformation Programme. Annual Assessment of ONS' Progress Towards an Administrative Data Census Post 2021. May 2016.. URL: https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusannualassessments [Accessed: 2-2-2017].

ONS (2016c). Mean House Price by Middle Layer Super Output Areas – House Price Statistics for Small Areas Dataset 3). URL: https://www.ons.gov.uk/peoplepopulationandcommunity/housing/datasets/hpssadataset3meanhousepricebymsoaquarterlyrollingyear [Accessed: 2-2-2017].

ONS (2016d). Research Outputs Estimating the Size of the Population in England and Wales: 2016 Release. URL: https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusresearchoutputs/sizeofthepopulation/researchoutputsestimatingthesizeofthepopulationinenglandandwales2016release [Accessed: 2-2-2017].

ONS (2016e). Rural Urban Classification (2011) of Middle Layer Super Output Areas in England and Wales. Last Modified: November 9, 2016. URL: http://ons.maps.arcgis.com/home/item.html?id=86fac76c60ed4943a8b94f64bff3e8b1 [Accessed: 2-2-2017].

ONS (2017). Annual Assessment of ONS' progress towards an Administrative Data Census post-2021. 20 June 2017.. URL: https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusannualassessments/annualassessmentofonssprogresstowardsanadministrativedatacensuspost2021#annex-b-update-on-acquisition-of-data [Accessed: 13-08-2017].

ONS. Social Survey Division (2012). Life Opportunities Survey: Wave One, 2009–2011 [data collection]. 3rd Edition, March 2012. SN: 6653. DOI: 10.5255/UKDA-SN-6653-2.

ONS. Social Survey Division, Northern Ireland Statistics and Research Agency. Central Survey Unit (2014a). Quarterly Labour Force Survey, April - June, 2011. [data collection] 2nd Edition. UK Data Service. SN: 6851. DOI: 10.5255/UKDA-SN-6851-2.

ONS. Social Survey Division, Northern Ireland Statistics and Research Agency. Central Survey Unit (2014b). Quarterly Labour Force Survey, January - March, 2010. [data collection] 3rd Edition. UK Data Service. SN: 6457. DOI: 10.5255/UKDA-SN-6457-3.

ONS. Social Survey Division, Northern Ireland Statistics and Research Agency. Central Survey Unit (2014c). Quarterly Labour Force Survey, July - September, 2010. [data collection] 2nd Edition. UK Data Service. SN: 6632. DOI: 10.5255/UKDA-SN-6632-2.

ONS. Social Survey Division, Northern Ireland Statistics and Research Agency. Central Survey Unit (2015). Quarterly Labour Force Survey, January - March, 2011. [data collection] 4th Edition. UK Data Service. SN: 6782. DOI: 10.5255/UKDA-SN-6782-4.

OPCS Social Survey Division (1973). *The General Household Survey: Introductory Report. An Inter-Departmental Survey Sponsored by the Central Statistical Office*. HMSO, London.

OPCS Social Survey Division (1975). *The General Household Survey 1972. An Inter-Departmental Survey Sponsored by the Central Statistical Office*. HMSO, London.

OPCS Social Survey Division (1977). *The General Household Survey 1976. An Inter-Departmental Survey Sponsored by the Central Statistical Office*. HMSO, London.

Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G., & Breidt, F. J. (2008). Non-Parametric Small Area Estimation Using Penalized Spline Regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**(1), 265–286. DOI: 10.1111/j.1467-9868.2007.00635.x.

Paech, D., Mernagh, P., & Weston, A. (2010). A Systematic Review of the Literature a Systematic Review of Economic Evaluations for Tobacco Control Programs Acknowledgements.

Paez, A., Mercado, R. G., Farber, S., Morency, C., & Roorda, M. (2010). Accessibility to Health Care Facilities in Montreal Island: An Application of Relative Accessibility Indicators from the Perspective of Senior and Non-Senior Residents. *International Journal of Health Geographics*, **9**(1), 52. DOI: 10.1186/1476-072X-9-52.

Pal, S. & Lahiri, P. (2017). Mapping of Arsenic Contamination in Ground Water: A New Hierarchical Bayesian Method. In *Presentation at the 2017 Small Area Conference, 61st International Statistical Institute World Statistics Congress Satellite Meeting, 10–12 July 2017, UniversitÃ© Pierre et Marie Curie, Paris, France*, Paris. URL: http://sae2017.ensai.fr/presentation/mapping-arsenic-contamination-ground-water-new-hierarchical-bayesian-method/ [Accessed: 2-8-2017].

Pascutto, C., Wakefield, J. C., Best, N. G., Richardson, S., Bernardinelli, L., Staines, A., & Elliott, P. (2000). Statistical Issues in the Analysis of Disease Mapping Data. *Statistics in Medicine*, **19**(17-18), 2493–519. URL: http://www.ncbi.nlm.nih.gov/pubmed/10960868 [Accessed: 2-2-2017].

Paul-Shaheen, P., Williams, D., & Clark, J. D. (1987). Small Area Analysis: A Review and Analysis of the North American Literature. *Journal of Health Politics, Policy and Law*, **12**(4), 741–809. DOI: 10.1215/03616878-12-4-741.

Payne, J. N., Coy, J., Patterson, S., & Milner, P. C. (1994). Is Use of Hospital Services a Proxy for Morbidity? A Small Area Comparison of the Prevalence of Arthritis, Depression, Dyspepsia, Obesity, and Respiratory Disease with Inpatient Admission Rates for These Disorders in England. *Journal of Epidemiology & Community Health*, **48**(1), 74–78. DOI: 10.1136/jech.48.1.74.

Pearce, D., Clark, A., & Baird, G. (1988). The 1987 Census Test. *Population Trends*, **53**, 22–26.

Pearce, D. & Thomas, F. (1990). The 1989 Census Test. *Population Trends*, **61**, 24–30.

Peña, V. S., Miravitlles, M., Gabriel, R., Jiménez-Ruiz, C. a., Villasante, C., Masa, J. F., Viejo, J. L., & Fernández-Fau, L. (2000). Geographic Variations in Prevalence and Underdiagnosis of COPD: Results of the Iberpoc Multicentre Epidemiological Study. *Chest*, **118**(4), 981–9. URL: http://www.ncbi.nlm.nih.gov/pubmed/11035667 [Accessed: 2-2-2017].

Pereira, L. N., Mendes, J. M., & Coelho, P. S. (2013). Model-Based Estimation of Unemployment Rates in Small Areas of Portugal. *Communications in Statistics – Theory and Methods*, **42**(7), 1325–1342. DOI: 10.1080/03610926.2012.749989.

Person-Based Resource Allocation Team (2010). Developing a Person-Based Resource Allocation Formula for Allocations to General Practice in England. URL: http://nuffield.dh.bytemark.co.uk/sites/files/nuffield/document/Developing_a_person-based_resource_allocation_formula_REPORT.pdf [Accessed: 2-2-2017].

Pfarr, C., Schmid, A., & Schneider, U. (2012). Reporting Heterogeneity in Self-Assessed Health Among Elderly Europeans. *Health Economics Review*, **2**(1), 21. DOI: 10.1186/2191-1991-2-21.

Pfeffermann, D. (2002). Small Area Estimation–New Developments and Directions. *International Statistical Review*, **70**(1), 125–143. DOI: 10.1111/j.1751-5823.2002.tb00352.x.

Pfeffermann, D. (2013). New Important Developments in Small Area Estimation. *Statistical Science*, **28**(1), 40–68. DOI: 10.1214/12-STS395.

Pfeffermann, D., Eltinge, J. L., & Brown, L. D. (2015). Methodological Issues and Challenges in the Production of Official Statistics: 24th Annual Morris Hansen Lecture. *Journal of Survey Statistics and Methodology*, **3**(4), 425–483. DOI: 10.1093/jssam/smv035.

Pfeffermann, D., Sikov, A., & Tiller, R. (2014). Single- and Two-Stage Cross-Sectional and Time Series Benchmarking Procedures for Small Area Estimation. *TEST*, **23**(4), 631–666. DOI: 10.1007/s11749-014-0398-y.

Pfeffermann, D. & Sverchkov, M. (2007). Small-Area Estimation under Informative Probability Sampling of Areas and within the Selected Areas. *Journal of the American Statistical Association*,

102(480), 1427–1439. DOI: 10.1198/016214507000001094.

Pfeffermann, D. & Tiller, R. (2006). Small-Area Estimation with State-Space Models Subject to Benchmark Constraints. *Journal of the American Statistical Association*, **101**(476), 1387–1397. DOI: 10.1198/016214506000000591.

Pickering, K., Scholes, S., & Bajekal, M. (2004). Synthetic Estimation of Healthy Lifestyles Indicators: Stage 2 Report Prepared for the Department of Health.

Pickett, K. E. & Wilkinson, R. G. (2008). People Like Us: Ethnic Group Density Effects on Health. *Ethnicity & Health*, **13**(4), 321–34. DOI: 10.1080/13557850701882928.

Pickett, W., Koushik, A., Faelker, T., & Brown, K. S. (2000). Estimation of Youth Smoking Behaviours in Canada. *Chronic Diseases in Canada*, **21**(3), 119–127.

Plummer, M. (2008). Penalized Loss Functions for Bayesian Model Comparison. *Biostatistics*, **9**(3), 523–539. DOI: 10.1093/biostatistics/kxm049.

Pollock, A. M., Macfarlane, A., & Godden, S. (2012). Dismantling the Signposts to Public Health? NHS Data Under the Health and Social Care Act 2012. *British Medical Journal*, **344**(apr26 2), e2364–e2364. DOI: 10.1136/bmj.e2364.

Porter, A. T., Holan, S. H., & Wikle, C. K. (2015a). Bayesian Semiparametric Hierarchical Empirical Likelihood Spatial Models. *Journal of Statistical Planning and Inference*, **165**, 78–90. DOI: 10.1016/j.jspi.2015.04.002.

Porter, A. T., Holan, S. H., & Wikle, C. K. (2015b). Multivariate Spatial Hierarchical Bayesian Empirical Likelihood Methods for Small Area Estimation: Multivariate Semiparametric Small Area Estimation. *Stat*, **4**(1), 108–116. DOI: 10.1002/sta4.81.

Porter, A. T., Wikle, C. K., & Holan, S. H. (2015c). Small Area Estimation Via Multivariate Fay-Herriot Models with Latent Spatial Dependence. *Australian & New Zealand Journal of Statistics*, **57**(1), 15–29. DOI: 10.1111/anzs.12101.

Prasad, N. G. N. & Rao, J. N. K. (1990). The Estimation of the Mean Squared Error of Small-Area Estimators. *Journal of the American Statistical Association*, **85**(409), 163. DOI: 10.2307/2289539.

Pratesi, M. & Salvati, N. (2008). Small Area Estimation: The EBLUP Estimator Based on Spatially Correlated Random Area Effects. *Statistical Methods & Applications*, **17**(1), 113–141. DOI: 10.1007/s10260-007-0061-9.

Public Health England (2014a). Elective Hospital Admissions (All Causes), Indirectly Age Standardised Ratio, All Ages, Persons. April 2008 to March 2013 (Inclusive). URL: http://www.localhealth.org.uk/Spreadsheets/Electivehospitaladmissionsforallcauses_July2014.xls [Accessed: 2-2-2017].

Public Health England (2014b). Emergency Hospital Admissions (All Causes), Indirectly Age Standardised Ratio, All Ages, Persons. April 2008 to March 2013 (Inclusive). URL: http://www.localhealth.org.uk/Spreadsheets/Emergencyhospitaladmissionsforallcauses_July2014.xls [Accessed: 2-2-2017].

Public Health England (2014c). Technical Document for the Chronic Kidney Disease Prevalence Model. URL: http://www.yhpho.org.uk/resource/view.aspx?RID=204689 [Accessed: 2-2-2017].

Public Health England (2015). New Tool Models Youth Smoking Variation across England. 27 January 2015. URL: https://www.gov.uk/government/news/new-tool-models-youth-smoking-variation-across-england [Accessed: 2-2-2017].

Public Health England (2016). Child Poverty – Index of Multiple Deprivation (IMD) 2015 Income Deprivation Affecting Children Index (IDACI). URL: http://localhealth.org.uk/Spreadsheets/Deprivation%20-%20IMD%202015%20income%20domain_March2016.xlsx [Accessed: 2-2-2017].

Public Health Evaluation and Impact Assessment Consortium (2013). Evaluation of the Use and Impact of the European Community Health Indicators (ECHI) by Member States. Final Report. August 2013. URL: https://ec.europa.eu/health/sites/health/files/indicators/docs/echi_report_v20131031.pdf [Accessed: 2-2-2017].

R Core Team (2014). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria*. URL: http://www.r-project.org/ [Accessed: 2-2-2017].

Racz, M. J. & Sedransk, J. (2010). Bayesian and Frequentist Methods for Provider Profiling Using Risk-Adjusted Assessments of Medical Outcomes. *Journal of the American Statistical Association*, **105**(489), 48–58. DOI: 10.1198/jasa.2010.ap07175.

Raghunathan, T. E., Xie, D., Schenker, N., Parsons, V. L., Davis, W. W., Dodd, K. W., & Feuer, E. J. (2007). Combining Information from Two Surveys to Estimate County-Level Prevalence Rates of Cancer Risk Factors and Screening. *Journal of the American Statistical Association*, **102**(478), 474–486. DOI: 10.1198/016214506000001293.

Rahman, A. & Harding, A. (2016). Microsimulation Modeling Technology for Small Area Estimation. In *Small Area Estimation and Microsimulation Modeling*, chapter 6, pages 119–146. Chapman and Hall/CRC, Boca Raton, FL.

Ralf, M., Burgard, J. P., Gabler, S., Ganninger, M., & Kolb, J.-P. (2011). Small Area Estimation for Population Counts in the German Census 2011. *Statistics in Transition new series and Survey Methodology*, **17**(1 Joint Issue: Small Area Estimation 2014), 181–190.

Ralphs, M. & Tutton, P. (2011). Beyond 2011: International Models for Census Taking: Current Processes and Future Developments. Version 1.0, June 2011. URL: http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes---projects/beyond-2011/news/reports-and-publications/early-reports-and-research-papers/international-models-for-census-taking.pdf [Accessed: 2-2-2017].

Rampaso, R. C., de Souza, A. D. P., & Flores, E. F. (2016). Bayesian Analysis of Spatial Data Using Different Variance and Neighbourhood Structures. *Journal of Statistical Computation and Simulation*, **86**(3), 535–552. DOI: 10.1080/00949655.2015.1022549.

Rao, J. & Molina, I. (2015). *Small Area Estimation*. John Wiley & Sons, Inc, Hoboken, NJ, USA. DOI: 10.1002/9781118735855.

Rao, J. N. K. (1999). Some Recent Advances in Model- Based Small Area Estimation. *Survey Methodology*, **25**(2), 175–186.

Rao, J. N. K. (2011). Impact of Frequentist and Bayesian Methods on Survey Sampling Practice: A Selective Appraisal. *Statistical Science*, **26**(2), 240–256. DOI: 10.1214/10-STS346.

Rao, J. N. K. (2014). Inferential Issues in Model-Based Small Area Estimation: Some New Developments. *Statistics in Transition new series and Survey Methodology*, **16**(4), 513–526.

Rao, J. N. K. & Choudhry, G. H. (2011). Small-Area Estimation: Overview and Empirical Study. In Cox, B. G., Binder, D. A., Chinnappa, B. N., Christianson, A., Colledge, M. J., & Kott, P. S., editors, *Business Survey Methods*, pages 527–542. John Wiley & Sons, Inc., Hoboken, NJ, USA. DOI: 10.1002/9781118150504.ch27.

Rao, J. N. K. & Yu, M. (1994). Small-Area Estimation by Combining Time-Series and Cross-Sectional Data. *Canadian Journal of Statistics*, **22**(4), 511–528. DOI: 10.2307/3315407.

Raudenbush, S. W. (1997). Statistical Analysis and Optimal Design for Cluster Randomized Trials. DOI: 10.1037/1082-989X.2.2.173.

Rees, P. (2013). Exploring the Feasibility of a Rolling Census Design for England Using Administrative Data and Modelling to Optimise the Socio-Demographic System. In *Keynote Presentation, Beyond 2011 Research Conference, 30 April–1 May 2013, University of Southampton, Southampton*, Southampton.

Rehm, J., Kehoe, T., Gmel, G., Stinson, F., Grant, B., & Gmel, G. (2010). Statistical Modeling of Volume of Alcohol Exposure for Epidemiological Studies of Population Health: The US Example. *Population Health Metrics*, **8**(1), 3. DOI: 10.1186/1478-7954-8-3.

Richardson, S., Abellan, J. J., & Best, N. (2006). Bayesian Spatio-Temporal Analysis of Joint Patterns of Male and Female Lung Cancer Risks in Yorkshire (UK). *Statistical Methods in Medical Research*, **15**(4), 385–407. DOI: 10.1191/0962280206sm458oa.

Richardson, S. & Monfort, C. (2000). Clustering, Cluster Detection, and Spatial Variation in Risks. In Elliott, P., Wakefield, J. C., Best, N. G., & Briggs, D. J., editors, *Spatial Epidemiology: Methods and Applications*, pages 205–220.

Richardson, S., Thomson, A., Best, N., & Elliott, P. (2004). Interpreting Posterior Relative Risk Estimates in Disease-Mapping Studies. *Environmental Health Perspectives*, **112**(9), 1016–1025. DOI: 10.1289/ehp.6740.

Robine, J.-M., Cambois, E., Nusselder, W., Jeune, B., Oyen, H. V., & Jagger, C. (2013). The Joint Action on Healthy Life Years (JA:EHLEIS). *Archives of Public Health*, **71**(1), 2. DOI: 10.1186/0778-7367-71-2.

Rodrigues, E. & Assunção, R. (2012). Bayesian Spatial Models with a Mixture Neighborhood Structure. *Journal of Multivariate Analysis*, **109**, 88–102. DOI: 10.1016/j.jmva.2012.02.017.

Ross, M. & Wakefield, J. (2015). Bayesian Hierarchical Models for Smoothing in Two-Phase Studies, with Application to Small Area Estimation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **178**(4), 1009–1023. DOI: 10.1111/rssa.12103.

Rotondi, M. S. & Donner, A. (2009). Sample Size Estimation in Cluster Randomized Educational Trials: An Empirical Bayes Approach. *Journal of Educational and Behavioral Statistics*, **34**(2), 229–237. DOI: 10.3102/1076998609332756.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**(2), 319–392. DOI: 10.1111/j.1467-9868.2008.00700.x.

Rushworth, A., Lee, D., & Mitchell, R. (2014). A Spatio-Temporal Model for Estimating the Long-Term Effects of Air Pollution on Respiratory Hospital Admissions in Greater London. *Spatial and Spatio-Temporal Epidemiology*, **10**, 29–38. DOI: 10.1016/j.sste.2014.05.001.

Ryan, L. (2009). Spatial Epidemiology: Some Pitfalls and Opportunities. *Epidemiology*, **20**(2), 242–244. DOI: 10.1097/EDE.0b013e318198a5fb.

Saei, A. & Chambers, R. (2003a). Small Area Estimation : A Review of Methods Based on the Application of Mixed Models. *S3RI Methodology Working Paper M03/16*.

Saei, A. & Chambers, R. (2003b). Small Area Estimation under Linear and Generalized Linear Mixed Models with Time and Area Effects. URL: http://eprints.soton.ac.uk/8165/ [Accessed: 2-2-2017].

Saei, A. & Chambers, R. (2005). Empirical Best Linear Unbiased Prediction for Out of Sample Areas. *S3RI Methodology Working Papers M05/03*.

Saei, A. & Chambers, R. (2011). Out of Sample Estimation for Small Areas Using Area Level Data. *S3RI Methodology Working Paper M05/11*. URL: http://eprints.soton.ac.uk/14327/1/14327-01.pdf [Accessed: 2-2-2017].

Saei, A. & Taylor, A. (2012). Labour Force Status Estimates under a Bivariate Random Components Model. *Journal of the Indian Society of Agricultural Statistics*, **66**(1), 187–201.

Sahu, S. K. & Smith, T. M. F. (2006). A Bayesian Method of Sample Size Determination with Practical Applications. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **169**(2), 235–253. DOI: 10.1111/j.1467-985X.2006.00408.x.

Salvati, N., Chandra, H., Giovanna Ranalli, M., & Chambers, R. (2010). Small Area Estimation Using a Nonparametric Model-Based Direct Estimator. *Computational Statistics & Data Analysis*, **54**(9), 2159–2171. DOI: 10.1016/j.csda.2010.03.023.

SAMPLE (2011). SAMPLE Project. – Small Area Methods for Poverty and Living Condition Estimates Project. URL: http://sample-project.eu/ [Accessed: 2-2-2017].

Särndal, C. E. (2010a). Models in Survey Sampling. In Carlson, M., Nyquist, H., & Villani, M., editors, *Official Statistics – Methodology and Applications in Honour of Daniel Thorburn*, pages 15–27. Stockholm.

Särndal, C.-E. (2010b). Models in Survey Sampling. *Statistics in Transition new series*, **11**(3), 539–554. URL: http://index.stat.gov.pl/repec/files/csb/stintr/csb_stintr_v11_2010_i3_n6.pdf [Accessed: 2-2-2017].

Särndal, C.-E., Swensson, B., & Wretman, J. H. (1992). *Model Assisted Survey Sampling*. Springer series in statistics. Springer-Verlag, New York.

Scarborough, P., Allender, S., Rayner, M., & Goldacre, M. (2009). Validation of Model-Based Estimates (Synthetic Estimates) of the Prevalence of Risk Factors for Coronary Heart Disease for Wards in England. *Health & Place*, **15**(2), 596–605. DOI: 10.1016/j.healthplace.2008.10.003.

Schenker, N. & Raghunathan, T. E. (2007). Combining Information from Multiple Surveys to Enhance Estimation of Measures of Health. *Statistics in Medicine*, **26**(8), 1802–11. DOI: 10.1002/sim.2801.

Schneider, K., Lapane, K., Clark, M. A., & Rakowski, W. (2009). Using Small-Area Estimation to Describe County-Level Disparities in Mammography. *Preventing Chronic Disease*, **6**(4), A125. URL: http://www.cdc.gov/pcd/issues/2009/oct/08_0210.htm [Accessed: 2-2-2017].

Scholes, S., Pickering, K., & Deverill, C. (2008). Healthy Lifestyle Behaviours: Model Based Estimates for Local Authorities in England, 2003-2005: Stage 3 Report Prepared for the NHS. URL: https://catalogue.ic.nhs.uk/publications/public-health/population/neig-la-mode-esti-heal-beha-03-05/neig-la-mode-estil-beha-03-05-rep-3.pdf [Accessed: 2-2-2017].

Sen, A. (2002). Health: Perception Versus Observation. *British Medical Journal*, **324**, 860–861. DOI: 10.1136/bmj.324.7342.860.

Senior, M. L. (1998). Area Variations in Self-Perceived Limiting Long Term Illness in Britain, 1991: Is the Welsh Experience Exceptional? *Regional Studies*, **32**(3), 265–280. DOI: 10.1080/713693455.

Shah, S. N., Russo, E. T., Earl, T. R., & Kuo, T. (2014). Measuring and Monitoring Progress toward Health Equity: Local Challenges for Public Health. *Preventing Chronic Disease*, **11**, E159. DOI: 10.5888/pcd11.130440.

Shouls, S., Congdon, P., & Curtis, S. (1996a). Geographic Variation in Illness and Mortality: The Development of a Relevant Area Typology for Sar Districts. *Health & Place*, **2**(3), 139–155. DOI: 10.1016/1353-8292(96)00002-0.

Shouls, S., Congdon, P., & Curtis, S. (1996b). Modelling Inequality in Reported Long Term Illness in the UK: Combining Individual and Area Characteristics. *Journal of Epidemiology & Community Health*, **50**(3), 366–376. DOI: 10.1136/jech.50.3.366.

Silva, D. B. N. & Clarke, P. (2008). Some Initiatives on Combining Data to Support Small Area Statistics and Analytical Requirements at ONS-UK. In *IAOS 2008 Conference on Reshaping Official Statistics*, Shanghai. URL: http://www.stats.gov.cn/english/specialtopics/iaos/Papers/CS_4_3_Silva_Clarke.pdf [Accessed: 2-2-2017].

Simpson, D. P., Rue, H., Martins, T. G., Riebler, A., & Sørbye, S. H. (2015). Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors. (v4, 6 Aug 2015). URL: http://arxiv.org/abs/1403.4630 [Accessed: 2-2-2017].

Sinharay, S. & Stern, H. S. (2003). Posterior Predictive Model Checking in Hierarchical Models. *Journal of Statistical Planning and Inference*, **111**(1-2), 209–221. DOI: 10.1016/S0378-3758(02)00303-8.

Skinner, C. (2007). Cross Validation in Small Area Estimation: Full Research Report. ESRC End of Award Report, RES-000-22-1798. URL: http://www.researchcatalogue.esrc.ac.uk/grants/RES-000-22-1798/outputs/read/55724d38-f71f-4a36-a4bd-9e70ff8f9842 [Accessed: 2-2-2017].

Smith, D. M., Pearce, J. R., & Harland, K. (2011). Can a Deterministic Spatial Microsimulation Model Provide Reliable Small-Area Estimates of Health Behaviours? an Example of Smoking Prevalence in New Zealand. *Health & Place*, **17**(2), 618–624. DOI: 10.1016/j.healthplace.2011.01.001.

Smith, M. P., Olatunde, O., White, C., & Statistics, N. (2010). Update to the Methodology Used to Calculate Health Expectancies for the UK and Constituent Countries. *Health Statistics Quarterly*, **45**, 81–99.

Smith, M. P. & White, C. (2009). An Investigation into the Impact of Question Change on Estimates of General Health Status and Healthy Life Expectancy. *Health Statistics Quarterly*, **41**, 28–41. URL: http://www.ons.gov.uk/ons/rel/hsq/health-statistics-quarterly/no--41--spring-2009/an-investigation-into-the-impact-of-question-change-on-estimates-of-general-health-status-and-hea pdf [Accessed: 2-2-2017].

Smith, P. F., Hadler, J. L., Stanbury, M., Rolfs, R. T., & Hopkins, R. S. (2013). "Blueprint Version 2.0": Updating Public Health Surveillance for the 21st Century. *Journal of Public Health Management and Practice*, **19**(3), 231–239. DOI: 10.1097/PHH.0b013e318262906e.

Smith, S. J. & Easterlow, D. (2005). The Strange Geography of Health Inequalities. *Transactions of the Institute of British Geographers*, **30**(2), 173–190. DOI: 10.1111/j.1475-5661.2005.00159.x.

Snijders, T. A. B. (2005). Power and Sample Size in Multilevel Modeling. *Encyclopedia of Statistics in Behavioral Science. Volume 3.*, **3**, 1570–1573.

Söhl, K., Schulz, R., & Kuhn, J. (2016). Ratenbildung bei KV-Daten mit GKV-Versicherten auf Kreisebene – Ein empirisches Schätzmodell auf der Basis des Mikrozensus. *Das Gesundheitswesen*. DOI: 10.1055/s-0042-100617.

Soljak, M., Samarasundera, E., Indulkar, T., Walford, H., & Majeed, A. (2011). Variations in Cardiovascular Disease Under-Diagnosis in England: National Cross-Sectional Spatial Analysis. *BMC Cardiovascular Disorders*, **11**(1), 1–11. DOI: 10.1186/1471-2261-11-12.

Song, L., Mercer, L., Wakefield, J., Laurent, A., & Solet, D. (2016). Using Small-area Estimation to Calculate the Prevalence of Smoking by Subcounty Geographic Areas in King County, Washington, Behavioral Risk Factor Surveillance System, 2009–2013. *Preventing Chronic Disease*, **13**(E59), 1545–1151. DOI: 10.5888/pcd13.150536.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **64**(4), 583–616. DOI: 10.1111/1467-9868.00353.

Spiegelhalter, D. J. & Freedman, L. S. (1986). A Predictive Approach to Selecting the Size of a Clinical Trial, Based on Subjective Clinical Opinion. *Statistics in Medicine*, **5**(1), 1–13. DOI: 10.1002/sim.4780050103.

Srebotnjak, T., Mokdad, A. H., & Murray, C. J. (2010). A Novel Framework for Validating and Applying Standardized Small Area Measurement Strategies. *Population Health Metrics*, **8**, 26. DOI: 10.1186/1478-7954-8-26.

Stafford, M., Becares, L., & Nazroo, J. (2009). Objective and Perceived Ethnic Density and Health: Findings from a United Kingdom General Population Survey. *American Journal of Epidemiology*, **170**(4), 484–493. DOI: 10.1093/aje/kwp160.

Steorts, R. & Ghosh, M. (2013). On Estimation of Mean Squared Errors of Benchmarked Empirical Bayes Estimators. *Statistica Sinica*. DOI: 10.5705/ss.2012.053.

Stern, H. S. & Cressie, N. (2000). Posterior Predictive Model Checks for Disease Mapping Models. *Statistics in Medicine*, **19**(17-18), 2377–2397. DOI: 10.1002/1097-0258(20000915/30)19:17/18<2377::AID-SIM576>3.0.CO;2-1.

Stone, R., Holzhauer, Barrie, J., Lowe, D., Searle, L., Skipper, E., Welham, S., & Roberts, C. (2015). COPD: Who Cares Matters. National Chronic Obstructive Pulmonary Disease (COPD) Audit Programme: Resources and Organisation of Care in Acute NHS Units in England and Wales 2014. URL: https://www.rcplondon.ac.uk/projects/national-copd-audit-programme-starting-2013 [Accessed: 2-2-2017].

Subramanian, S. V., Jones, K., & Duncan, C. (2003). Multilevel Methods for Public Health Research. In *Neighborhoods and Health*, pages 65–111. Oxford University Press, Oxford. DOI: 10.1093/acprof:oso/9780195138382.003.0004.

Sullivan, D. F. (1971). A Single Index of Mortality and Morbidity. *Health Services and Mental Health Administration Health Reports*, **86**(4), 347. DOI: 10.2307/4594169.

Sun, D., Tsutakawa, R. K., Kim, H., & He, Z. (2000). Spatio-Temporal Interaction with Disease Mapping. *Statistics in Medicine*, **19**(15), 2015–2035. DOI: 10.1002/1097-0258(20000815)19:15<2015::AID-SIM422>3.0.CO;2-E.

Sutton, M., Carr-Hill, R., Gravelle, H., & Rice, N. (1999). Do Measures of Self-Reported Morbidity Bias the Estimation of the Determinants of Health Care Utilisation? *Social Science & Medicine*, **49**(7), 867–878. DOI: 10.1016/S0277-9536(99)00169-0.

Szatkowski, L., Fahy, S. J., Coleman, T., Taylor, J., Twigg, L., Moon, G., & Leonardi-Bee, J. (2015). Small Area Synthetic Estimates of Smoking Prevalence during Pregnancy in England. *Population Health Metrics*, **13**(1), 34. DOI: 10.1186/s12963-015-0067-8.

Tarozzi, A. & Deaton, A. (2009). Using Census and Survey Data to Estimate Poverty and Inequality for Small Areas. *Review of Economics and Statistics*, **91**(4), 773–792. DOI: 10.1162/rest.91.4.773.

Tassone, E. C., Waller, L. A., & Casper, M. L. (2009). Small-Area Racial Disparity in Stroke Mortality. *Epidemiology*, **20**(2), 234–241. DOI: 10.1097/EDE.0b013e3181935aee.

Taylor, J., Twigg, L., & Moon, G. (2014). The Convergent Validity of Three Surveys As Alternative Sources of Health Information to the 2011 UK Census. *Social Science & Medicine*, **116**, 187–192. DOI: 10.1016/j.socscimed.2014.06.017.

Teutsch, S. M. (2010). Considerations in Planning a Surveillance System. In *Principles & Practice of Public Health Surveillance*, pages 18–31. Oxford University Press. DOI: 10.1093/acprof:oso/9780195372922.003.0002.

Thacker, S. B. & Berkelman, R. L. (1988). Public Health Surveillance in the United States. *Epidemiology Review*, **10**, 164–190.

Thayn, J. B. & Simanis, J. M. (2013). Accounting for Spatial Autocorrelation in Linear Regression Models Using Spatial Filtering with Eigenvectors. *Annals of the Association of American Geographers*, **103**(1), 47–66. DOI: 10.1080/00045608.2012.685048.

Thompson, E. J. (1995). The 1991 Census of Population in England and Wales. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **158**(2), 203. DOI: 10.2307/2983290.

Tiller, R. B. (1992). Time Series Modeling of Sample Survey Data from the U.s. Current Population Survey. *Journal of Official Statistics*, **8**(2), 149–166. URL: http://www.jos.nu/Articles/abstract.asp?article=82149 [Accessed: 2-2-2017].

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, **46**, 234. DOI: 10.2307/143141.

Toson, B. & Baker, A. (2003). *Life Expectancy at Birth: Methodological Options for Small Populations. National Statistics Methodological Series No.33*. HMSO, London. URL: http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/guide-method/method-quality/specific/gss-methodology-series/gss-methodology-series--33--life-expectancy-at-birth--methodological-options-for-small-populatior pdf [Accessed: 2-2-2017].

Tower Hamlets JSNA Reference Group (2011). Chronic Obstructive Pulmonary Disease (COPD): Factsheet. Tower Hamlets Joint Strategic Needs Assessment 2010–11. URL: http://democracy.towerhamlets.gov.uk/documents/s24101/COPD%20JSNA%20Factsheet.pdf [Accessed: 2-2-2017].

Tranmer, M., Steel, D., & Fieldhouse, E. (2003). Exploring Small Area Population Structure with Census Data. In Courgeau, D., editor, *Methodology and Epistemology of Multilevel Analysis*, pages 121–156. DOI: 10.1007/978-1-4020-4675-9_5.

Tranmer, M. & Steel, D. G. (1998). Using Census Data to Investigate the Causes of the Ecological Fallacy. *Environment & Planning A*, **30**(5), 817–831. URL: http://www.ncbi.nlm.nih.gov/pubmed/12293872 [Accessed: 2-2-2017].

Tranmer, M. & Steel, D. G. (2001). Ignoring a Level in a Multilevel Model: Evidence from UK Census Data. *Environment and Planning A*, **33**(5), 941–948. DOI: 10.1068/a3317.

Tudor Hart, J. (1971). The Inverse Care Law. *The Lancet*, **297**(7696), 405–412. DOI: 10.1016/S0140-6736(71)92410-X.

Twigg, L. & Moon, G. (2002). Predicting Small Area Health-Related Behaviour: A Comparison of Multilevel Synthetic Estimation and Local Survey Data. *Social Science & Medicine*, **54**(6), 931–937. DOI: 10.1016/S0277-9536(01)00065-X.

Twigg, L., Moon, G., & Jones, K. (2000). Predicting Small-Area Health-Related Behaviour: A Comparison of Smoking and Drinking Indicators. *Social Science & Medicine*, **50**(7-8), 1109–1120. DOI: 10.1016/S0277-9536(99)00359-7.

Twigg, L., Moon, G., & Walker, S. (2004). *The Smoking Epidemic in England*. NHS Health Development Agency, London. URL: https://researchportal.port.ac.uk/portal/en/publications/the-smoking-epidemic-in-england(57b43cc3-0499-42a6-b069-bc79275f42cb).html [Accessed: 2-2-2017].

Tzavidis, N., Ranalli, M. G., Salvati, N., Dreassi, E., & Chambers, R. (2015). Robust Small Area Prediction for Counts. *Statistical Methods in Medical Research*, **24**(3), 373–395. DOI: 10.1177/0962280214520731.

Tzavidis, N., Salvati, N., Pratesi, M., & Chambers, R. (2008). M-Quantile Models with Applic-

ation to Poverty Mapping. *Statistical Methods and Applications*, **17**(3), 393–411. DOI: 10.1007/s10260-007-0070-8.

Ugarte, M. D., Goicoa, T., Militino, A. F., & Sagaseta-López, M. (2009). Estimating Unemployment in Very Small Areas. *SORT*, **33**(1), 49–70. URL: https://eudml.org/doc/43040 [Accessed: 2-2-2017].

UK Administrative Data Taskforce (2012). The UK Administrative Data Research Network: Improving Access for Research and Policy. URL: http://www.esrc.ac.uk/files/research/administrative-data-taskforce-adt/improving-access-for-research-and-policy/ [Accessed: 2-2-2017].

UK Cabinet Office (2016). Better Use of Data in Government: Consultation a Government Summary of Responses. June 2016. URL: https://www.gov.uk/government/consultations/better-use-of-data-in-government [Accessed: 2-2-2017].

UK Cabinet Office. Prime Minister's Strategic Unit (2005). Improving the Life Chances of Disabled People. – Final Report January 2005. – A Joint Report with Department for Work and Pensions, Department of Health, Department for Education and Skills, Office of the Deputy Prime Minister. URL: http://webarchive.nationalarchives.gov.uk/+/http:/www.cabinetoffice.gov.uk/media/cabinetoffice/strategy/assets/disability.pdf [Accessed: 2-2-2017].

UK Department for Business Innovation & Skills (2012). 2011 Skills for Life Survey : Small Area Estimation Technical Report. *BIS Research Paper Number 81C*, (81).

UK Department for Communities and Local Government (2013). English Housing Survey, 2010–2011: Household Data. [data collection]. SN: 7040. 2nd edition, January 2013. Colchester, Essex: UK Data Archive [distributor]. DOI: 10.5255/UKDA-SN-7040-2.

UK Department for Communities and Local Government (2015). English Housing Survey, 2008-2014: Secure Access. [data Collection]. 5th Edition.. DOI: 10.5255/UKDA-SN-6923-5.

UK Department for Health and Social Services (1976). Sharing Resources for Health in England. Report of the Resource Allocation Working Party.

UK Department of Health (1989). *Working for Patients (Cm. 155)*. HMSO, London.

UK Department of Health (1995). Variations in Health: What Can the Department of Health and the National Health Service Do? A Report from the Variations Subgroup of the Chief Medical Officer's Health of the Nation Working Group.

UK Department of Health (1999). Saving Lives: Our Healthier Nation (Cm. 4386).

UK Department of Health (2003). Tackling Health Inequalities: A Programme for Action (Cm. 6374). URL: http://webarchive.nationalarchives.gov.uk/20130107105354/http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_4019362.pdf [Accessed: 2-2-2017].

UK Department of Health (2010). Equity and Excellence: Liberating the NHS (Cm. 7881).

UK Department of Health (2011a). An Outcomes Strategy for Chronic Obstructive Pulmonary Disease (COPD) and Asthma in England. URL: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/216139/dh_128428.pdf [Accessed: 2-2-2017].

UK Department of Health (2011b). Public Health in Local Government: Commissioning Responsibilities. URL: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/216712/dh_131901.pdf [Accessed: 2-2-2017].

UK Department of Health (2012). Public Health Surveillance: Towards a Public Health Surveillance Strategy for England. URL: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/213339/Towards-a-Public-Health-Surveillance-Strategy.pdf [Accessed: 2-2-2017].

UK Department of Health (2013). Improving Outcomes and Supporting Transparency. âĂŞ- Part 2: Summary Technical Specifications of Public Health Indicators. Updated November 2013. URL: https://www.gov.uk/government/publications/healthy-lives-healthy-people-improving-outcomes-and-supporting-transparency [Accessed: 2-2-2017].

UK Department of Health (2016). Improving Outcomes and Supporting Transparency. – Part 2: Summary Technical Specifications of Public Health Indicators. – Updated Au-

gust 2016. URL: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/545605/PHOF_Part_2.pdf [Accessed: 2-2-2017].

UK Department of Health and Social Security (1980). Inequalities in Health: Report of a Research Working Group.

UK House of Commons (2016). Digital Economy Bill (HC Bill 45). URL: https://www.publications.parliament.uk/pa/bills/cbill/2016-2017/0045/cbill_2016-20170045_en_1.htm [Accessed: 2-2-2017].

UK House of Commons. Treasury Committee (2008). *Counting the Population. Eleventh Report of Session 2007-08. Volume I. — Report, Together with Formal Minutes, Oral and Written Evidence Ordered by the House of Commons to Be Printed 14 May 2008. HC 183-I*. The Stationary Office Limited, London. URL: http://www.publications.parliament.uk/pa/cm200708/cmselect/cmtreasy/183/183.pdf [Accessed: 2-2-2017].

UK Information Commissioner's Office (2012). Anonymisation: Managing Data Protection Risk. Code of Practice. URL: https://ico.org.uk/media/1061/anonymisation-code.pdf [Accessed: 2-2-2017].

UK National Audit Office (2015). Report by the Comptroller and Auditor General. General Practice Extraction Service – Investigation (HC 265, 2015–16, 2 July 2015). URL: https://www.nao.org.uk/wp-content/uploads/2015/07/General-Practice-Extraction-Service-Investigation.pdf [Accessed: 2-2-2017].

UK Statistics Authority (2009). Code of Practice for Official Statistics. January 2009. URL: https://www.statisticsauthority.gov.uk/wp-content/uploads/2015/12/images-codeofpracticeforofficialstatisticsjanuary2009_tcm97-25306.pdf [Accessed: 2-2-2017].

UK Statistics Authority (2016a). Delivering Better Statistics for Better Decisions. Data Access Legislation. March 2016. URL: https://www.statisticsauthority.gov.uk/publication/delivering-better-statistics-for-better-decisions-data-access-legislation-march-2016/ [Accessed: 2-2-2017].

UK Statistics Authority (2016b). Health and Care Statistics in England – The Statistics Authority's Direction of Travel. 22 March 2016. URL: https://www.statisticsauthority.gov.uk/publication/health-and-care-statistics-in-england-the-statistics-authoritys-direction-of-travel/ [Accessed: 2-2-2017].

US Census Bureau (2008). *A Compass for Understanding and Using American Community Survey Data: What General Data Users Need to Know*. U.S. Government Printing Office, Washington, DC. URL: https://www.census.gov/content/dam/Census/library/publications/2008/acs/ACSGeneralHandbook.pdf [Accessed: 2-2-2017].

US Census Bureau (2013). US Census Bureau Statistical Quality Standards. Reissued July 2013. URL: https://www.census.gov/content/dam/Census/about/about-the-bureau/policies_and_notices/quality/statistical-quality-standards/Quality_Standards.pdf [Accessed: 2-2-2017].

US Census Bureau (2015). 2020 Census Operational Plan. The Most Innovative Operational Design since 1790. Issued October 6, 2015. URL: https://www.census.gov/content/dam/Census/programs-surveys/decennial/2020-census/pmr100615/2020%20census%20operational%20plan%20final.pdf [Accessed: 2-2-2017].

US Department of Health and Human Services, Health Resources and Services Administration, Bureau of Health Workforce (2016). Area Health Resources Files. 2015–2016. Rockville, MD. URL: http://ahrf.hrsa.gov/ [Accessed: 2-2-2017].

US National Center for Health Statistics (1968). *Synthetic State Estimates of Disability. PHS Publication No. 1759*. Government Printing Office, Washington, DC.

Vaida, F. & Blanchard, S. (2005). Conditional Akaike Information for Mixed-Effects Models. *Biometrika*, **92**(2), 351–370. DOI: 10.1093/biomet/92.2.351.

Vallejo-Torres, L., Morris, S., Carr-Hill, R., Dixon, P., Law, M., Rice, N., & Sutton, M. (2009). Can Regional Resource Shares Be Based Only on Prevalence Data? An Empirical Investigation of

the Proportionality Assumption. *Social Science & Medicine*, **69**(11), 1634–1642. DOI: 10.1016/j.socscimed.2009.09.020.

Vandendijck, Y., Faes, C., Kirby, R. S., Lawson, A., & Hens, N. (2016). Model-Based Inference for Small Area Estimation with Sampling Weights. *Spatial Statistics*, **18**, 455–473. DOI: 10.1016/j.spasta.2016.09.004.

Voutilainen, A. & Sherwood, P. R. (2015). Explaining State-Level Differences in Brain Cancer Mortality in the United States by Population Ancestries: A Spatial Approach. *Journal of Epidemiological Research*, **1**(1). DOI: 10.5430/jer.v1n1p12.

Wakefield, J. C., Kelsall, J. E., & Morris, S. E. (2000). Clustering, Cluster Detection, and Spatial Variation in Risks. In Elliott, P., Wakefield, J. C., Best, N. G., & Briggs, D. J., editors, *Spatial Epidemiology: Methods and Applications*, pages 128–152.

Walford, H., Kearns, B., & Barron, S. (2011). *Prevalence Modelling*. Technical Briefing 8. Association of Public Health Observatories, York. URL: http://www.apho.org.uk/resource/item.aspx?RID=100181 [Accessed: 2-2-2017].

Walford, H. & Ramsay, L. (2011). COPD Prevalence Modelling Briefing Document. URL: http://www.apho.org.uk/resource/item.aspx?RID=111137 [Accessed: 2-2-2017].

Wall, M. M. (2004). A Close Look at the Spatial Structure Implied by the Car and Sar Models. *Journal of Statistical Planning and Inference*, **121**(2), 311–324. DOI: 10.1016/S0378-3758(03)00111-3.

Walsh, M. G. & Haseeb, M. A. (2014). Small-Area Estimation of the Probability of Toxocariasis in New York City Based on Sociodemographic Neighborhood Composition. *PLoS ONE*, **9**(6). DOI: 10.1371/journal.pone.0099303.

Wang, Y., Ponce, N. A., Wang, P., Opsomer, J. D., & Yu, H. (2015). Generating Health Estimates by Zip Code: A Semiparametric Small Area Estimation Approach Using the California Health Interview Survey. *American Journal of Public Health*, **105**(12), 2534–2540. DOI: 10.2105/AJPH.2015.302810.

Ware, J. E. & Sherbourne, C. D. (1992). The MOS 36-ltem Short-Form Health Survey (SF-36). *Medical Care*, **30**(6), 473–483. DOI: 10.1097/00005650-199206000-00002.

Webber, L., Mytton, O. T., Briggs, A. D. M., Woodcock, J., Scarborough, P., McPherson, K., & Capewell, S. (2014). The Brighton Declaration: The Value of Non-Communicable Disease Modelling in Population Health Sciences. *European Journal of Epidemiology*, **29**(12), 867–870. DOI: 10.1007/s10654-014-9978-0.

Wennberg, J. & Gittelsohn, A. (1973). Small Area Variations in Health Care Delivery: A Population-Based Health Information System Can Guide Planning and Regulatory Decision-Making. *Science*, **182**(4117), 1102–1108. DOI: 10.1126/science.182.4117.1102.

Wennberg, J. E. (1987). Population Illness Rates Do Not Explain Population Hospitalization Rates. *Medical Care*, **25**(4), 354–359. DOI: 10.1097/00005650-198704000-00008.

White, H. L., Matheson, F. I., Moineddin, R., Dunn, J. R., & Glazier, R. H. (2011). Neighbourhood Deprivation and Regional Inequalities in Self-Reported Health Among Canadians: Are We Equally at Risk? *Health & Place*, **17**(1), 361–369. DOI: 10.1016/j.healthplace.2010.11.016.

White, P. (1990). A Question on Ethnic Group for the Census: Findings from the 1989 Census Test Post-Enumeration Survey. *Population Trends*, **59**, 11–20.

Whitehead, F. (1988). How the 1991 Census Should Improve Government Statistics. *Population Trends*, **53**, 18–21.

Whitehead, M., Townsend, P., & Davidsen, N. (1992). *Inequalities in Health: The Black Report: The Health Divide*. Penguin.

Whittle, P. (1954). On Stationary Processes in the Plane. *Biometrika*, **41**(3-4), 434–449. DOI: 10.1093/biomet/41.3-4.434.

Whitworth, A. (2013). Methodological Review Paper Evaluations and Improvements in Small Area Estimation Methodologies. *Discussion Paper. NCRM.*. URL: http://eprints.ncrm.ac.uk/3210/ [Accessed: 2-2-2017].

Williams, T., Dinsdale, H., Eayres, D., & Tahzib, F. (2005). Technical Report: Calculating Life Expectancy in Small Areas. URL: http://www.thehealthwell.info/node/31640 [Accessed: 2-2-2017].

World Health Organisation (1948). The Constitution.

World Health Organisation (2001). International Classification of Functioning, Disability and Health. URL: http://apps.who.int/classifications/icfbrowser/ [Accessed: 2-2-2017].

World Health Organisation (2009). *Quantification of the Disease Burden Attributable to Environmental Risk Factors*. URL: http://www.who.int/quantifying_ehimpacts/summaryEBD_updated.pdf%0A [Accessed: 2-2-2017].

World Health Organisation (2011). International Statistical Classification of Diseases and Related Health Problems, Tenth Revision. Edition 2010. URL: http://apps.who.int/classifications/icd10/browse/2010/en#/ [Accessed: 2-2-2017].

Wright, J., Williams, R., & Wilkinson, J. R. (1998). Health Needs Assessment: Development and Importance of Health Needs Assessment. *British Medical Journal*, **316**(7140), 1310–1313. DOI: 10.1136/bmj.316.7140.1310.

Xie, D. (2004). *Combining Information from Multiple Surveys for Small Area Estimation: Bayesian Approaches*. Unpublished doctoral thesis, University of Michigan. URL: http://hdl.handle.net/2027.42/124596 [Accessed: 2-2-2017].

Xie, D., Raghunathan, T. E., & Lepkowski, J. M. (2007). Estimation of the Proportion of Overweight Individuals in Small Areas – A Robust Extension of the Fay-Herriot Model. *Statistics in Medicine*, **26**, 2699–2715. DOI: 10.1002/sim.

Xu, R. (2003). Measuring Explained Variation in Linear Mixed Effects Models. *Statistics in Medicine*, **22**(22), 3527–3541. DOI: 10.1002/sim.1572.

Ybarra, L. M. R. & Lohr, S. L. (2008). Small Area Estimation When Auxiliary Information Is Measured with Error. *Biometrika*, **95**(4), 919–931. DOI: 10.1093/biomet/asn048.

You, J., Datta, G. S., & Maples, J. J. (2014). Modeling Disability in Small Areas: An Area-level Approach of Combining Two Surveys. *U.S. Census Bureau Center for Statistical Research & Methodology Research Report Series (Statistics #2014-11)*. URL: https://www.census.gov/srd/papers/pdf/RRS2014-11.pdf [Accessed: 2-2-2017].

You, Y. & Rao, J. N. K. (2002). Small Area Estimation Using Unmatched Sampling and Linking Models. *Canadian Journal of Statistics*, **30**(1), 3–15. DOI: 10.2307/3315862.

You, Y. & Zhou, Q. M. (2011). Hierarchical Bayes Small Area Estimation under a Spatial Model with Application to Health Survey Data. *Survey Methodology*, **37**(1), 25–37. URL: http://www5.statcan.gc.ca/olc-cel/olc.action?ObjId=12-001-X201100111445&ObjType=47 [Accessed: 2-2-2017].

Zandbergen, P. A. (2014). Ensuring Confidentiality of Geocoded Health Data: Assessing Geographic Masking Strategies for Individual-Level Data. *Advances in Medicine*, **2014**, 1–14. DOI: 10.1155/2014/567049.

Zhang, L.-C. (1999). SMAREST. – A Survey of SMall ARea ESTimation. *Department for Coordination and Development Document 99/9*, **Statistics**. URL: http://www.ssb.no/a/histstat/doc/doc_199909.pdf [Accessed: 2-2-2017].

Zhang, X., Holt, J. B., Lu, H., Wheaton, A. G., Ford, E. S., Greenlund, K. J., & Croft, J. B. (2014). Multilevel Regression and Poststratification for Small-Area Estimation of Population Health Outcomes: A Case Study of Chronic Obstructive Pulmonary Disease Prevalence Using the Behavioral Risk Factor Surveillance System. *American Journal of Epidemiology*, **179**(8), 1025–1033. DOI: 10.1093/aje/kwu018.

Zhang, X., Holt, J. B., Yun, S., Lu, H., Greenlund, K. J., & Croft, J. B. (2015). Validation of Multilevel Regression and Poststratification Methodology for Small Area Estimation of Health Indicators from the Behavioral Risk Factor Surveillance System. *American Journal of Epidemiology*, **182**(2), 127–137. DOI: 10.1093/aje/kwv002.

Zhang, X., Onufrak, S., Holt, J. B., & Croft, J. B. (2013). A Multilevel Approach to Estimating Small Area Childhood Obesity Prevalence at the Census Block-Group Level. *Preventing Chronic Disease*, **10**(8), E68. DOI: 10.5888/pcd10.120252.

Zhang, Z. & Wang, L. (2009). Statistical Power Analysis for Growth Curve Models Using Sas. *Behavior Research Methods*, **41**(4), 1083–1094. DOI: 10.3758/BRM.41.4.1083.

Zhang, Z., Zhang, L., Penman, A., & May, W. (2011). Using Small-Area Estimation Method to Cal-

culate County-Level Prevalence of Obesity in Mississippi, 2007–2009. *Preventing Chronic Disease*, **8**(4), A85. URL: https://www.ncbi.nlm.nih.gov/pubmed/21672409 [Accessed: 2-2-2017].

Zou, K. H. & Normand, S. L. T. (2001). On Determination of Sample Size in Hierarchical Binomial Models. *Statistics in Medicine*, **20**(14), 2163–2182. DOI: 10.1002/sim.855.

Zou, K. H., Resnic, F. S., Gogate, A. S., & Ondategui-Parra, S. (2003). Efficient Bayesian Sample Size Calculation for Designing a Clinical Trial with Multi-Cluster Outcome Data. *Biometrical Journal*, **45**(7), 826–836. DOI: 10.1002/bimj.200390052.

zu Erbach-Schoenberg, E., Alegana, V. A., Sorichetta, A., Linard, C., Lourenço, C., Ruktanonchai, N. W., Graupe, B., Bird, T. J., Pezzulo, C., Wesolowski, A., & Tatem, A. J. (2016). Dynamic Denominators: The Impact of Seasonally Varying Population Numbers on Disease Incidence Estimates. *Population Health Metrics*, **14**(1), 35. DOI: 10.1186/s12963-016-0106-0.

Appendix A

Additional tables and figures

Appendix A. Additional tables and figures

Table A.1 | Detail of database search queries and number of results extracted on 25 June 2016

Database	Number of citations
Ovid Embase Classic + Embase 1947 to 2016 June 24 `(("small area estim*" OR "synthetic estimat*" OR "model-based estimat*" OR "model-assisted estimat*") AND (morbidity OR accident* OR alcohol OR casualt* OR contraceptive OR crash* OR depriv* OR disability OR disease OR "drink*" OR "drug*" OR "economic activity" OR "economic status" OR health OR illness OR life OR lifestyle OR poverty OR reproductive OR safe* OR "sick*" OR "smok*" OR substance OR "unemploy*" OR well-being).mp. [mp=title, abstract, heading word, original title, keyword]`	**116**
MEDLINE/PubMed `((small area estim* OR synthetic estimat* OR model-based estimat* OR model-assisted estimat*) AND (morbidity OR accident* OR alcohol OR casualt* OR contraceptive OR crash* OR depriv* OR disability OR disease OR drink* OR drug* OR economic activity OR economic status OR health OR illness OR life OR lifestyle OR poverty OR reproductive OR safe* OR sick* OR smok* OR substance OR unemploy* OR well-being))`	**396**
Scopus `TITLE-ABS-KEY (("small area estim*" OR "synthetic estimat*" OR "model-based estimat*" OR "model-assisted estimat*") AND (morbidity OR accident* OR alcohol OR casualt* OR contraceptive OR crash* OR depriv* OR disability OR disease OR "drink*" OR "drug*" OR "economic activity" OR "economic status" OR health OR illness OR life OR lifestyle OR poverty OR reproductive OR safe* OR "sick*" OR "smok*" OR substance OR "unemploy*" OR well-being))`	**624**
Web of Science™ Core Collection `TS=(("small area estim*" OR "synthetic estimat*" OR "model-based estimat*" OR "model-assisted estimat*") AND (morbidity OR accident* OR alcohol OR casualt* OR contraceptive OR crash* OR depriv* OR disability OR disease OR "drink*" OR "drug*" OR "economic activity" OR "economic status" OR health OR illness OR life OR lifestyle OR poverty OR reproductive OR safe* OR "sick*" OR "smok*" OR substance OR "unemploy*" OR well-being))` Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH.	**525**
BIOSIS Citation Index `TS=(("small area estim*" OR "synthetic estimat*" OR "model-based estimat*" OR "model-assisted estimat*") AND (morbidity OR accident* OR alcohol OR casualt* OR contraceptive OR crash* OR depriv* OR disability OR disease OR "drink*" OR "drug*" OR "economic activity" OR "economic status" OR health OR illness OR life OR lifestyle OR poverty OR reproductive OR safe* OR "sick*" OR "smok*" OR substance OR "unemploy*" OR well-being))` Indexes: BCI. Last updated: 2016-06-17.	**226**

Table A.2 | Percentage of the between-area variance explained by each model

		LLTI	SRH
LAD	**M1**	53%	61%
	M2	61%	72%
MSOA	**M1**	37%	38%
	M2	38%	40%

Table A.3 | Moran's *I* estimates of spatial autocorrelation in model residuals by spatial weight matrix (number of nearest neighbours)

				SRH
		3 neighbours	2 neighbours	1 neighbour
LAD	**M1**	0.3814	0.4653	0.5836
		$p < 0.001$	$p < 0.001$	$p = 0.006$
	M2	0.3217	0.4014	0.5378
		$p = 0.011$	$p = 0.019$	$p = 0.122$
MSOA	**M1**	0.2945	0.3814	0.5453
		$p < 0.001$	$p < 0.001$	$p < 0.001$
	M2	0.2919	0.3789	0.5443
		$p < 0.001$	$p < 0.001$	$p < 0.001$
				LLTI
		3 neighbours	2 neighbours	1 neighbour
LAD	**M1**	0.3573	0.4243	0.5879
		$p < 0.001$	$p = 0.003$	$p = 0.004$
	M2	0.3538	0.4356	0.5910
		$p < 0.001$	$p = 0.001$	$p = 0.003$
MSOA	**M1**	0.303	0.3843	0.5417
		$p < 0.001$	$p < 0.001$	$p < 0.001$
	M2	0.3015	0.3827	0.5404
		$p < 0.001$	$p < 0.001$	$p < 0.001$

Appendix A. Additional tables and figures

Table A.4 | Model coefficients: M1 – LAD

	SRH			LLTI		
	$\hat{\beta}$	S.E.	p	$\hat{\beta}$	S.E.	p
Intercept	–3.402	0.026	0.000	–2.057	0.017	0.000
(Age–40)	0.521	0.010	0.000	0.414	0.004	0.000
$(\text{Age–40})^2$	–0.028	0.003	0.000	0.012	0.002	0.000
LAD hosp. rate	1.571	0.106	0.000	1.073	0.076	0.000
London suburbs	0.117	0.087	0.179	–0.112	0.063	0.074
London cosmopolitan	0.420	0.109	0.000	0.095	0.081	0.239
South East prosperous	–0.255	0.068	0.000	–0.154	0.043	0.000
Coastal & countryside	0.025	0.063	0.692	0.041	0.043	0.348
σ^2	0.044	0.210		0.028	0.166	
AIC	27,220			50,998		
BIC	27,294			51,072		
Deviance	27,202			50,980		

Table A.5 | Model coefficients: M1 – MSOA

	SRH			LLTI		
	$\hat{\beta}$	S.E.	p	$\hat{\beta}$	S.E.	p
Intercept	–3.475	0.024	0.000	–2.102	0.014	0.000
(Age–40)	0.538	0.010	0.000	0.430	0.004	0.000
$(\text{Age–40})^2$	–0.029	0.003	0.000	0.013	0.002	0.000
MSOA hosp. rate	1.701	0.079	0.000	1.295	0.052	0.000
LAD hosp. rate	0.176	0.106	0.097	–0.043	0.068	0.528
MSOA hosp. rate × LAD hosp. rate	–1.414	0.249	0.000	–0.865	0.163	0.000
σ^2	0.217	0.466		0.122	0.349	
AIC	49,210			106,077		
BIC	49,278			106,146		
Deviance	49,196			106,063		

Table A.6 | Model coefficients: M2 – LAD

	SRH			LLTI		
	$\hat{\beta}$	S.E.	p	$\hat{\beta}$	S.E.	p
Intercept	–5.362	0.084	0.000	–3.282	0.039	0.000
Age 16–24	0.430	0.112	0.000	0.377	0.045	0.000
Age 25–34	1.072	0.096	0.000	0.594	0.043	0.000
Age 35–44	1.683	0.086	0.000	1.084	0.038	0.000
Age 45–49	2.106	0.089	0.000	1.389	0.041	0.000
Age 50–64	2.705	0.080	0.000	1.914	0.034	0.000
Age 65–74	2.973	0.081	0.000	2.356	0.035	0.000
Age 75–84	3.415	0.082	0.000	2.878	0.037	0.000
Age 85+	3.685	0.090	0.000	3.391	0.047	0.000
White	–	–	–	–	–	–
Asian	0.039	0.117	0.740	–0.264	0.075	0.000
Black	–0.529	0.226	0.020	–0.464	0.130	0.000
Mixed	0.193	0.148	0.194	0.198	0.080	0.013
Other	0.387	0.108	0.000	–0.176	0.079	0.025
London Centre	0.529	0.114	0.000	0.124	0.089	0.162
South East prosperous	–0.234	0.065	0.000	–0.145	0.041	0.001
Coastal & countryside	0.125	0.062	0.043	0.091	0.043	0.035
Regional centres	0.125	0.057	0.027	0.208	0.042	0.000
Mining & manufacturing	0.239	0.061	0.000	0.208	0.044	0.000
LAD hosp. rates	1.215	0.133	0.000	0.719	0.097	0.000
% not White	0.412	0.151	0.007	–0.018	0.110	0.873
% not White × Asian	0.565	0.320	0.078	0.614	0.212	0.004
% not White × Black	1.477	0.553	0.008	0.931	0.337	0.006
σ^2	0.031	0.176		0.023	0.152	
AIC	11,119			17,337		
BIC	11,275			17,493		
Deviance	11,073			17,291		

Table A.7 | Model coefficients: M2 – MSOA

	SRH			LLTI		
	$\hat{\beta}$	S.E.	p	$\hat{\beta}$	S.E.	p
Intercept	−5.385	0.079	0.000	−3.262	0.033	0.000
Age 16–24	0.426	0.112	0.000	0.378	0.046	0.000
Age 25–34	1.057	0.096	0.000	0.583	0.043	0.000
Age 35–44	1.702	0.086	0.000	1.106	0.038	0.000
Age 45–49	2.146	0.089	0.000	1.429	0.041	0.000
Age 50–64	2.756	0.080	0.000	1.968	0.034	0.000
Age 65–74	3.043	0.082	0.000	2.431	0.036	0.000
Age 75–84	3.486	0.083	0.000	2.963	0.038	0.000
Age 85+	3.769	0.091	0.000	3.495	0.048	0.000
White	–	–	–	–	–	–
Asian	0.002	0.065	0.974	−0.209	0.042	0.000
Black	−0.186	0.094	0.047	−0.298	0.059	0.000
Mixed	0.074	0.151	0.627	0.103	0.082	0.206
Other	0.268	0.111	0.015	−0.290	0.080	0.000
% not White	0.461	0.089	0.000	0.044	0.060	0.464
LAD hosp. rate	0.200	0.106	0.058	−0.061	0.068	0.372
MSOA hosp. rate	1.592	0.082	0.000	1.321	0.053	0.000
MSOA hosp. rate × LAD hosp. rate	−1.260	0.251	0.000	−0.939	0.164	0.000
σ^2	0.209	0.458		0.121	0.348	
AIC	37,688			74,475		
BIC	37,848			74,635		
Deviance	37,652			74,439		

Table A.8 | Regression coefficients: baseline models

M0	WHITE			BLACK			ASIAN			MIXED			OTHER		
	$\hat{\boldsymbol{\beta}}$	SE	p	$\hat{\boldsymbol{\beta}}$	SE	p	$\hat{\boldsymbol{\beta}}$	SE	p	$\hat{\boldsymbol{\beta}}$	SE	p	$\hat{\boldsymbol{\beta}}$	SE	p
(Intercept)	–4.046	0.025	<.001	–4.179	0.034	<.001	–4.493	0.059	<.001	–3.911	0.054	<.001	–4.239	0.049	<.001
Male	–0.080	0.001	<.001	–0.054	0.007	<.001	–0.248	0.005	<.001	0.138	0.009	<.001	–0.089	0.012	<.001
Aged 0–15	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Aged 16–49	0.914	0.003	<.001	0.816	0.012	<.001	0.658	0.009	<.001	0.980	0.012	<.001	0.924	0.023	<.001
Aged 50–64	2.036	0.003	<.001	1.975	0.013	<.001	2.466	0.009	<.001	2.342	0.015	<.001	2.417	0.024	<.001
Aged 65+	3.179	0.003	<.001	3.232	0.013	<.001	3.623	0.009	<.001	3.201	0.016	<.001	3.353	0.025	<.001
Wales	0.274	0.047	<.001	0.191	0.080	0.017	–0.028	0.079	0.727	0.269	0.056	<.001	0.181	0.080	0.024
Lond. centre	–0.255	0.076	0.001	0.491	0.097	<.001	–0.019	0.115	0.872	0.069	0.085	0.416	0.165	0.103	0.108
Prospering	–0.341	0.029	<.001	–0.219	0.040	<.001	–0.337	0.056	<.001	–0.351	0.041	<.001	–0.296	0.047	<.001
Coastal	–0.181	0.040	<.001	0.147	0.069	0.033	–0.316	0.079	<.001	–0.075	0.059	0.206	–0.097	0.077	0.208
Mining	0.163	0.037	<.001	0.008	0.054	0.876	0.034	0.069	0.620	0.088	0.053	0.093	0.024	0.061	0.698
% same ethnicity							1.887	0.297	<.001	–6.016	1.329	<.001	6.477	1.634	<.001
σ^2	0.041	0.202		0.062	0.248		0.092	0.303		0.042	0.205		0.051	0.227	
Shapiro–Wilks	0.996		0.519	0.994		0.148	0.995		0.362	0.997		0.724	0.989		0.009
Moran's I	0.511		<.001	0.431		<.001	0.550		<.001	0.503		<.001	0.394		<.001
AIC	97,530			15,517			21,760			16,112			12,298		

Table A.9 | Regression coefficients: testing models with LAD-level mortality SMRs as predictors of LAD-level prevalence of LLTI

M1	WHITE			BLACK			ASIAN			MIXED			OTHER		
	$\hat{\boldsymbol{\beta}}$	SE	p	$\hat{\boldsymbol{\beta}}$	SE	p	$\hat{\boldsymbol{\beta}}$	SE	p	$\hat{\boldsymbol{\beta}}$	SE	p	$\hat{\boldsymbol{\beta}}$	SE	p
(Intercept)	–4.046	0.024	<.001	–4.189	0.035	<.001	–4.305	0.037	<.001	–4.210	0.024	<.001	–4.167	0.039	<.001
Male	–0.095	0.001	<.001	–0.054	0.007	<.001	–0.248	0.005	<.001	0.137	0.009	<.001	–0.090	0.012	<.001
Aged 0–15	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Aged 16–49	0.914	0.003	<.001	0.814	0.012	<.001	0.658	0.009	<.001	0.979	0.012	<.001	0.934	0.023	<.001
Aged 50–64	2.036	0.003	<.001	1.974	0.013	<.001	2.466	0.009	<.001	2.344	0.015	<.001	2.431	0.024	<.001
Aged 65+	3.179	0.003	<.001	3.231	0.013	<.001	3.624	0.009	<.001	3.204	0.016	<.001	3.362	0.025	<.001
Wales	0.270	0.046	<.001	0.184	0.080	0.022	–0.088	0.075	0.243	0.248	0.046	<.001	0.132	0.082	0.106
Lond. centre	–0.253	0.074	0.001	0.500	0.097	<.001	0.090	0.109	0.412	0.063	0.064	0.325	0.392	0.094	<.001
Prospering	–0.330	0.028	<.001	–0.195	0.045	<.001	–0.348	0.048	<.001	–0.052	0.029	0.071	–0.290	0.046	<.001
Coastal	–0.173	0.039	<.001	0.153	0.069	0.026	–0.477	0.066	<.001	0.169	0.040	<.001	–0.168	0.073	0.021
Mining	0.162	0.036	<.001	–0.012	0.055	0.822	–0.295	0.056	<.001	0.137	0.034	<.001	–0.125	0.058	0.032
SMR‡				0.028	0.021	0.180	0.191	0.021	<.001	0.180	0.013	<.001	0.091	0.022	<.001
Male SMR‡× Male	0.033	0.002	<.001												
Female SMR‡× Female	–0.014	0.002	<.001												
σ^2	0.039	0.198		0.061	0.247		0.081	0.285		0.024	0.155		0.056	0.237	
Moran's I	0.512		<.001	0.419		<.001	0.456		<.001	0.439		<.001	0.354		<.001
AIC	95,594			15,517			21,726			15,981			12,225		
Reduction in AIC (M0)	1,936			0			34			131			73		

Note: SMR: directly standardised mortality ratio. – ‡ Variable Z-standardised.

Table A.10 | Regression coefficients: final models predicting LAD-level prevalence of LLTI

M2	WHITE			BLACK			ASIAN			MIXED			OTHER		
	$\hat{\beta}$	SE	p	$\hat{\beta}$	SE	p	$\hat{\beta}$	SE	p	$\hat{\beta}$	SE	p	$\hat{\beta}$	SE	p
Intercept	–4.121	0.019	<.001	–4.187	0.035	<.001	–4.528	0.053	<.001	–4.086	0.045	<.001	–4.398	0.050	<.001
Male	–0.095	0.001	<.001	–0.054	0.007	<.001	–0.233	0.006	<.001	0.137	0.009	<.001	–0.089	0.012	<.001
Aged 0–15	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Aged 16–49	0.888	0.003	<.001	0.814	0.012	<.001	0.589	0.011	<.001	0.981	0.012	<.001	0.954	0.023	<.001
Aged 50–64	1.968	0.003	<.001	1.947	0.014	<.001	2.353	0.011	<.001	2.304	0.016	<.001	2.466	0.025	<.001
Aged 65+	3.150	0.003	<.001	3.181	0.014	<.001	3.554	0.010	<.001	3.183	0.016	<.001	3.381	0.025	<.001
Wales	0.344	0.034	<.001	0.222	0.081	0.007	–0.050	0.072	0.491	0.275	0.047	<.001	0.178	0.079	0.024
Lond. Centre	–0.104	0.056	0.062	0.526	0.096	<.001	0.100	0.101	0.321	0.140	0.063	0.028	0.180	0.096	0.060
Prospering	–0.135	0.024	<.001	–0.173	0.046	<.001	–0.114	0.054	0.034	–0.106	0.036	0.003	–0.091	0.052	0.078
Coastal	–0.018	0.030	0.549	0.184	0.070	0.008	–0.201	0.072	0.005	0.103	0.049	0.035	–0.002	0.075	0.977
Mining	0.092	0.027	<.001	–0.012	0.054	0.819	–0.066	0.061	0.278	0.043	0.041	0.292	–0.008	0.058	0.886
% same ethnicity							1.898	0.260	0.000	–3.493	1.029	0.001	8.926	1.598	<.001
ISAR†	0.004	0.001	<.001	0.003	0.002	0.068	0.002	0.002	0.302	0.003	0.001	0.001	0.005	0.002	0.005
ISAR†× Male	0.001	0.000	<.001				–0.003	0.001	<.001						
ISAR†× 16–49	0.004	0.000	<.001				0.003	0.001	<.001						
ISAR†× 50–64	0.010	0.000	<.001				0.002	0.001	0.015				–0.004	0.001	<.001
ISAR†× 65+	0.004	0.000	<.001	0.002	0.001	0.040									
SMR‡				–0.062	0.029	0.030	0.061	0.028	0.032	0.098	0.018	<.001	0.147	0.030	<.001
SMR‡× Male							0.037	0.008	<.001						
SMR‡× 16–49							0.028	0.012	0.020	0.119	0.011	<.001	–0.152	0.012	<.001
SMR‡× 50–64				0.092	0.010	<.001	0.178	0.012	<.001	0.086	0.013	<.001			
SMR‡× 65+				0.121	0.013	<.001	0.101	0.010	<.001						
Male SMR‡× Male	0.027	0.002	<.001												
Female SMR‡× Female	–0.010	0.002	<.001												
σ^2	0.021	0.146		0.059	0.243		0.068	0.260		0.022	0.147		0.042	0.205	
Shapiro–Wilks	0.994		0.192	0.996		0.423	0.995		0.314	0.997		0.863	0.987		0.003
Moran's I	0.555		<.001	0.412		<.001	0.449		<.001	0.410		<.001	0.344		<.001
AIC	85,779			15,239			20,925			15,837			12,040		

Note: ISAR: indirectly standardised emergency admission ratio. – SMR: directly standardised mortality ratio. – † Variable centred around 1.00. – ‡ Variable Z-standardised.

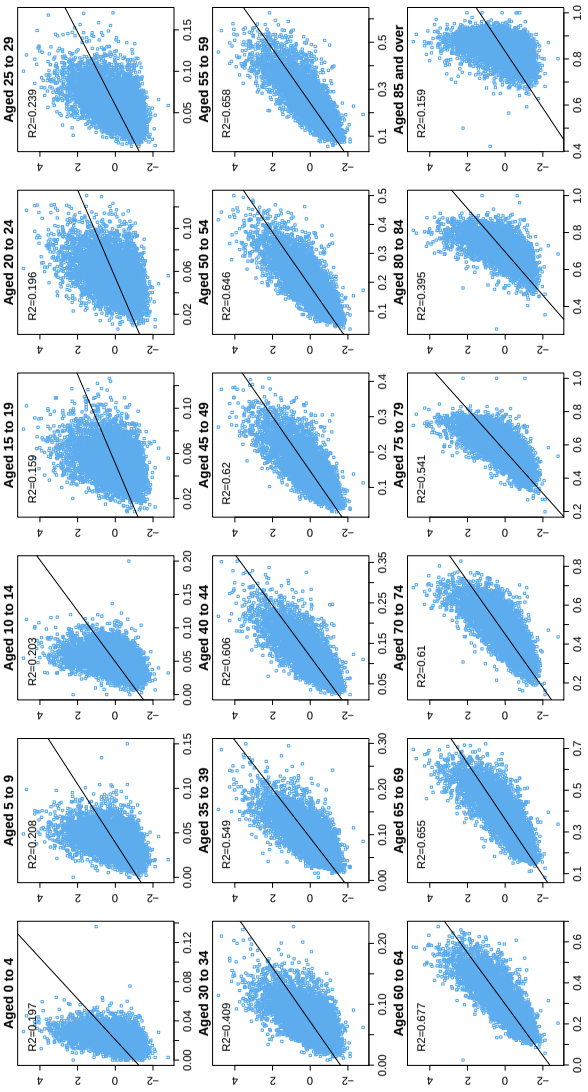


Figure A.1 | Plot of ISARs against prevalence of LLTI by 18 age groups (MSOA-level)

Appendix B

R syntax generating interaction matrices based on subnational migration flows

The below R syntax contains commands to download Office for National Statistics internal migrations datasets, process them and compute an interaction matrix of size 348 × 348 following methods ·.C*k* and ·.D*k* (see Table 6.1). The output can be used as a spatial weights matrix in spatial analyses.

```
### R code for migration-based adjacency matrices
### Example using 2009-2011 data from the Office for National Statistics
### Date:   08/10/2014
### Author: Peter Dutey-Magni
### Email:  p.dutey-magni@soton.ac.uk

#####################################################
### Downloading and unzipping
### between-district migration files
#####################################################

# Setting local working directory
setwd("C:/local/")
install.packages(c("maptools", "spdep", "sqldf", "RCurl"))
library(maptools)
library(sqldf)
library(spdep)
library(RCurl)

# 2009 data
```

Appendix B. R syntax generating interaction matrices based on subnational migration flows

```
temp <- tempfile()
download.file("http://www.ons.gov.uk/ons/rel/migration1/internal-migration-by-
local-authorities-in-england-and-wales/research-series--years-ending-june-2009
-to-june-2011/rft-detailed-estimates--part-1.zip", temp, mode = "wb")
migtemp <- unz(temp, filename = "Internal migration detailed estimates file -
research series year ending June 2009 - part 1 of 2.csv")
mig09 <- read.table(file = migtemp, sep = ",", colClasses = c(rep("character",
  2), rep("integer", 2), "numeric"), header = T)

download.file("http://www.ons.gov.uk/ons/rel/migration1/internal-migration-by-
local-authorities-in-england-and-wales/research-series--years-ending-june-2009
-to-june-2011/rft-detailed-estimates--part-2.zip", temp, mode = "wb")
migtemp <- unz(temp, filename = "Internal migration detailed estimates file -
research series year ending June 2009 - part 2 of 2.csv")
mig09 <- rbind(mig09, read.table(file = migtemp, sep = ",",
  colClasses = c(rep("character", 2), rep("integer", 2), "numeric"),
  header = T))

# 2010 data
download.file("http://www.ons.gov.uk/ons/rel/migration1/internal-migration-by-
local-authorities-in-england-and-wales/research-series--years-ending-june-2009
-to-june-2011/rft-detailed-estimates--june-2010-part-1.zip", temp, mode = "wb")
migtemp <- unz(temp, filename = "Internal migration detailed estimates file -
research series year ending June 2010 - part 1 of 2.csv")
mig10 <- read.table(file = migtemp, sep = ",", colClasses = c(rep("character",
  2), rep("integer", 2), "numeric"), header = T)

download.file("http://www.ons.gov.uk/ons/rel/migration1/internal-migration-by-
local-authorities-in-england-and-wales/research-series--years-ending-june-2009
-to-june-2011/rft-detailed-estimates--june-2010-part-2.zip", temp, mode = "wb")
migtemp <- unz(temp, filename = "Internal migration detailed estimates file -
research series year ending June 2010 - part 2 of 2.csv")
mig10 <- rbind(mig10, read.table(file = migtemp, sep = ",",
  colClasses = c(rep("character", 2), rep("integer", 2), "numeric"),
  header = T))

# 2011 data
download.file("http://www.ons.gov.uk/ons/rel/migration1/internal-migration-by-
local-authorities-in-england-and-wales/research-series--years-ending-june-2009
-to-june-2011/rft-detailed-estimates--june-2011-part-1.zip", temp, mode = "wb")
migtemp <- unz(temp, filename = "Internal migration detailed estimates file -
research series year ending June 2011 - part 1 of 2.csv")
mig11 <- read.table(file = migtemp, sep = ",",
  colClasses = c(rep("character", 2), rep("integer", 2), "numeric"),
```

```
  header = T)

download.file("http://www.ons.gov.uk/ons/rel/migration1/internal-migration-by-
local-authorities-in-england-and-wales/research-series--years-ending-june-2009
-to-june-2011/rft-detailed-estimates--june-2011-part-2.zip", temp, mode = "wb")
migtemp <- unz(temp, filename = "Internal migration detailed estimates file -
research series year ending June 2011 - part 2 of 2.csv")
mig11 <- rbind(mig11, read.table(file = migtemp, sep = ",",
  colClasses  =  c(rep("character", 2), rep("integer", 2), "numeric"),
  header = T))

# District population data
download.file("http://www.nomisweb.co.uk/api/v01/dataset/nm_144_1.bulk.csv?time
=latest&measures=20100&rural_urban=total&geography=TYPE464", temp, mode = "wb")
ks1010ew <- read.csv(file = temp, stringsAsFactors=F)
ks1010ew <- ks1010ew[ , c("geography.code", "geography",
  "Variable..All.usual.residents..measures..Value")]
names(ks1010ew) <- c("LAD13CD", "LAD13NM", "pop")

closeAllConnections()
rm(temp, migtemp)

# Pooling years together; calculating the annual mean migation flows
migall <- aggregate(formula  =  Flow ~ OutLA + InLA + Age + Sex,
                data     =  rbind(mig11, mig10, mig09),
                FUN      =  sum)
migall$Flow <- migall$Flow / 3

# Creating a new combined Origin district/Destination district migration code
migall$Comb <- as.factor(paste(migall$OutLA, migall$InLA))

# Removing cross-border migrations to/from Scotland and Northern Ireland
migall <- migall[grep("^[EW]", migall$InLA), ]
migall <- migall[grep("^[EW]", migall$OutLA), ]


#####################################################
### Amending disctrict ONS geography codes
### to reflect changes incurred in 2012
#####################################################

# District boundary changes came into force on 1 April 2013; see
# SI 2013 No. 596 The East Hertfordshire and Stevenage (Boundary Change) Order
#   2013;
```

Appendix B. R syntax generating interaction matrices based on subnational migration flows

```
# SI 2013 No. 595 The Gateshead and Northumberland (Boundary Change) Order
#   2013.
# For consistency, we restore 2012 ONS geography codes used in the migration
#   datasets.
ks1010ew$LAD12CD <- ks1010ew$LAD13CD
ks1010ew$LAD12CD[ks1010ew$LAD13CD == 'E07000242'] <- 'E07000097'
ks1010ew$LAD12CD[ks1010ew$LAD13CD == 'E07000243'] <- 'E07000101'
ks1010ew$LAD12CD[ks1010ew$LAD13CD == 'E08000037'] <- 'E08000020'
ks1010ew$LAD12CD[ks1010ew$LAD13CD == 'E06000057'] <- 'E06000048'

# Note: Any versions of the Nomis census table after 13 November 2015 may
# reflect ulterior boundary changes and therefore codes. The commands below
# check that local authority district codes match in both data sources.
# Boundary changes can be reviewed on the ONS Code History Database on the
# Geoportal:
# https://geoportal.statistics.gov.uk/geoportal/rest/find/document?
#    searchText="code history database"

# The result of this overall test must be TRUE
all(sort(unique(migall$InLA)) == sort(unique(ks1010ew$LAD12CD)))

# one-by-one test in case the overall test is not TRUE
data.frame("migall.code" = sort(unique(migall$InLA)),
  "pop.code" = sort(unique(ks1010ew$LAD12CD)),
  "match" = sort(unique(migall$InLA)) == sort(unique(ks1010ew$LAD12CD)))



#####################################################
### Plot: Examining the age profile
### of migrations in 2009, 2010 and 2011
#####################################################

c09 <- NULL
for (i in 0:112){
c09[i+1] <- sum(mig09$Flow[mig09$Age == i])/1000}
c10 <- NULL
for (i in 0:112){
c10[i+1] <- sum(mig10$Flow[mig10$Age == i])/1000}
c11 <- NULL
for (i in 0:112){
c11[i+1] <- sum(mig11$Flow[mig11$Age == i])/1000}

col <- c('plum2', 'violetred2', 'purple3')
cex <- .8
```

```
pch <- 16
lwd <- 2

# Intensity of annual intra-national migrations by age
# in England and Wales 2009-2011
plot(0:112, c09, pch = pch, cex = cex, col = col[1], ylim = c(0, 160),
     xlab = "Age",
     ylab = "Intra-national migallrants per annum (in thousands)",
     type  =  'l' , lwd = lwd)
abline(h = seq(0,200,10), col = "grey80")
abline(v = seq(0,120,5), col = "grey80")
points(0:112, c10, pch = pch, cex = cex, col = col[2], type = 'l',
       lwd = lwd)
points(0:112, c11, pch = pch, cex = cex, col = col[3], type =  'l',
       lwd = lwd)
legend("topright", legend = c("2009", "2010", "2011"), fill = col,
       bg = "white")


#####################################################
### Producing the adjacency matrix
#####################################################


# Adding census population estimates to the origin-destination data

mig <- sqldf("select d.*, c.pop as popOutLA
              from (select a.*, b.pop as popInLA
                    from migall a left join ks1010ew b
                    on a.InLA = b.LAD12CD) d
              left join ks1010ew c
              on d.OutLA = c.LAD12CD")

ODadjacency <- function(data = mig, popdata = ks1010ew[ , c("LAD12CD", "pop")],
                        minage = 0, maxage = 120,
                        nbneighbours = 3, ranked = T){
  # Computes an origin destination matrix of size 348 x 348 in which each cell
  # indicates either the rank (ranked = T) or the standardised weight
  # (ranked = F) of the pair of district, where the maximum number of pairs
  # is set by parameter nbneighbours.

  library(spdep)
  # Restricting the data to predefined age bounds
  # (for instance to exclude student/young people)
```

Appendix B. R syntax generating interaction matrices based on subnational migration flows

```
data <- data[data$Age >= minage & data$Age <= maxage, ]
data <- aggregate(formula = Flow ~ OutLA + InLA + Comb + popOutLA + popInLA,
                  data = data,
                  FUN = sum)
# Calculating flow intensity
# (relative to the population of the origin district)
data$rate <- data$Flow/data$popOutLA
# Ordering by district of destination (alphabetically),
# and by flow intensity (descending order) respectively
data <- data[order(data$InLA, -data$rate), ]

# By district of destination: giving weight 5 to the district
# of origin of the biggest flow of immigrants, 4 to the
# second biggest, ... 1 to the fifth biggest, and 0 to all the other district
data$rankneigh <- unlist(tapply(X=data$rate, INDEX = data$InLA,
                          FUN = function(x){
    if(length(x) > nbneighbours){
        return(c(nbneighbours:1, rep(0, length(x)-nbneighbours)))
    } else {
        return(c(nbneighbours:1)[1:length(x)])
    }}), use.names = F)

#Taking all pairwise combinations of LADs to prepare the neighbourhood matrix
LADs <- popdata[,1]
LADs <- LADs[order(LADs)]
LADs <- expand.grid(LADs, LADs, stringsAsFactors = F)
names(LADs) <- c("OutLA", "InLA")
LADs <- LADs[order(LADs[,1], LADs[,2]), ]

# Producing the final matrix
LADs <- merge(LADs, data[,c("OutLA", "InLA", "rankneigh")],
              by = c("OutLA", "InLA"), all.x = T )
# removing the missing ranks (occuring either because no migration or data
# issue, we assume their rank is zero because these is typically small LAs)
LADs$rankneigh[is.na(LADs$rankneigh)] <- 0

# Generating a dichotomous matrix
LADs$binw <- 0
LADs$binw[LADs$rankneigh>0] <- 1

# Returning the required matrix
if(ranked == F){
    message("Returning a binary adjacency matrix")
    matbin <- matrix(LADs$binw, length(unique(popdata[,1])), byrow = F)
```

```
      row.names(matbin) <- as.character(unique(LADs$InLA))
      return(mat2listw(matbin, style = "W"))
  } else {
      message("Returning a ranked adjacency matrix")
      matweights <- matrix(LADs$rankneigh, 348, 348, byrow = F)
      dimnames(matweights) <- list(as.character(unique(LADs$InLA)),
                                   as.character( unique(LADs$OutLA)))
      row.names(matweights) <- as.character(unique(LADs$InLA))
      return(mat2listw(matweights, style = "W"))
  }
}

test <- ODadjacency(minage = 0, nbneighbours = 3, ranked = T)
test25 <- ODadjacency(minage = 25, nbneighbours = 10, ranked = T)


#####################################################
### Mapping the resulting adjacency structures
#####################################################

# Downloading district boundary shapefile from the ONS geoportal

bin <- getBinaryURL("https://geoportal.statistics.gov.uk/Docs/Boundaries/Local_
authority_district_(GB)_2013_Boundaries_(Generalised_Clipped).zip",
  ssl.verifypeer=FALSE)
con <- file("LADboundary.zip", open = "wb")
writeBin(bin, con)
close(con)
rm(bin, con)
unzip("LADboundary.zip")

# Loading shapefile
shp <- readShapePoly("LAD_DEC_2013_GB_BGC.shp")
# Removing Scotland
shp <- shp[grep("^[EW]",shp@data$LAD13CD), ]
# Recovering 2012 ONS codes
shp@data$LAD12CD <- as.character(shp@data$LAD13CD)
shp@data$LAD12CD[shp@data$LAD13CD == 'E07000242'] <- 'E07000097'
shp@data$LAD12CD[shp@data$LAD13CD == 'E07000243'] <- 'E07000101'
shp@data$LAD12CD[shp@data$LAD13CD == 'E08000037'] <- 'E08000020'
shp@data$LAD12CD[shp@data$LAD13CD == 'E06000057'] <- 'E06000048'

# Must ensure districts in the shapefile are alphabetically ordered by ONS code
shp <- shp[order(shp@data$LAD12CD), ]
```

```
pdf(file = "checkcontiguitymat.pdf" )
par(omi = c(0,0,0,0))
plot(shp,border = 'grey40', lwd = 0.2)
plot(test, coordinates(shp), lwd = 0.1, add = T, col = 2, pch = ".",
     cex = .000000000001 )
dev.off()
```

Algorithm 3 | R syntax generating interaction matrices based on subnational migration flows

Appendix C

Sensitivity testing of age cut-offs for migration interaction matrices

This section contains outputs of Algorithm 3 and reports on sensitivity analysis performed in order to ensure that the intensity of sub-national migration flows for younger residents does not have an excessive influence on the way local authorities are matched with one another to form a neighbourhood matrix based on the number of residents exchanged. Figure C.1 below illustrates the spike of migration for individuals aged 16 to 34 years.

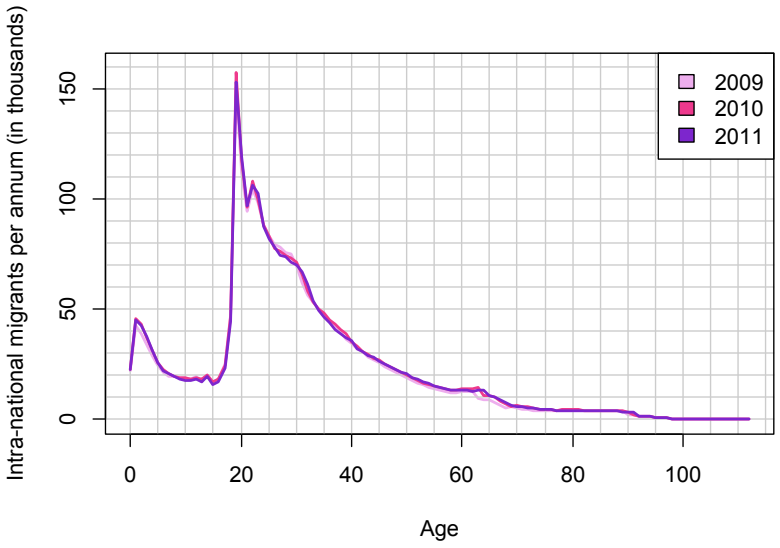


Figure C.1 | Intensity of annual intra-national migrations by age in England and Wales 2009–11 (source: ONS 2013h)

In order to examine how the migration peak influences the ranking of pairwise migration intensity between districts, we generated these ranking separately for residents aged 30+, then 25+, 20+ and finally 15+ years. Results are reproduced in

Appendix C. Sensitivity testing of age cut-offs for migration interaction matrices

Table C.1 below for the City of Manchester, the City of Nottingham and the London Borough of Barnet. After manual reviews of those matrices we concluded that changing the minimum age only has a moderate effect on the ranking of neighbours.

Table C.1 | Sensitivity analyses for neighbours allocated to two metropolitan councils and one city council LADs under different age scenarios

Lower age bound	30 years	25 years	20 years	15 years
		Manchester		
1st neighbour	Salford	Salford	Salford	Trafford
2nd neighbour	Trafford	Trafford	Trafford	Salford
3rd neighbour	Stockport	Stockport	Stockport	Stockport
4th neighbour	Oldham	Oldham	Bury	Bury
5th neighbour	Bury	Bury	Tameside	Tameside
		Barnet		
1st neighbour	Camden	Camden	Camden	Camden
2nd neighbour	Haringey	Haringey	Haringey	Haringey
3rd neighbour	Enfield	Brent	Brent	Brent
4th neighbour	Brent	Enfield	Enfield	Enfield
5th neighbour	Hertsmere	Hertsmere	Harrow	Hertsmere
		Nottingham		
1st neighbour	Rugby	Rugby	Erewash	Erewash
2nd neighbour	Erewash	Erewash	Rugby	Rugby
3rd neighbour	N. Warwickshire	N. Warwickshire	N. Warwickshire	N. Warwickshire
4th neighbour	Malvern Hills	Lancaster	Lancaster	Lancaster
5th neighbour	Lancaster	Malvern Hills	Derby	Craven

Appendix D

Procedure to compute auxiliary UK population data broken down by age and ethnicity

Annual population estimates are for all usual residents, including residents of communal establishments (CEs).To approximate auxiliary data for our study population (residents of private households), three 2011 census tables (ONS, 2013b) were used to calculate the population headcounts exposed to risk by age and ethnicity. On the one hand, tables DC1104EW and DC1117EW were used to produce breakdowns of population headcounts by single year of age. On the other hand, population headcounts by age (9 groups) and ethnicity (5 groups) was computed by subtracting CE residents in table DC2117EW to all residents enumerated in table at the LAD level. This was more difficult to achieve for the MSOA level. This is because census table DC2117EW (Ethnic group by sex by age - Communal establishment residents) is not published for super output areas out of concern for respondent confidentiality. We approximately reconstructed this auxiliary data from 2011 census tables by multiplying the total population figures by age and ethnicity for MSOAs (table DC2101EW) by the proportion of residents of communal establishments estimated

using the formula below as per Bayes' theorem:

$$\begin{aligned}
&\Pr(\text{residing in a private household}|\text{ethnicity} = j) \\
&\quad = 1 - \Pr(\text{residing in a private household}|\text{ethnicity} = j) \\
&\quad = 1 - \Pr(\text{ethnicity} = j|\text{residing in a private household}) \\
&\qquad \times \Pr(\text{residing in a private household}) \\
&\qquad \div \Pr(\text{ethnicity} = j)
\end{aligned} \tag{D.1}$$

The proportions are only approximations because although the last two factors of equation (D.1) can be directly calculated from tables DC1104EW and DC2101EW respectively, the first factor is taken from table DC2117EW under the assumption that all MSOAs in a given LAD have the same ethnic breakdown for residents of communal establishments. They are computed for every age group in each MSOA separately.