Drivers of Mortality: Risk Factors and Inequality

Andrew J.G. Cairns[a], [b] Torsten Kleinow[a] and Jie Wen[a]

This version: December 14, 2019

To be presented to the Institute and Faculty of Actuaries, London, 6 January 2020.

Supplementary plots and data can be found on the research programme website:

`www.macs.hw.ac.uk/~andrewc/ARCresources`

Abstract

This paper takes a detailed look at socio-economic variation in mortality across England. All-cause mortality is analysed at a neighbourhood level (Lower Layer Super Output Areas) with mortality rates linked to a number of socio-economic predictive variables that determine the character of a neighbourhood. We find that income and employment deprivation are key determinants of mortality, but also that urban-rural class and the presence of care homes in a neighbourhood have an important role to play in assessing underlying mortality rates relative to national mortality. Residual spatial/regional variation in mortality is found to be much less important than socio-economic variation. Based on these results, we propose the $LIFE$ index (Longevity Index For England) and $LIFE$ deciles as a new tool for assessing all-cause mortality.

In the second part of the paper we look at cause-of-death mortality by region and income-deprivation decile. By working with 34 causes of death we are able to gain significant further insights into mortality inequalities and how these are linked to controllable risk factors such as smoking, poor diet, lack of exercise and alcohol consumption.

Keywords: Mortality inequality; Lower Layer Super Output Area; $LIFE$ indices; Regional mortality variation; Predictive variables; Local linear regression; Cause of death; Controllable risk factors.

[a]Maxwell Institute for Mathematical Sciences, and Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh, EH14 4AS, UK.

[b]Corresponding author: A.J.G.Cairns@hw.ac.uk

1 Introduction

1.1 Socio-economic differences in mortality

It is well known that socio-economic status and associated variables are strongly correlated with high and low mortality. The evidence base for this has been growing over a number of decades as data become available with sufficient detail to be able to investigate the dependency between particular predictive variables and mortality. Socio-economic status is not always easily available in the same form but typical information about individuals or groups includes income, income deprivation or affluence (e.g. Chetty et al., 2016 [US], Villegas and Haberman, 2014, Longevity Science Panel, 2018 [UK], Cairns et al., 2019 [Denmark], Wen et al., 2019, 2020 [UK, Canada]) and education (e.g. Mackenbach et al., 2003, 2016).

This paper seeks to delve more deeply into the the links between different measures of socio-economic status, geographic location and mortality. We do this by exploiting a large dataset for England built up from multiple datasets obtained from the UK's Office for National Statistics (ONS).

A key modelling tool in the section on all-cause mortality is *local linear regression*. As a statistical method, a version was first proposed by Cleveland (1979) and is often referred to as LOWESS (Locally Weighted Scatterplot Smoothing). As we demonstrate in this paper, it is an effective tool to handle large datasets with many covariates. Compared to e.g. Generalised Linear Models (see, for example, Macdonald et al., 2018) the model fits the data effectively without having to specify in advance the functional relationship between predictive variables (including interactions) and the outcome.

1.2 Questions addressed in this paper

This paper contains a wide ranging analysis of all-cause and cause-of-death data as part of this we seek to address the following questions:

- What are the most significant socio-economic factors that influence mortality rates?
- What other factors push mortality rates up or down at the level of small geographical or regional level?
- Does it make a difference if a neighbourhood is in an urban or rural area?
- Where are care homes located and what impact do they have on mortality?
- Can regional differences in mortality be explained entirely by differences in the socio-economic mix and other non-spatial predictive variables?

- After socio-economic and non-spatial effects have been filtered out, what remains in terms of spatial or regional variation in mortality across England?
- How much inequality is there in mortality rates at different ages?
- Have mortality inequalities been widening over the last 16 years?
- What is the difference between controllable and non-controllable risk factors?
- Which causes of death have significant controllable risk factors?
- Which causes of death have significant levels of mortality inequality?
- What are the contributors to the slowdown in mortality improvements since 2011?

In answering some of these questions, we develop a new index: the Longevity Index For England, or $LIFE$. $LIFE$ is, in fact, a group of indices based on different age groups. It can be adopted as a continuous index that actuaries and others can use as a predictive variable alongside other variables that are available at the individual level such as pension level and geodemographic profiling (e.g. Richards, 2008). Additionally, it can be used to divide the population up into deciles instead of using the Index of Multiple Deprivation or income deprivation.

1.3 Outline of the paper

In Section 2 we introduce the several datasets that we will use in our various analyses, with further details in Appendix A. Section 3 sets the scene for what is to come with some introductory analysis of the data. This also introduces the problem of regional variation in mortality. We then take a deep dive into all-cause mortality at the small-neighbourhood level in Section 4, followed by cause-of-death mortality by region and income-deprivation decile in Section 5. Section 6 concludes.

2 Data

Data used in this study have been sourced from the UK's Office for National Statistics (ONS). Data are available at either the regional level in England (cause of death data) or by Lower Layer Super Output Areas (LSOA; small, socially-homogeneous, geographical areas with, on average, 1500 people) and are of the following types:

- mid-year population data by LSOA;
- all-cause death counts by LSOA;

- death counts by cause of death by region and income-deprivation decile;
- predictive variables by LSOA.

We outline key elements of the data below, with further details to be found in Appendix A.

2.1 Geographical data

Data are available geographically at three levels:

- English national level;
- region level (9 regions; see Figure 1);
- LSOA level (32844 areas).

2.2 Population data

For each LSOA, we have mid-year population estimates from the ONS by single year of age and single calendar years from 2001-2016, $E(g, i, t, x)$, where i is the LSOA index, g is the gender, t is the calendar year, and x is the age last birthday. As with standard practice we equate this with the central exposed to risk for age x last birthday across the whole of year t.[1]

Exposures can be aggregated into specified deciles, regions or the whole of England as required.

2.3 All cause mortality

For all cause mortality, the ONS make available death counts by gender, LSOA, single age and single year, $D(g, i, t, x)$.

Death rates at the LSOA level are then

$$m_L(g, i, t, x) = \frac{D(g, i, t, x)}{E(g, i, t, x)}$$

and at the regional level for region r

$$m_R(g, r, t, x) = \frac{\sum_{i \in R(r)} D(g, i, t, x)}{\sum_{i \in R(r)} E(g, i, t, x)}$$

[1] We have not adjusted exposures at this stage as proposed by Cairns et al. (2016). However, even if systematic, cohort-based errors exist at the total population level it is unlikely that this would have an impact on our assessments of relative risk that are the focus of this paper.

Figure 1: Map showing the nine English regions, plus Northern Ireland, Scotland and Wales. Source: www.democraticaudit.com.

where the set $R(r)$ contains all LSOA's, i, that lie within region r.

Similar expressions can be written down for decile-based death rates, $m_D(g, j, t, x)$, for decile j, and national mortality, $m(g, t, x)$.

2.4 Cause of death mortality

For this study, we obtained a customised dataset from the ONS. Apart from the addition of the cause of death, these data were less detailed than the all-cause mortality data, and, as the data contain sensitive information, they are not available publicly. Death counts, $D(r, i, c, g, t, y)$, were divided by

- r: region;
- i: income deprivation decile;
- c: cause of death according to Table 1;

1	Infectious diseases	2	Cancer: oesophagus, stomach
3	Cancer: bowel, gut	4	Cancer: larynx
5	Cancer: trachea	6	Cancer: lung, bronchus
7	Cancer: breast	8	Cancer: uterus, cervix
9	Cancer: ovary	10	Cancer: other female genital
11	Cancer: prostate, testicular	12	Cancer: other male genital
13	Cancer: skin, bones and certain organs	14	Cancer: lymphatic
15	Benign tumours	16	Diseases: blood
17	Diabetes	18	Mental illness
19	Diseases of nervous system excl. Alzh.	20	Alzheimers
21	Blood pressure + rheumatic fever	22	Ischaemic heart diseases
23	Other heart diseases	24	Cerebrovascular diseases
25	Circulatory diseases	26	Lungs, breathing diseases
27	Digestive diseases (excl. alcohol)	28	Alcohol-related liver disease (*)
29	Diseases: skin, bone, tissue	30	Urine, kidney and related diseases
31	Road/other accidents	32	Other causes of death
33	Suicide (*)	34	Accidental poisonings (*)

Table 1: Cause of death groups used in the study. (*) identifies causes of death that have been categorised as deaths of despair.

- g: gender;
- t: calendar year;
- y: 5-year age groups 20-24, 25-29, ..., 85-89, 90+.

LSOA-based exposures were used to construct exposures, $E(r, i, g, t, x)$ that matched the cause-of-death groupings.[2]

We deliberately chose a detailed list of causes of death (Table 1) rather than the more usual large groups (e.g. all cancers, cardiovascular, etc.). The reason for this is that we wish to explore the links between controllable risk factors and inequality. In particular, some causes of death have well known controllable risk factors while others do not, and we wish to be able to differentiate between the two. We also want to investigate *deaths of despair* (marked with a (*) in Table 1) as highlighted by Case and Deaton (2015).[3]

[2]Note that $E(r, i, g, t, x)$ has no cause of death index, c. Instead the same $E(r, i, g, t, x)$ applies for each $D(r, i, c, g, t, x)$ for $c = 1, \ldots, 34$.

[3]Case and Deaton (2015) identified that so-called deaths of despair were on the rise amongst middle-aged, white non-hispanic, working-class Americans. Deaths of despair include deaths from suicide, deaths resulting from excessive alcohol consumption, and deaths resulting from an accidental overdose of prescription drugs. The latter has given rise to the concept of the *opiod epidemic*.

2.5 Migration

It needs to be noted that the exposures data do not contain information about migration between LSOAs or between regions[4]. This means that caution needs to be exercised when viewing death rates across age groups as there might be some variation in the underlying population mix.

2.6 Predictive variables

The focus of this paper is to identify socio-economic and other variables that have significant predictive power in estimating mortality rates.

Starting with the Index of Multiple Deprivation (IMD) the following variables were available from the ONS or related government websites, all with individual or multiple values for each LSOA. Further details of these can be found in Appendix A.

- IMD: Domains of the IMD:
  - ID: income deprivation
    - IDO: income deprivation amongst the old
  - EMP: employment deprivation
  - EDU: education deprivation
  - CR: crime
  - barriers to housing and services
    - WID: wider barriers (housing affordability, overcrowding and homelessness)
    - GEO: geographical barriers (distance to key services)
  - LIV: living environment
- BED: Average number of bedrooms
- EDL+: Education level
- OCC+: Occupation group
- UKB: Proportion born in the UK
- COM: Communal establishment populations
  - COM1: Care home population with nursing

[4]However, it is possible to assess net migration by comparing changes in exposures by cohort with death counts but not determine where individuals are coming from or going to.

  - COM0: Care home population without nursing

- UR: Urban-rural classification
  - UR1: Conurbation: non London
  - UR2: City or town
  - UR3: Rural town
  - UR4: Rural village and dispersed
  - UR5: Conurbation: London

- REG: Region

(+) Education level and occupation group have many levels and age groups and therefore have many sub-indices rather than just one.

3 Preliminary analysis of all-cause mortality

Before we move into a more detailed analysis on the various datasets, it is worthwhile summarising the main patterns that exist in socio-economic data. Building on the recommendations of the Longevity Science Panel (LSP 2018) we will, here, use income deprivation rather than the main IMD to divide the LSOAs into deciles. Income deprivation is, of course, not a causal effect, but it was found by LSP to be more strongly *correlated* with high and low mortality than other domains of the IMD.[5]

In Figure 2 we have plotted age-specific death rates in 2016 for each of the income-deprivation deciles. For both males and females we can make the following observations:[6]

- Even though the data contain a certain amount of sampling variation, the deciles are clearly ranked from decile 1 (most deprived, highest mortality) down to decile 10 (least deprived, lowest mortality).

- There are very significant differences between the deciles at ages 40 to 60 (the mortality inequality gap) before gradually narrowing as the population gets older. This narrowing is very typical for mortality differences between socio-economic groups using different measures and in different countries (see, for

[5]Note that the IMD itself includes a health domain. As with LSP (2018) we do not consider the health domain. Instead, we seek to identify socio-economic covariates that are predictors of poor health and increased mortality.

[6]For a more detailed analysis of English deciles, see Wen et al. (2019).

example, Cairns et al., 2019, Wen et al., 2020, Redondo Loures and Cairns, 2019).

At ages 40 to 50, death rates in the most deprived group are around 4 times (males) and 3 times (females) the corresponding death rates in the least deprived group. Both narrow to about 1.4 times at age 89.

The importance of income deprivation is investigated further in Figure 3. In this figure we consider centiles rather than deciles, and use Age Standardised Mortality Rates (ASMRs) over ages 65 to 89 rather than crude age-specific death rates to dampen sampling variation.[7]

We see that income deprivation has a clear impact across all centiles, and, indeed, steepens towards both edges.

- At the right hand end (high deprivation and high mortality, the steeper curve might reflect the possibility that prolonged ill-health drives some people into more deprived areas.
- At the left hand end for the least deprived (which we interpret as most affluent), the rationale for a steepening curve is less clear, although (speculating) it might be that higher levels of wealth might facilitate better health care in old age. This pattern at the level of centiles is also very evident in the US (Chetty et al., 2016).

[7]For an overview of ASMRs, see Appendix C.

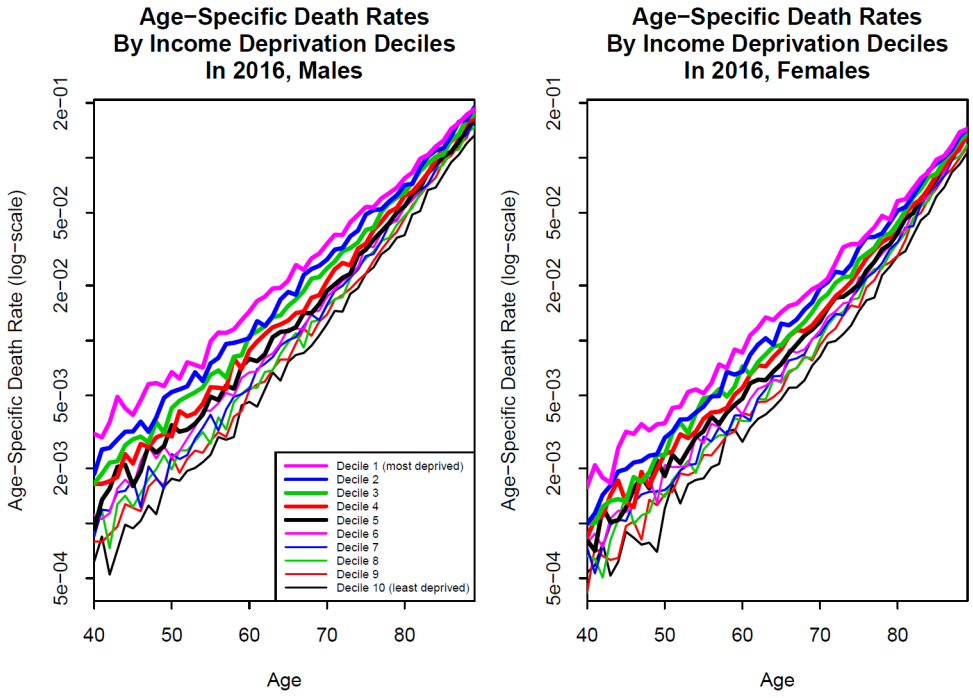

Figure 2: Age-specific death rates for English males and females in 2016, by income-deprivation deciles.

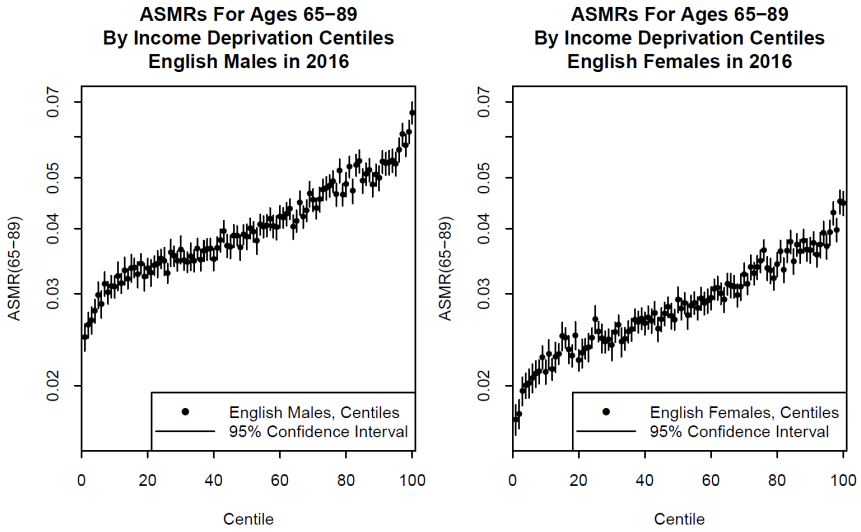

Figure 3: Estimated ASMRs by income-deprivation centile for ages 65 to 89. Centile 1: least deprived. Centile 100: most deprived. Bars show approximate 95% confidence intervals for ASMR estimates.

In Figure 4, we illustrate how the ASMRs for the deciles have changed over time for two age groups: 40 to 64 and 65 to 89. These plots reveal the following features:

- Similar to Figure 2, the gap in ASMRs between groups 1 and 10 is wider in the upper plots for younger ages and that females' mortality is lower.
- We can see a widening gap between groups 1 and 10. This is more marked for females and for the older age group.
- Mortality improvements can be seen to have slowed down since 2010 or 2011 in all four sub-plots. However, assessing the impact on deciles 1 to 10 needs care. The slowdown is more marked in the older age group. And it is also more marked in the most deprived groups, even after taking account of the fact that they had been experiencing slower improvements since 2001.

A clear takeaway from Figure 4 concerns the setting of future mortality improvement rates in actuarial valuations. Specifically, short-term improvement rates should be different for different socio-economic groups, with higher short-term improvement rates from 2017 for the least deprived groups.

Lastly, we consider how mortality rates vary from region to region across England. In Table 2 we give the ratio of actual versus expected deaths by region using English national mortality for expected deaths. Corresponding values by income-deprivation decile are given in Table 3 for comparison.

We can see significant differences between regions leading to the well-publicised North/South divide. But we can also see that differences by region are dwarfed by differences between income-deprivation deciles.

This leads to a key objective of this paper:
to what extent are regional differences explained by socio-economic differences between the regions?

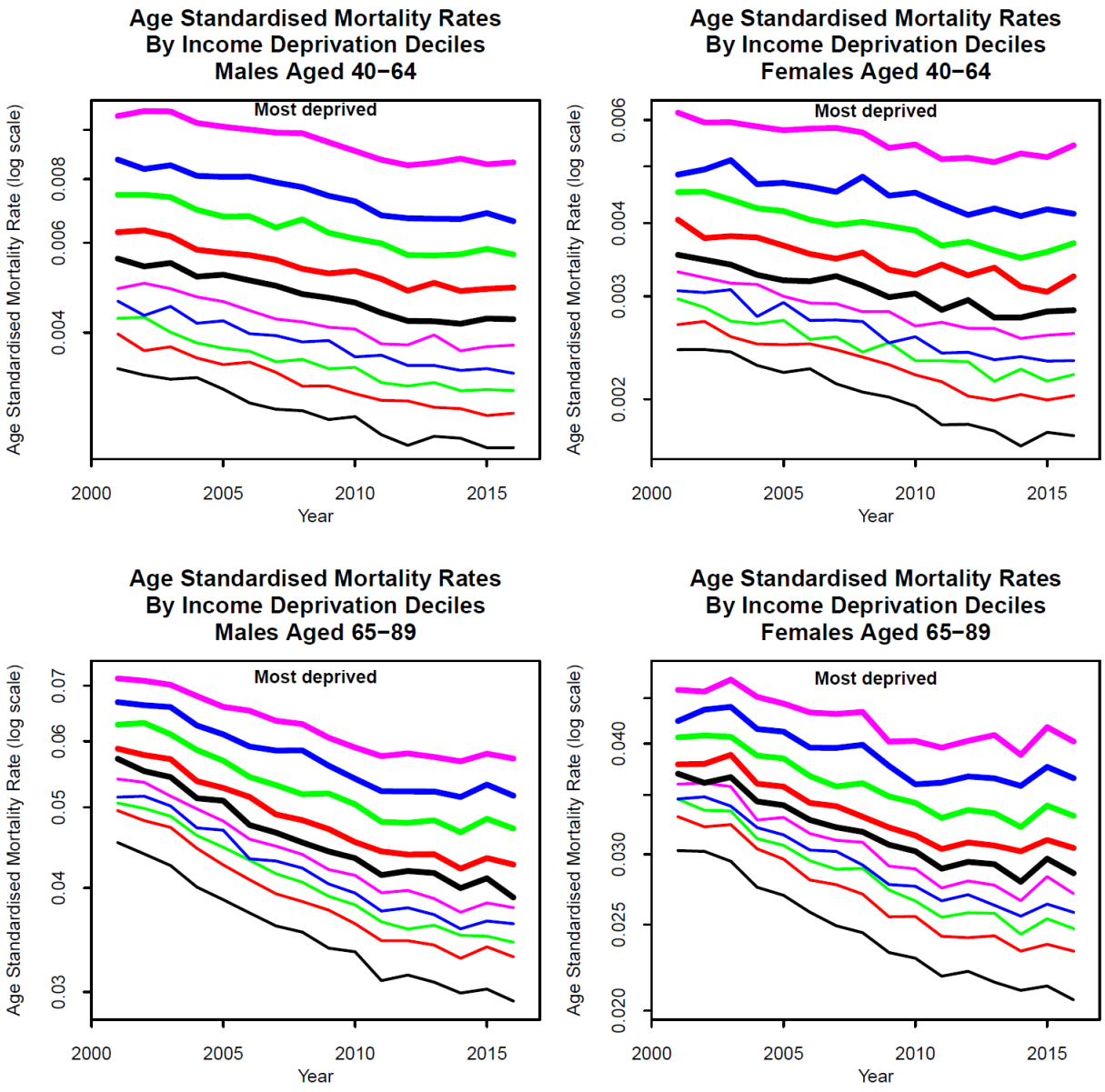

Figure 4: Estimated ASMRs by income-deprivation deciles over the period 2001 to 2016 for males and females and for age groups 40 to 64 (top) and 65 to 89 (bottom).

Region	Regional Relative To National Mortality Males	Females
North East	117.7	118.7
North West	116.2	118.3
Yorkshire and The Humber	107.1	107.7
East Midlands	98.5	101.3
West Midlands	105.3	102.9
East	87.5	89.6
London	104.7	98.2
South East	88.6	89.2
South West	86.5	86.4

Table 2: Comparison of regional death counts over the age range 60 to 69, and 2001-2015 versus expected deaths using English national mortality.

Income Deprivation Decile	Decile Relative To National Mortality Males	Females
1	174.9	167.2
2	146.7	141.6
3	127.3	124.5
4	113.1	111.5
5	101.1	101.0
6	90.5	90.9
7	82.4	84.6
8	76.7	79.1
9	69.9	73.3
10	61.9	65.2

Table 3: Comparison of mortality rates for individual income-deprivation deciles over the age range 60 to 69, and 2001-2015 versus English national mortality.

4 All-cause mortality by LSOA

This section will consider *relative risk*: a scaling factor that can be applied to standard mortality rates that takes account of specified predictive variables.

For each LSOA we have, as outlined previously:

- deaths and exposures, $D(i,t,x)$ and $E(i,t,x)$;
- a vector of predictive variables, $X(i)$.

Our challenges are:

- Which predictive variables are the best at predicting high or low mortality, and at which ages?
- How many predictive variables do we need to get a reasonable model?
- Can we identify an effective approach to estimate relative risks using these predictive variables?

We will first outline a new, non-parametric method for assessing the relative risk for a specific LSOA based on the method of *local linear regression*. We will then use the method to discuss which covariates are most useful, and what purpose they play. And finally we will look at the results, including a look at residual regional effects.

4.1 Local linear regression

We begin with a stylised example, not based on mortality, to illustrate the method. In Figure 5, we show an example in which there is a single predictive variable, X, and a set of observations (grey dots) that are of the form $Y = f(X) + \epsilon$ where ϵ is an error term and $f(x)$ is a non-linear function. We seek to estimate $f(x)$. Because $f(x)$ is non-linear, standard linear regression, or a standard Generalised Linear Model, will not deliver a good fit.

- We observe $(X(i), Y(i))$ for $i = 1, \ldots, n$.
- We seek to estimate $\hat{Y}(i) = E[Y(i)|X(i)] = f(X(i))$.
- For each $i = 1, \ldots, n$ we seek to minimise, over $a \equiv a(i)$ and $b \equiv b(i)$, the weighted least squares function:

$$S_i(a,b) = \sum_{j=1}^{n} w(i,j) \left(Y(j) - (a + bX(j))\right)^2$$

where, for each i, the weights $w(i,j)$ tend to zero the further $X(j)$ is from $X(i)$.[8] In other words, we fit a straight line through points near $X(i)$.

In Figure 5 (top), the size of the black dots illustrate how the weights vary around $X(i)$, larger dots represent higher weights.

- Minimisation gives us local parameter estimates $\hat{a}(i)$ and $\hat{b}(i)$, and our local estimator is
$$\hat{Y}(i) = f(X(i)) = \hat{a}(i) + \hat{b}(i)X(i).$$
We only make use of this single point estimate and discard other information about a and b.

  In Figure 5 (top), the straight red line shows the result of the local linear regression around $(X(i), Y(i))$.

- We then move on to the next i and repeat the local linear regression using different weights.

  In Figure 5 (bottom), the red dots illustrate the complete set of point estimates for $i = 1, \ldots, n$.

- For comparison Figure 5 (bottom) also shows estimates (blue dots) using kernel smoothing (each dot being a local weighted average using the same weight function). Local linear regression and kernel smoothing give quite similar results in the main body of the data, but kernel smoothing can be seen to perform poorly near the edges because the data have a clear underlying slope, and because near the edges of the data, most of a point's neighbours are towards the middle of the data.

4.2 Advantages of local linear regression

A key reason for using local linear regression rather than, for example, kernel smoothing is that it captures the local slope in the data (see the last bullet in Section 4.1). This is particularly the case where the response variable (e.g. the relative risk) is believed to be increasing or decreasing as the underlying predictive variables change. This is particularly useful for any data point i whose socio-economic neighbours are mostly to one side rather than evenly distributed round about: e.g. near the edges of the data on the far left or right of Figure 5.

Local linear regression offers an approach that can be easily extended to multiple dimensions. This is in contrast to some alternatives such as P-splines[9] which can be numerically very intensive or challenging as the number of dimensions increases.

[8] Additionally, if we set $w(i,i) = 0$ (exclude data from LSOA i itself), then each observation $D(i)$ is out-of sample, allowing us to implement an out-of-sample analysis.

[9] With P-splines the number of basis functions grows exponentially with the number of dimensions (i.e. the number of predictive variables).

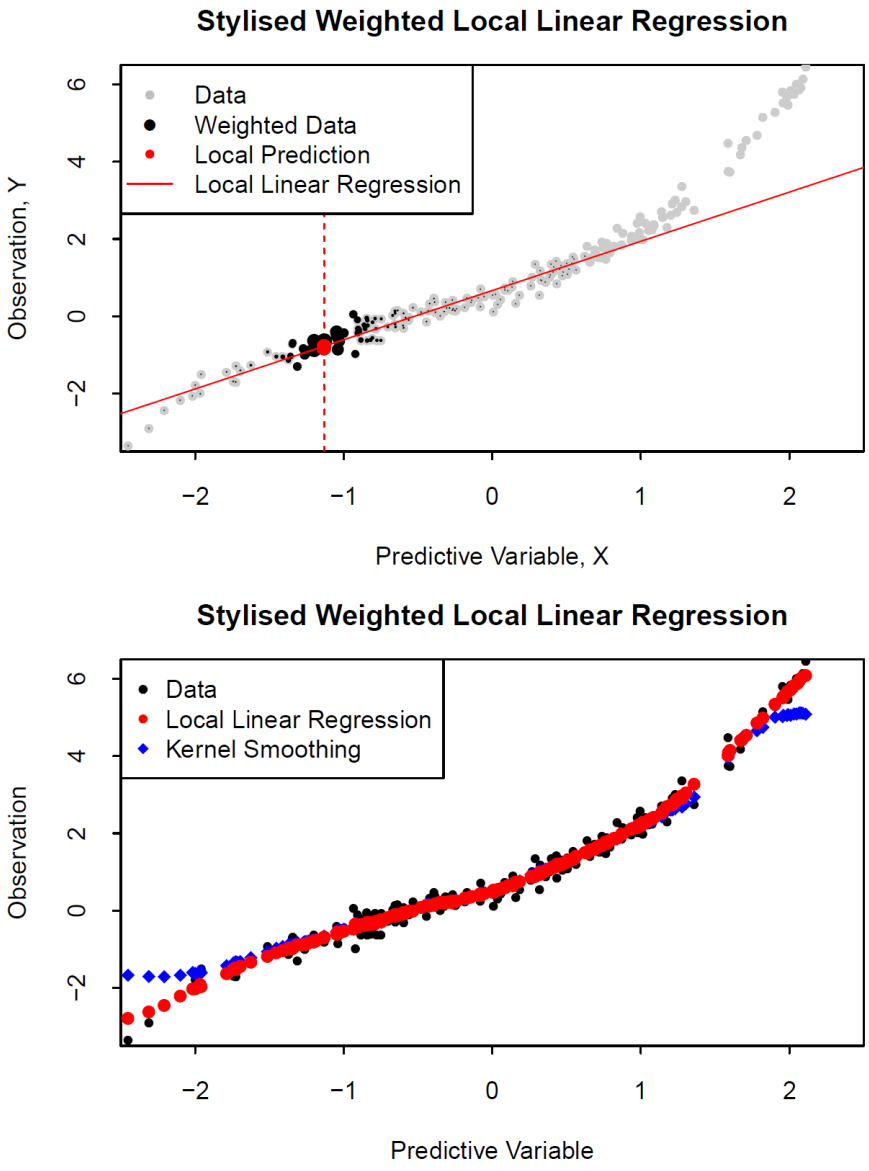

Figure 5: Stylised example of a local linear regression with a single predictive variable. Top: A single local linear regression with the size of black dots representing the weight attached to each point. Bottom: red dots show the results of all of the individual local linear regressions. Blue dots show the equivalent results using simpler kernel smoothing.

An additional advantage of local linear regression is that estimates of relative risk are not especially sensitive to non-linear transformations of the predictive variable, X (e.g. a log-transform). The exception to this might be at the edges of the data when those points with an appreciable weight, $w(i, j)$, are spread out over a wider range of values of X.[10] This contrasts, e.g. with Generalised Linear Models (GLMs), where, by design, the relative risk is log-linear in X: so X needs to be scaled and transformed accordingly before model fitting to get a good fit.

Finally, local linear regression automatically captures interactions between different predictive variables in multiple dimensions. With GLMs, interactions must be investigated in a systematic way, that could be very challenging computationally in multiple dimensions.

[10] However, most methods will be less reliable at the edges of the data.

4.3 Local linear regression and socio-economic mortality

We are now going to move from the stylised example to the mortality setting.

Our general approach is as follows:

First define $m(t,x)$ to be the *crude* age-specific national death rate in year t, age x. We now seek to model the *underlying* LSOA-specific death rates, $m(i,t,x)$, in LSOA i linking to a vector, $X(i)$, of predictive variables. The death rate satisfies the usual assumption that

$$E[D(i,t,x)] = m(i,t,x)E(i,t,x).$$

Our general model is then that, *over the limited age range* (x_0, x_1) *and range of years* (t_0, t_1),

$$m(i,t,x) = m(t,x)F_1(i)F_2(i)$$

where

- $F_1(i)$ is the relative risk due to the socio-economic (and other) characteristics of LSOA i;
- $F_2(i)$ is a relative risk that captures residual spatial effects once the $F_1(i)$ have been fitted.

Having a constant $F_1(i)$ and $F_2(i)$ over a range of ages and years then means that if we define

$$D(i) = \sum_{t=t_0}^{t_1} \sum_{x=x_0}^{x_1} D(i,t,x)$$

and

$$\hat{D}(i) = \sum_{t=t_0}^{t_1} \sum_{x=x_0}^{x_1} E[D(i,t,x)] = \sum_{t=t_0}^{t_1} \sum_{x=x_0}^{x_1} m(i,t,x)E(i,t,x)$$

then

$$E[D(i)] = F_1(i)F_2(i)\hat{D}_0(i)$$

where

$$\hat{D}_0(i) = \sum_{t=t_0}^{t_1} \sum_{x=x_0}^{x_1} m(t,x)E(i,t,x).$$

In taking this approach,

- $\hat{D}_0(i)$ represents the baseline expected deaths with no allowance for socio-economic or other effects.
- $\hat{D}_0(i)F_1(i)$ represents our best estimate of the expected deaths based on socio-economic drivers alone.
- $\hat{D}_0(i)F_1(i)F_2(i)$ represents our best estimate of the expected deaths based on soecio-economic and spatial drivers.

4.4 Stage 1: estimate the socio-economic relative risk, $F_1(i)$

To fit this model we, first, calculate the actual-over-expected based on the baseline model that there are no socio-economic effects:

$$R_0(i) = D(i)/\hat{D}_0(i).$$

We then use the $R_0(i)$ to derive estimates of the $F_1(i)$ making use of the socio-economic predictive variable $X(i)$.

Unadjusted predictive variables take different ranges: some are on the scale $(0, 1)$, some $(0, 100)$, some $(-\infty, \infty)$. To aid comparison and computation of the local-linear-regression weights, the predictive variables are (unless stated otherwise) standardised as follows:

- Suppose that the LSOAs are indexed by $i = 1, \ldots, L$, and that the predictive variables are indexed by $j = 1, \ldots, n_P$.
- Let $P(i, j)$ be the unadjusted predictive variable.
- Define $X(i, j) = (P(i, j) - \mu_j)/\sigma_j$ where μ_j and σ_j are the empirical mean and standard deviation of $P(1, j), \ldots, P(L, j)$.

Hence, the $X(i, j)$ all have mean 0 and variance 1.

Variables that we chose not to standardise include

- Urban-Rural classification and region (which are categorical rather than ordinal variables);
- The proportions in communal establishments (including care homes with and without nursing) in a particular LSOA (which remain as proportions on a $(0, 1)$ scale).[11]

In the setting where we intend to use several predictive variables (so a vector rather than a scalar X), the idea behind local linear regression of the $R_0(i)$ versus $X(i)$ is that we fit a local n_P-dimensional sheet around $X(i)$ to obtain an estimator for $F_1(i)$. This is achieved as follows:
For each LSOA, i, fit the linear (in the vector x) function

$$F(i, \mathbf{x}) = a(i) + b(i)^T x$$

[11] From a statistical perspective, and in this analysis, care homes can be considered to be nuisance parameters. Specifically, the presence of a care home in an LSOA should not impact on the socio-economic characteristics and mortality experience of the non-care-home population in the same LSOA.

by minimising, over the scalar $a(i)$ and vector $b(i)$, the weighted sum of squares

$$S(a(i), b(i)) = \sum_j w_1(i,j) \left(R_0(j) - a(i) - b(i)^T X(j)\right)^2,$$

where the $w_1(i,j)$ are weights that are discussed below. We then set

$$\hat{F}_1(i) \;=\; \hat{a}(i) + \hat{b}(i)^T X(i), \tag{1}$$

and update estimated deaths

$$\hat{D}_1(i) \;=\; \hat{D}_0(i) F_1(i) \tag{2}$$

and the actual-over-expected

$$R_1(i) = D(i)/\hat{D}_1(i) = D(i)/\hat{D}_0(i)\hat{F}_1(i).$$

4.4.1 Stage 1 weights

The calculations above need the weights to be specified. Broadly speaking, the weights are large when two LSOAs, i and j, share similar characteristics $X(i)$ and $X(j)$, and weights tend to zero as $X(i)$ and $X(j)$ get further apart. The idea here is that LSOAs with similar socio-economic characteristics should have similar relative risks.

A number of different ways could be used to generate the weights. Here we start with the socio-economic distance between two LSOA's

$$d_1(i,j) \;=\; \begin{cases} ||X(i) - X(j)||_S & \text{for } j \neq i,\ u(j) = u(i) \\ \infty & \text{for } i = j \\ \infty & \text{for } u(j) \neq u(i) \end{cases}$$

where $u(j)$ is the urban-rural class of LSOA j, the distance measure is

$$||X(i) - X(j)||_S \;=\; \sum_{k \in S} (X(i,k) - X(j,k))^2 \tag{3}$$

and S is a subset of $(1, \ldots, n_P)$. The reason why S might not be the full set of predictive variables $1, \ldots, n_P$, is to allow for "nuisance" variables such as care-home proportions that do not contribute to the underlying socio-economic characteristics of an LSOA but do influence mortality rates.

We next transpose the distances into ranks and then weights:

$$\begin{aligned} r_1(i,j) \;&=\; \text{rank of } d_1(i,j) \text{ out of } d_1(i,1), \ldots, d_1(i,L) \\ w_1(i,j) \;&=\; \begin{cases} \exp\left[-\beta_1 r_1(i,j)\right] \hat{D}_0(j) & \text{if } d_1(i,j) < \infty \\ 0 & \text{if } d_1(i,j) = \infty \end{cases} \end{aligned}$$

There are three elements in the weight function: the use of ranks; the choice of exponential function; and the use of expected deaths, $hatD_0(j)$.

The reason for using ranks, $r(i,j)$, is that they automatically adjust for LSOAs that lie in a sparse or dense region socio-economically.[12] In a sparse region, a standard distance-based weighting system might only give significant weight to a very small number of LSOA's resulting in higher standard errors.

Second, the use of the the exponential is a pragmatic choice out of a range of possibilities that gives good weight to nearest neighbours, but also allows more distant neighbours to contribute in a lesser way. A large value of β reduces potential bias if there is some curvature in the relative risk function F_1 as a function of the predictive variables $X(i)$, but it also increases sampling variation. A smaller value of β_1 reduces sampling variation by including more weight to more LSOAs but increases potential bias. The β_1 parameter (here $\beta_1 = 0.001$) was chosen after some experimentation, by looking at the variance of standardised errors at the LSOA level.

Third, we scale the weights by $\hat{D}_0(j)$. The purpose of this is to give greater weight to LSOAs that have larger numbers of people in the age range of interest: larger numbers means that sampling variation in a particular LSOA is lower.

4.5 Stage 2: estimate the residual spatial relative risk

In this second stage, we assess the residual spatial relative risk based on the updated actual-over-expected, $R_(i)$, using only the longitude and latitude, $Y(i) = (Y_1(i), Y_2(i))$, of each LSOA, i, and the urban-rural class.

Unlike the socio-economic predictive variables, there is no *ex ante* reason why we might expect any systematic trends in relative risk as we move from west to east or north to south. We will, therefore use the simpler method of kernel smoothing to generate estimates of the relative risk:

$$\hat{F}_2(i) = \frac{\sum_j w_2(i,j) R_1(j)}{\sum_j w_2(i,j)}$$

where the weights depend on the physical distance between LSOAs i and j, with adjustment for the urban-rural class of each.

[12] That is, LSOA i only has a small number of LSOAs, j, where $X(j)$ is close to $X(i)$.

4.5.1 Stage 2 weights

We first calculate the adjusted distance between LSOAs i and j as follows:

$$\begin{aligned} d_2(i,j) &= \begin{cases} \left(Y_1(i) - Y_1(j)\right)^2 + \left(Y_2(i) - Y_2(j)\right)^2 \big/ \phi(u(i))\phi(u(j)) & \text{if } j \neq i \\ \infty & \text{if } j = i \end{cases} \\ r_2(i,j) &= \text{rank of } d_2(i,j) \text{ out of } d_2(i,1), \ldots, d_2(i,L) \\ w_2(i,j) &= \exp\left[-\beta_2 r_2(i,j)\right] \hat{D}_1(i). \end{aligned}$$

The construction of the $w_2(i,j)$ is similar to $w_1(i,j)$ but with the addition of the $\phi(u(i))$ terms. As with the $w_1(i,j)$ we allocate greater weight to LSOAs with higher expected deaths. The $\phi(u)$ scaling parameters are dependent of the urban-rural class $u = 1, \ldots, 5$. The rationale for this is that, e.g., rural LSOAs tend to be much further apart than urban LSOAs. Thus a rural neighbour 10km away might carry more weight than a city-based neighbour 1km away. Here, we have used $\phi = (1.5, 1.5, 5, 25, 1)$.[13] Also (similar to $w_1(i,j)$), the use of ranks adjusts weights for LSOA's that, after taking account of urban-rural scaling, are still in a more dispersed or dense area. $\beta_2 = 0.01$ seemed to give acceptable results.

4.6 The core model

We now present results for our core model, using the data from 2001 to 2015. We experimented with different groups of predictive variables and settled on those listed in Table 4 on the basis of (a) lower variance of randomised probability-transformed residuals (see subsection 4.13) and (b) lower variance of the residual spatial relative risk. In aiming to minimise the residual spatial relative risk *we are seeking to explain as much as possible of the observed variation in mortality using socio-economic variables.*

The zeros in the weight function, $w_1(i,j)$, mean that the model for each urban-rural class is fitted independently of other urban-rural classes. Within each UR class we have 8 predictive variables, out of which care home proportions are to be treated as nuisance parameters rather than socio-economic parameters. Hence, in equation (3), $S = \{1, 2, 3, 4, 5, 6\}$.

The analysis produces 32,844 estimates each for the $F_1(i)$ and $F_2(i)$. Empirical distributions (CDFs) for these are plotted in Figure 6. The two plots show CDFs for each of the five age groups. First, consider the left hand plot: socio-economic relative risk. We can see that the wide spread for the age 40-49 age group gradually narrows with age (consistent with Figure 2). For 40-49, the plot reveals how large the mortality inequality gap is between the top and bottom groups: about 7% of LSOAs

[13] Since we use the ranks of the adjusted distances, ϕ can be scaled without having an impact on the weights. So we choose $\phi(5) = 1$ (London) as a reference point.

	Predictive Variable	Scaling	Short Name
1	income deprivation (older people)	standardised	IDO
2	employment deprivation	standardised	EMP
3	average number of bedrooms	standardised	BED
4	living environment deprivation	standardised	LIV
5	wider barriers deprivation	standardised	WID
6	high educated 65+	standardised	EDL7
7	care home *with* nursing, 60+	$[0, 1]$ proportion	COM1
8	care home *without* nursing, 60+	$[0, 1]$ proportion	COM0
9	urban-rural class	categories 1 to 5	UR

Table 4: Predictive variables and scalings used in the core analysis. (See appendix A for a more description of the predictive variables.)

have a mortality that is less than half of the national rate, while 6% have mortality that is more than double the national rate. The factor of $4\times$ is approximately equivalent to an effective age difference of 14 years.

Second, consider the right-hand plot: residual spatial relative risk. The most striking feature of these CDFs is that they are much narrower than the CDFs for socio-economic risk. Indeed, the empirical variance of $F_2(i)$ is in the range 1.3% to 2.8% of the variance of $F_1(i)$ for the same age group. This indicates that, with the right choice of socio-economic variables, socio-economic status by far outweighs geographical location as a determinant of mortality.

Figure 7 (males, ages 40-49) plots geographically the estimated values of the spatial relative risk, $\hat{F}_2(i)$ (top row), and the combined relative risk, $\hat{F}_1(i)\hat{F}_2(i))$ (bottom row), plotting values using a limited colour range.

The top left shows how $\hat{F}_2(i)$ varies across England. We can see that there are patches of oranges/reds and greens/blues (higher and lower mortality than predicted by the socio-economic model only), but there is no systematic north/south or other divide. The top right plot shows the same data but zoomed in on London. The left hand plot has an interesting arc of higher mortality to the west of London, and, in London, there is a substantial red hotspot in west London around and to the north of Heathrow airport. The reasons for these patterns is currently not clear.

The lower plots in Figure 7 show the combined relative risk, $\hat{F}_1(i)\hat{F}_2(i)$, for England and London. For England, there is apparently much more blue, but blues tend to be in less densely populated LSOAs, sometimes rural, and therefore much larger in area. The oranges and reds, representing high relative risk, tend to be in high-density inner-city areas in London, and northern cities.

Equivalent plots for the higher age groups and for females can be seen in Appendix D. In terms of residual spatial relative risk, there is is a gradual shift in the geographical

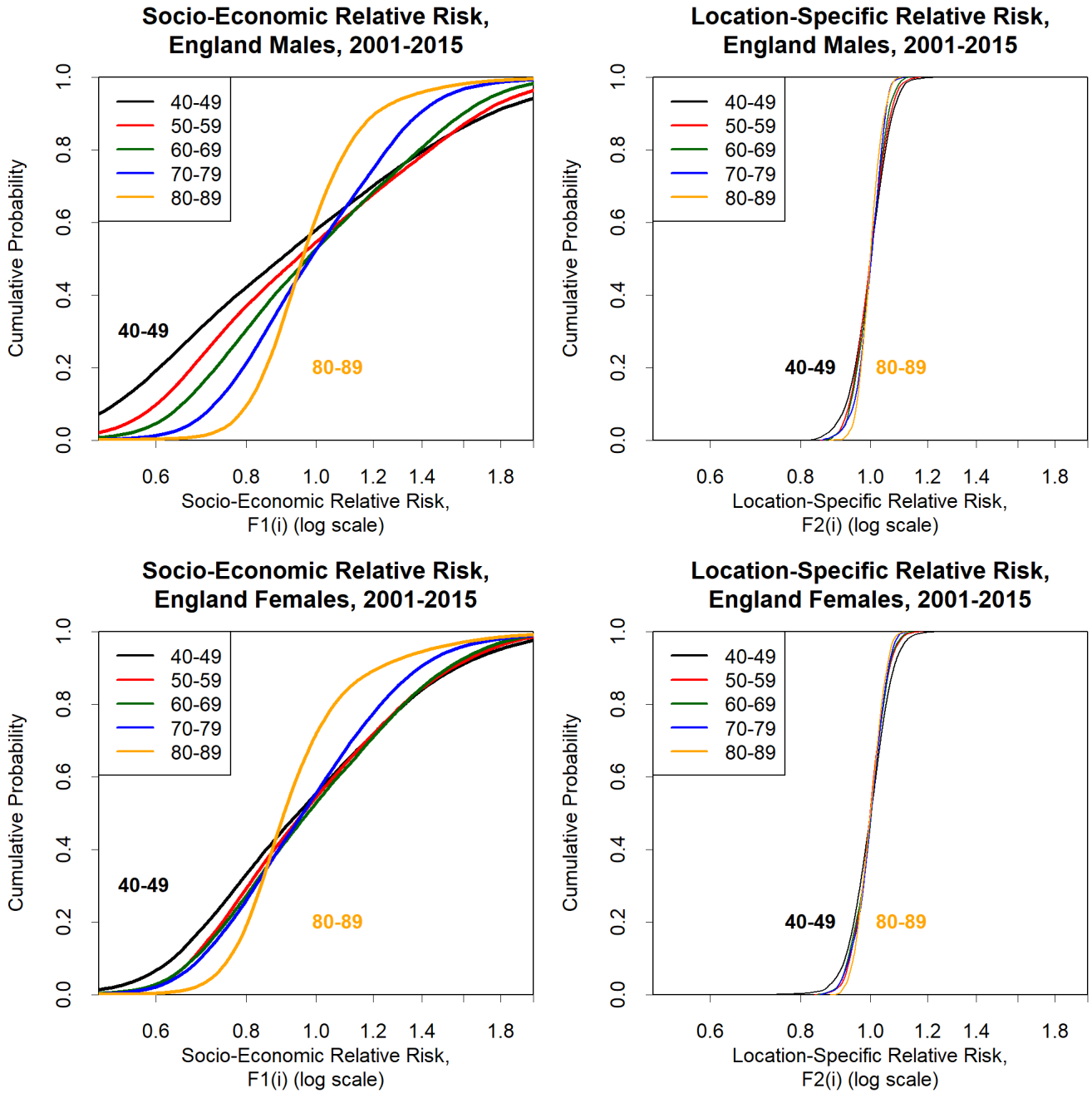

Figure 6: Left: Empirical cumulative distribution functions for the estimated socio-economic relative risk $F_1(i)$ for age groups 40-49, ..., 80-89, each covering years 2001-2015. Right: Empirical cumulative distribution functions for the estimated residual spatial relative risk $F_2(i)$ for age groups 40-49, ..., 80-89.

distribution as well as a narrowing of the range of values and the emergence of a split in the form of south coast and east (low) versus the rest of the country, and persistent higher mortality in west London. For the combined relative risk (lower plots) the pictures are more consistent with Figure 7 but with a gradually diminishing range.

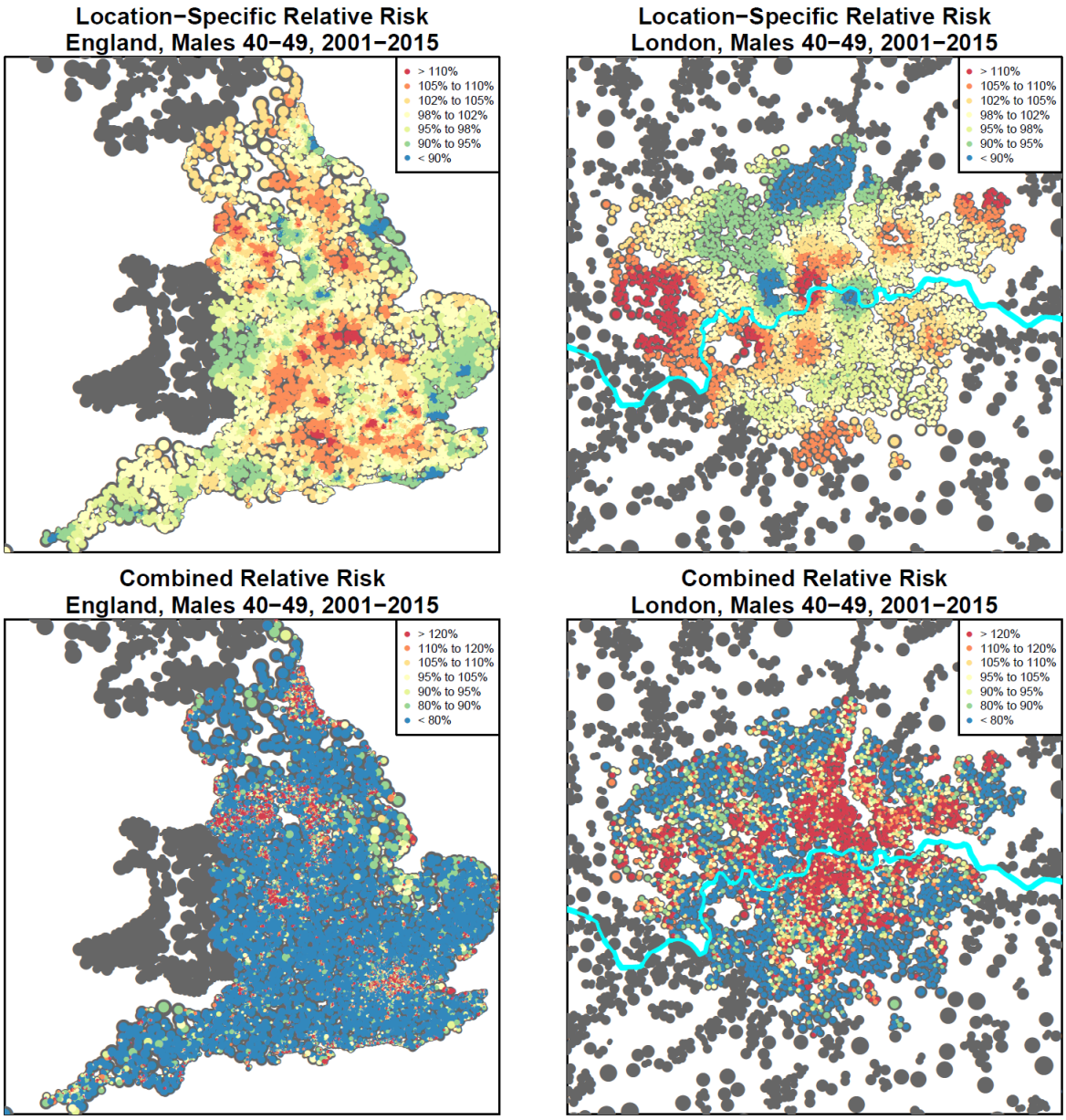

Figure 7: Top row: estimated spatial relative risk, $F_2(i)$, by LSOA for England (left) and London (right) for males, ages 40-49. Bottom row: combined relative risk, $F_1(i)F_2(i)$. Dot sizes reflect the physical size of each LSOA.

4.7 Care homes

The presence of a care home in an LSOA can clearly have an impact on observed mortality. For example, an affluent LSOA that we would expect to have low mortality might have higher than expected deaths if there is a large care home in the neighbourhood. As remarked above, therefore, care homes with and without nursing have been included as predictive variables, but, as they are additionally nuisance variables, they do not influence the weight function, $w_1(i, j)$.

In the data, there are 23,464 LSOAs with no care home, 6,023 have care homes without nursing only, 2,058 have care homes with nursing only, and 1,299 have care homes of both types. In 2011, there were 177,530 persons above age 60 in a care home without nursing[14] and 131,158 persons in a care home with nursing.[15]

Care home data for each LSOA (proportions in the 60+ age group in care homes with and without nursing). It is therefore of interest to consider the following questions: are care homes concentrated more in either less or more deprived areas; can we estimate mortality rates for the healthier non-care-home population; what do mortality rates look like if each LSOA has an average sized care home population?

The first question is addressed in Figure 8. On the left we plot the rank of the estimated relative risk, $\hat{F}_1(i)\hat{F}_2(i)$ for the 80-89 age group versus the proportion in the 60+ age group in a care home in the LSOA, $X_7(i) + X_8(i)$. As the relative risk includes the effect on mortality of having a care home, this scatterplot is skewed to the right. Indeed it is not surprising that the 5% of LSOAs with the highest relative risk all have significant proportions of elderly in care homes. On the right we plot the adjusted relative risk $\hat{F}_1^B(i)F_2(i)$ that removes the impact on mortality of excess care home mortality. Specifically, in Equation 1 we have

$$\hat{F}_1(i) = \hat{a}(i) + \sum_{j=1}^{8} \hat{b}(i, j)X(i, j).$$

Below, we refer to this as $\hat{F}_1^A(i)$ (case A below). The adjusted relative risk, $\hat{F}_1^B(i)$ (see case B below), estimates what the relative risk would be if there were no care homes in the LSOA: that is, as if $X(i, 7) = X(i, 8) = 0$. The Right-hand scatterplot is much more evenly distributed from left to right. This is an indication that the location of care homes is not significantly influenced by the socio-economic mix of the area.

In Figures 9 and 10 we compare the base case A with four hypothetical cases:

- case A: this is our base case using Equation (1).

[14] Average 18 persons in those LSOAs with a care home without nursing.
[15] Average 53 persons in those LSOAs with a care home with nursing.

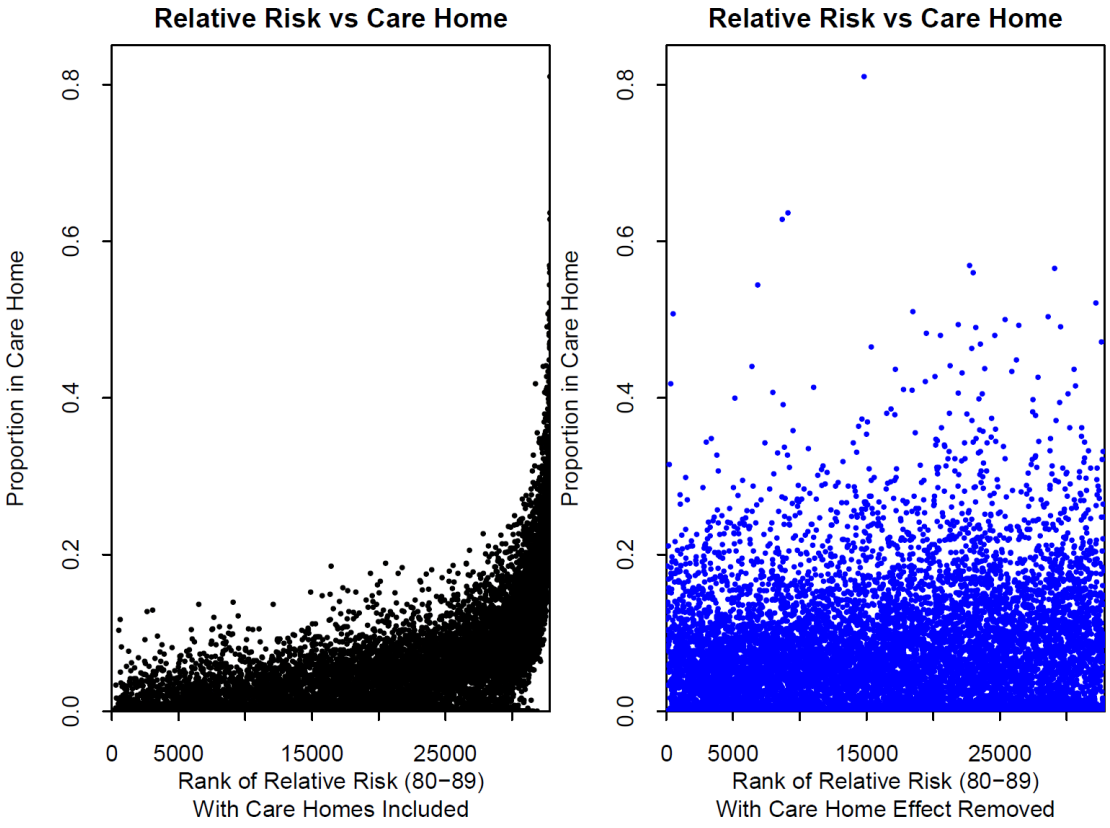

Figure 8: Left: scatterplot of the rank of $\hat{F}_1^A(i)\hat{F}_2(i)$ for the 80-89 age group versus $X_7(i)+X_8(i)$. Right: scatterplot of the rank of the relative risk of the non-care-home population relative risk $\hat{F}_1^B(i)\hat{F}_2(i)$ for the 80-89 age group versus $X_7(i) + X_8(i)$.

- case B: we assume that none of the population is resident in a care home with or without nursing:

$$\hat{F}_1^B = \hat{a}(i) + \sum_{j=1}^{6} \hat{b}(i,j)X(i,j). \quad (4)$$

- case C:[16] we assume that 100% of the 80-89 population reside in a care home without nursing,

$$\hat{F}_1^C = \hat{a}(i) + \sum_{j=1}^{6} \hat{b}(i,j)X(i,j) + \hat{b}(i,8).$$

- case D:[17]we assume that 100% of the 80-89 population reside in a care home with nursing,

$$\hat{F}_1^D = \hat{a}(i) + \sum_{j=1}^{6} \hat{b}(i,j)X(i,j) + \hat{b}(i,7).$$

- case E: we assume that the average proportions $\bar{X}_7$ and $\bar{X}_8$ of the 80-89 population reside in a care home with and without nursing,

$$\hat{F}_1^E = \hat{a}(i) + \sum_{j=1}^{6} \hat{b}(i,j)X(i,j) + \hat{b}(i,7)\bar{X}_7 + \hat{b}(i,7)\bar{X}_8.$$

The second question also uses the adjusted relative risk, $\hat{F}_1^B(i)$. This is illustrated in Figure 9 where, for ages 80-89, we compare relative risks with and without inclusion of the care home effect (Case A in the figure, black dots). All LSOAs with no care homes lie on the 1x diagonal. Those with a care home lie above the 1x diagonal. The ratio with to without is affected by three factors: the number of people in a care home with or without nursing; and the estimated "impact" on individuals of being in a care home (the size of the $\hat{b}(i,7)$ and $\hat{b}(i,8)$ parameters). As we can see, in some LSOAs the presence of care homes more than doubles the estimated mortality rate.

In Figure 9 the red cluster is for case C, and the orange cluster is for case D. We can see that there is a lot of variation in the estimated impact of a care home. This might reflect a number of things: variation between care homes in terms of the state of health of individuals at the time of admission; variation within the 60+ age

[16]Note that cases C and D implicitly assume that the care home proportions apply evenly to age groups 60-69, 70-79 and 80-89. The extent to which this is incorrect will bias the results for cases C and D. In contrast, case B implicitly assumes that the age profile of the care home populations is similar in each LSOA.

range of the age profile of care home residents; and variation between the urban-rural classes. For example, a more detailed look at the orange cluster, case D, (not shown here) reveals that class 5 (London) has relatively low relative risk values while class 1 (other conurbations) has high values. The large difference points to timing differences between the classes in terms of admission to a care home with nursing as well as, potentially, differences in the quality of care.

In Figure 10 we contrast the two cases A and E versus the adjusted case with no care homes. In case E, the impact of having the average care home population is closer to a parallel shift than a proportional adjustment to the adjusted case with no care homes. In aggregate, case E (as with case A) should correspond approximately to national mortality, whereas case B will have lower aggregate mortality than the national population as it excludes the effect of care homes.

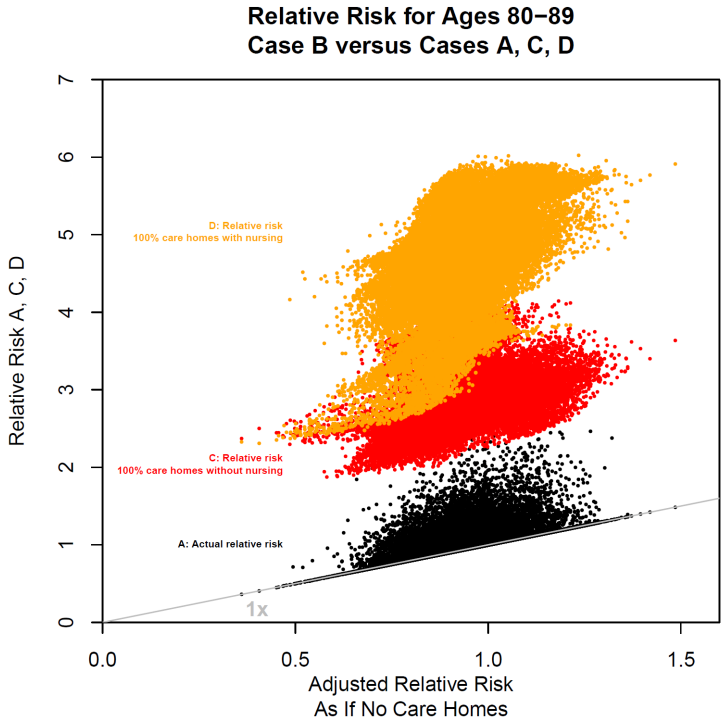

Figure 9: Scatterplot of the adjusted relative risk $\hat{F}_1^B(i)\hat{F}_2(i)$ (case B) versus the three cases A, C and D. A (black dots): actual relative risk $\hat{F}_1^A(i)\hat{F}_2(i)$. C (red dots): relative risk assuming 100% in a care home without nursing, $\hat{F}_1^C(i)\hat{F}_2(i)$. D (orange dots): relative risk assuming 100% in a care home with nursing, $\hat{F}_1^D(i)\hat{F}_2(i)$.

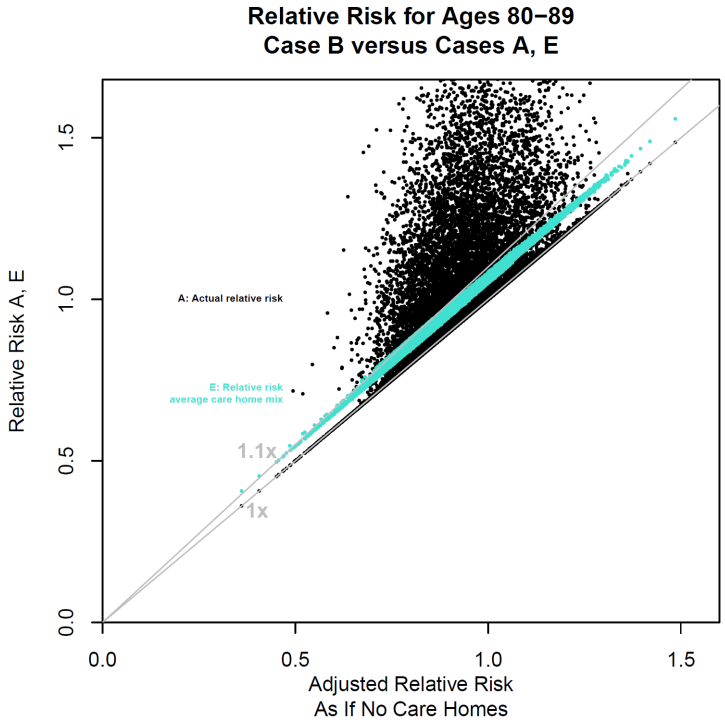

Figure 10: As Figure 9 but for case A (black dots) and case E (blue dots): relative risk assuming average proportion in care homes with and without nursing.

4.8 Impact of urban-rural class

It is interesting to compare the baseline results with two cases:

- case F: urban-rural class plays no role in the local linear regression. This produces relative risks $\hat{F}_1^F(i)$ as well as $\hat{F}_2^F(i)$.
- case G: all LSOAs benchmarked against urban-rural class 2 (cities and towns).[18] This produces relative risks $\hat{F}_1^G(i)$ as well as $\hat{F}_2^G(i)$.

In case F, we simply treat each LSOA as though they are all in the same urban-rural class. In Figure 11 we show a scatterplot of $\hat{F}_1^F(i)$ versus $\hat{F}_1^A(i)$ coloured by urban-rural class for ages 50-59. If there was no urban-rural effect then all points would lie on the $x = y$ diagonal. In reality, for example, the blue dots (UR class 4: very rural) are less steep indicating that this class is less sensitive to changes in the socio-economic predictive variables than other urban-rural classes. Rural towns (green) and London (orange) can also be seen to be slightly less steep, with notable differences between the with and without cases towards the top right.

When we consider the combined relative risk, $\hat{F}_1(i)\hat{F}_2(i)$, we find that cases A and F produce similar results in terms of goodness of fit in each age group. However, the influence of $\hat{F}_2(i)$ is quite different. Specifically, the variance of $\hat{F}_2^F(i)$ is significantly higher than $\hat{F}_2^A(i)$. This violates our objective to minimise variation in the residual spatial relative risk in order to explain as much of the observed variation as possible using socio-economic and non-geographical predictive variables.

In case G, we benchmark all LSOA's against their nearest neighbours socio-economically, in urban-rural class 2. Thus we use weights derived from the modified distances:

$$d_1(i,j) \;=\; \begin{cases} ||X(i) - X(j)||_S & \text{for } j \neq i,\ u(j) = u(i) \\ \infty & \text{for } j = i \\ \infty & \text{for } u(j) \neq 2 \end{cases}$$

that is, we only assign a non-zero weight to LSOAs in urban-rural class 2. Results are plotted in Figure 12. Here, we plot $\hat{F}_1^G(i)$ versus $\hat{F}_1^A(i)$. Red dots for class 2 lie on a straight line as there is no change in how class 2 is fitted. Urban-rural classes 4 (rural, blue dots) and 5 (London, orange dots) stand out as being well below the the main diagonal. This suggests that, on a like for like basis (i.e. similar socio-economic predictive variables) LSOAs in class 2 (cities and towns) have significantly higher mortality than very rural areas and in London, particularly at the more deprived end of the spectrum. Systematic differences by urban-rural class highlighted here diminish with age.

[18] Class 2 is the largest of the five classes and has the widest range of predictive variables, allowing LSOAs in other classes to be sensibly matched with LSOAs with similar predictive variables in class 2.

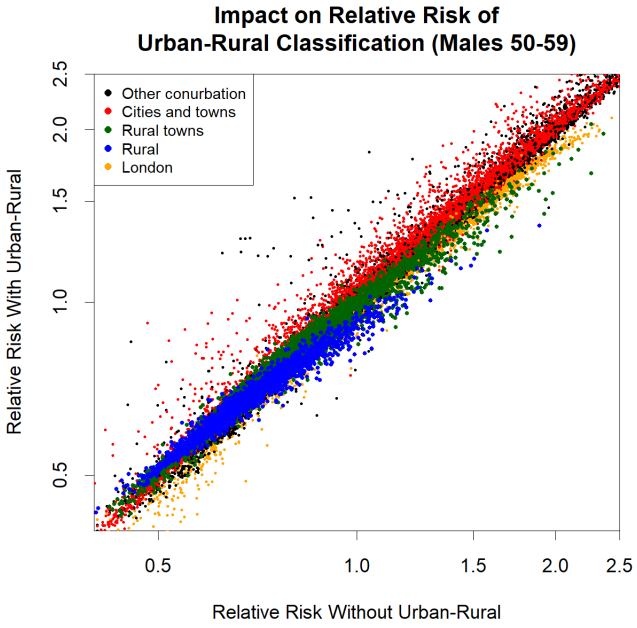

Figure 11: Comparison of socio-economic relative risk, $\hat{F}_1(i)$ without using urban-rural classes and with urban-rural classes.

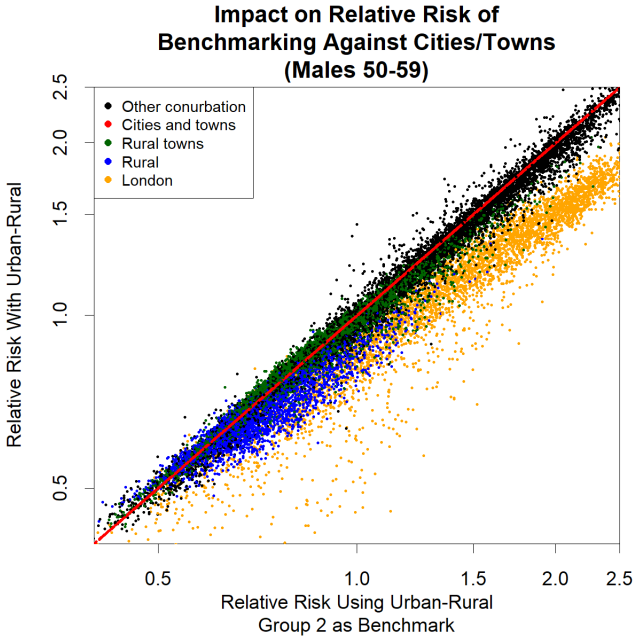

Figure 12: Comparison of socio-economic relative risk, $\hat{F}_1(i)$ where, in case G, urban-rural class 2 is used as a benchmark versus case A, where each LSOA is benchmarked against its own urban-rural class. Scatterplot of $(\hat{F}_1^G(i), \hat{F}_1^A(i))$.

4.9 Summary of the role of predictive variables

The nine predictive variables (including urban-rural class) play different roles.

- Income deprivation (elderly) and employment deprivation are the principal *socio-economic* drivers. Employment deprivation is the main driver for mortality in younger age groups. Income deprivation for the elderly is the main driver for older age groups.
- Urban-rural class is also an important driver, particularly in minimising residual spatial relative risk.
- The proportion of an LSOA in care homes is also very important in terms of its influence over an LSOAs relative risk. However, this is an effect that we wish to remove when we seek to assess the underlying mortality characteristics of each LSOA.
- Average bedrooms, living environment deprivation, wider barriers deprivation, and high education in the 65+ age group are all statistically significant but play a lesser role. Additionally, the importance of each depends on the age group.

By way of example, Figure 13 illustrates how the relative risk with care homes removed, $F_1^B(i)$, depends on given pairs of predictive variables. Predictive variables are expressed as ranks to spread the points out. Individual points (LSOAs) are coloured according to which decile $F_1^B(i)$ falls in.[19] In the upper plot (males 70-79), the coloured bands are close to vertical indicating the dominance of income-deprivation amongst the old (IDO) over high education amongst the elderly (EDL7). However, the slope of the bands indicates that higher levels of education do result in lower mortality rates. Additionally, the bands in the lower right are a bit steeper than the bands towards the upper left indicating a non-linear relationship between the two predictive variables in terms of their impact on mortality in the males 70-79 group. In the lower plot (males 50-59) we look at the relationship between employment deprivation, wider barriers and relative risk. Employment is the dominant effect, but we can see that increased values for the wider barriers variable also increases relative risk.

[19]The extent to which the coloured bands are fuzzy or lack crisp divisions is an indication that the two predictive variables do not, on their own, provide a complete picture: there is additional information in the other variables.

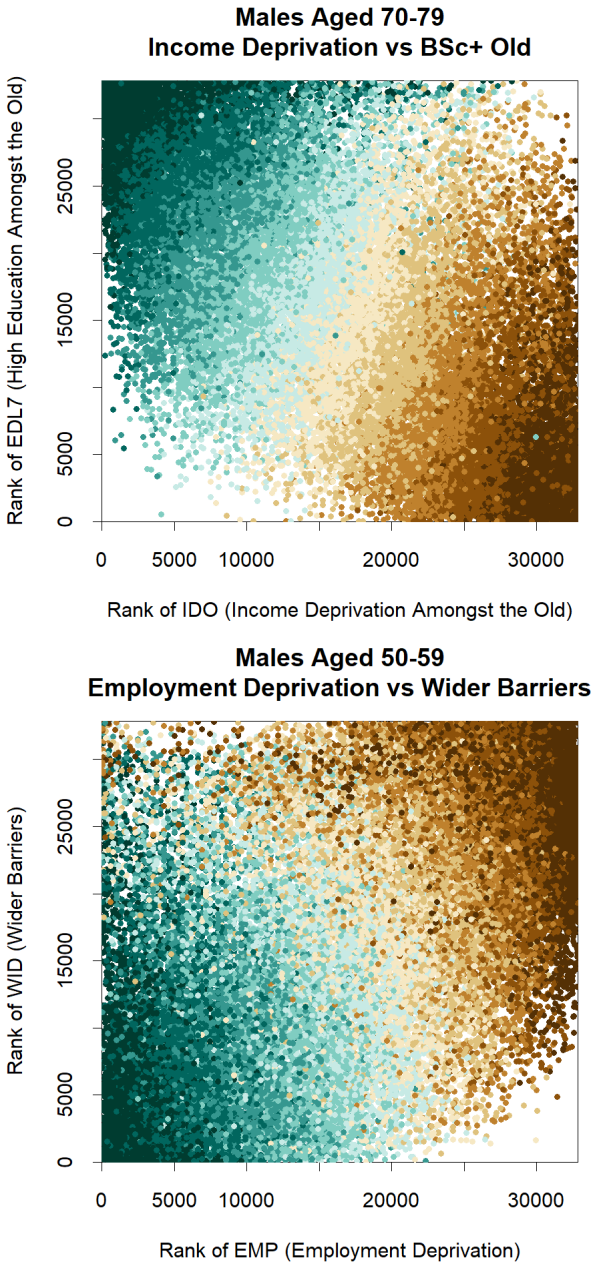

Figure 13: Scatterplots showing the relationship between pairs of predictive variables and socio-economic relative risk, $F_1^B(i)$, with the care home effect removed. Top: males aged 70-79, income-old deprivation (rank) versus proportion amongst the old with a high education (rank). Bottom: males aged 50-59, employment deprivation versus wider barriers. Dots: individual LSOAs. Colours: dark blue/green – 10% lowest relative risk $F_1^B(i)$; dark brown 10% highest relative risk, $F_1^B(i)$.

4.10 The $LIFE$ indices

We are now in a position to consider how the results of local linear regression can be used as an alternative to IMD or income-deprivation deciles or centiles.

In choosing an index, it is appropriate to strip out the impact of care homes to focus on the underlying mortality of each LSOA. To this effect, we propose the use of the socio-economic relative risk, $\hat{F}_1^B(i)$ (equation 4). Further, we have the choice of which age group to use 40-49, ..., 80-89, with the five options being labelled, respectively, as the *Longevity Indices For England*: $LIFE^4$, $LIFE^5$, $LIFE^6$, $LIFE^7$ and $LIFE^8$.

In what follows, we will use the $LIFE^7$ index as this gives the best results for ASMRs over the age range 40-89. However, the other $LIFE$ indices give similar results. The $LIFE^7$ deciles are calculated as follows:

- Decile 1 is the group of LSOAs with the 10% highest values of the $LIFE^7$ index, $\hat{F}_1^B(i)$, for age group 70-79.
- Decile 2 is the group of LSOAs with the next 10% highest values of the $LIFE^7$ index.
- ...
- Decile 10 is the group of LSOAs with the 10% lowest values of the $LIFE^7$ index.

Decile 1 will have the highest mortality and decile 10 the lowest.

The use of the relative risk $\hat{F}_1^B(i)$ with care home bias removed, in combination with the results illustrated in Figure 8, means that each of the deciles will have roughly similar proportions of people in care homes.

The $LIFE$ indices can also be used as additional covariates when analysing life insurance and pension scheme data. That is, it does not need to be used directly as a relative risk: simply that it might be used as a predictor of higher or lower mortality alongside other covariates such as pension amount and geodemographic profiling to enhance model fit. In particular, the $LIFE$ index for a particular LSOA is indicative of the relative risk that would apply to an *average* person in that LSOA (i.e. if a male was chosen at random from the 50-59 age group). On the other hand, if we have further information about a specific individual within a given LSOA then this might mark them out as being likely to have higher or lower mortality. For example, it might be known that they are a smoker, or that they are an active member of an occupational pension scheme.

	Regional Relative To National Mortality			
	Males		Females	
Region	Unadjusted	Adjusted	Unadjusted	Adjusted
North East	117.7	101.2	118.7	99.8
North West	116.2	103.2	118.3	104.9
Yorkshire and The Humber	107.1	100.0	107.7	100.2
East Midlands	98.5	99.9	101.3	100.5
West Midlands	105.3	99.4	102.9	97.5
East	87.5	95.9	89.6	96.5
London	104.7	99.9	98.2	100.3
South East	88.6	101.7	89.2	101.6
South West	86.5	94.9	86.4	94.6

Table 5: Comparison of regional death counts over the age range 60 to 69, and 2001-2015 versus expected deaths using English national mortality, without (unadjusted) and with (adjusted) allowance for socio-economic relative risk.

4.11 Regional and decile mortality

In Table 2 earlier we noted how much variation there was in regional mortality versus national mortality. We now compare expected deaths at the regional level before and after adjusting for the socio-economic relative risk $\hat{F}_1(i)$. These are reported in Table 5. As can be seen, much of the regional variation that we observed previously (Table 5, males/females "unadjusted") can be explained by socio-economic effects (males/females "adjusted"). However, some differences still remain with the south-west and north-west standing out.

In Figure 14 we plot Age Standardised Mortality Rates (ASMRs) for males for the age range 40-89 based on two sets of deciles:

- deciles based on income deprivation (ID);
- deciles based on the $LIFE^7$ index, $\hat{F}_1^B(i)$, using the 70-79 age group.

We can see that the two plots have broadly similar patterns: improving mortality and a widening gap. However, the $LIFE^7$ deciles produce a slightly wider spread, consistent with greater predictive power.

In Figure 15 we plot Age and Deprivation Standardised Mortality Rates (ADSMRs) where we standardise using either income deprivation deciles or $LIFE^7$ deciles.[20]

[20]See Appendix C.3 for definitions and a description of ADSMRs, and how ASDMRs mitigate the differences between ASMRs that are simply caused by the fact that some regions are more deprived than others.

We can make the following observations:

- Use of the $LIFE^7$ deciles results in a much narrower range of ADSMRs than the income deprivation deciles: again an indication that the new approach in this paper explains much more of the differences between regions compared to income deprivation as a sole predictive variable. If London is excluded, then the gap between regions is almost halved.
- London is a clear outlier amongst the nine regions with much stronger improvements over the 16 year period than all other regions.
- All regions have seen a slowdown in mortality improvements since 2010/11. For London, this is a slowdown relative to its previously faster rate of improvement.
- For females (not plotted) the patterns are quite similar, but rates are lower, although regional differences are a bit bigger.

Explaining the London effect is beyond the scope of this paper, but it clearly needs to be better understood. Possible reasons are: changing demographics/gentrification[21]; a widening gap in NHS spending; more effective use of NHS and public-health funding. Note, also, that the predictive variables are single values covering the whole of the period 2001-2016 rather than time dependent, so it is possible that any drift in the predictive variables (e.g. income or employment deprivation) over time that is significantly different in London from other regions might explain the London effect.

[21] E.g. faster growth in London than elsewhere of GDP or higher levels of education, or patterns of migration within England and from outside its borders might have benefitted London mortality (e.g. the *healthy immigrant effect*; see Vang et al., 2017, and Wen et al., 2020).

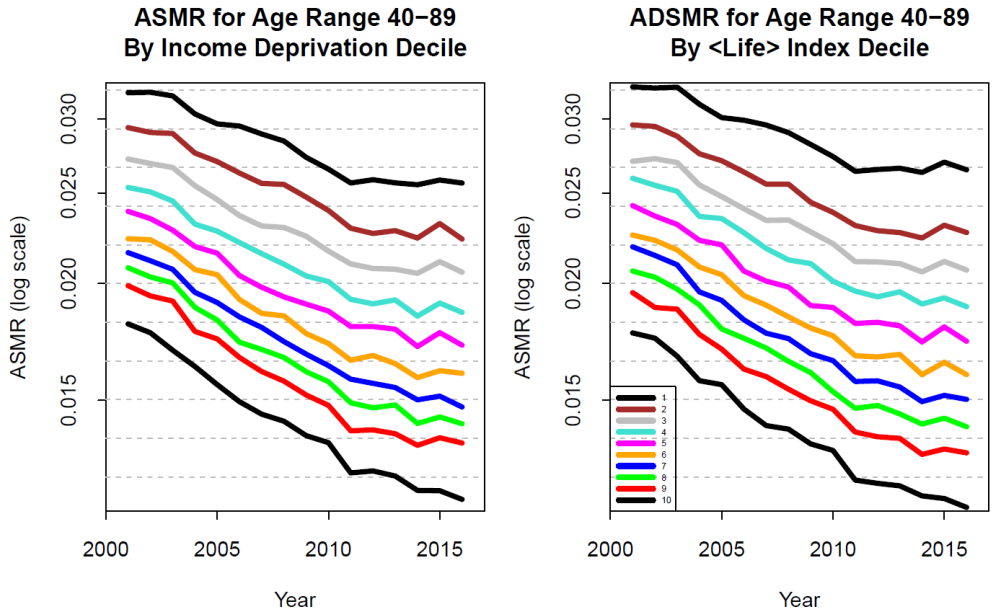

Figure 14: ASMRs for males for the age range 40-89 for each of 10 deciles. Left: deciles based on income deprivation deciles. Right: deciles based on the $LIFE^7$ index.

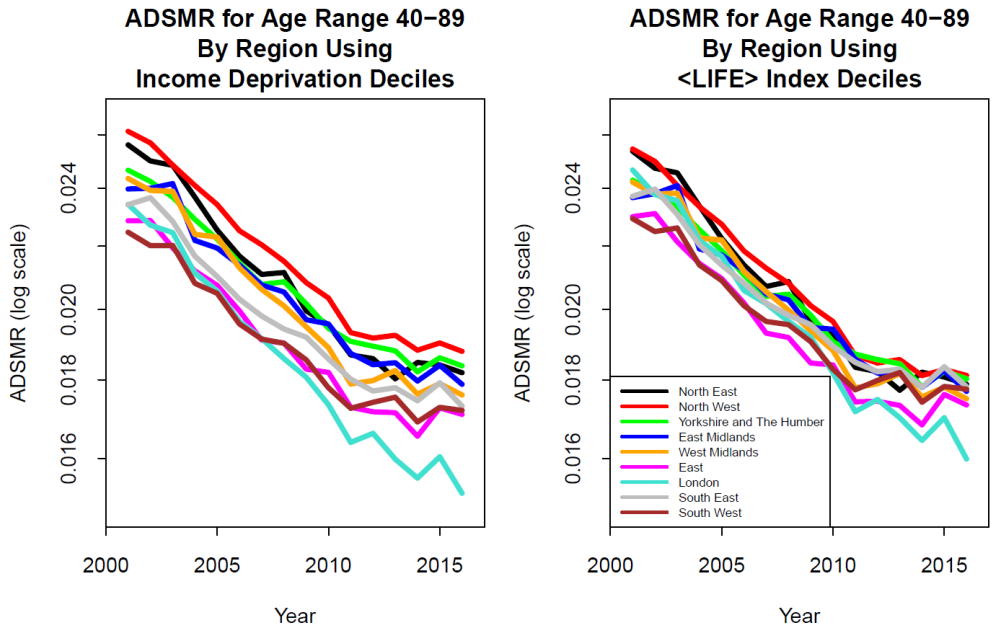

Figure 15: ADSMRs for males for the age range 40-89 for each region. Left: deprivation standardisation using income deprivation. Right: deprivation standardisation using the $LIFE^7$ index.

4.12 Time dependency

It is of interest to investigate how much the relative risks vary through time. Here, we do this by fitting the model again to the non-overlapping time periods 2001-2008 and 2009-2016.

Figure 16 shows how the socio-economic relative risk changes between the two periods. As expected, there is a very high correlation between estimates for the two periods. The scatterplots also highlight differences between the different urban-rural classes.

- For the 40-49 age group, most points lie close to the main diagonal. This is an indication that levels of inequality have remained roughly similar. Much of the "fuzziness" around the main diagonal will be due to sampling variation in the death counts. However, for London (orange dots) the dots clearly lie on a flatter slope, and this is an indication that levels of inequality in this age group have fallen between the two periods in London.
- For the 80-89 age group, the scatterplot is clearly steeper than the main diagonal: a clear indication that levels of inequality have rise sharply in this age group in recent years.

In Figure 17 we mimic the earlier Figure 10 where we illustrate the impact of different levels of care home proportions. We can note the following points:

- Case A (black dots): with current levels of care home provision we see that the relative risks are a bit higher in 2009-2016, indicating that those in care homes have worse prospects now compared to the earlier period.
- Cases C and D (red and orange dots): both of these are significantly higher also pointing to worsening prospects.

Possible reasons (amongst others) for this worsening might be:

- Provision of care home places is not keeping up with demand for places, and so people are even less healthy at the point of entry than they were in the past, pushing up death rates.
- Our data only tell us about the total population above age 60 in care homes. They do not tell us about the age profile of the care home populations. A changing age profile could have an impact on cases C and D.[22]

[22]For example, if there is a shift towards a greater proportion of the care home population being in their 80's, then this could result in a rise in the orange cluster D in the scatterplot from 2001-2008 to 2009-2016.

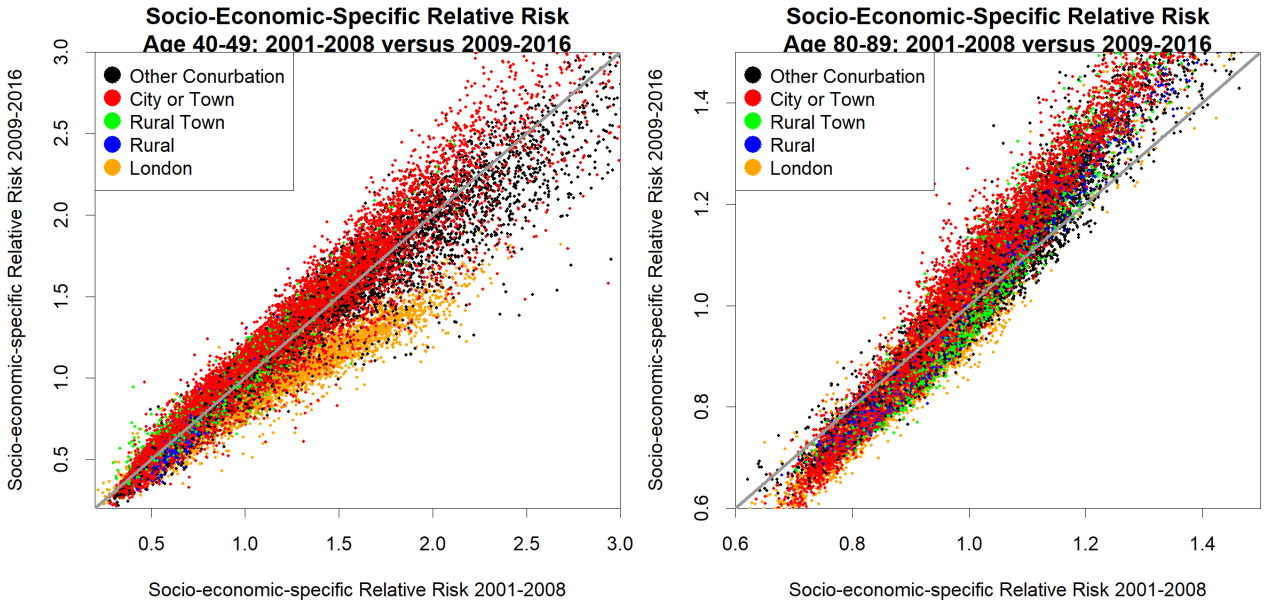

Figure 16: Socio-economic relative risk for males, ages 40-49 and 80-89. 2001-2008 estimates versus 2009-2016 estimates.

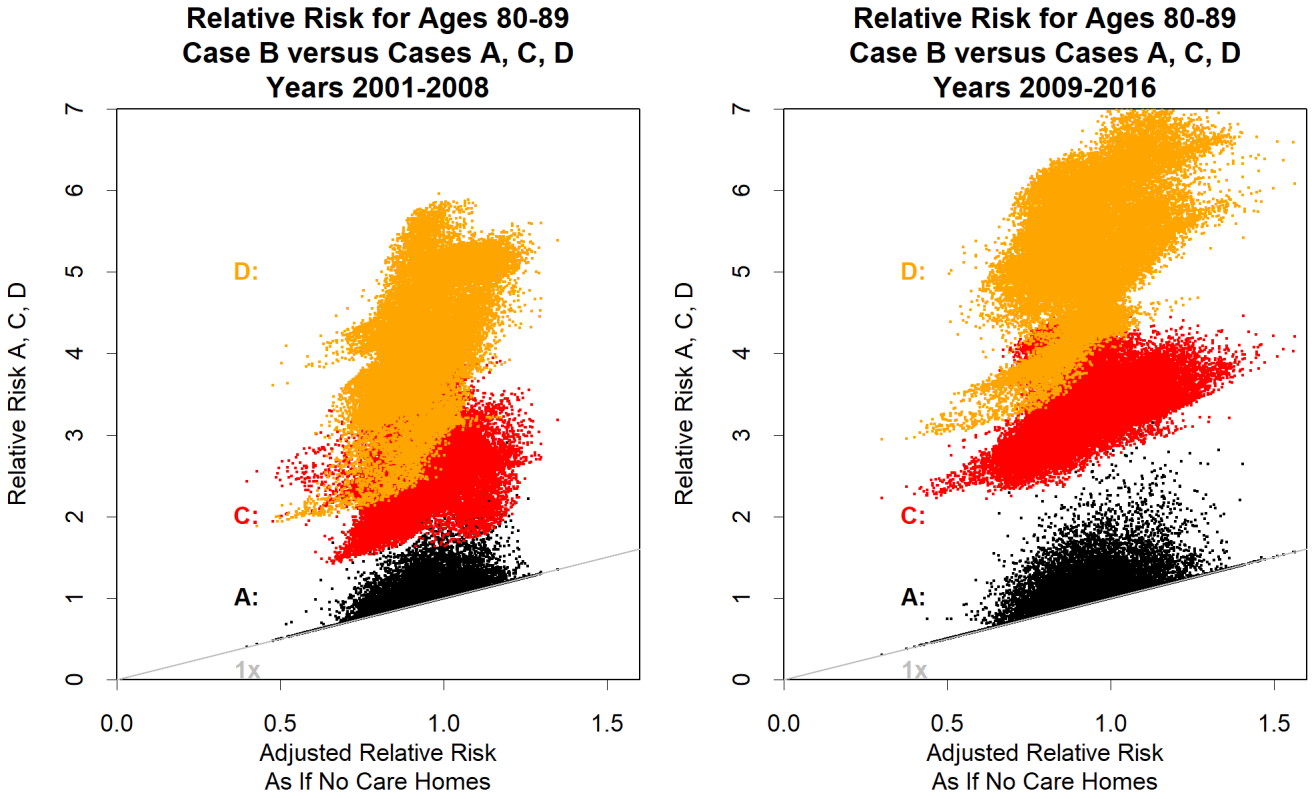

Figure 17: Impact of care homes by time period for case B versus cases A, C and D (see subsection 4.7).

4.13 Analysis of residuals

Local linear regression itself does not require any assumption about the distribution of the deaths, simply that $E[D(i)] = \hat{D}_0(i)F_1(i)F_2(i)$. In our analysis of the residuals we refine this further by investigating if the death counts have a Poisson distribution.

We propose here the use of *randomised probability-transformed residuals*.[23] These are calculated as follows:

- Suppose we have observations $y_1, \ldots, y_n$ of the random variables $Y_1, \ldots, Y_n$ with the hypothesis that $Y_i \sim \text{Poisson}(\lambda_i)$.
- For each $i = 1, \ldots, n$, let

$$\begin{aligned} q_{0i} &= Pr(Y_i < y_i|\lambda_i) \\ q_{1i} &= Pr(Y_i \leq y_i|\lambda_i) \end{aligned}$$

  where the probabilities are calculated under the Poisson assumption with mean λ_i. If $y_i = 0$ then $q_{0i} = 0$.
- Simulate the *randomised probability-transformed residual* $U_i \sim U(q_{0i}, q_{1i})$.[24]

If the Poisson null hypothesis is true then the U_i will be independent and identically distributed uniform$(0, 1)$ random variables. In our mortality setting, each U_i is derived from the observed deaths, $D(i)$, and the estimated mean, $\hat{D}_0(i)F_1(i)F_2(i)$. This provides us with the potential for a number of graphical diagnostics. These could include, for example, QQ plots and histograms of the U_i. However, here we consider two types of scatterplots.

- Plot the (V_i, U_i) where V_i is the rank of the expected deaths $\hat{D}_2(i) = \hat{D}_0(i)\hat{F}_1(i)\hat{F}_2(i)$ out of $\hat{D}_2(1), \ldots, \hat{D}_2(L)$. This is illustrated in Figure 18 for three of the age ranges.
- Plot the (W_i, U_i) where W_i is the rank of the combined relative risk $\hat{F}(i) = \hat{F}_1(i)\hat{F}_2(i)$ out of $\hat{F}(1), \ldots, \hat{F}(L)$. This is illustrated in Figure 19.

In both cases, if the Poisson hypothesis is true then the scatterplots should look uniform and random with no clustering[25].

[23] Randomised probability-transformed residuals are a generalised version of probability-transformed residuals for continuous random variables. If Z is a continuous random variable with cumulative distribution function $F(z)$, then the probability-transformed residual $U = F(Z)$ has a uniform$(0, 1)$ distribution. Here randomisation is required because the random variables of interest are discrete valued rather than continuous.

[24] That is, uniformly distributed on the interval (q_{0i}, q_{1i}).

[25] Note that the V_i and W_i are evenly distributed on the integers $1, \ldots, L$ rather than randomly distributed.

In both Figures 18 and 19, the scatterplots do look reasonably uniform and random for age groups 40-49 and 60-69. But in both cases the scatterplots are less uniform for age group 80-89 indicating that there are further effects that have not been captured in the model.

In Figure 18 we can see some clustering in the top left and bottom right for ages 80-89. In Figure 19 the clustering is more evenly distributed along the top and bottom for ages 80-89: an indication of overdisperion of some sort. The differences in clustering in the two Figures are consistent with the fact that exposures are *estimated* exposures and, therefore, subject to estimation error. The pattern in Figure 18 is simply telling us that if exposures are underestimated then expected deaths are low (so towards the left of the scatterplot) and deaths tends to be higher than estimated (so a high value of U_i). Similarly, if the exposures are overestimated then expected deaths are high (so towards the right of the scatterplot) and deaths tends to be lower than estimated (so a low value of U_i). This illustrates the challenge that the ONS faces when estimating the population between censuses in each LSOA. And Figures 18 and 19 suggest that the problem is only significant at higher ages.[26]

[26]Note, the the presence of overdispersion has no direct impact on the estimation methodology, which only relies on the expected number of deaths rather than the distribution around that mean. Methods such as Generalised Linear Models would, on the other hand, would need to specify how to handle overdispersion: e.g. by replacing the Poisson assumption above with the negative binomial distribution.

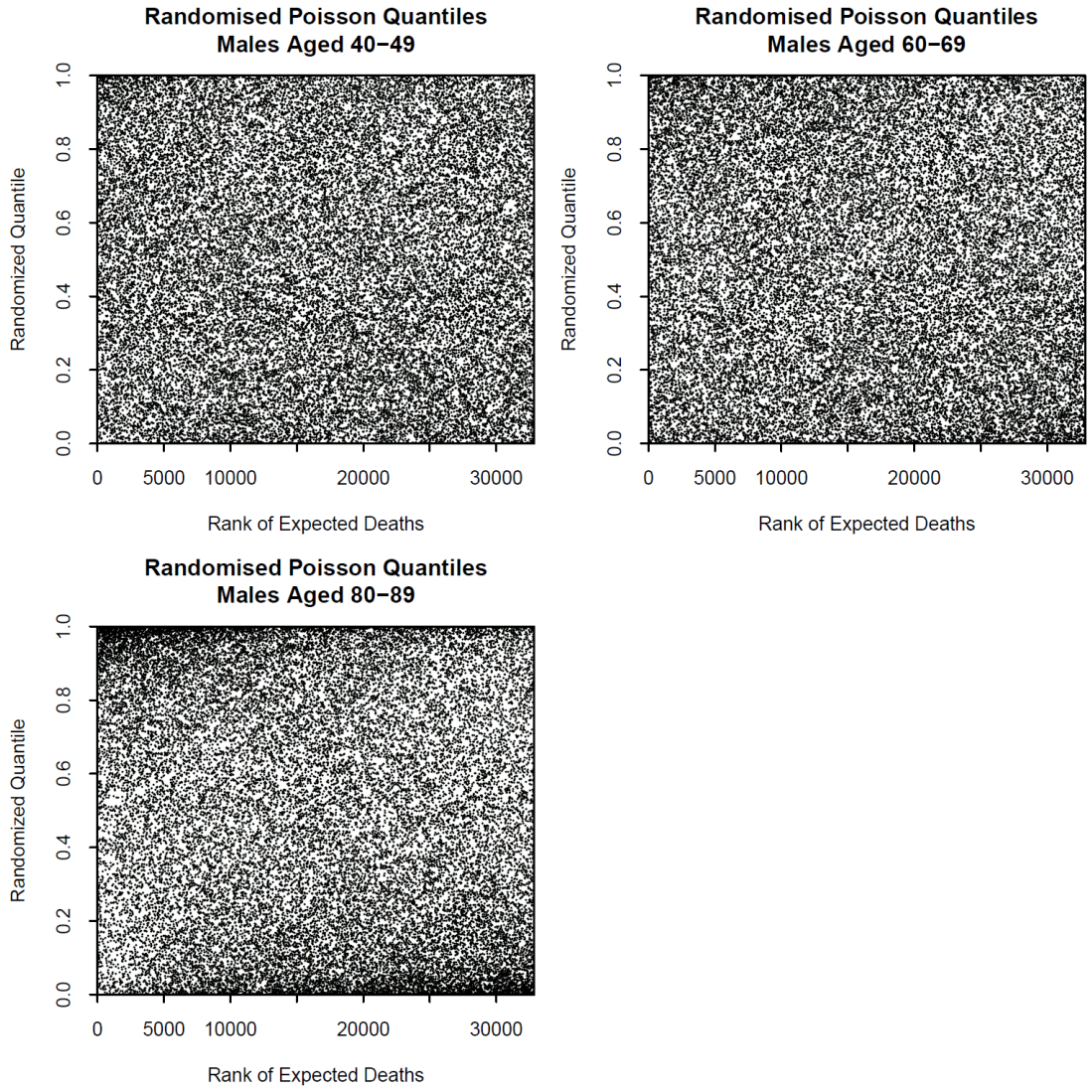

Figure 18: Scatterplots of the rank of the expected deaths, $\hat{D}_2(i)$ versus the randomised probability-transformed residual, U_i, for three age groups.

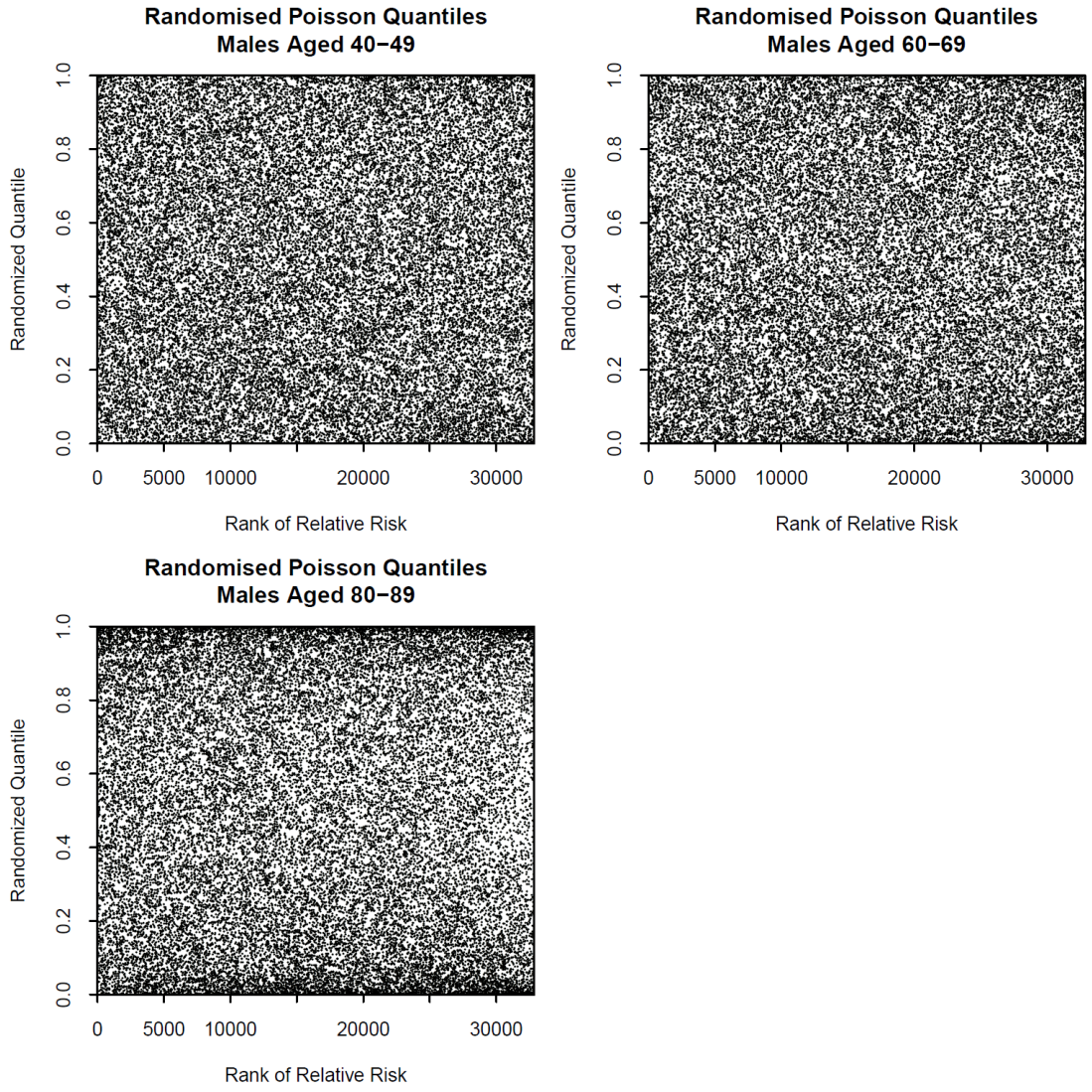

Figure 19: Scatterplots of the rank of the relative risk, $\hat{F}_1(i)\hat{F}_2(i)$ versus the randomised probability-transformed residual, U_i, for three age groups.

5 Cause of death analysis

5.1 Introduction

As remarked earlier, cause of death data have been provided as (confidential) user-requested data in the form of death counts which can be compared against the exposures data used earlier in this paper. Death counts are in the form *region* × *income deprivation decile* × *cause of death* × *gender* × *year* × *5-year age group*. [27] See Table 1 for a list of the 34 cause of death groups.

By way of introduction, we have clustered the 34 causes of death into seven higher-level groupings summarised in Table 5.1.

1	Infectious diseases	COD 1
2	Cancers	COD 2-15
3	Circulatory	COD 21-25
4	Respiratory	COD 26
5	Mental, nervous system and Alzheimer's	COD 18-20
6	External causes (incl. self harm)	COD 31, 33, 34
7	Other causes	COD 16, 17, 27-30, 32

Table 6: High-level cause-of-death groupings. Details for COD 1-34 as in Table 1.

Figures 20 and 21 show a breakdown of total death rates into the seven cause-of-death groups. The figures show how death rates themselves and their decomposition have changed over time by income deprivation decile (with the two deciles on the same scale to aid comparison), by gender and by 5-year age group.

We can make the following observations, some well known, but others perhaps not.

- In all plots, circulatory and cardiovascular diseases have seen substantial falls, leaving less room for further reductions in *absolute* levels of mortality compared to other cause-of-death groups.
- Cancer, in contrast, has only seen modest improvements. By 2016 it was the most significant cause of death in all categories and, therefore, offers the greatest potential for improvements in the future.
- For females, cancer is, and has been, the major cause of death across the three age groups apart from the early years in the 75-79 age group.
- In the earlier sections of this paper we saw how death rates are much higher in decile 1 compared to decile 10. We now see that these differences are repeated

[27] Note that data by deprivation decile only go as far back as 2001. The period 2001-2016 is entirely covered by ICD-10 (International Classification of Diseases).

at the cause of death level. For example, for females aged 55-59, for the four largest causes of death, the coloured bands are all much wider for decile 1 compared to decile 10.

- We see particularly significant differences (in proportional terms) in respiratory diseases, especially at ages 55-59.
- For ages 75-79, we can see increasing levels of mortality for group 5 (at these ages dementia and Alzheimer's).
- Returning to cancers, the differences between deciles 1 and 10 point to a very significant role for public-health strategy alongside future medical advances in reducing death rates from cancer: public health strategy is the key to altering people's lifestyles and reducing the prevalence of the controllable risk factors that we discuss further below.

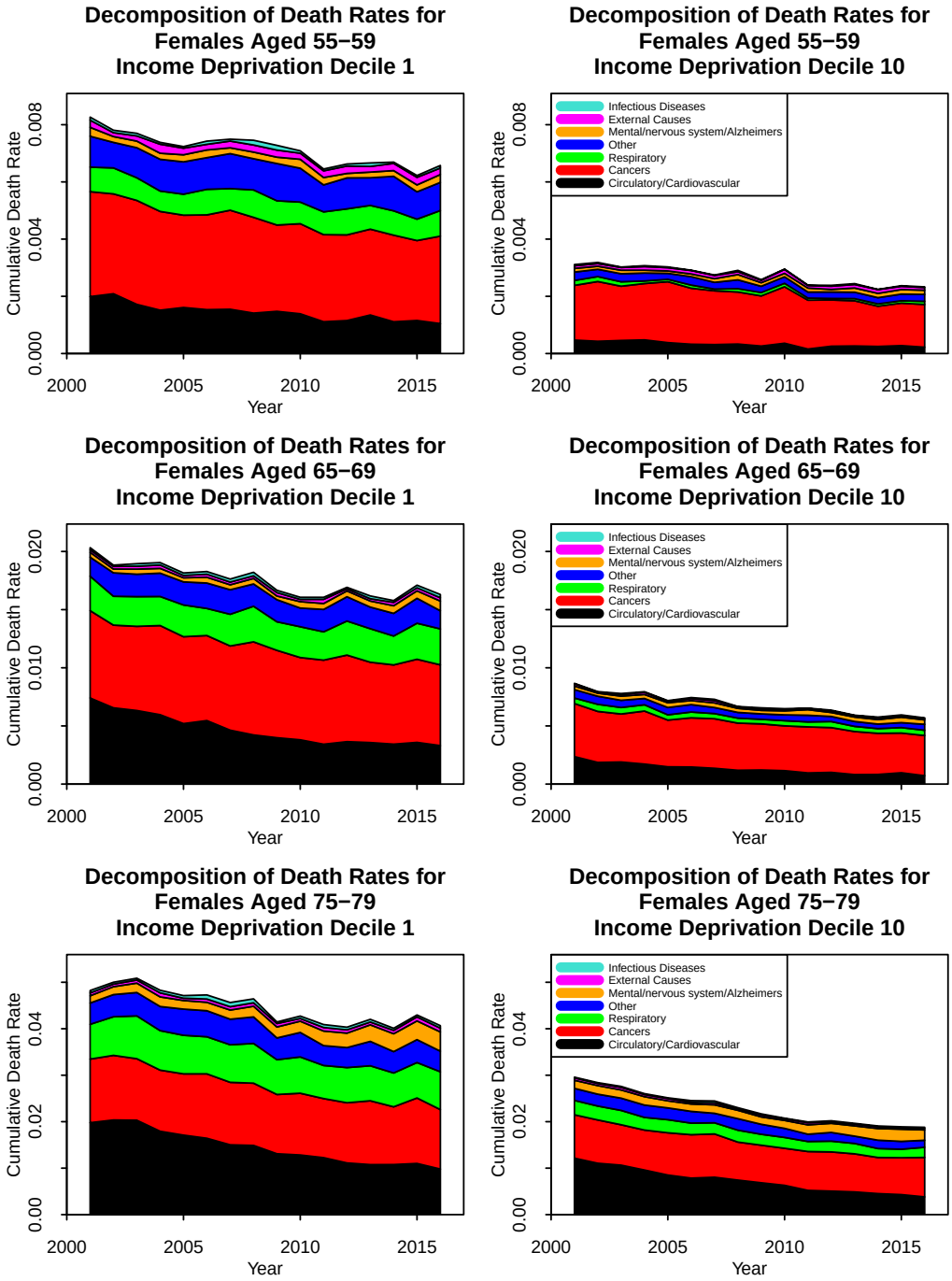

Figure 20: Decomposition of total death rates for females for 5-year age groups into seven cause of death categories. Left: income deprivation decile 1. Right: income deprivation decile 10.

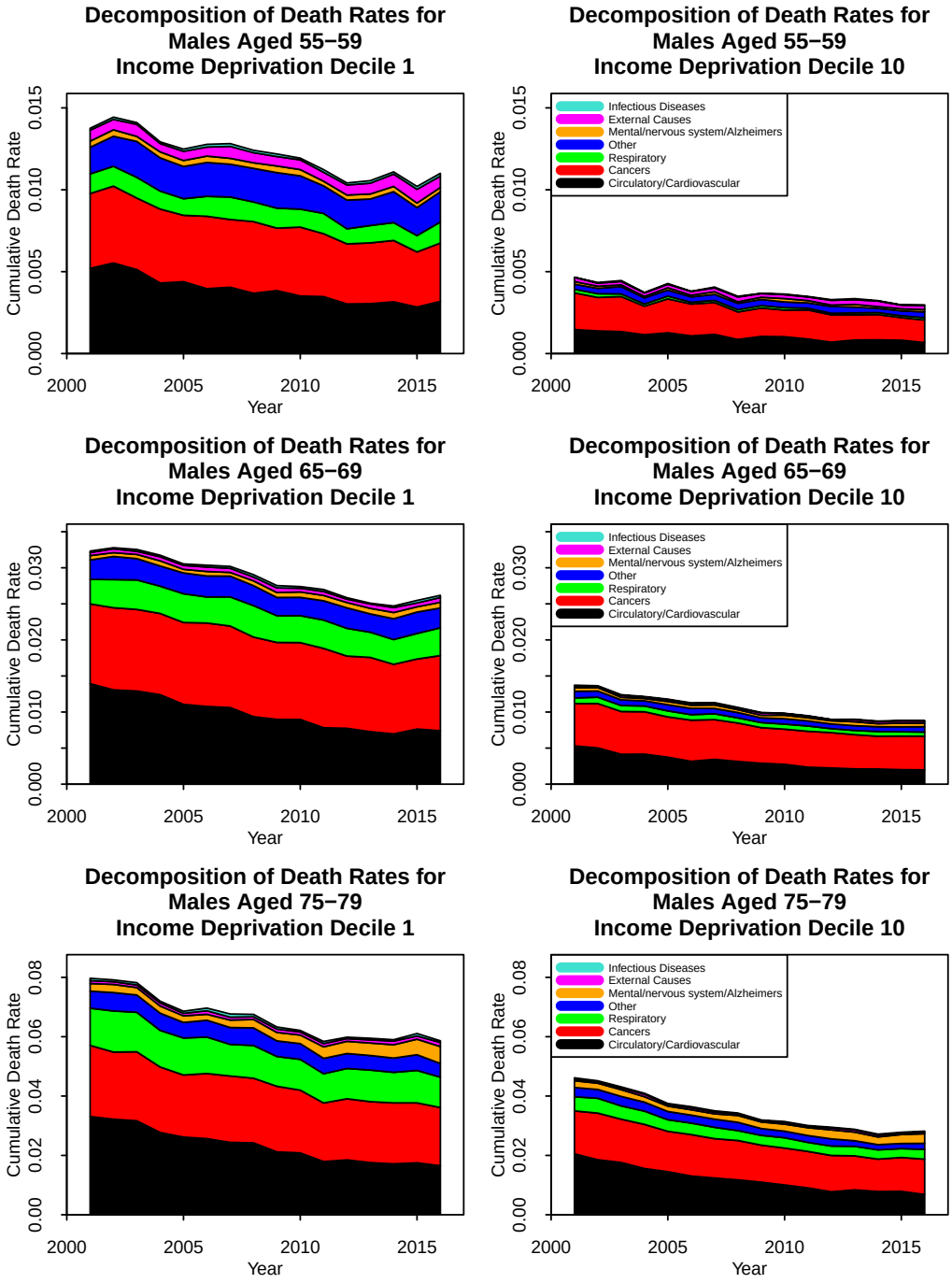

Figure 21: Decomposition of total death rates for males for 5-year age groups into seven cause of death categories. Left: income deprivation decile 1. Right: income deprivation decile 10.

5.2 Detailed cause-of-death analysis

We now move from the broad cause-of-death groupings in Section 5.1 to look in more depth at the more specific causes of death listed in Table 1. The reason for looking at this detailed list is that we wish to differentiate between CoDs that have significant *controllable risk factors* and others that have no significant controllable risk factors.

Examples of controllable risk factors include smoking, diet, exercise, alcohol, sun, drugs, hygiene, risky sex, stress, and risk factors linked to occupation such as the environment in which people live and work. These risk factors have a known significant impact on certain causes of death: e.g. smoking is the major controllable risk factor for lung cancer and chronic obstructive pulmonary disease (COPD). These controllable risk factors also lead to intermediate conditions such as obesity, diabetes, high blood pressure etc. which have, additionally, been identified as risk factors for many causes of death.

A similar concept to controllable risk factors is that of *preventable risk factors* (see, for example, `https://www.cancerresearchuk.org/`). This is a slightly wider class that includes risk factors that are not easily controllable by the individual, but are preventable. An example of a risk factor that is preventable but not easily controllable is Human Papillomavirus (HPV) which is strongly linked to cervical cancer and other diseases. HPV can be *prevented* through vaccination programmes.

Examples of risk factors that are not easily controllable nor preventable (or not controllable at all) are:

- age (which we will generally take as a given in any discussions below!);
- gender;
- racial factors (e.g. some groups are genetically more likely to suffer from certain diseases);
- other genetic factors (e.g. family history);
- personality traits (e.g. *conscientiousness* is known to be the trait that is most strongly associated with low or high mortality; see, for example, Kern and Friedman, 2008, and Deary, Weiss and Batty, 2010); [28]

[28] In the field of psychology, *conscientiousness* is one of the five major character or personality traits. Of these, the personality trait that we are most interested in is conscientiousness because it is the most strongly associated with life expectancy (see, for example, Kern and Friedman, 2008, and Deary, Weiss and Batty, 2010). In the mortality context, conscientious individuals are more likely to: adhere to a healthier diet; exercise (even if this is considered boring!); visit the doctor early when they develop symptoms related to ill health; and follow doctor's orders when given a diagnosis.

- affluence and education (arguably these are slightly controllable but perhaps only early in life).

In our discussions below we will contrast causes of death with one or more significant, controllable risk factors, with causes of death that have no controllable risk factors. Examples of the former include lung cancer and ischaemic heart disease. Causes of death with no or limited controllable risk factors include prostate cancer, breast cancer and ovarian cancer. But they do, in some cases, have non-controllable (and non-preventable) risk factors such as race (e.g. prostate cancer; Cheng et al., 2009, and Taitt, 2018).

This returns us to our reason for using a detailed list of CoDs: we wish to investigate how patterns of inequality differ between CoDs with and without controllable risk factors. Additionally, our focus on the detailed list allows us better prospects for analysing the underlying controllable risk factors: for example, lung cancer trends and inequalities give us insight into the differences in smoking prevalence in different sections of the population. To a large extent, inequalities will reflect differences in underlying risk-taking behaviours in the different groups (i.e. the prevalence of controllable risks such as smoking).

Figure 22 is the first of a sequence of figures that plot Age Standardised Mortality Rates (ASMRs; left hand plot) over the period 2001-2016 for the 10 income deprivation deciles and Age and Deprivation Standardised Mortality Rates (ADSMRs; right hand plot) for the 9 regions. ASMRs and ADSMRs are over the age range 40-89. (A full set of plots for all causes of death can be found on the ARC programme website: `www.macs.hw.ac.uk/~andrewc/ARCresources`.)

Figure 22 shows lung-cancer mortality. For lung cancer, the main controllable risk factor is smoking,[29] making the disease highly cohort dependent (e.g. linked to how fashionable smoking is during each cohort's early adult years).

We can make the following observations and comments.

- Levels of inequality are very high between deprivation deiles: rates are mostly over three times higher in decile 1 compared to decile 10.

- Since smoking is overwhelmingly the main risk factor for lung cancer, with smokers at least 20 times more likely to die from lung cancer than non-smokers (e.g. Pesch et al., 2012, and Malhotra et al., 2016), differences between deprivation deciles are most likely due to differences between deciles in smoking prevalence. In particular, from Figure 22 we infer that smoking prevalence is much higher amongst the more deprived groups.

[29] See, for example, `www.cancerresearchuk.org`. More generally the Cancer Research UK website offers an invaluable resource for actuaries to identify risk factors for most cancers.

- We can see an improving trend in lung cancer mortality. Over the same period there has been relatively little in the way of significant improvements in the treatment of lung cancer: no cures and limited improvements in treatments (Jones and Baldwin, 2018). Putting these two together indicates that it is more likely that falls in lung cancer deaths amongst males is more likely to be due to falls in smoking prevalence than medical advances.

  But there has also been a slight widening of the gap between deciles 1 and 10 indicating that smoking prevalence amongst males in less deprived areas has been falling at a faster rate than in more deprived areas, in response to government health warnings.

- The right-hand plot shows that there are significant differences between regions, even after adjusting for deprivation, of up to 40% for males and 70% for females. One explantion for this is that patterns of smoking, as a controllable risk factor, vary considerably from region to region. Alternatively, differences might be due to variations between regions in lung cancer treatment (slowing progress of the disease).

- Amongst the regions, London stands out as having experienced much stronger improvements.

The smoking theme continues in Figures 23, 24 and 25 which show mortality from respiratory diseases for males and females and lung cancer for females. The major cause of death under respiratory diseases is chronic obstructive pulmonary disease (COPD) and COPD also has smoking as the major risk factor.[30] Consequently, for both males and females, we see a very similar pattern to lung cancer including levels of mortality inequality. For females, the slight upward trend is quite different from males, indicating that smoking prevalence amongst females has been rising over time (at least relative to males). For females, also, we see more widening in the inequality gap compared to males, indicating that females in less deprived areas are listening more to government health warnings than those in the most deprived areas.

For both males and females, the ADSMRs by region for lung cancer and respiratory diseases have a consistent ordering, again providing indirect evidence that smoking is the key controllable risk factor and that there are variations between regions.

Ischaemic heart disease is another cause of death that ranks alongside lung cancer and respiratory diseases in the top few causes of death (indeed, the number one cause of death at some ages and deprivation groups). However, in this case there are multiple controllable risk factors rather than just one, including smoking, lack of exercise and a poor diet (see, for example, `https://www.bhf.org.uk/informationsupport/conditions/coronary-heart-disease`). Family history

[30] See for example, `https://www.mayoclinic.org/diseases-conditions/copd/symptoms-causes/syc-20353679` and Forey, Thornton and Lee (2011).

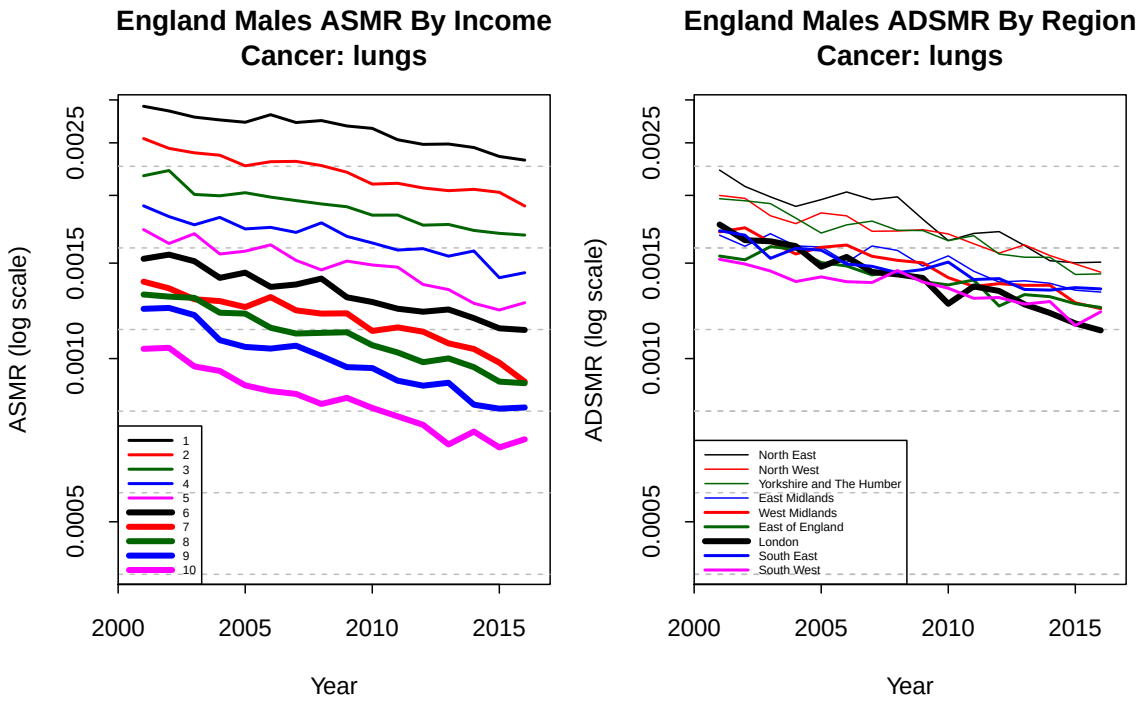

Figure 22: England Males CoD 6: Lung cancer and cancer of the bronchus.

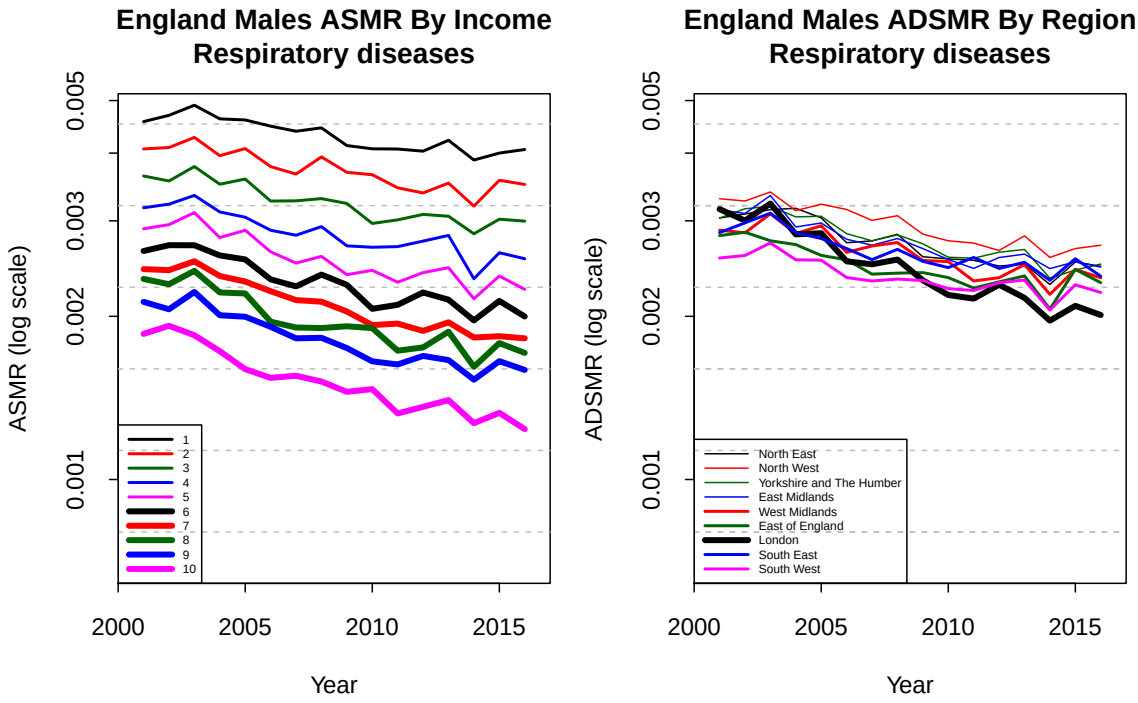

Figure 23: England Males CoD 26: Respiratory diseases including Chronic Obstructive Pulmonary Disease (COPD).

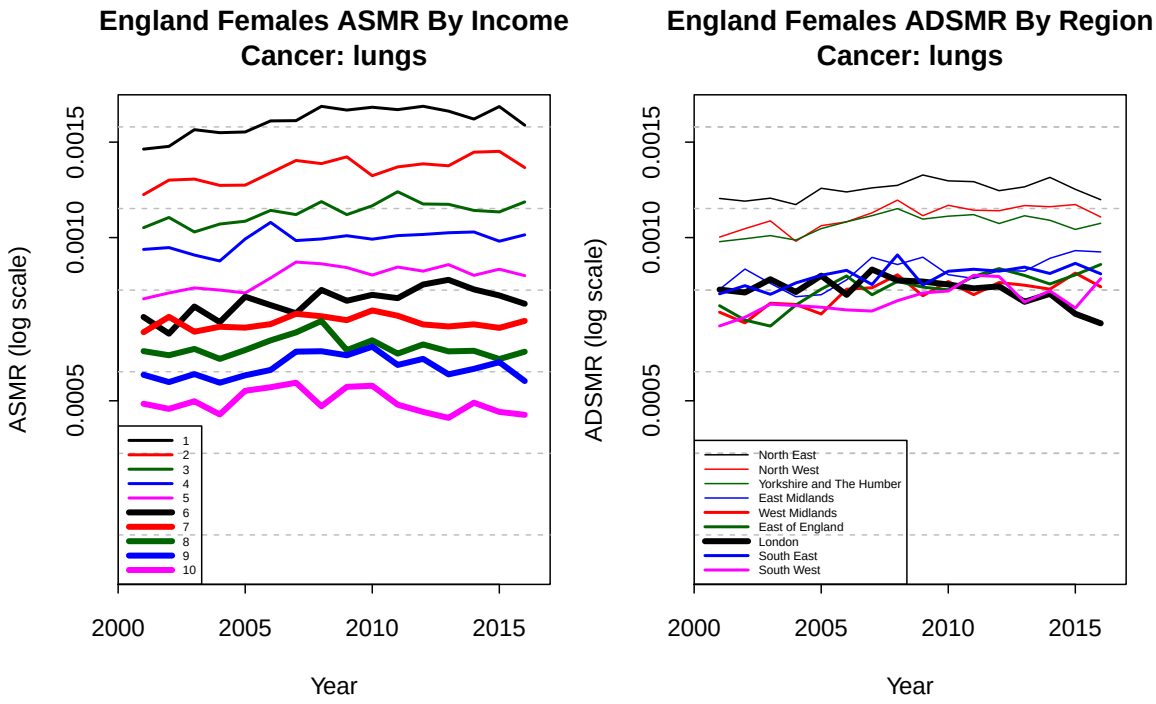

Figure 24: England Females CoD 6: Lung cancer and cancer of the bronchus.

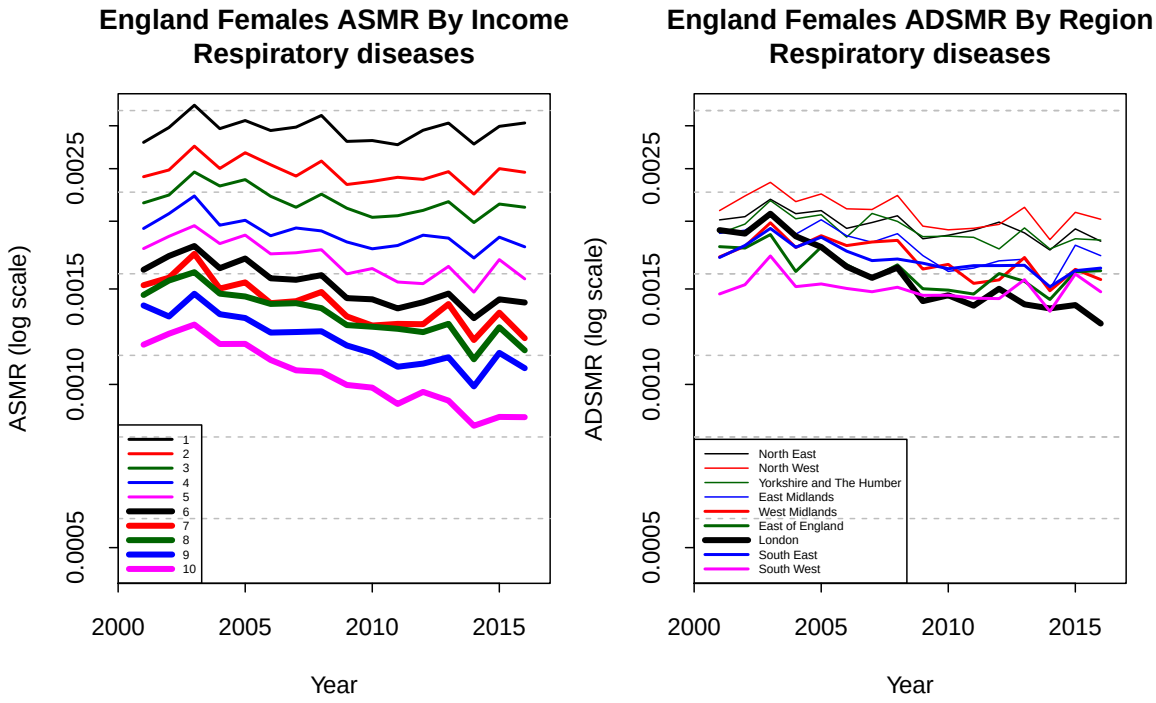

Figure 25: England Females CoD 26: Respiratory diseases including Chronic Obstructive Pulmonary Disease (COPD).

and ethnic background are additional, non-controllable risk factors. Lack of exercise and poor diet link up with e.g. high cholesterol, obesity and diabetes as further specified risk factors.[31] Figures 26 and 27 show the ASMRs and ADSMRs for ischaemic heart disease.

- Mortality inequality is again very significant, although not quite as wide as lung cancer and the respiratory diseases.
- As is well known, there have been significant reductions in mortality from ischaemic heart disease: something that is very evident in the plots for males and females. However, for both males and females we see a widening gap between income deprivation deciles 1 and 10. And this widening seems to be more significant that that observed for e.g. lung cancer. This, then, points to a more significant widening of the gap between the 10 deciles in controllable risk factors such as diet, exercise or statin uptake compared to the widening gap in smoking prevalence noted previously.
- For males and more so for females, the least income-deprived groups have seen steady improvements over the whole of 2001-2016, while, for the most deprived, improvements have slowed in recent years. This, again, is consistent with a widening gap in controllable risk factors as noted above as well as uptake of statins and adherence to prescriptions (see, for example, Gitsels et al., 2016).

  So while we have generally seen a slowdown in all-cause mortality improvements in recent years (see, e.g., Figure 14) in all deciles, the slowdown cannot be uniformly attributable to all causes of death. Here there has been no slowdown in improvement rates for ischaemic heart disease in the most affluent deciles.

  For females, improvements in the ADSMRs in the right-hand plot also do not slow down.

[31] A further controllable risk factor acting in a positive direction is the use of statins. The timing of the initial prescription is, in some cases, controllable (e.g. seeking early advice), but the more controllable part concerns adherence to the prescription (also a manifestation of conscientiousness).

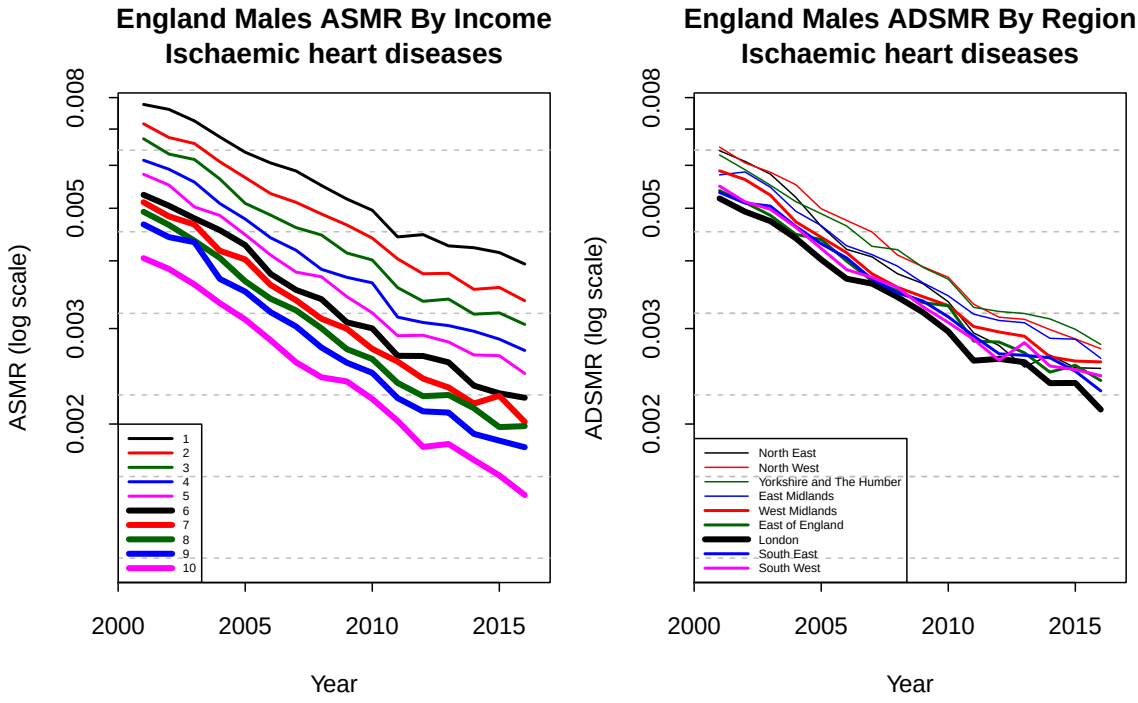

Figure 26: England Males CoD 22: Ischaemic heart diseases

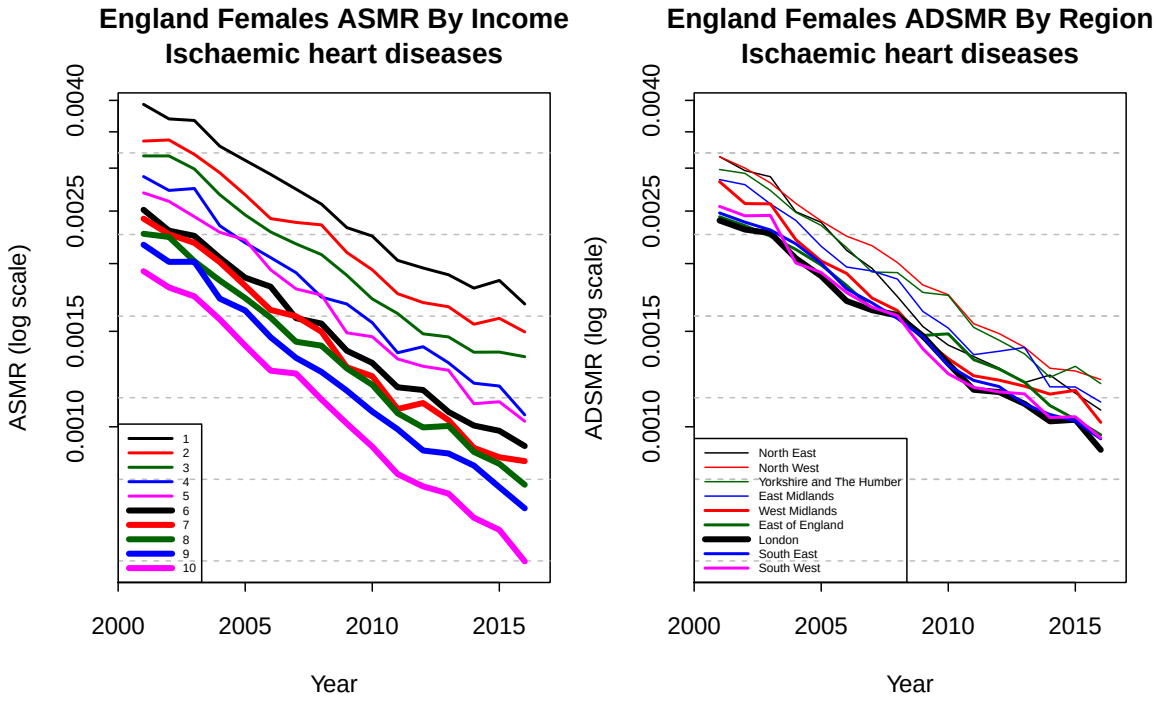

Figure 27: England Females CoD 22: Ischaemic heart diseases

In Figure 28 we show ASMRs and ADSMRs for diabetes for females. Here we see a slightly wider spread than ischaemic heart disease, and a similar widening of the spread. And, in contrast to heart disease, we see more evidence for a slowdown since 2010/2011. The situation here, though, is more complicated, as diabetes itself is a risk factor for other causes of death. The changing pattern in Figure 28 might, in part, reflect changes in reporting practice: with diabetes more likely to be recorded as the cause of death now than before. Diabetes has similar controllable and non-controllable risk factors compared to ischaemic heart disease, with the exception of smoking.

We move next to look at some causes of death that have no significant controllable risk factors: prostate, breast and ovarian cancer. Figure 29 shows ASMRs for prostate cancer. We can, first, see that there is an improving trend which (because there are no controllable risk factors) must be due to improvements in the treatment of the cancer: both cures and slowdown in the progression of the cancer. And these improvements seem to be accruing equally to all deciles. For the ASMRs by income deprivation it is hard to see any differences between the 10 deciles other than that due to sampling variation. For breast and ovarian cancer, the picture is the same for income deprivation deciles (Figures 30 and 31): no significant differences between income deprivation deciles.

However, a different picture emerges when we look at ADSMRs by region in the three plots. For breast cancer, the regions all look quite similar. However, for prostate cancer and, even more so, ovarian cancer we can see significant differences by region. For prostate cancer, ADSMRs for London are on average about 20% lower in the later years. For ovarian cancer, London, the North East and Yorkshire & Humber do better (and with a slightly larger and more persistent gap than prostate cancer), while the East and West Midlands are persistently poor. The reasons for these regional differences are not clear, especially when we observe that there are no significant differences between income deprivation groups. Possible reasons include regional differences in the effectiveness of screening programmes, and differences in the quality of the treatment following diagnosis.

For these causes of death it is, perhaps, helpful to consider the following stages and landmarks:

1. healthy;
2. T_1 =time of onset of disease X;
3. T_2 =time when symptoms emerge;
4. T_3 =time of diagnosis of disease X;
5. T_4 =time when treatment commences;

6. T_5 =time of either death from X, death from another cause, or cured of disease X.

Out of these, and for a cause of death with no controllable risk factors, the time of onset, T_1, might be the same for all groups. But, thereafter, behavioural differences and regional variations in treatment and care might differ, meaning that T_2, T_3, T_4 and T_5 might all be different including the final outcome.

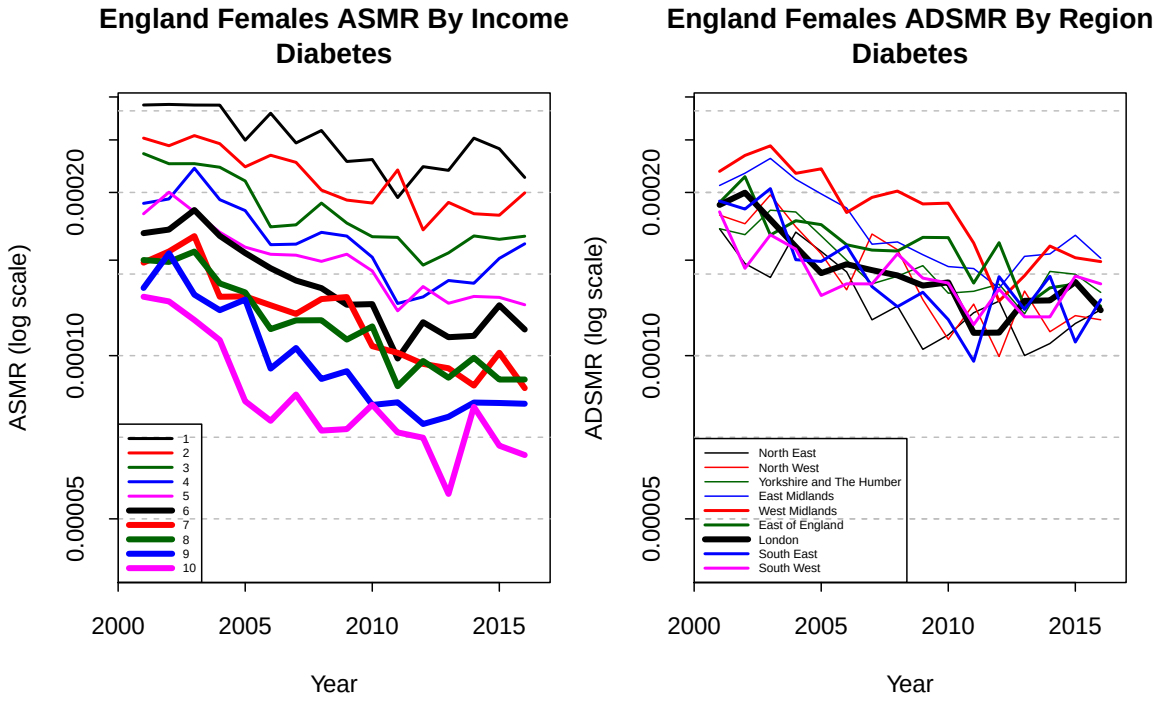

Figure 28: England Females CoD 17: Diabetes

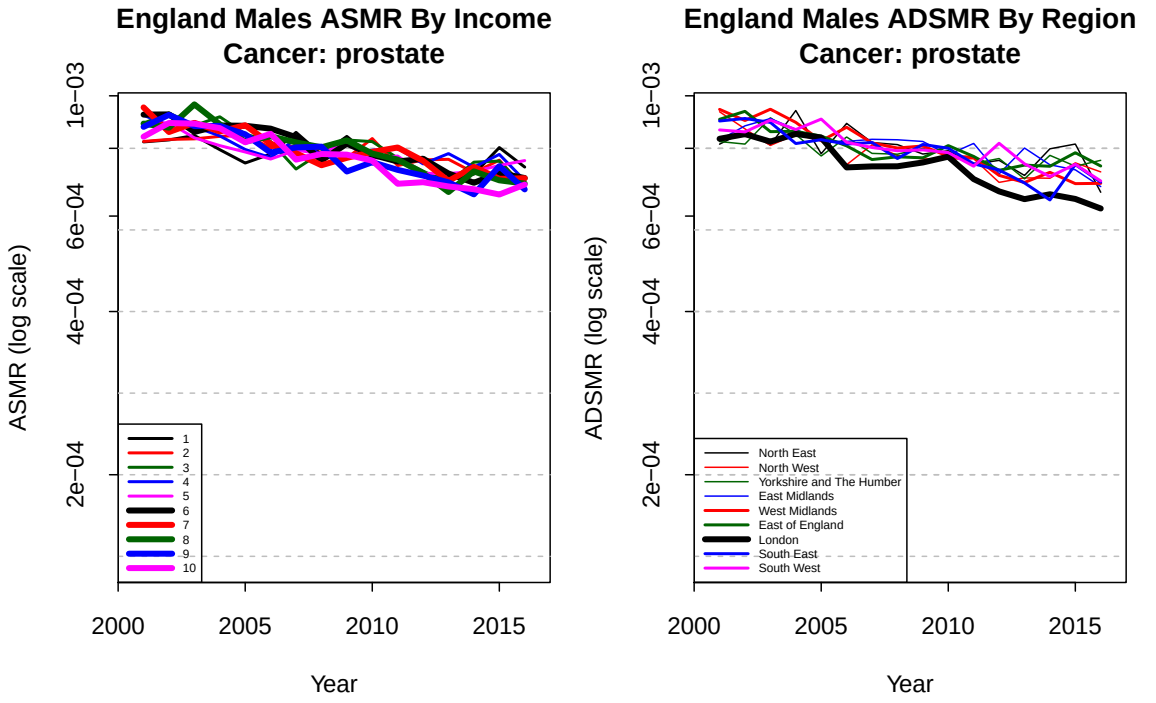

Figure 29: England Males CoD 11: Cancer: prostate

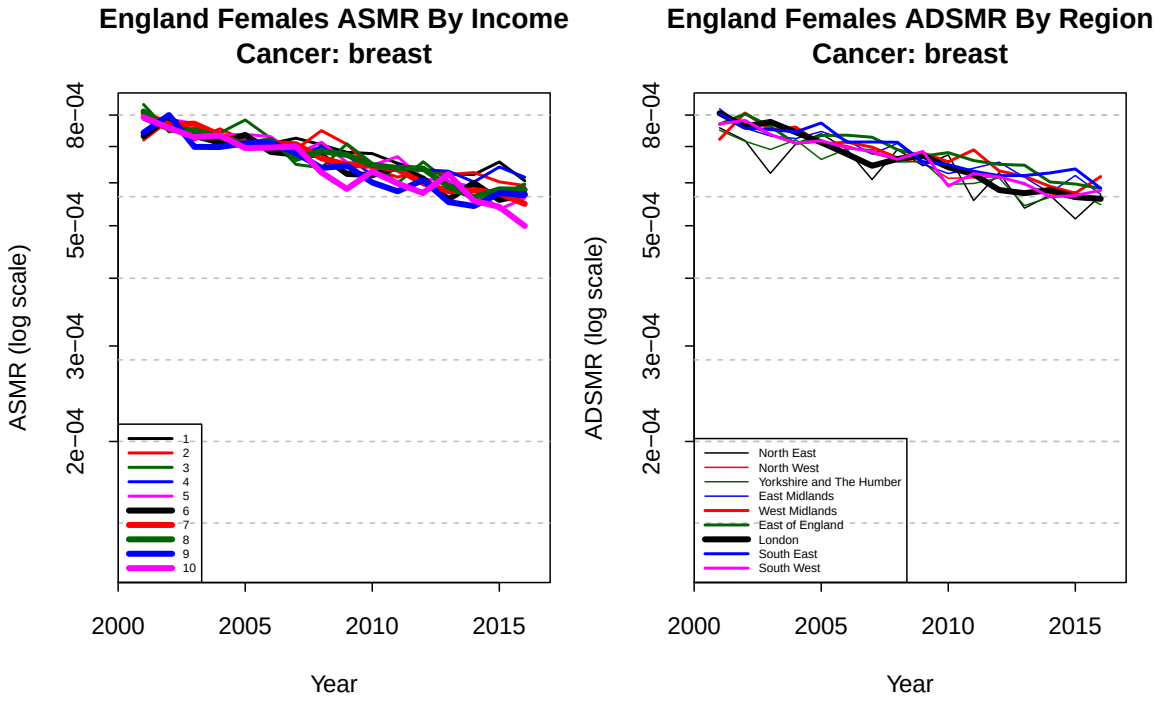

Figure 30: England Females CoD 7: Cancer: breast

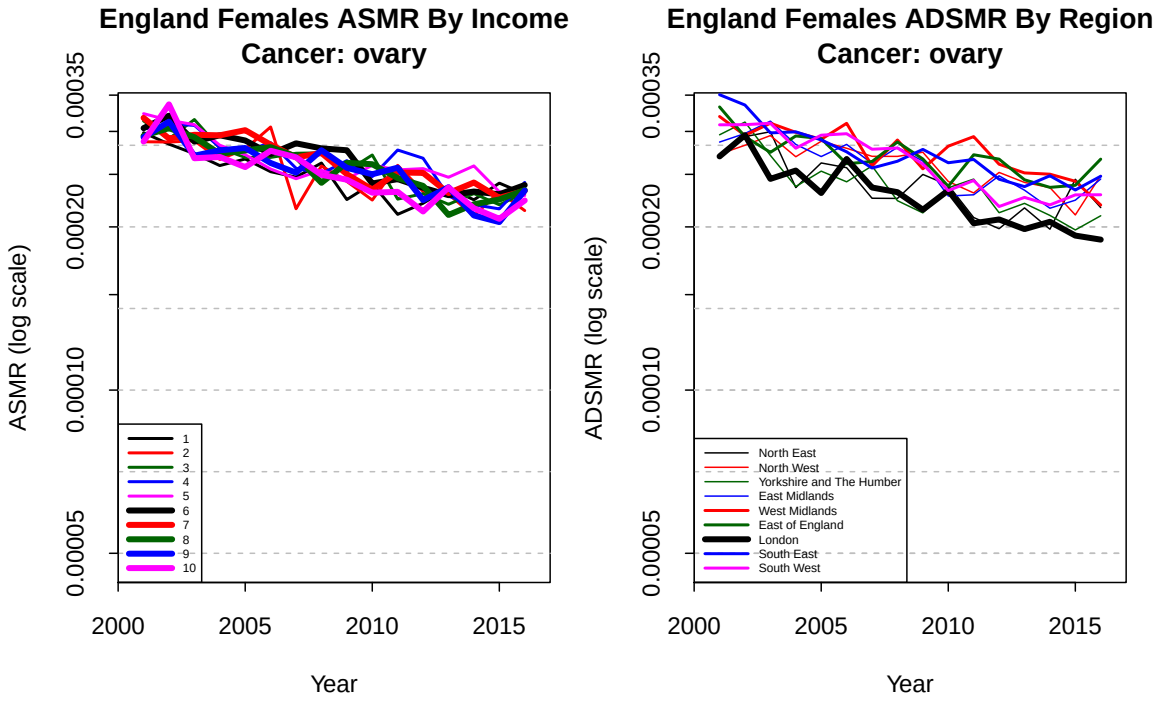

Figure 31: England Females CoD 9: Cancer: ovary

We next look at dementia, Alzheimer's and related diseases. In Figure 32 we look at CoD group 18 for females which includes vascular dementia as the major cause of death in older age. Vascular dementia has similar controllable risk factors to heart disease, and so we see significant levels of mortality inequality and a widening gap. On the other hand the gap is narrower than ischaemic heart disease (suggesting that the controllable risk factors have less impact) and has worsened over time. There is a clear jump in ASMRs in 2011 and this is due to changes in the coding of cause of death (Office for National Statistics, 2011). This jump up is matched by a jump down in deaths from cerebrovascular diseases (Figure 33). The connection between the two is that vascular dementia can be characterised by a series of mini strokes. So a patient now being recorded as having died from vascular dementia might previously have been recorded has having died from a stroke. The general rise in mortality from vascular dementia might be linked to underlying risk factors. For example, a prior history of heart attack raises the subsequent risk of vascular dementia (and Alzheimer's). The rise in deaths from vascular dementia might, in part, be due to a growing proportion of the elderly population being survivors of heart attacks.[32] In other words, gains elsewhere (reduced cardiovascular mortality) might be causing worsening mortality elsewhere.

ASMRs for Alzheimer's disease are plotted in Figures 34 and 35. There is some debate about risk factors for Alzheimer's disease, but the narrow spread of ASMRs by income deprivation decile suggest that the impact of controllable risk factors might be relatively small.[33] The ADSMRs by region paint a different picture that is similar to ovarian cancer, with a much wider inequality spread than by income deprivation. This might point to differences in long-term healthcare provision by region. However, some of the regional differences might be due to where people move to when they move into a care home. To investigate this, we can use the care home data discussed in Section 4.7 and compare the age 60+ care home population in 2011 with the general population in 2011 by region. For the general population we could also use ages 60+, but to reduce the impact of migration on retirement we take the population aged 70+.[34] This tells us that London (also, to a lesser extent, the East, and the West Midlands) has a much lower proportion of its population that are resident in care homes. A possible conclusion is that people suffering from Alzheimer's (as well as other conditions requiring long-term care) move to care

[32]This concerns the frailty of persons at a particular age. Age standardisation removes the impact of the fact that more people are surviving into older age as a result of medical advances affecting survivorship at younger ages. But if, as a result of more people surviving heart attacks, a greater *proportion* of people at, say, age 75 are more frail now than 10 or 20 years ago with respect to vascular dementia, then death rates from vascular dementia at age 75 will rise.

[33]That is, by considering ASMRs for other causes of death, there is strong, but implicit, evidence that, on average, people who live in more deprived deciles are more likely to adopt lifestyles that increase their risk of death from many other causes such as heart disease, lung cancer etc.. If these had a significant impact on Alzheimer's then we would see a wider inequality gap by income deprivation.

[34]Most of the care-home population, additionally, will be 70+.

homes outside of London (where costs might be cheaper), consequently lowering the London death rates from Alzheimer's.

The other key feature of Figures 34 and 35 is the rising trend after 2005, similar to vascular dementia. Again this might be due to the increasing frailty of those who survive to higher ages, or it might be due to the increasing focus of general practitioners on Alzheimer's as an illness and cause of death (i.e. individual sufferers from Alzheimer's might die from a related cause and there might be a shifting balance between the two reported causes of death).

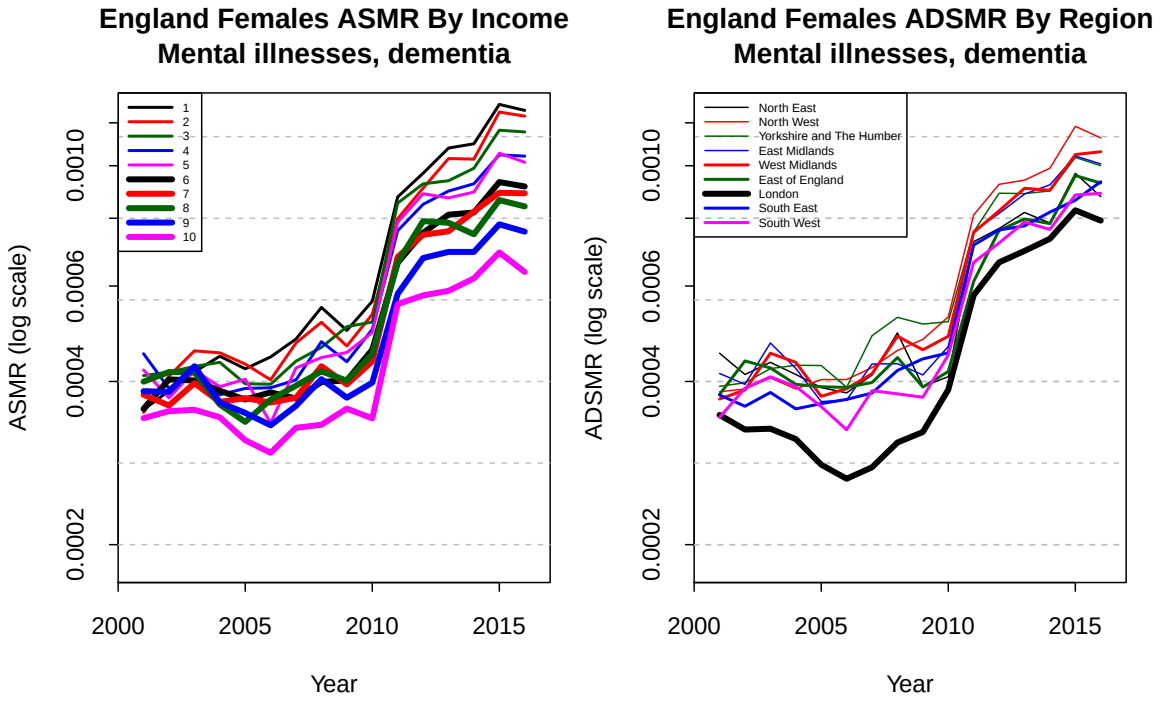

Figure 32: England Females CoD 18: Mental illnesses, dementia

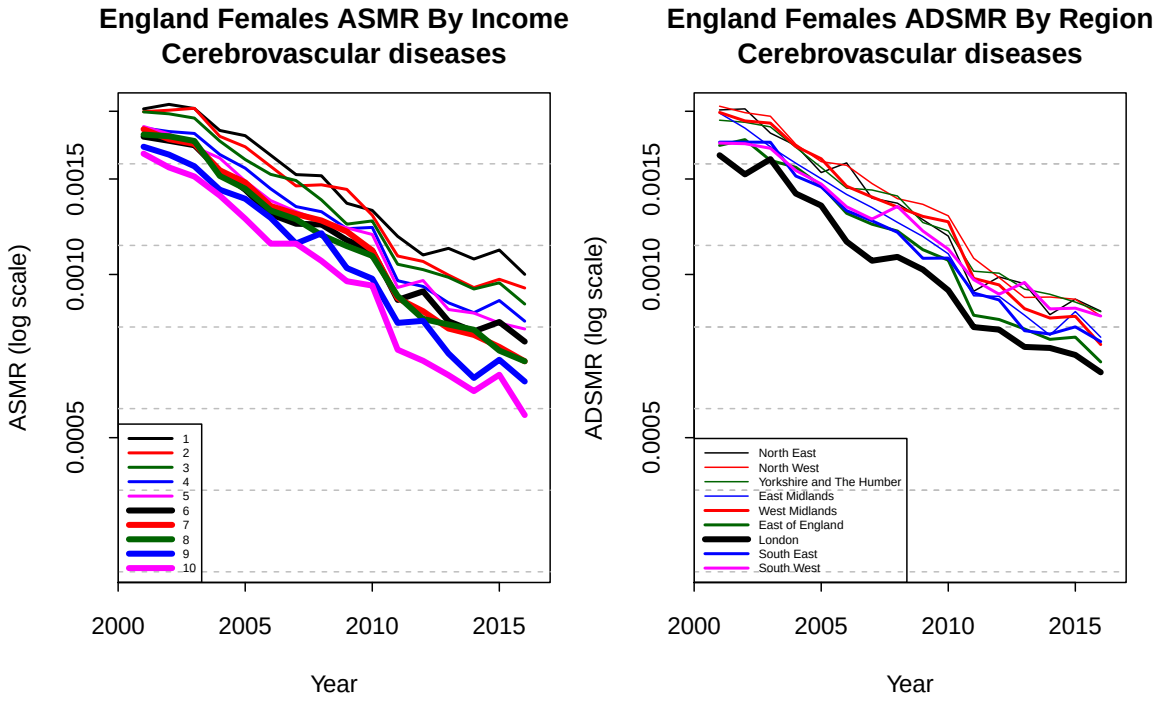

Figure 33: England Females CoD 24: Cerebrovascular diseases

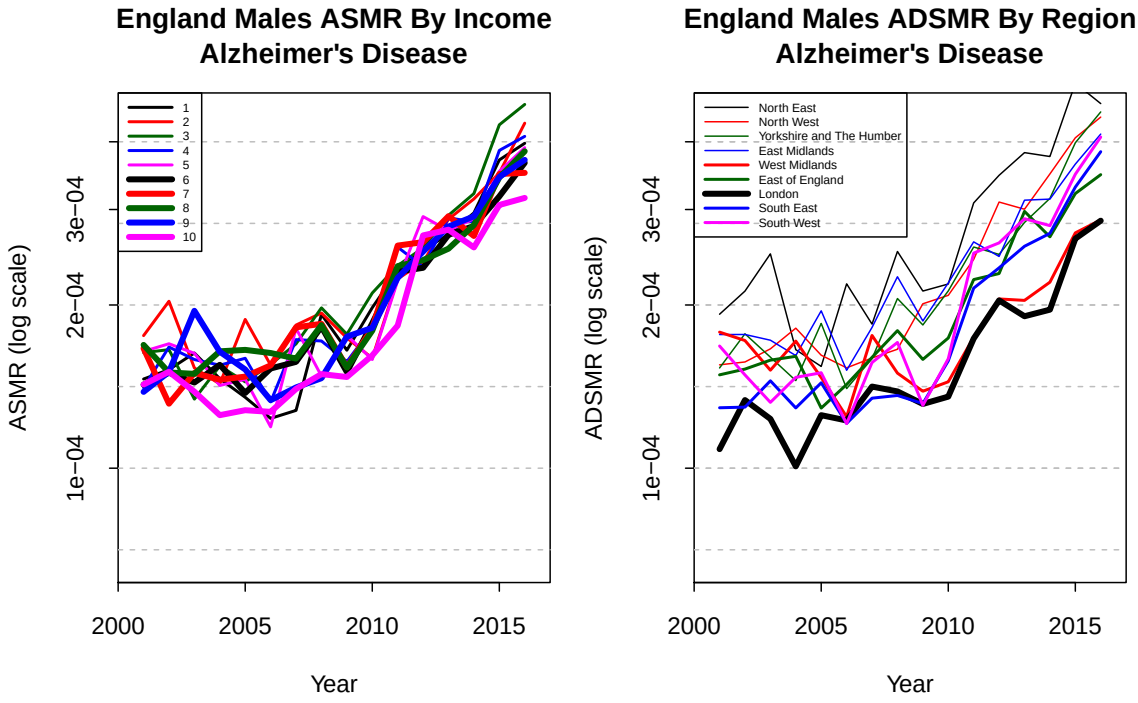

Figure 34: England Males CoD 20: Alzheimer's Disease

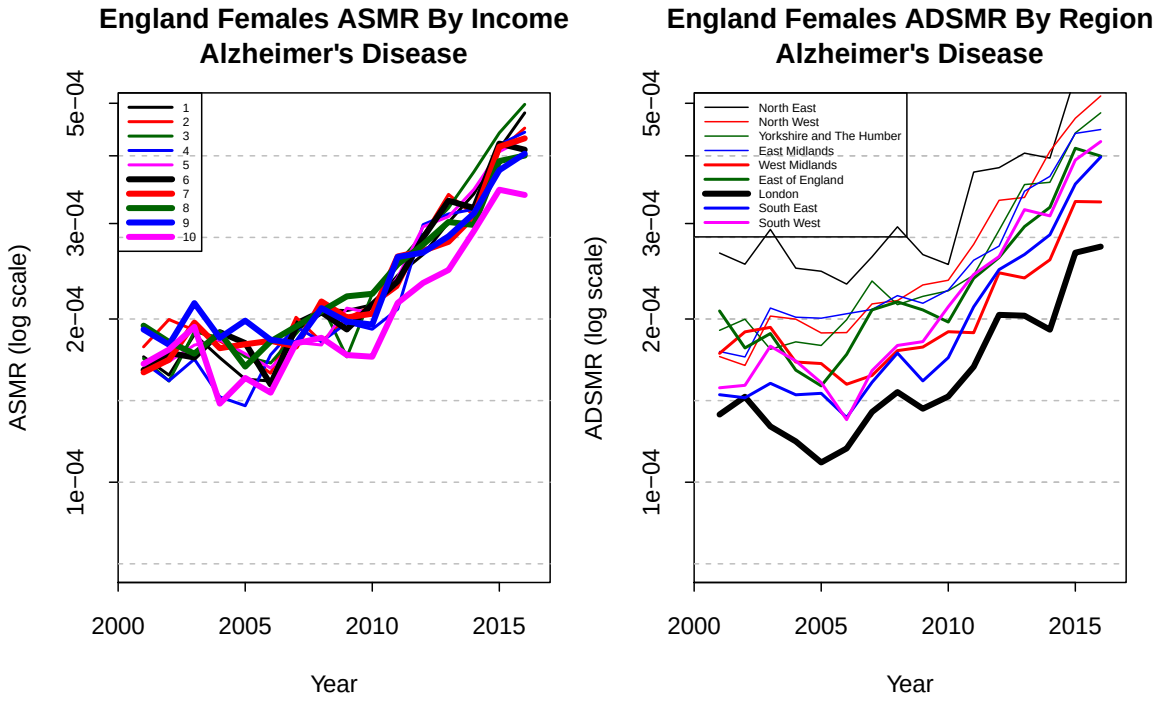

Figure 35: England Females CoD 20: Alzheimer's Disease

The final causes of death we will look at are two of the so-called *deaths of despair*, alcohol-related liver disease and accidental poisoning (including accidental overdose of prescribed pain killers and anti-depressants). This group has been highlighted by Case and Deaton (2015) in the US. Although numbers of deaths are relatively small, the deaths of despair are of major concern as levels of inequality are very high (see Figures 36 and 37)[35] and mortality rates are rising very fast (Figure 37). The rise in death rates from accidental poisoning is very significant.[36] The relentless risk of accidental poisonings contrasts with the flat profile of alcohol-related liver disease, pointing more to more of a pressing need to address the so-called opiod epidemic than alcoholism.

For females (not illustrated here), the picture for alcohol-related liver disease is quite similar to males (but lower). For accidental poisoning, we also see a generally steep increase. However, deciles 9 and 10 stay much more flat than for males and so the inequality gap has widened very significantly.

[35]Note that Figures 36 and 37 are on different scale from other causes of death: being 16× and 32× from top to bottom on the log-scale rather than 8× in the other figures to allow us to capture the much wider spread and the steep increase in accidental poisonings.

[36]In currently unpublished research on Danish data by both education and affluence, we have found that low affluence is hugely more important as a predictor of high mortality rates from deaths of despair than education level. A possible inference is that low income or income deprivation leads to low self esteem leading to alcoholism etc.. In contrast, people who have a low level of education might still be able get good, rewarding work.

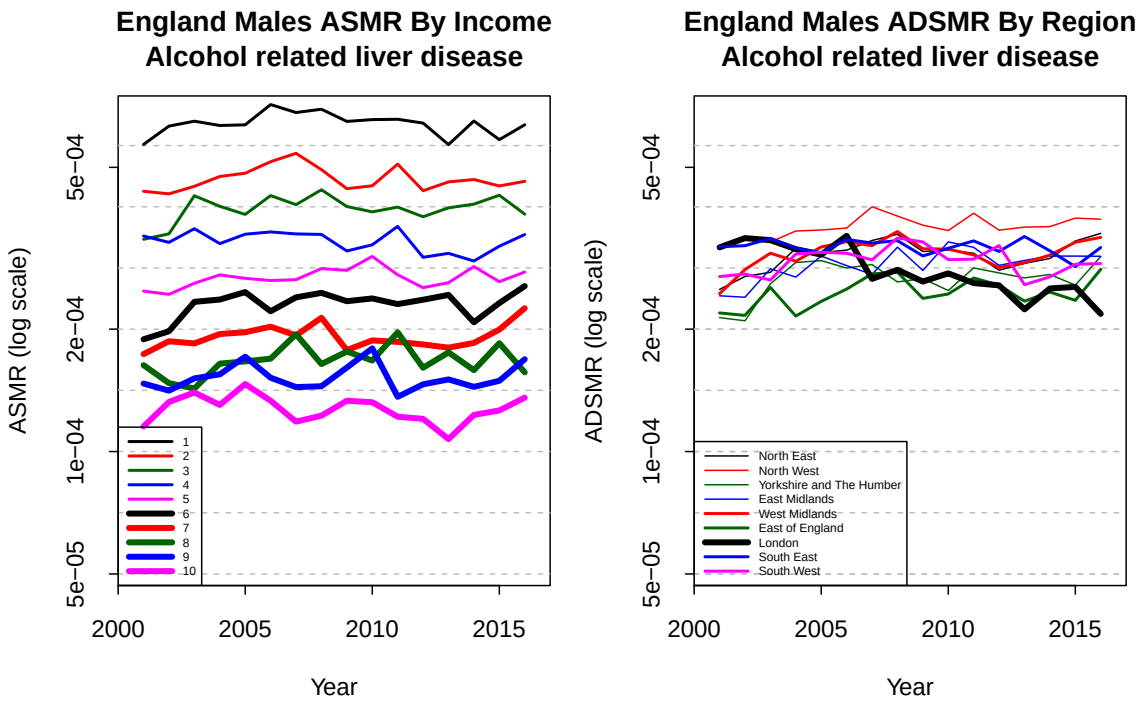

Figure 36: England Males CoD 28: Alcohol related liver disease

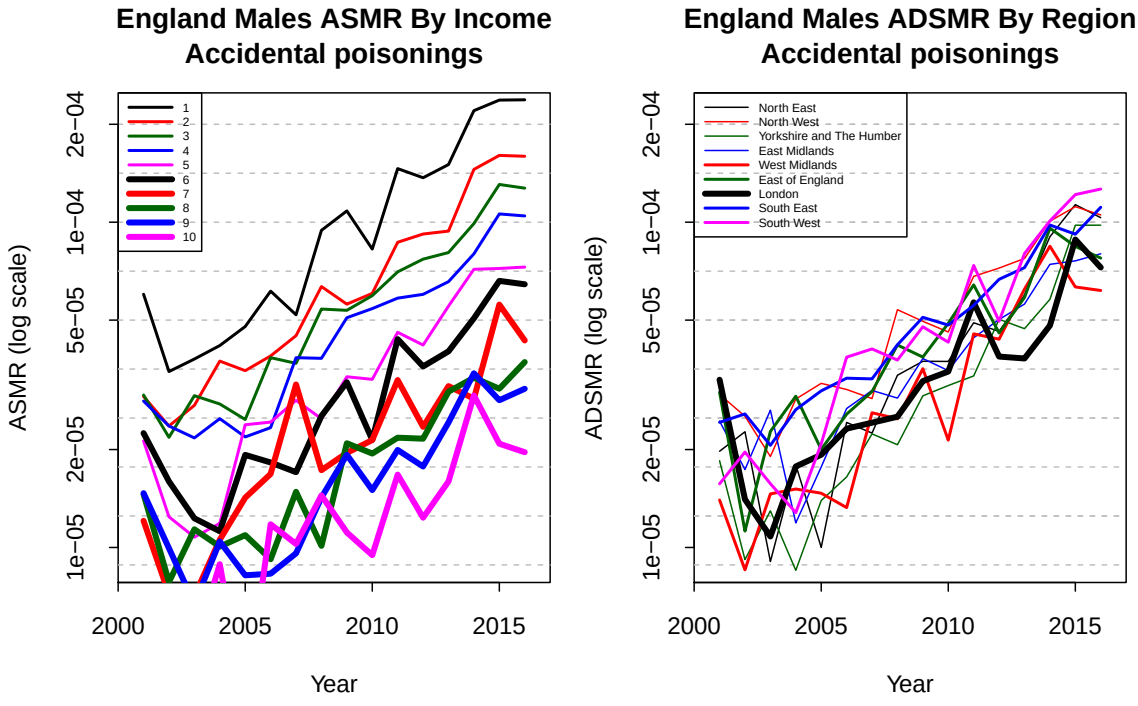

Figure 37: England Males CoD 34: Accidental poisonings

5.3 Contributors to widening inequality

It is of interest to identify which causes of death have contributed most to widening inequality at the all-cause level (Figure 14). Figures 38 and 39 attempt to illustrate this in single plots by charting how mortality in deciles 1 and 10 by cause of death has changed over time. Inequality is measured by looking at how far below the main diagonal each cause of death is. Widening inequality can be inferred when the arrow is moving further away from the main diagonal as we go from 2001 to 2016.

For most of the more-significant causes of death towards the upper right the arrows are moving further away from the main diagonal (grey line). This indicates that over, a sustained period of time, there have been differences in trends in the prevalence of controllable risks. For example, as discussed earlier, for males, smoking prevalence has been falling at a faster rate from an already lower level in the least deprived groups.[37]

5.4 Contributors to mortality improvements

Figures 38 and 39 in conjunction with other plots also allow us to identify which causes of death have been improving and which deteriorating. For males:

- significant improvers include cardiovascular diseases (22, 23, 24, 25), lung cancer and respiratory disease (6, 26), and (for decile 10 but not decile 1) diabetes (17) and other causes of death (32);
- significant worsening can be seen for dementias (18, 20) and accidental poisoning (34).

For females, the picture is similar except for breast cancer (7) as a further improver, and no improvements in lung cancer and respiratory diseases (6, 26).

5.5 Contributors to the slowdown in mortality improvements

We have also touched on the slowdown earlier in the paper. Identifying contributors to the slowdown is not so straightforward especially for less common causes of death with more noisy data. For males, the main contributors seem to be heart diseases (but decile 1 much more than decile 10), respiratory diseases (marginal slowdown), and mental illnesses including dementia.

[37]Since ASMRs cover a wide age range, they slightly mask bigger changes in smoking prevalence by cohort that can be inferred from looking at death rates over narrower age ranges.

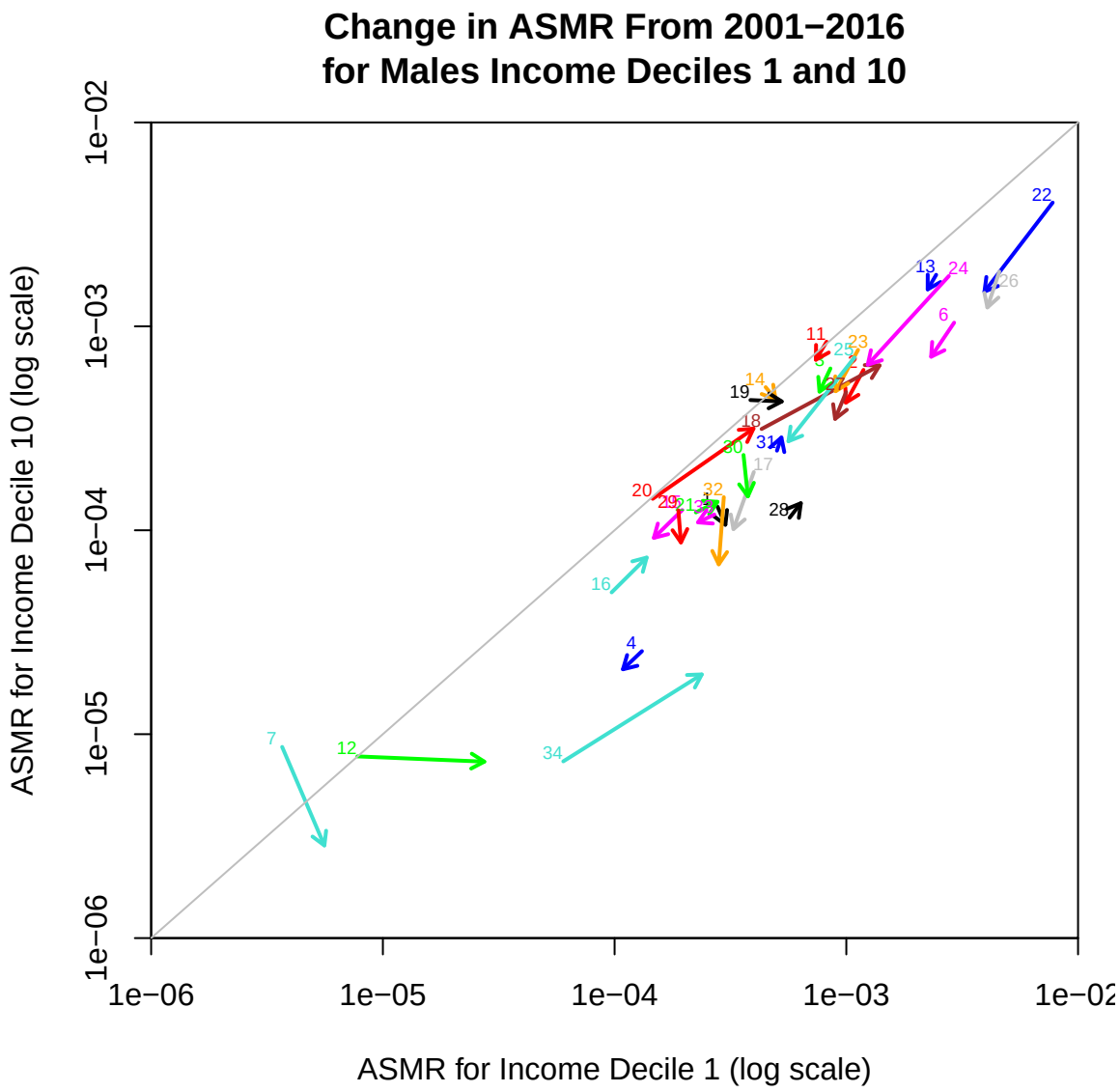

Figure 38: Chart showing ASMRs for income deprivation decile 1 versus decile 10 in 2001 and 2016. Arrows show the direction of travel from 2001 to 2016. Numbering is for causes of death as in Table 1. Colours help match numbers to arrows.

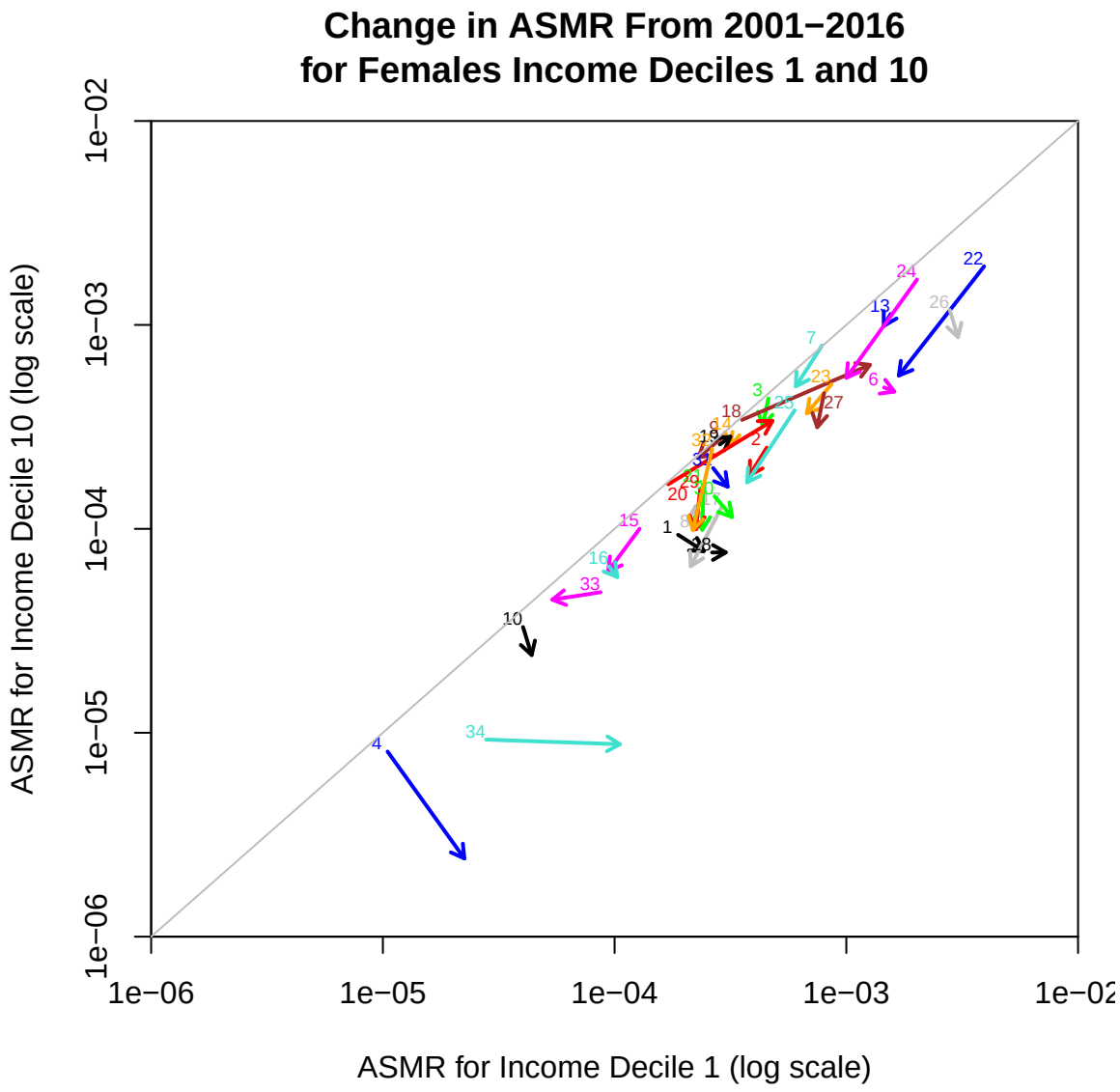

Figure 39: As Figure 39 but for females.

For other causes of death that contribute significantly to all-cause mortality there is no evidence for a change in trend around 2010/11.

5.6 Cause of death rates by age

To complement the time trends in the previous section, we now look at death rates by age group using the year 2016 by way of example. A selection of plots are contained in Figures 40 and 41.[38] The causes of death included are intended to be representative of the more significant causes of death, and exhibit a number of specific patterns. As might be expected, causes of death with significant levels of inequality for both males and females exhibit similar patterns of convergence to Figure 2 at high ages.

For each cause of death, the broad shape by age tends to fall into one of four groups.

- Group 1: curves that, at high ages, are broadly aligned with the Gompertz curve (i.e. log mortality has a similar slope) with rate 0.1 (Figure 40, top two rows; Figure 41, top row), some with high levels of inequality, others with very little.
- Group 2: mainly cancers, with less steep curves at higher ages, but with less consistency than Group 1 in terms of shape and slope (Figure 40, third row; Figure 41, second and third rows).
- Group 3: curves that are much steeper than those in Group 1 (Figure 40, fourth row).
- Group 4: (not plotted) causes of death that either stay flat or affect younger ages more.

It is striking that so many, quite-different, causes of death align with each other in old age in Group 1. The groupings 1, 2 and 3 more generally suggest that there might be different biological mechanisms causing increasing mortality with age: for example, Group 1 being linked to the general ageing process, frailty and failure of organs; Group 2 being linked to the processes that lead to cell mutations; Group 3 being linked to general damage within the brain.

It is, perhaps, notable that cerebrovascular diseases (COD 24) fall into Group 3 rather than with Group 1 where we see the other circulatory and cardiovascular diseases. This allocation to Group 3 might be due to the link between stroke and vascular dementia discussed earlier. It is also notable that road and other accidents (COD 31; males plotted but also females) apparently falls into Group 1, indicating

[38] Again, a full set of plots for all causes of death can be found in the supplementary files on the ARC programme web page `www.macs.hw.ac.uk/~andrewc/ARCresources`.

that increasing frailty in old age goes hand in hand with increased risk of falls and other accidents.

The data here extend only up to age group 85-89. Above this age we might be particularly interested in what happens to Group 3 causes of death given their steep gradient below age 90. We have looked at the Human Cause-of-Death Database (`www.causesofdeath.org`) where, for the much larger US population, there are death rates by cause for 90-94 and 95-99. Below age 90, US rates for vascular dementia and Alzheimer's grow at a fast rate similar to England (i.e. as Group 3), but this gradually slows above age 90: i,e, the gradient falls slowly from 0.2 towards the Group 1 gradient of around about 0.1.

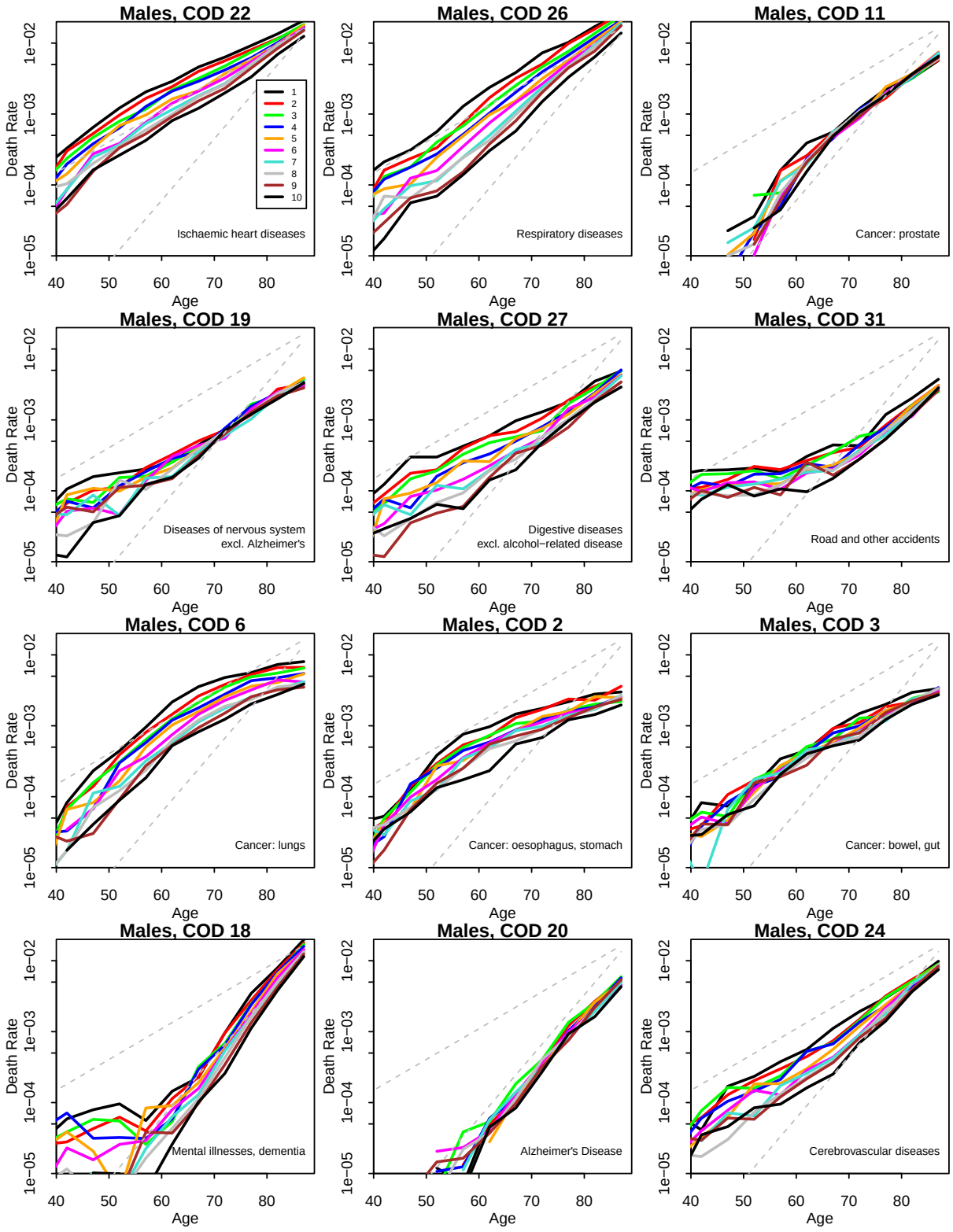

Figure 40: Death rates (log scale) by age and income deprivation decile for twelve causes of deaths for males in 2016. Grey dotted lines are Gompertz curves with rates 0.1 and 0.2. COD numbers as in Table 1.

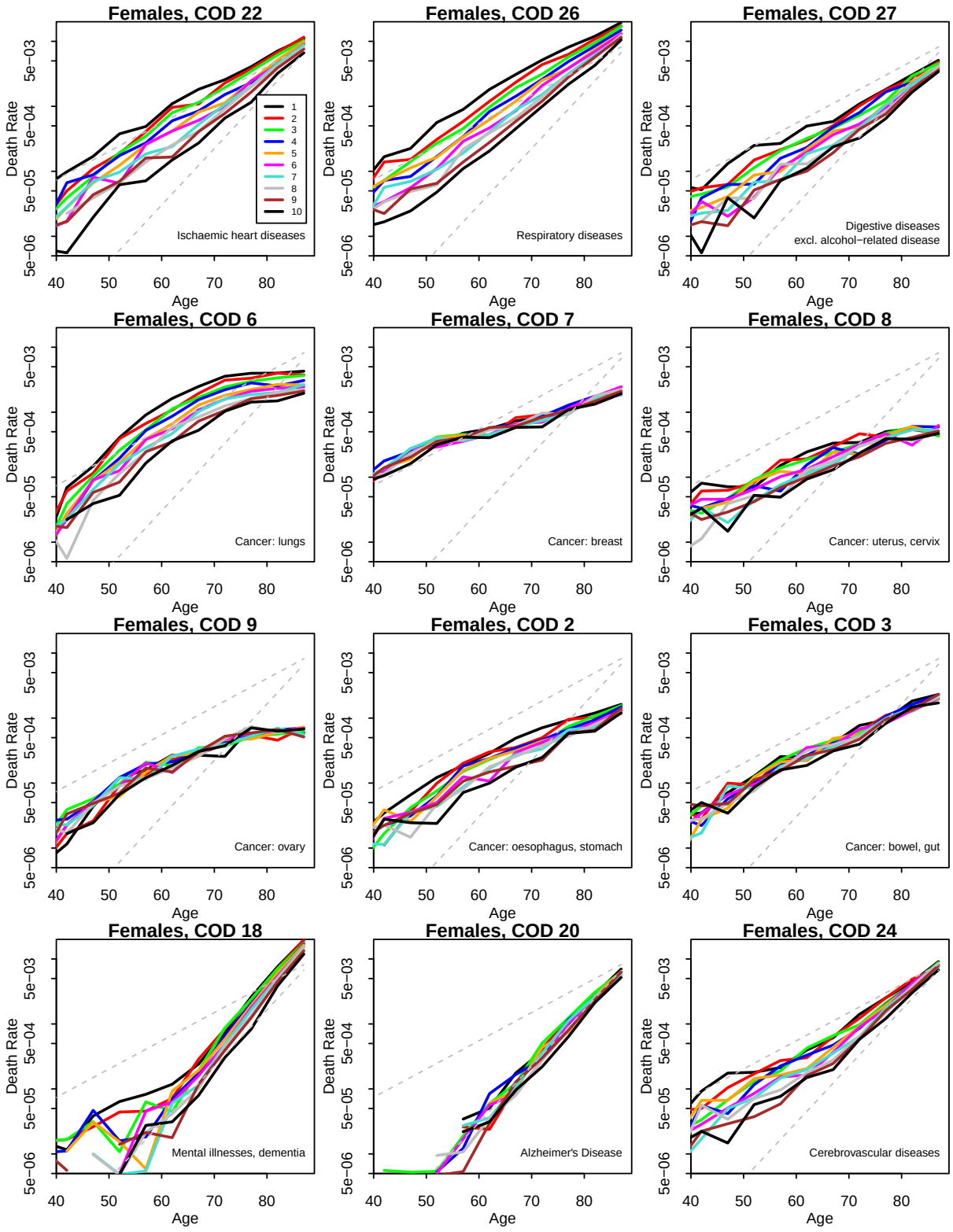

Figure 41: Death rates by age and income deprivation decile for twelve causes of deaths for females in 2016. Grey dotted lines are Gompertz curves with rates 0.1 and 0.2. COD numbers as in Table 1.

6 Conclusions

The increasing availability of large mortality-related datasets opens up the possibility for more detailed analyses of the key drivers of death rates.

In this paper we have conducted a detailed analysis of mortality inequalities in England, using all-cause mortality data at the level of LSOAs and cause-of-death data by region and income-deprivation decile. For all-cause mortality we have used the non-parametric method of local linear regression to quantify more accurately the very significant mortality inequalities that exist across England, particularly at younger ages. The method is very well suited to our large dataset and can handle, in a straightforward way, the inclusion of several predictive variables, and not particularly sensitive to non-linear transformations of the predictive variables. In particular, the method automatically captures any potential interactions between predictive variables.

Amongst all of the available predictive variables, income and employment deprivation were found to have the strongest predictive power. But we also found that urban-rural class and the presence of care homes within a neighbourhood were also important predictors. Once socio-economic effects have been filtered out, residual spatial relative risk was found to be quite small in comparison, countering the headline differences between English regions. Perhaps this is not surprising, but it emphasizes that, on a like for like basis, there is no intrinsic disadvantage in terms of life expectancy to living in the north rather than the south.

The all-cause data also confirm that inequality between different socio-economic groups has been rising including at high ages where, in the past, there has been a ganerally narrower inequality gap.

The methodology leads us to the proposal of the $LIFE$ indices (Longevity Index for England). These can potentially be used in three ways. First, in their "raw" form directly as relative risks. Second, the $LIFE$ indices can be used to group LSOAs into deciles (with a clear improvement over income deprivation as a predictor of high or low mortality). Lastly, the $LIFE$ indices can be used as predictive variables in their own right (on a continuous scale) in the assessment of the mortality of life insurance and pensions portfolios alongside other predictive variables such as pension amount and geodemographic grouping.

In the second part of this paper, we have carried out a detailed empirical analysis of cause of death data. The data being subdivided by region, income deprivation decile and 5-year age groups rather than by LSOA and single ages for all-cause data. We used 34 causes of death to allow us to focus on the links between controllable risk factors (such as smoking) and inequalities in death rates by deprivation decile and region for individual causes of death. We found that:

- whenever a cause of death has a significant controllable risk factor, the higher

the relative risk, the higher the level of mortality inequality;

- whenever a cause of death has no significant controllable risks, levels of mortality inequality tend to be very low *between income deprivation deciles*;
- in three cases (prostate cancer, ovarian cancer and Alzheimer's) with no significant controllable risks, we did find significant and unexplained differences *between regions*: differences that need further investigation.

Mortality inequality can be seen in almost all causes of death. Implicitly this is the result of the link between most causes of death and controllable risk factors. This provides indirect evidence that the most deprived deciles tend to have higher numbers of people who smoke, follow a poor diet, don't get enough exercise, drink too much alcohol etc.. We also found indirect evidence for variation in the prevalence of controllable risks between regions even after adjusting for differences in the deprivation profile of individual regions. However, the rankings of the regions varied from one cause of death to another. This indicates that there is no consistent variation in the prevalence of individual controllable risk factors between regions: that is, for example, regions with the worst levels of smoking might not be the worst in terms of poor diet or exercise.

The implicit link between inequality by cause of death and controllable risk factors points to a key role for public health policy makers alongside researchers pursuing medical advances in delivering future reductions in mortality. Specifically, the data point to how much potential there is in persuading people in more deprived areas to adopt healthier lifestyles.

We also looked at variation in cause of death rates by age. For many common causes of death, death rates in deprived groups are 2 to 4 times rates in less deprived groups in the younger age groups. For so-called "deaths of despair", the gap is much wider. The shapes of mortality curves by age for individual causes of death were found to fall into one of four groups: Group 1 containing causes of death whose mortality curves are approximately Gompertz with rate 0.1 at high ages; Group 2 (mostly cancers) for flatter but still increasing mortality curves; Group 3 Gompertz with rate approximately 0.2; and Group 4 peaking at younger ages. As greater proportions of people reach older ages it is, perhaps, Group 3 that actuaries need to pay attention to (dementias, Alzheimer's and cerebrovascular diseases) as these have mortality curves that are much steeper than causes of death in groups 1 and 2 and become the dominant causes in very old age. Additionally, Group 3 causes of death are on the rise, perhaps because of increased risk of dementia amongst people with a history of cardiovascular events.

Acknowledgements

The authors would like to thank a scrutineer of this paper for their helpful comments.

The authors gratefully acknowledge funding from the Actuarial Research Centre of the Institute and Faculty of Actuaries, the Society of Actuaries and the Canadian Institute of Actuaries through the "Modelling Measurement and Management of Longevity and Morbidity Risk" research programme (see `www.actuaries.org.uk/arc`).

References

Cairns, A.J.G., Kallestrup-Lamb, M., Rosenskjold, C.P.T., Blake, D., and Dowd, K., (2019) Modelling Socio-Economic Differences in the Mortality of Danish Males Using a New Affluence Index. *ASTIN Bulletin*, 49: 555-590.

Cairns, A.J.G., Blake, D., Dowd, K., and Kessler, A.R., (2016) Phantoms Never Die: Living with Unreliable Population Data. *Journal of the Royal Statistical Society, Series A*, 179: 975-1005.

Case, A., and Deaton, A. (2015) Rising morbidity and mortality in midlife among white non-hispanic americans in the 21st century. *Proceedings of the National Academy of Sciences*, 112: 15078-15083.

Cheng, I., Witte, J.S., McClure, L.A., Shema, S.J., Cockburn, M.G., John, E.M., and Clarke, C.A., (2009) Socioeconomic status and prostate cancer incidence and mortality rates among the diverse population of California. *Cancer Causes & Control*, 20: 1431-1440.

Chetty, R., Stepner, M., Abraham, S., Lin, S., Scuderi, B., Turner, N., Bergeron, A, and Cutler, D. (2016) The association between income and life expectancy in the United States, 2001-2014. *Journal of the American Medical Association,* 315: 1750-1766.

Cleveland, W.S. (1979) Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association* 74: 829–836.

Deary, I.J., Weiss, A., and Batty, G.D. (2010) *Intelligence and personality as predictors of illness and death: How researchers in differential psychology and chronic disease epidemiology are collaborating to understand and address health inequalities.* Monograph, University of Edinburgh.

Eurostat (2013) *Revision of the European standard population. Report of Eurostat's task force. 2013 edition.* Luxembourg: Publications Office of the European Union.

Forey, B.A., Thornton, A.J., and Lee, P.N. (2011) Systematic review with meta-analysis of the epidemiological evidence relating smoking to COPD, chronic bron-

chitis and emphysema. *BMC Pulmonary Medicine*, 11: 36.

Gitsels, L.A., Kulinskaya, E., Steel, N., (2016) Survival benefits of Statins for primary prevention: A cohort study. *PLoS ONE*, 11 (11): art. no. e0166847.

HMD (2018) Human Mortality Database: University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de

Jones, G. S., and Baldwin, D. R. (2018) Recent advances in the management of lung cancer. *Clinical Medicine (London)*, 18(Suppl 2): s41–s46.

Kern, M.L., and Friedman, H.S. (2008) Do conscientious individuals live longer? A quantitative review. *Health Psychology* 27: 505-512.

Longevity Science Panel (2018) Life expectancy: Is the socio-economic gap narrowing?
`www.longevitypanel.co.uk/viewpoint/life-expectancy-is-the-socio-economic-gap-narrowing/`
(Accessed 10/12/2019)

Macdonald, A.S., Richards, S.J., and Currie, I.D. (2018) *Modelling mortality with actuarial applications*. Cambridge University Press, Cambridge.

Mackenbach, J.P., Bos, V., Andersen, O., Cardano, M., Costa, G., Harding, S., Reid, A., Hemström, Ö., Valkonen, T., and Kunst, A.E. (2003). Widening socioeconomic inequalities in mortality in six Western European countries. *International Journal of Epidemiology,* 32: 830-837.

Mackenbach, J.P., et al. (2016) Trends in inequalities in premature mortality: a study of 3.2 million deaths in 13 European countries. *Journal of Epidemiology and Community Health,* 69: 207–217.

Malhotra, J., Malvezzi, M., Negri, E., La Vecchia, C., and Boffetta, P. (2016) Risk factors for lung cancer worldwide, *European Respiratory Journal*, 48: 889–902.

Office for National Statistics (2011) Results from the ICD–10 v2010 bridge coding study. *Statistical Bulletin*, 1 February 2011.

Office for National Statistics (2015) English indices of deprivation 2015.
`https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015` (Accessed 3/11/2019)

Office for National Statistics (2018) Population estimates by output areas, electoral, health and other geographies, England and Wales: mid 2017. *Statistical Bulletin*, 25 October 2018.

Pesch, B., Kendzia, B., Gustavsson, P., et al. (2012) Cigarette smoking and lung cancer–relative risk estimates for the major histological types from a pooled analysis of case-control studies. *International Journal of Cancer*, 131: 1210–1219.

Plat, R. (2009). On stochastic mortality modelling. *Insurance: Mathematics and*

Economics, 45: 393-404.

Redondo Lourés, C., and Cairns, A.J.G. (2019) Mortality In The US By Education Level. To appear in *Annals of Actuarial Science*.

Richards, S. J. (2008) Applying Survival Models to Pensioner Mortality Data. *British Actuarial Journal*, 14: 257–303.

Taitt, H.E. (2018) Global Trends and Prostate Cancer: A Review of Incidence, Detection, and Mortality as Influenced by Race, Ethnicity, and Geographic Location. *American Journal of Men's Health.* 12: 1807–1823.

Vang, Z.M., Sigouin, J., Flenon, A., and Gagnon, A. (2017) Are immigrants healthier than native-born Canadians? A systematic review of the healthy immigrant effect in Canada. *Ethnicity and Health*, 22: 209-241.

Villegas, A.M., and Haberman, S. (2014) On the Modeling and Forecasting of Socioeconomic Mortality Differentials: An Application to Deprivation and Mortality in England. *North American Actuarial Journal*, 18: 168-193.

Wen, J., Kleinow, T., and Cairns, A.J.G. (2020) Trends in Canadian Mortality By Pension Level: Evidence From the CPP and QPP. To appear in *North American Actuarial Journal*.

Wen, J., Cairns, A.J.G., and Kleinow, T., (2019) Fitting Multi-Population Mortality Models to Socio-Economic Groups. Working paper, Heriot-Watt University.

A Datasets

Data for England are available at the level of small geographical areas known as Lower Layer Super Output Areas (LSOAs). Each area has typically between 1,000 and 3,000 persons, with an average of about 1,500, across all ages.

There are 32,844 LSOAs at present. New LSOAs are created from time to time in response to growth in housing. Data relating to the LSOAs can be found on the ONS (Office for National Statistics) website www.ons.gov.uk.

A.1 Deaths and exposures

For each LSOA we have:

- data from 2001-2016:
  - death counts, $D(g, i, t, x)$, where g is the gender, i is the LSOA, t is the

year, and x is the age last birthday;[39] [40]

– central exposed to risk, $E(g, i, t, x)$, equated to the mid-year population estimates for 2001-2016 available from the ONS;[41]

- One off (i.e. not observed through time) *predictive variables* for each LSOA that might be associated with higher or lower than average rates of mortality.

A.2 Potential predictive variables and related data

1. LSOA index.

   - LSOA codes are of the form "E010xxxxx" where the LSOA index *xxxxx* ranges from 00001 to 33768.
   - Only 32,844 indexes are currently in use and, therefore, some codes are missing. These are codes that would have been used previously. However, if an LSOA has grown substantially, then it would be split, the old LSOA code deleted, and the two new LSOAs given new codes not yet used. And some LSOAs have shrunk and will have been merged and allocated a new index.

2. The Index of Multiple Deprivation (IMD)[42]

   This is the official composite measure of relative deprivation in England, with a single value for each LSOA. A higher value indicates a higher level of deprivation. The IMD has seven domains:

   - income deprivation;
   - employment deprivation;
   - education, skills and training;
   - health deprivation and disability;

[39] User-requested deaths data used in this study can be found at `https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/adhocs/007807deathsbylowersuperoutputareaageandsexenglandandwales2001to2016`

[40] Death counts are *registrations* in calendar year t rather than *occurrences*. The advantage to the ONS and users of using registrations is that the tables can be produced in a much more timely manner. Death counts by year of occurrence can be delayed by the very small number of deaths that require an inquest.

[41] Mid-year population estimates can be found at `https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/lowersuperoutputareamidyearpopulationestimates`. Note that data for 2012-2016 have been revised slightly since our original download by the ONS to account for revisions made to local authority population estimates (ONS, 2018). Unrevised files for 2015 and 2016 are available on the same web page. The authors have verified that use of the revised data make very little difference to estimates of relative risk, $F_1(i)$ and $F_2(i)$.

[42] For further details, see Office for National Statistics (2015).

- crime;
- barriers to housing and services;
- living environment.

Some of these have further sub-domains (which we discuss below) that we consider to be useful to refine predictions of mortality.

3. Income deprivation (a domain of the Index of Multiple Deprivation (IMD)):
    - this measures the proportion of the population in each LSOA who are receiving benefits from the state because they are on a low income;
    - the data are in a vector of length 32,844: one entry for each LSOA;
    - sub-domains include *income deprivation affecting older people*, which measures income deprivation amongst people aged 60 and older.
4. Employment deprivation (a domain of the IMD)
    - this measures the proportion of the *working* population in each LSOA who are unemployed;
    - the data are in a vector of length 32,844: one entry for each LSOA corresponding to the vector of 5-digit LSOA codes above.
5. Living environment deprivation (a domain of the IMD)
    - this measures the quality of the living environment (indoors and outdoors);
    - indoors: (poor) quality of housing;
    - outdoors: e.g. (poor) air quality and traffic accidents;
    - the data are in a vector of length 32,844: one entry for each LSOA.
6. Barriers to housing and services (a domain of the IMD)
    - like living environment deprivation, this measures a number of different things;
    - this measures 'wider barriers' and 'geographical barriers';
    - wider barriers includes overcrowding in households and homelessness;
    - geographical barriers measures distance to key services;
    - although a higher value for geographical barriers implies more 'deprived', it can also be associated with lower mortality; for example, greater distances to services might indicate that the LSOA is more affluent or rural with housing more spaced out; in fact, the geographical barriers variable is negatively correlated with income deprivation;

- the data are in a vector of length 32,844: one entry for each LSOA;
- and data are available separately for wider barriers and geographical barriers.

7. Average number of bedrooms

   - this measures the average number of bedrooms per household in the LSOA
   - the data vector has been standardised to a $N(0, 1)$ distribution;
   - in contrast to the deprivation indices, a high value (more bedrooms) is likely to be associated with lower mortality;
   - the data are in a vector of length 32,844: one entry for each LSOA.

8. Highest level of qualification:

   - this gives the proportion of a particular group within the LSOA who have attained a particular level of education
   - data are in the form of a 3-dimensional array for males and females combined

     lsoa x age-group x education level $(32,844 \times 6 \times 8)$
   - 6 age groups: All; 16 to 24; 25 to 34; 35 to 49; 50 to 64; 65 plus;
   - 8 education groups:
     (a) All categories: Highest level of qualification
     (b) No qualifications
     (c) Level 1 qualifications
     (d) Level 2 qualifications
     (e) Apprenticeship
     (f) Level 3 qualifications
     (g) Level 4 qualifications and above
     (h) Other qualifications
   - see

     `www.gov.uk/what-different-qualification-levels-mean/list-of-qualification-levels`;
   - you can use the education data to construct a vector of predictive variables: e.g.
     - the proportion of people in the LSOA aged 50-64, who have no qualification or level 1 only;
     - an average level of educational attainment in a particular age group;

9. Occupation group proportions

- gives the proportion of a particular group within the LSOA who have a particular type of occupation
- data are in the form of a 4-dimensional array
  gender x lsoa x age-group x occupation group $(2 \times 32,844 \times 14 \times 9)$
- 14 age groups: All; 16-19; 20-24; 25-29; 30-34; 35-39; 40-44; 45-49; 50-54; 55-59; 60-64; 65-69; 70-74; 35-64
- most age groups are small, so there will be a lot of sampling variation, weakening their predictive ability. This is less of a problem for the 35-64 age group.
- 9 occupation groups
  (a) Higher managerial, administrative and professional occupations
  (b) Lower managerial administrative and professional occupations
  (c) Intermediate occupations
  (d) Small employers and own account workers
  (e) Lower supervisory and technical occupations
  (f) Semi-routine occupations
  (g) Routine occupations
  (h) Never worked, long-term unemployed and full-time students
  (i) Total: NS-SeC
- you can use the occupation data to construct a vector of predictive variables: e.g.
  – the proportion of people in the LSOA aged 35-64, who fall into the "higher managerial" group;

10. Urban-Rural Classification

    1 Conurbation: non London
    2 City or town
    3 Rural town
    4 Rural village and dispersed
    5 Conurbation: London

    - the data are in a vector of length 32,844: one entry for each LSOA.

11. Region

    1 North East
    2 North West
    3 Yorkshire and Humber

4 East Midlands

5 West Midlands

6 East

7 London

8 South East

9 South West

- the data are in a vector of length 32,844: one entry for each LSOA.

12. Communal establishments

- This element of the data (a user-requested dataset from the ONS) record the number of persons in each LSOA in a communal establishment at the time of the 2011 census.
- The data count the number of persons, $C(i, g, y, \tau)$ where
  - i is the LSOA index;
  - g is gender;
  - y is the age group 0-59, and 60+;
  - τ is the type of communal establishment:
    1 Care home: Private or local authority, with nursing;
    2 Care home: Private or local authority, without nursing;
    3 Remainder of medical and care establishments;
    4 Other communal establishments.

B ICD10 codes

	Informal description: primary cause of death	ICD10 Codes
1	Infectious diseases (including TB and HIV/AIDS)	A00-B99
2	Cancer in mouth, gullet, and stomach	C00-C16
3	Cancer in gut or rectum	C17-C21
4	Cancer in larynx	C32
5	Cancer in trachea	C33
6	Cancer in bronchus or lungs	C34
7	Breast cancer	C50
8	Cancer in femal genital organs: uterus, cervix and neighbourin	C53-C55
9	Cancer in femal genital organs: ovary	C56
10	Cancer in femal genital organs: other	C51-C52, C57-C58
11	Cancer in male genital organs: prostate cancer	C61
12	Cancer in male genital organs: other	C60, C62-63
13	Cancer in bones, skin, and other locations	C22-C31, C37-C39, C40-C49, C64-C80, C97, D00-D09
14	Cancer in lymphatic or blood-forming tissues	C81-C96
15	Benign tumours or tumours without specification	D10-D49
16	Diseases in blood and blood-forming organs	D50-99, E00-E07, E15-E90
17	Diabetes	E08-E14
18	Mental illnesses	F00-F99
19	Diseases in the nervous system, excluding Alzheimers	G00-G29, G32-G99, H00-H99
20	Alzheimers and related degenerative diseases	G30-G31
21	Increased blood pressure or rheumatic fever	I00-I16
22	Ischaemic heart diseases	I20-I25
23	Other heart diseases	I26-I28, I30-I52
24	Cerebrovascular diseases	I60-I69
25	Circulatory diseases	I70-I99
26	Lungs and breathing diseases	J00-J99
27	Digestive diseases excluding alcohol liver disease and cirrhosis	K00-K69, K71-K72, K75-K99
28	Alcohol related liver disease and cirrhosis	K70, K73-K74
29	Diseases in skin, bones, and connective tissue	L00-L99, M00-M99
30	Urine, kidney, genital organs, and breast gland diseases	N00-N99
31	Road and other accidents	V00-V99, W00-W99, X00-X39, X50-X59, Y40-Y86, Y87.1-Y89, Y90-Y99
32	Other causes of death	O00-O99, P00-P99, Q00-Q99, R00-R99, U00-U99, X85-X99, Y00-Y09
33	Suicide	X60-X84, Y10-Y34, Y87.0
34	Accidental poisonings	X40-X49
35	Other factors not used as the primary cause of death	S00-S99, T00-T99, Z00-Z99

Figure 42: Table of ICD codes for the 35 cause of death groupings. Group 35 is for completeness, but not used as a cause of death.

C Age Standardised Mortality Rate (ASMR)

C.1 Basic definition

The purpose of the ASMR is to facilitate comparison of mortality rates in different populations. In particular, if the age profiles of different populations are different then some measures (e.g. deaths per 100,000 population) might simply reflect differences in the age profile even if death rates at individual ages are identical. The ASMR avoids this by using a standard population rather than the actual age profile.

The basic definition, ignoring other indices, over the age range (x_0, x_1) is

$$ASMR(x_0, x_1) = \frac{\sum_{x=x_0}^{x_1} m(x)ES(x)}{\sum_{x=x_0}^{x_1} ES(x)}$$

where $ES(x)$ is the standard population at age x[43], and $m(x)$ is the death rate at age x.

C.2 Further development

In this paper we make use of a number of variants of the ASMR. Various age ranges are considered: e.g. 40-49, 40-64, 65-89, etc. We also calculate ASMR's be region and by deprivation or other deciles; and for all-cause mortality and by cause-of-death.

Suppressing indices for the age range, (x_0, x_1), the cause of death, c,[44] calendar year, t, and gender, g, we start with the following death rates:

- $m(x)$ = national death rate at age x,
- $m_I(i, x)$ = income-deprivation decile i death rate,
- $m_R(r, x)$ = region r death rate,
- $m_{RI}(r, i, x)$ = death rate at age x in region r, deprivation decile i.

[43] Here we use the European Standard Population, 2013 (Eurostat, 2013). Comparative results in this paper for different populations are unlikely to be sensitive to the choice of standard population.

[44] For ease of notation, we consider $c = 0$ to be all cause mortality, and the formulae for the different ASMRs by cause of death apply equally well.

Each of these has corresponding ASMRs:

$$\begin{aligned} ASMR \equiv ASMR(t) &= \frac{\sum_{x=x_0}^{x_1} m(x)ES(x)}{\sum_{x=x_0}^{x_1} ES(x)} \\ ASMR_I(i) &= \frac{\sum_{x=x_0}^{x_1} m_I(i,x)ES(x)}{\sum_{x=x_0}^{x_1} ES(x)} \\ ASMR_R(r) &= \frac{\sum_{x=x_0}^{x_1} m_R(r,x)ES(x)}{\sum_{x=x_0}^{x_1} ES(x)} \\ ASMR_{RI}(r,i) &= \frac{\sum_{x=x_0}^{x_1} m_{RI}(r,i,x)ES(x)}{\sum_{x=x_0}^{x_1} ES(x)} \end{aligned}$$

For all-cause mortality ($c = 0$) we will be summing over single ages. For cause-of-death mortality ($c > 0$), we will be summing over 5-year age groups: for example if the stated age range is 40 to 49, then we are summing over two age groups: 40-44 and 45-49.

C.3 The Age and Deprivation Standardised Mortality Rate (ADSMR)

Now, we can develop the formula for $ASMR_R(r)$ as follows:

$$ASMR_R(r) = \frac{\sum_{x=x_0}^{x_1} ES(x) \sum_{i=1}^{10} m_{RI}(r,i,x) w_{RI}(r,i,x)}{\sum_{x=x_0}^{x_1} ES(x)} \tag{5}$$

where $w_{RI}(r,i,x) = E_{RI}(r,i,x) / \sum_j E_{RI}(r,j,x)$ represents the weight carried by decile i amongst deciles 1 to 10, in region r at age x (so, for each r and x, $\sum_i w_{RI}(r,i,x) = 1$).

We then note than some regions have greater proportions of more deprived areas than other regions. The resulting differences in weights then means that some regions will have naturally higher ASMRs even if there are no differences in death rates at the level of income deprivation between regions (i.e. $m_{RI}(r,i,x) = m_I(i,x)$ for all $r = 1, \ldots, 9$).

To remedy this we propose the ADSMR as an alternative to the regional $ASMR_R$s. Specifically we replace the weights $w_{RI}(r,i,x)$ in 5 by $\tilde{w}_{RI}(r,i,x) = 0.1$. Hence

$$\begin{aligned} ADSMR(r) &= \frac{\sum_{x=x_0}^{x_1} \sum_{i=1}^{10} m_{RI}(r,i,x) \tilde{w}_{RI}(r,i,x) ES(x)}{\sum_{x=x_0}^{x_1} ES(x)} \\ &= \frac{\frac{1}{10} \sum_{i=1}^{10} \sum_{x=x_0}^{x_1} m_{RI}(r,i,x) ES(x)}{\sum_{x=x_0}^{x_1} ES(x)} \\ &= \frac{1}{10} \sum_{i=1}^{10} ASMR_{RI}(r,i). \end{aligned}$$

The use of the ADSMR allows us to filter out the impact of differences in deprivation levels. Any differences that remain need further investigation.

D Supplementary plots

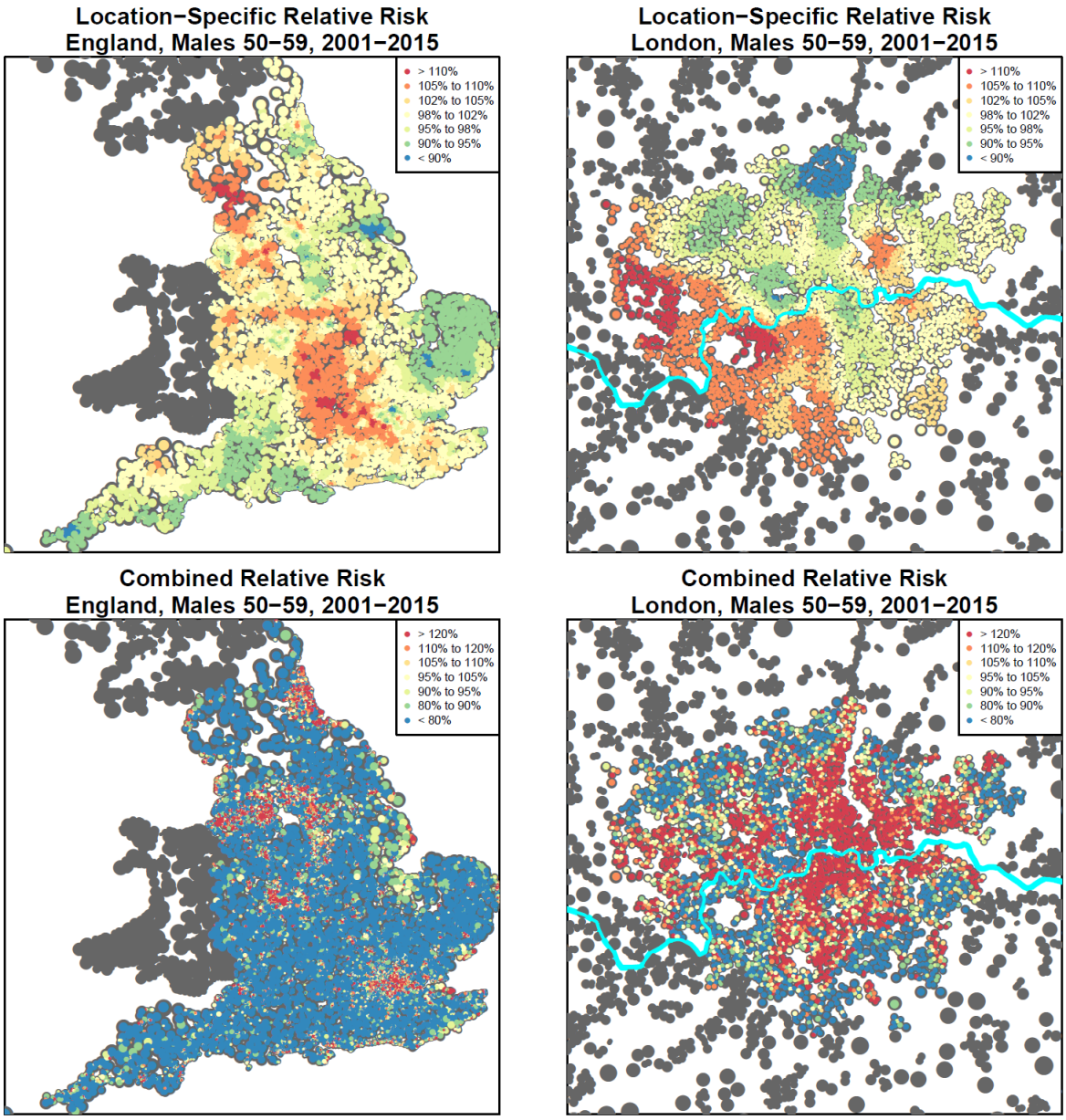

Figure 43: Top row: estimated spatial relative risk, $F_2(i)$, by LSOA for England (left) and London (right) for males, ages 50-59. Bottom row: combined relative risk, $F_1(i)F_2(i)$. Dot sizes reflect the physical size of each LSOA.

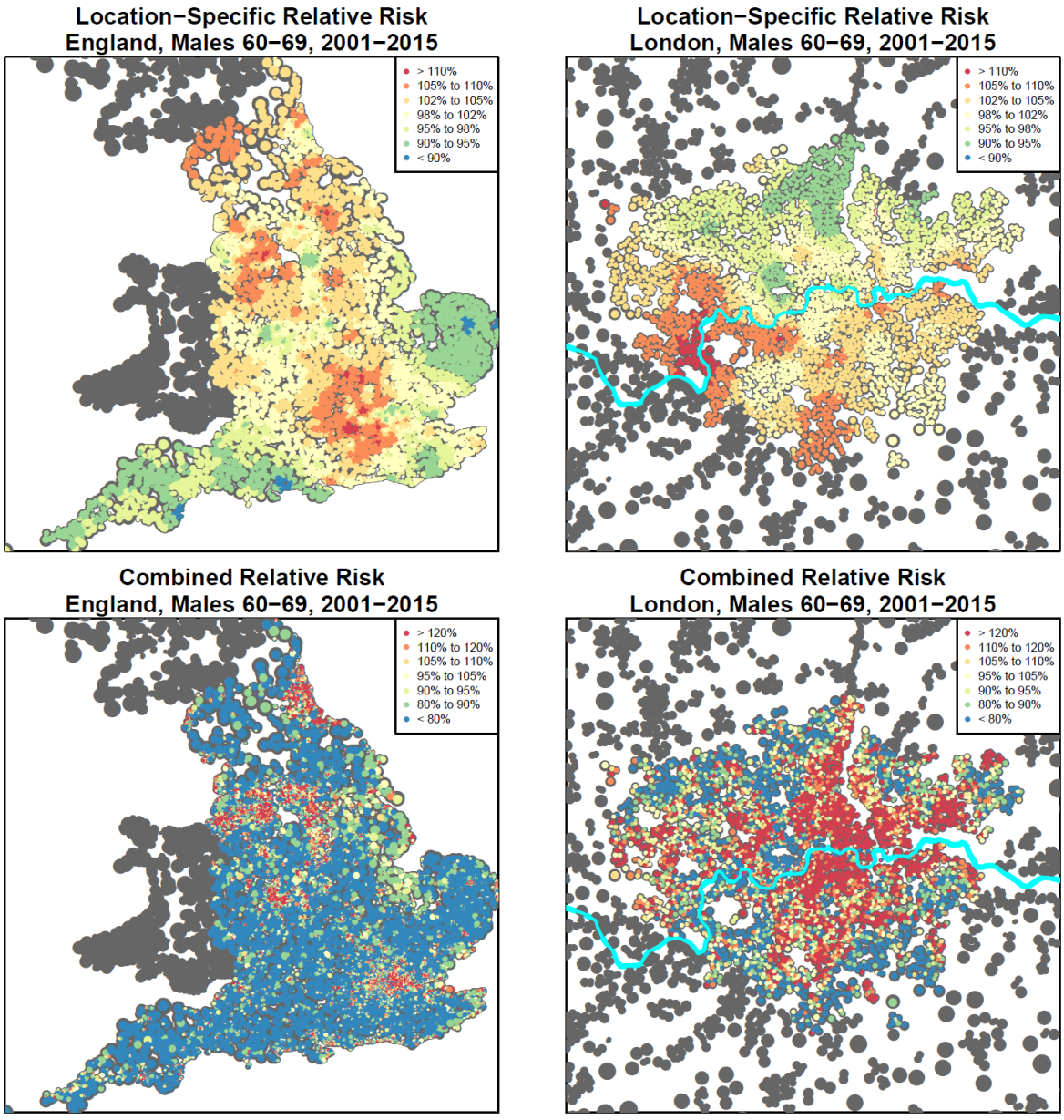

Figure 44: As Figure 43 but for ages 60-69.

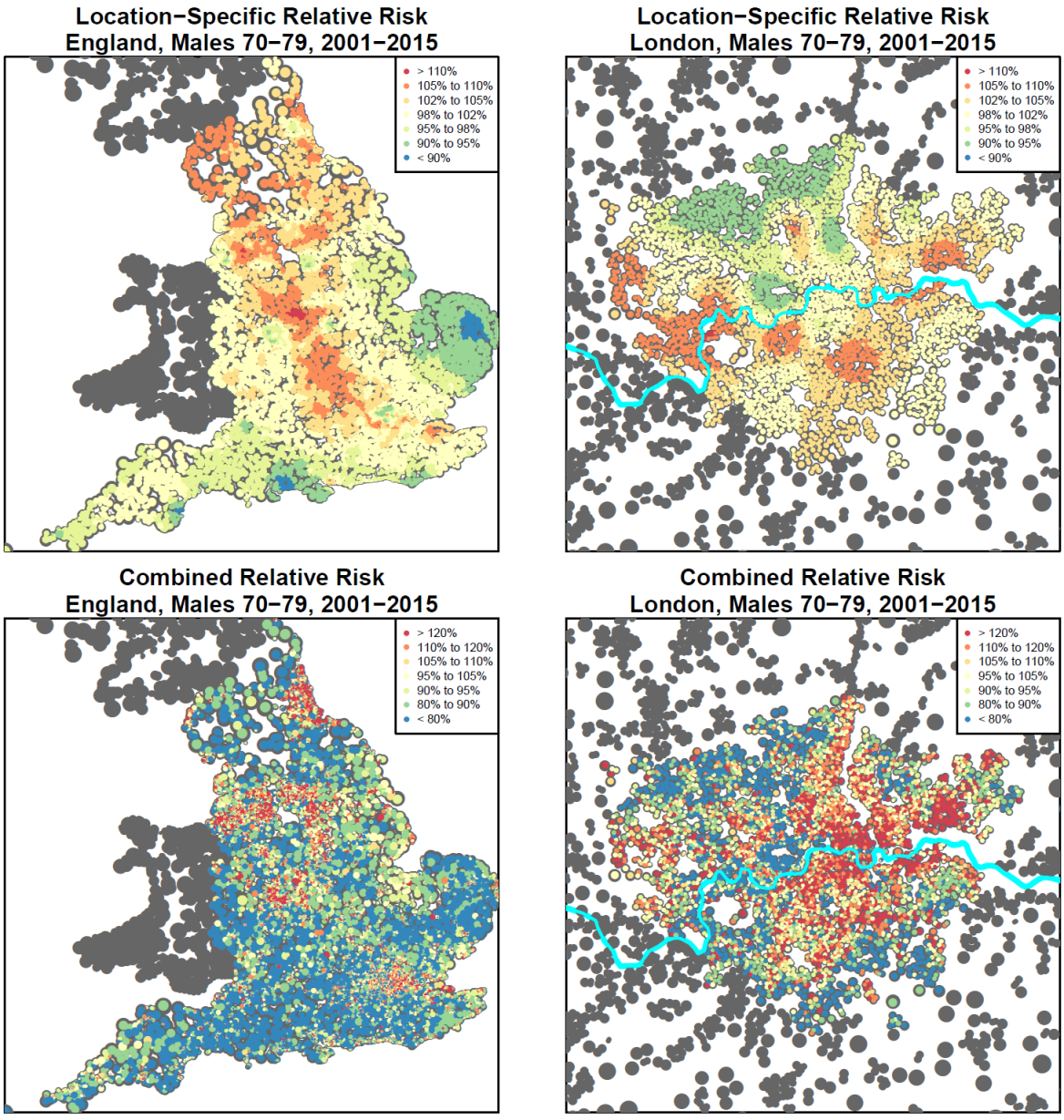

Figure 45: As Figure 43 but for ages 70-79.

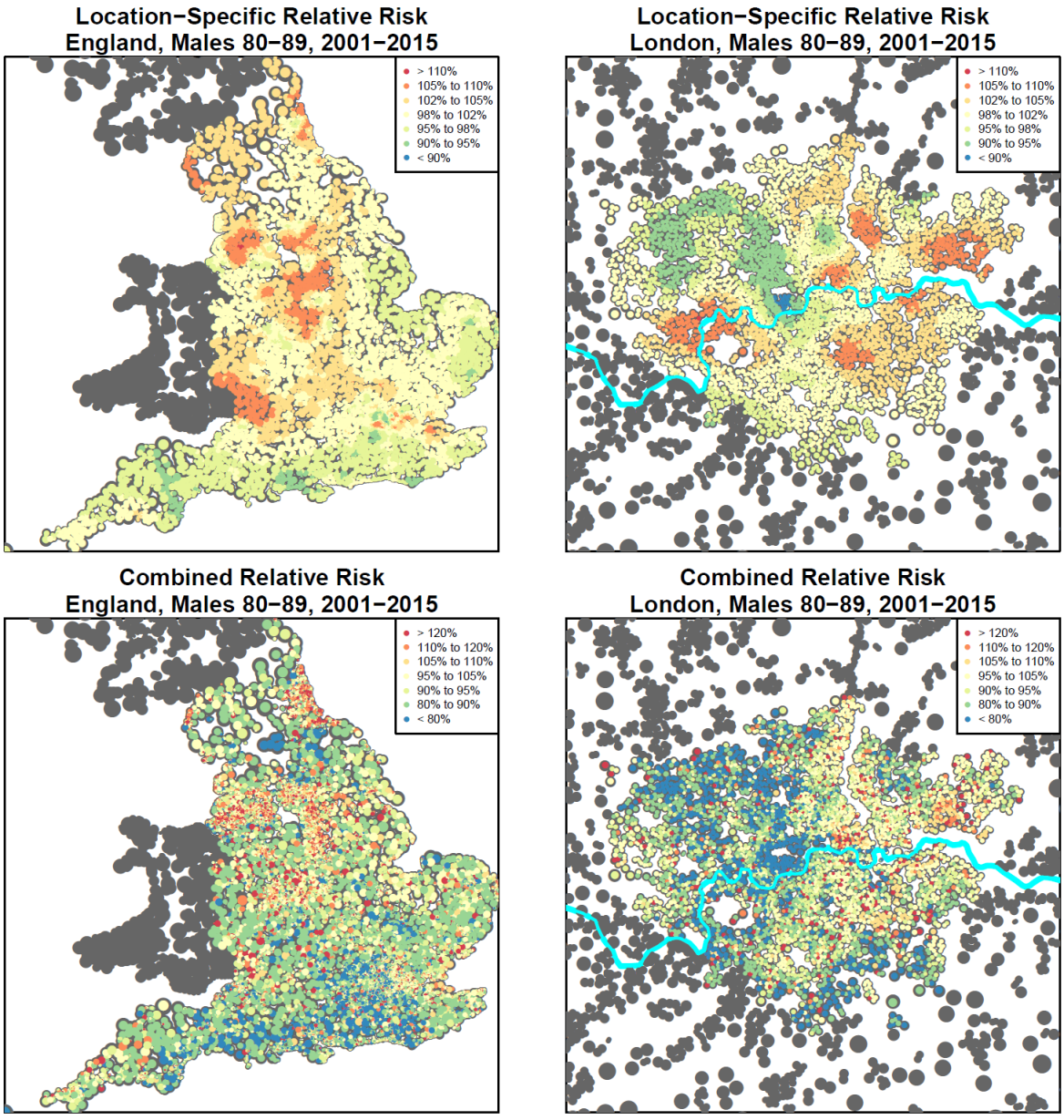

Figure 46: As Figure 43 but for ages 80-89.

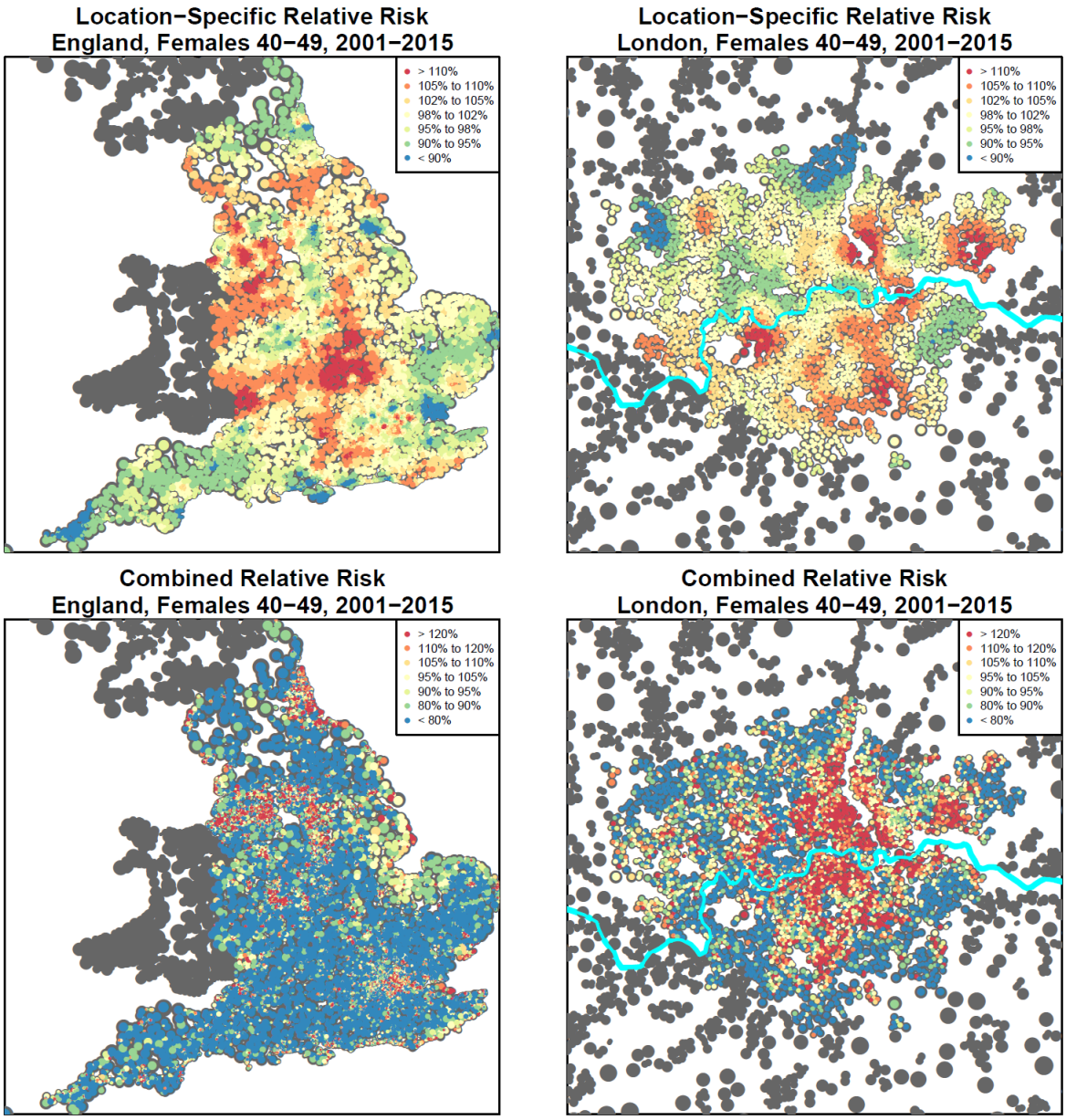

Figure 47: Top row: estimated spatial relative risk, $F_2(i)$, by LSOA for England (left) and London (right) for females, ages 40-49. Bottom row: combined relative risk, $F_1(i)F_2(i)$. Dot sizes reflect the physical size of each LSOA.

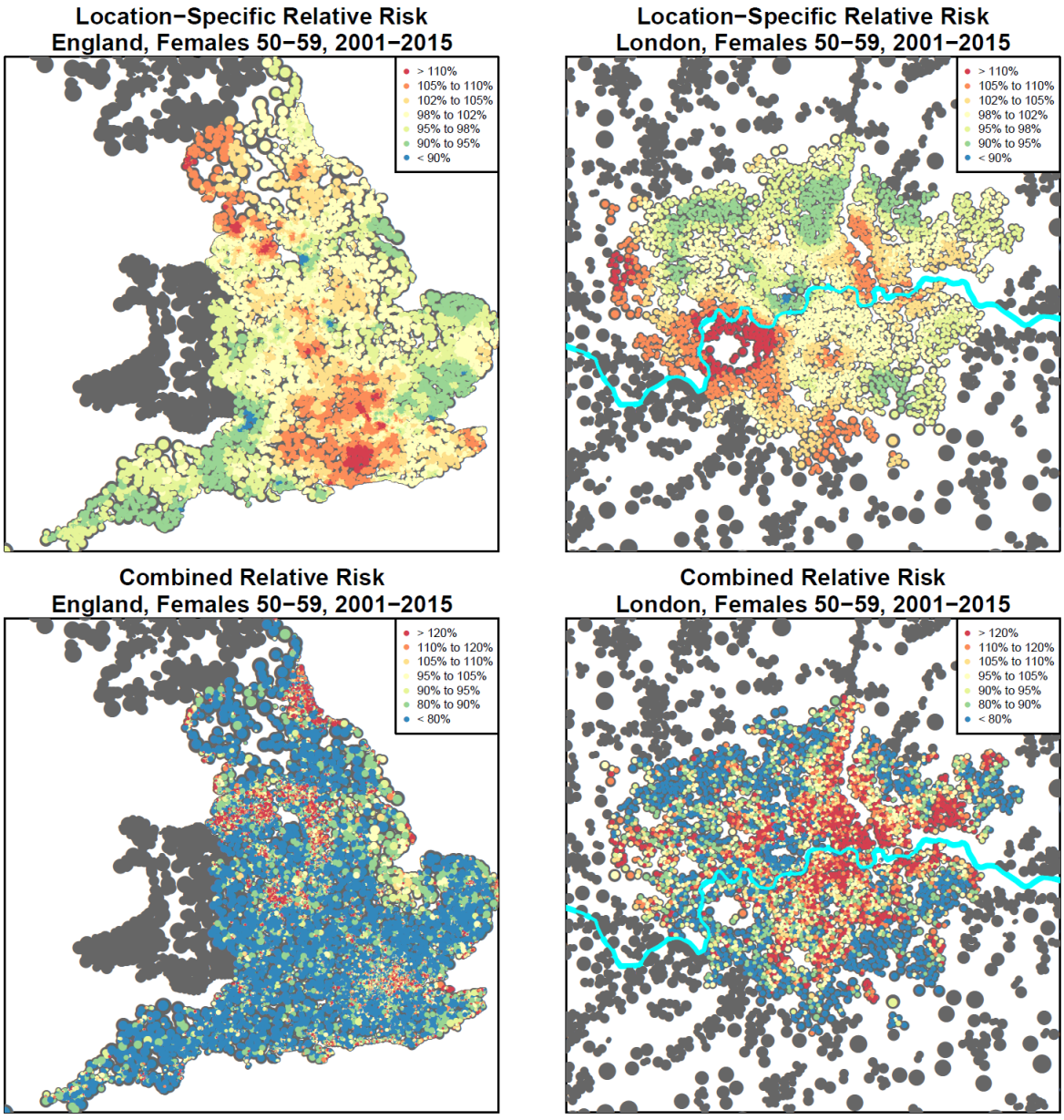

Figure 48: As Figure 47 but for ages 50-59.

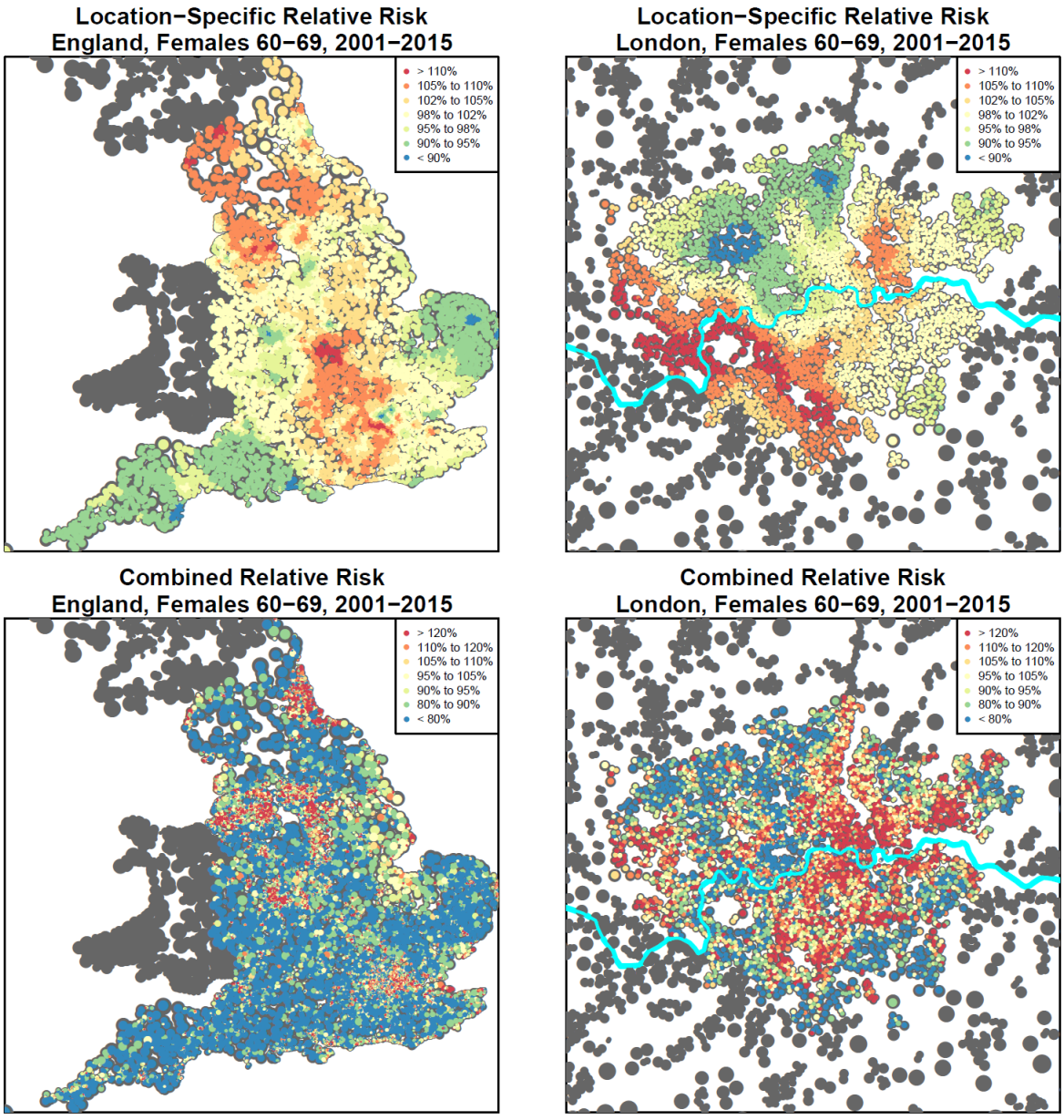

Figure 49: As Figure 47 but for ages 60-69.

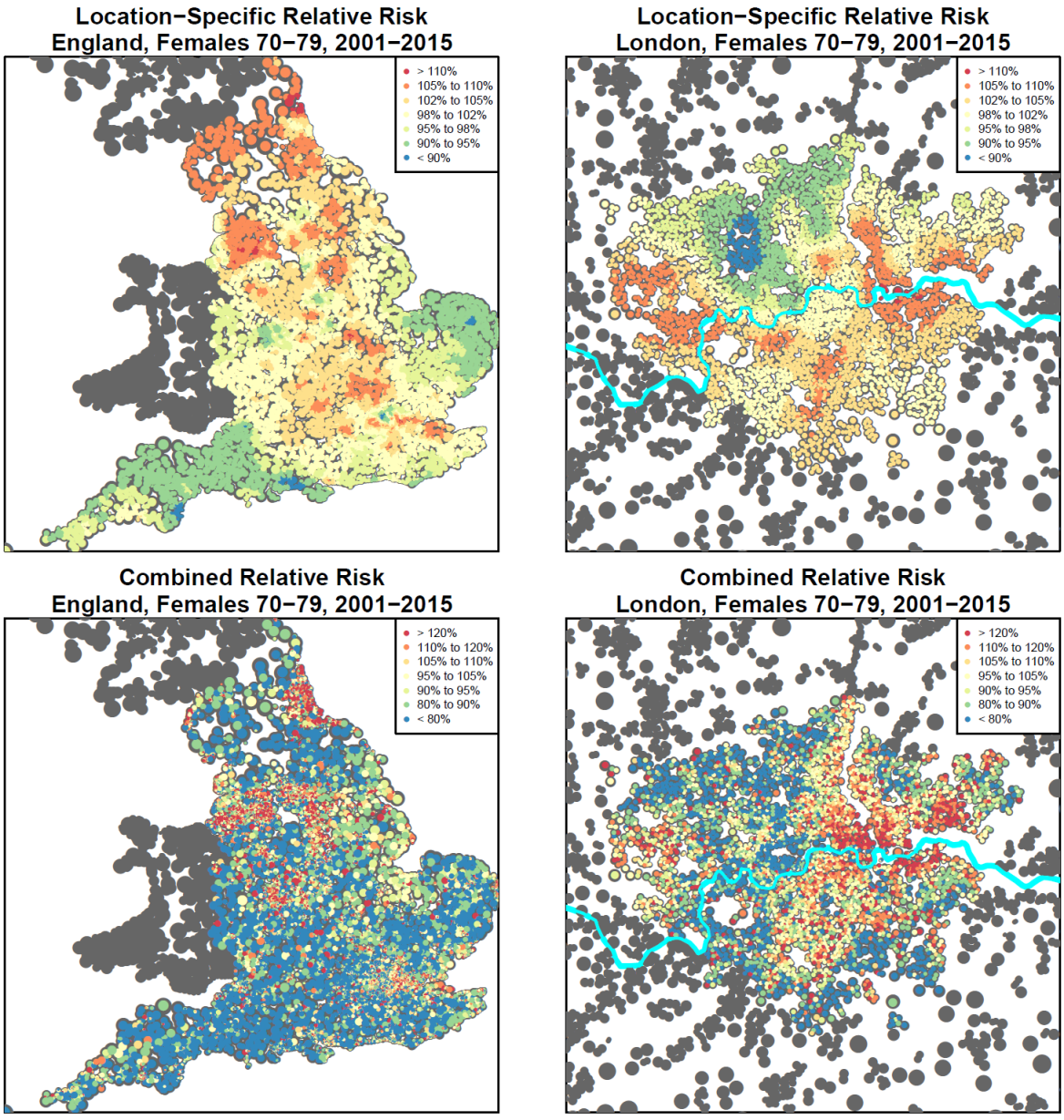

Figure 50: As Figure 47 but for ages 70-79.

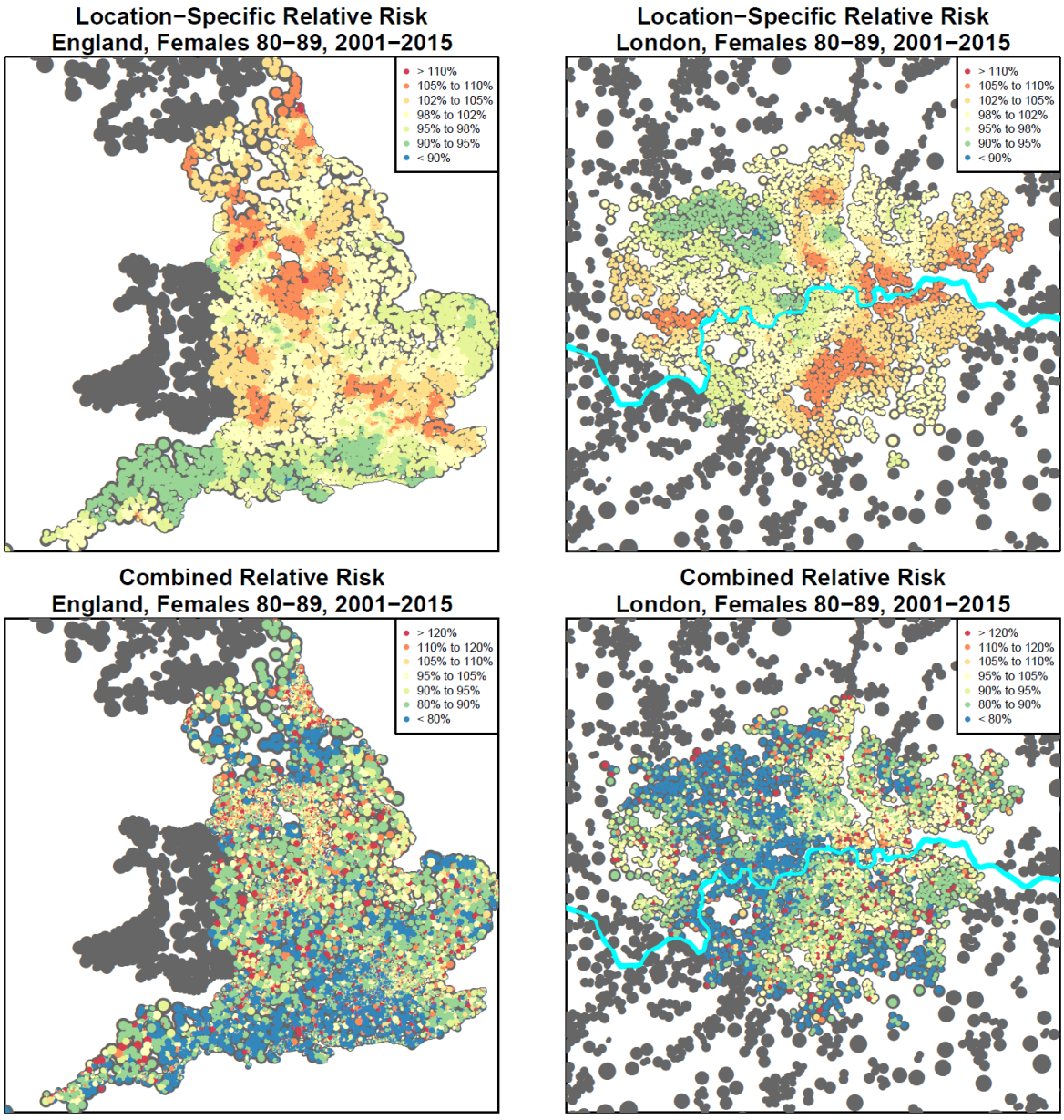

Figure 51: As Figure 47 but for ages 80-89.