WHY IS INEQUALITY INCREASING IN MORTALITY FROM RESPIRATORY DISEASE IN ENGLAND?

AN EXPLORATORY ANALYSIS OF PUBLIC HEALTH AND DEPRIVATION DATA USING VISUAL ANALYTIC TECHNIQUES

Veronica Tuffrey

Report submitted in partial fulfilment of the requirements for module INM433 Visual Analytics
MSc Data Science
City, University of London

December 2018

1. INTRODUCTION

1.1. Context and motivation: Prevalence rates of chronic non-communicable diseases (CNCDs) have reached epidemic proportions worldwide [1]. These diseases[1], including cancer, cardiovascular conditions (mainly heart disease and stroke), diabetes and respiratory disease, affect all ages, classes and nationalities. NCDs cost lives, cost dignity and increasingly place an economic burden on individuals and society.

In England there are huge health inequalities between the least and most deprived areas [2], which are especially manifest for CNCDs [3], and have been well documented for mortality [4, 5]. Behavioural risk factors, including unhealthy diets, physical activity and tobacco smoking, are often unevenly distributed between socioeconomic groups, and so may be important determinants of these health inequalities [6].

While overall mortality rates have improved in England, rates in the most deprived areas have not improved as quickly as in least deprived areas [5]. Stakeholders in the British Lung Foundation[2] are concerned that inequality in mortality rates of respiratory disease is increasing, in contrast to rates of cardiovascular and cancer, and requested that I examine the underlying drivers for these patterns. Recognising the significance of NCDs, and of understanding and addressing the causes of ill-health and mortality in order to feed into policy and interventions, I was happy to undertake this task.

1.2 Data and suitability

<75 mortality rates for respiratory disease (RD), cardiovascular disease (CVD) and cancer: Two sources

- NHS Outcomes Framework[3] (NHS OF) Indicators[4] 1.1 (CVD), 1.2 (RD) and 1.4 (Cancer) from https://digital.nhs.uk/data-and-information/publications/clinical-indicators/nhs-outcomes-framework/current

[1] Chronic non-communicable diseases can be defined as "diseases or conditions that occur in, or are known to affect, individuals over an extensive period of time, and for which there are no known causative agents that are transmitted from one affected individual to another" [1].
[2] A charity that promotes lung health and supports those affected by lung disease (https://www.blf.org.uk/)
[3] NHSOF is a set of indicators developed by the Department of Health and Social Care to monitor health outcomes of adults and children in England, so as to provide an overview of how the NHS is performing.
[4] Original source is Primary Care Mortality Database managed by NHS Digital and Office for National Statistics (ONS).

- Public Health Outcomes Framework[5] (PHOF) Indicators[6] 4.04i, 4.05i and 4.07i from https://fingertips.phe.org.uk/profile/public-health-outcomes-framework

The two datasets are complementary in that NHSOF offers disaggregation by age-group but disaggregates geographically only to regions, while PHOF disaggregates to the level of local authority district (LAD), and mortality data are smoothed over three years. The dataset was suitable due to its disaggregation by deprivation decile[7], and existence of data from 2003 to 2016.

Lifestyle factors: Indicators 2.11 to 2.14 provide data on the behavioural factors of diet and physical activity (2015 and 2016) and smoking from PHOF (2011 – 2017).

Deprivation indices[8] : from https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015
The dataset was suitable as it enables exploration of which domains of deprivation are associated with the mortality rates. The seven domains[9] are combined to produce the "Index of Multiple Deprivation" (IMD).

Pollution Data: Indicator 3.01: "Fraction of Mortality Attributable to Pollution" from PHOF[10] (2010-2016)

Geographical data: Geographical data callable online from within the visualisation software were used, from https://www.tableaumapping.bi/wdc/

1.3 Research Questions

- What are the patterns of variation in <75 y. mortality rates of respiratory disease in England with respect to deprivation, age, gender and time?
- How do the patterns for respiratory disease mortality differ from mortality patterns of cancer and cardiovascular disease?

2. TASKS AND APPROACH

SPSS Statistics for Windows was used for most analysis (V24.0, Armonk, NY: IBM Corp), together with the Python programming language (Python Software Foundation https://www.python.org/) and Tableau (https://www.tableau.com/). The analysis process incorporated computational methods to summarise and apply statistical tests to large quantities of data, and to generate visual interfaces that illustrated the output of the summaries and tests. Interpretation of the visual interfaces enabled design of the next stage of analysis. Human input to this process was essential to plan analysis and decipher outputs, while machine input was essential for enabling speedy computation.

---

[5] PHOF are indicators designed to help stakeholders, especially local authorities, to understand how well public health is being improved and protected.
[6] Original source is Public Health England, based on ONS source data.
[7] For mortality data from years 2009 -16 only.
[8] Original source is Department for Communities and Local Government.
The Indices of Deprivation 2015 are relative measures of deprivation for small areas (Lower-layer Super Output Areas) across England, based on seven domains covering a range of economic, social and housing issues.
The IMD is intended to offer multidimensional information on material living conditions in an area or neighbourhood based on a 'lack of' living necessities causing an unfulfilled social or economic need, relative to the rest of the country. The most recent data are from 2015 and relate to tax year 2012/13 [7].
[9] The domains are Income Deprivation; Employment Deprivation; Education, Skills and Training Deprivation; Health Deprivation and Disability; Crime; Barriers to Housing and Services, and Living Environment Deprivation.
[10] Ideally data would have been obtained from the Department for Environment, Food and Rural Affairs (https://uk-air.defra.gov.uk/), but their data had higher geographical granularity than the other datasets being used in this project.

2.1 Manipulation of data to normalise distribution shape

Visual techniques: Histograms and QQ plots were used to check the shape of variables' distributions. If they were noticeably not normal the variables were transformed using logarithmic or square root functions for positive skew, or square function for negative skew. Then the visual technique of plotting was applied again to check the transformation had corrected the skew.

2.2 Simple bivariate exploratory analysis

Visual techniques: Scatterplot matrices were used to examine cross-sectional bivariate associations between variables in order to check for outliers and identify collinearity. If outliers were identified, the dataset was interrogated to identify the identity of local authority area.

2.3 Bivariate and multivariate analysis to address research questions

To address the research questions, it was necessary to examine patterns of variation in mortality rates with respect to deprivation, age, gender and time. I decided also, if time allowed, to explore

- geographical variation, because the parts of England most affected by poor health needed to be identified[11] for the study findings to be most useful
- variation by lifestyle variables, to help explain variation by deprivation.

STAGE 1 - Visual techniques: Charts and maps

1) Charts were produced to examine trends in mortality rates over time, differences between age-groups, and differences between deprivation deciles. Ideally, I would have created visualisations with all four attributes in the same graphic, however the datasets did not enable this[12]. Charts quickly enabled comparison of patterns of RD, CVD and cancer mortality, and indicated next steps for the computational analysis, as follows -

- to merge the more substantial public health dataset (PHOF) with mortality data disaggregated by local authority area, with the deprivation dataset, in order to examine interaction of deprivation and gender, interaction of gender and time, and possibly even interaction of the three variables;
- to incorporate non-linear elements into modelling of changes over time.

2) Maps were produced by matching mortality data with geographical data, and this enabled geographical variation in mortality rates to be examined both cross-sectionally and longitudinally. For the latter, a decision was necessary on how best to show variation. Space- and time-referenced data can be viewed either as a "spatial arrangement of local behaviours over time", or as a "sequence of momentary behaviours over the territory" [7, p.9]. It proved impractical to produce and insert mini-time graphs for each region to show variation over time on a single map, so a series of 14 maps was produced for each mortality.

STAGE 2: Regression analysis – computational and visual techniques

Least squares linear regression was first used to model the variation in mortality over time, using first and second degree predictors, that is (Number of years since 2000) and (Number of years since 2000)squared).

[11] Since 2012 local authorities have been granted greater funding and responsibilities to protect the public health of their populations.

[12] The NHOF dataset includes disaggregated mortality data from all England by age-group and gender by year; from all England by deprivation decile and gender by year, and by region (9) by year.
The PHOF dataset includes smoothed mortality data from all England disaggregated by gender by time; from all England by deprivation decile and gender by year; by region (9) and gender by year, and by LAD (326) and gender by year.

Then variables for gender (coded 1 for male and 0 for female), deprivation (Logarithm of IMD), and interaction terms were added sequentially to the model, checking each time that the increase in R^2 was statistically significant (using P = 0.05 as the threshold value). QQ plots and histograms of residuals were derived to evaluate if assumptions underlying the regression method were being met. The residuals from the regression were saved and in turn regressed against lifestyle variables, to examine which lifestyle variables and deprivation domains best explained departures from the expected variation due to overall deprivation.

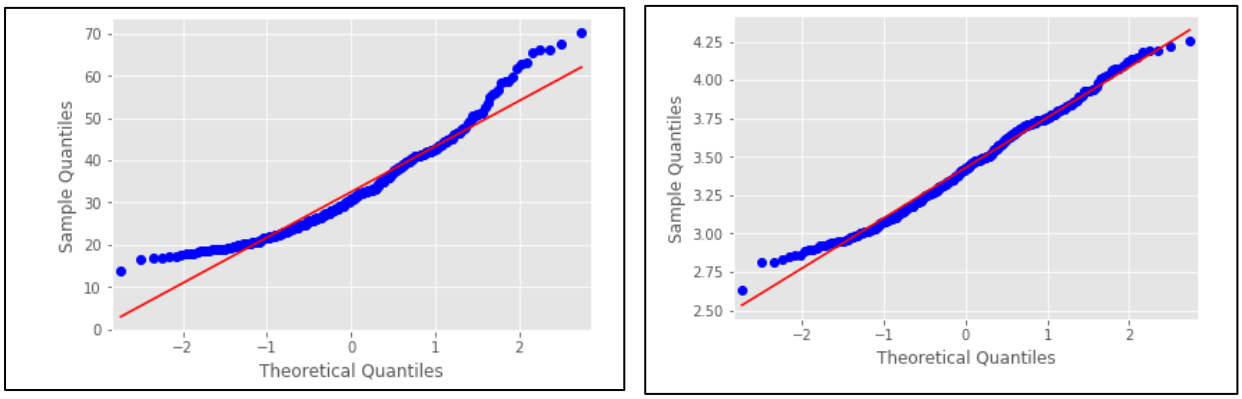
STAGE 3: Visual techniques - mapping

Residuals from the regression of mortality against IMD were mapped to examine whether local authority areas which departed from expected patterns were geographically clustered.

3. ANALYTICAL STEPS

This section includes examples of the outputs from the tasks described above.

3.1 Manipulation of data to normalise distribution shape



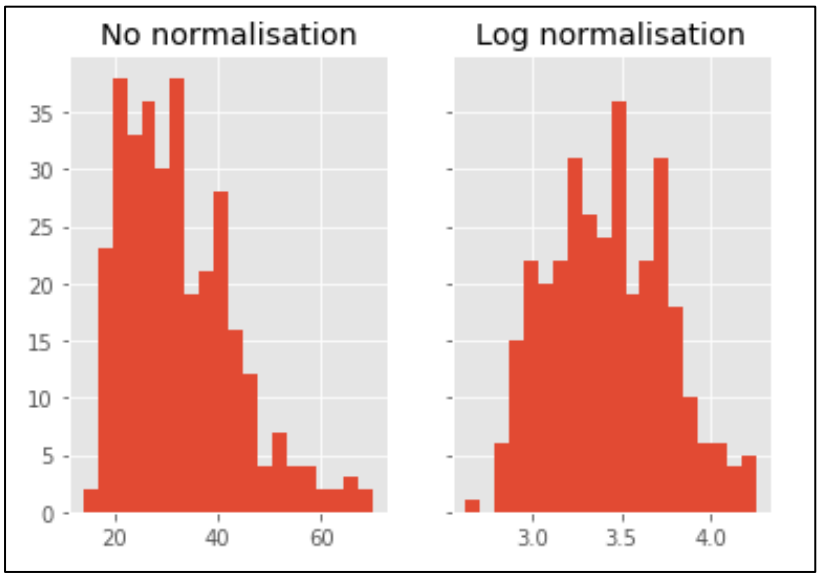
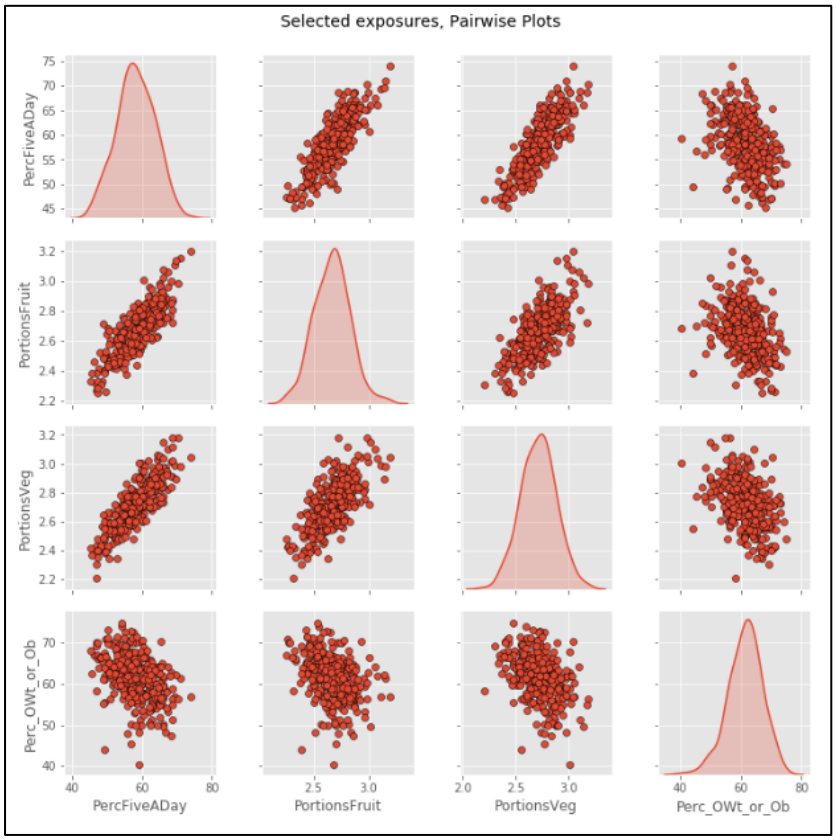


Figure 1 (above): QQ plots before and after log normalisation of respiratory disease mortality data.

Figure 2 (left): Histograms before and after log normalisation of variable respiratory disease mortality data.

Figures 1 and 2 illustrate how QQplots and histograms were used to check the shape of variables' distributions. In this case, distribution of RD mortality was observed to be positively skewed, so the log transformation was applied, and the distribution shape checked again.

3.2 Simple bivariate exploratory analysis

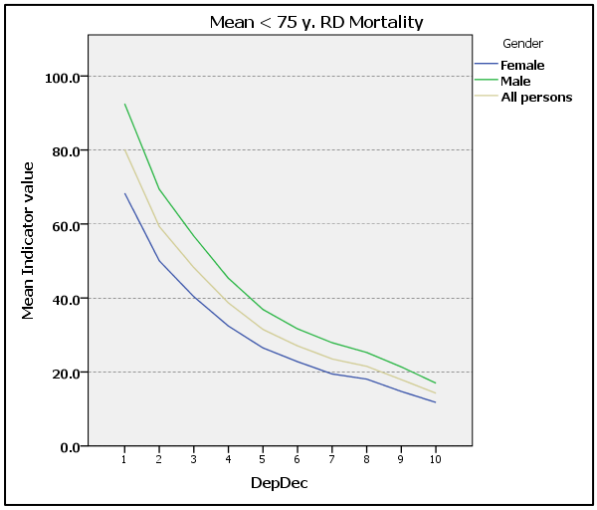


Scatterplot matrices were used to examine cross-sectional bivariate associations between variables as a secondary check for distribution shape, to check for outliers (if outliers were identified, the dataset was interrogated to identify the local authority area) and to see if certain variables were likely to cause issues of collinearity in the later regression analysis. For example, in Figure 1, the variables "PortionsFruit", "PortionsVeg" and "PercFiveADay" are highly correlated with each other. This would be helpful to know later if high values of the VIF (Variance Inflation Factor) were obtained during regression modeling.

Figure 3: Scatterplots of selected lifestyle variables

3.3 Bivariate and multivariate analysis to address research questions

3.3.1 Charts: The first research question necessitated examining variation in mortality rates of respiratory disease with respect to four variables - deprivation, age, gender and time.



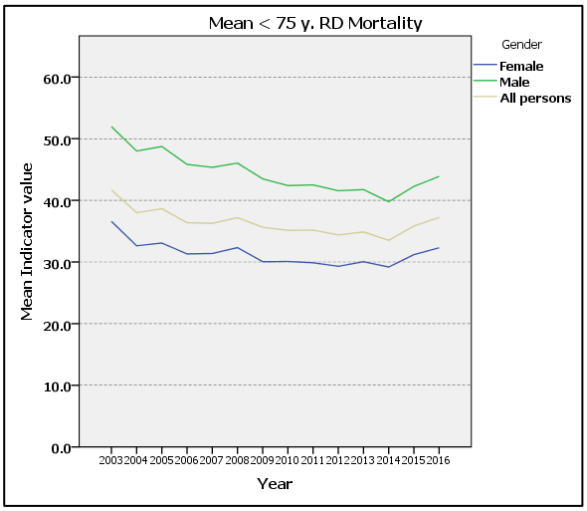


Figure 4 (left): < 75 y. RD mortality by year (2003 – 2016) and gender, from NHOF data
Figure 5 (right): < 75 y. RD mortality by deprivation decile and gender, from NHOF data

Figures 4 and 5 each show two variables, and Figure 6 shows three variables simultaneously. It was useful to produce both types because the first type provides an overview, and the second enables its interpretation.

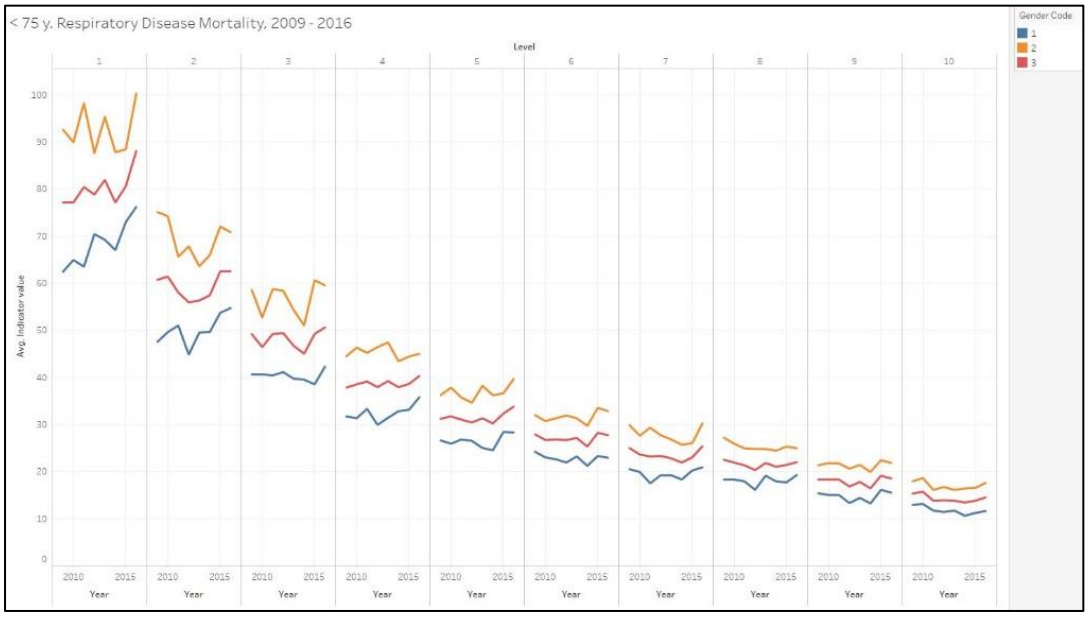


Figure 6: < 75 y. RD mortality by year (2003 – 2016), deprivation decile and gender, from NHOF data

The fourth plot, Figure 6 also shows three variables, illustrating both how RD mortality varies by age group, and how changes in RD mortality rate between 2003 – 2016 (earlier plotted in Figure 5) varies by age group.

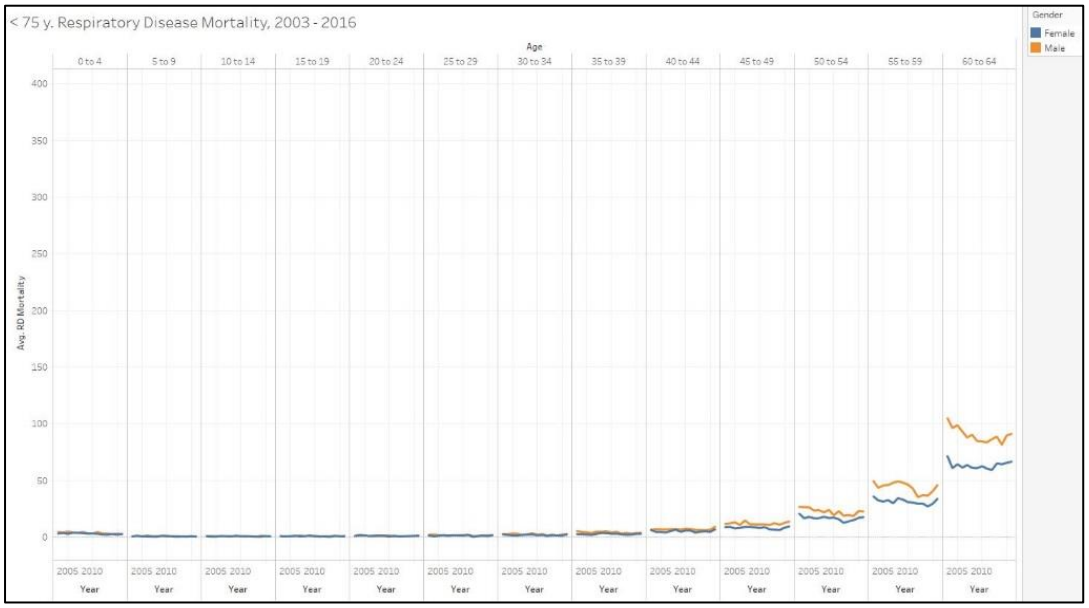


Figure 7: < 75 y. RD mortality by year (2003 – 2016), age group and gender

Findings from this simple visual analysis provided valuable information to feed into the next stage of analysis

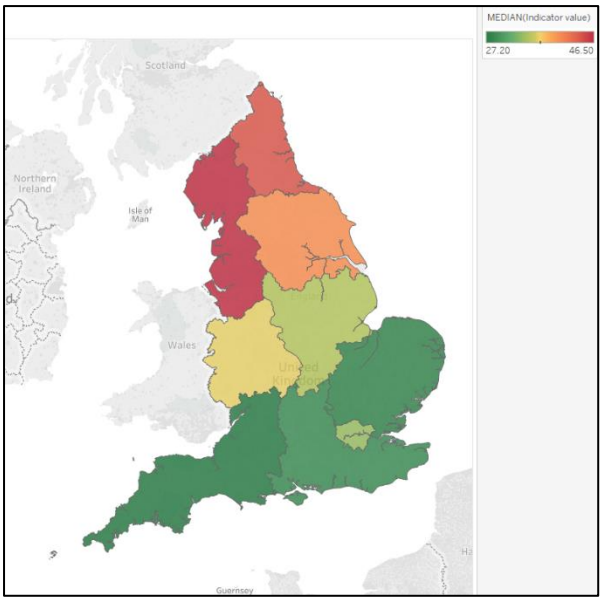
- Figure 4 shows a strong association (curvilinear) of RD mortality with increasing deprivation
- Figure 5 shows that overall there has been only relatively small decrease in RD mortality since 2003;
- Figure 6 shows RD mortality in the lowest deprivation decile has actually increased over time for women;

- Figure 7 shows RD mortality in the oldest age group has decreased for men but not for women.

One can infer that, since RD mortality affects older age groups, it is older women in the lowest deprivation decile who account for the increasing inequality in RD mortality.

As explained in Section 2.3, these findings could not be further examined using the NHOF or PHOF data alone, however by merging the mortality data in the PHOF dataset at local authority level with the dataset of deprivation indices, it was possible to test for interactions later using regression modelling.

3.3.2 Maps



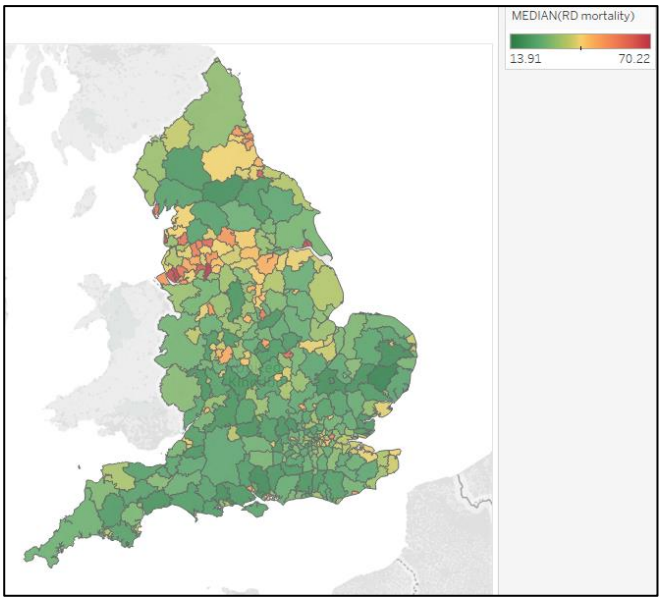


Figure 8 (left): < 75 y. RD mortality rate by region (Median value 2009-2016)
Figure 9 (right): <75 y. RD mortality rate by local authority area

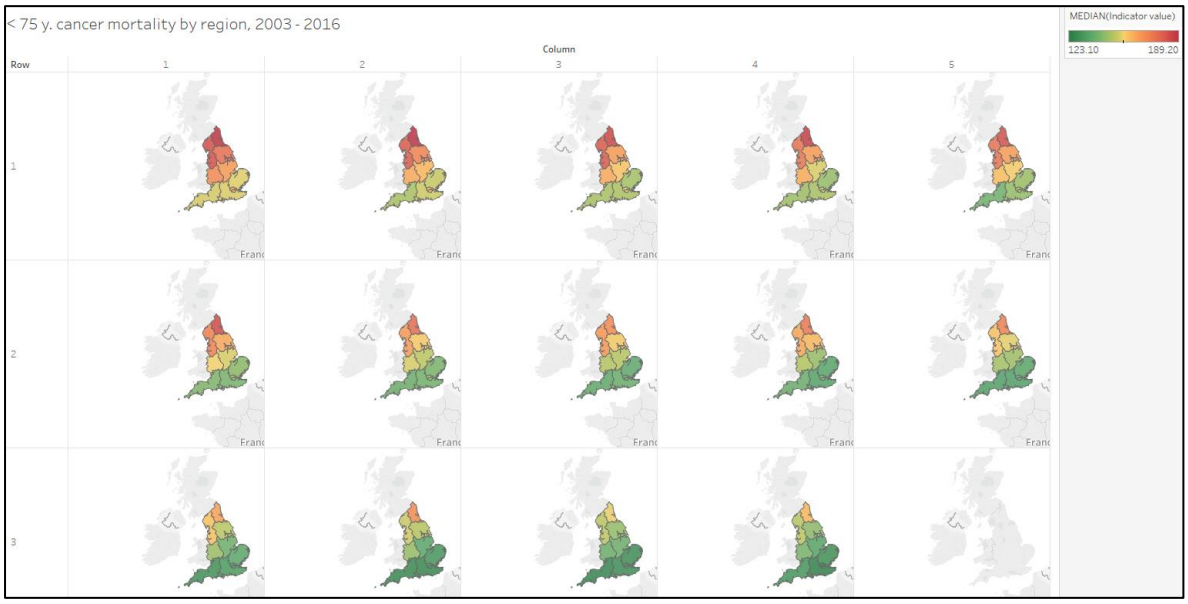


Figure 10: Sequence of mini-maps showing changes in < 75 y. mortality rate from cancer over time

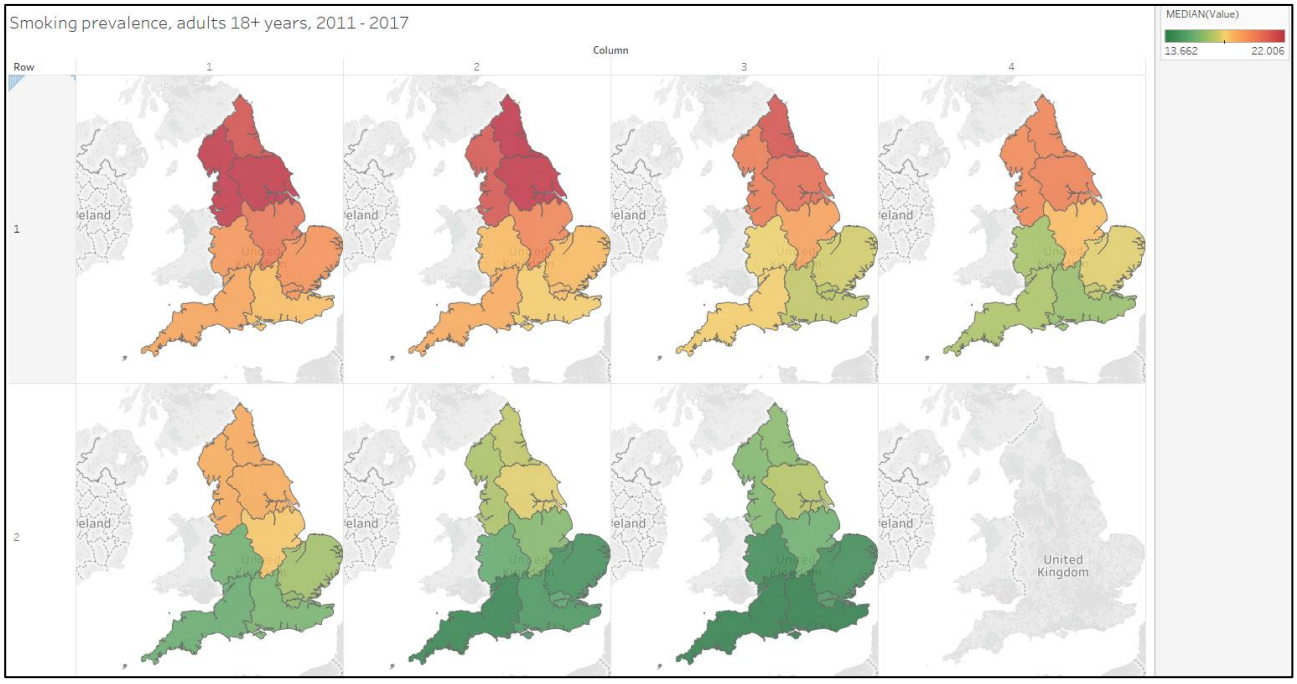


Figure 11: Mini-maps to illustrate change in smoking prevalence over time by region (2011 – 2017)

Figures 10 and 11 show how I used maps to examine time trends in both outcome and predictor variables. This helped identify those variables it was most essential to include in regression models.

I produced many maps showing disaggregation both at regional and LAD level. One issue was that very high values in one LAD could skew the scale used. For example, for RD mortality above, very high values in certain LADs in the north-west (Figure 9) mean that most of the map is coloured green. From that point onwards I manually amended the scales to help interpretation.

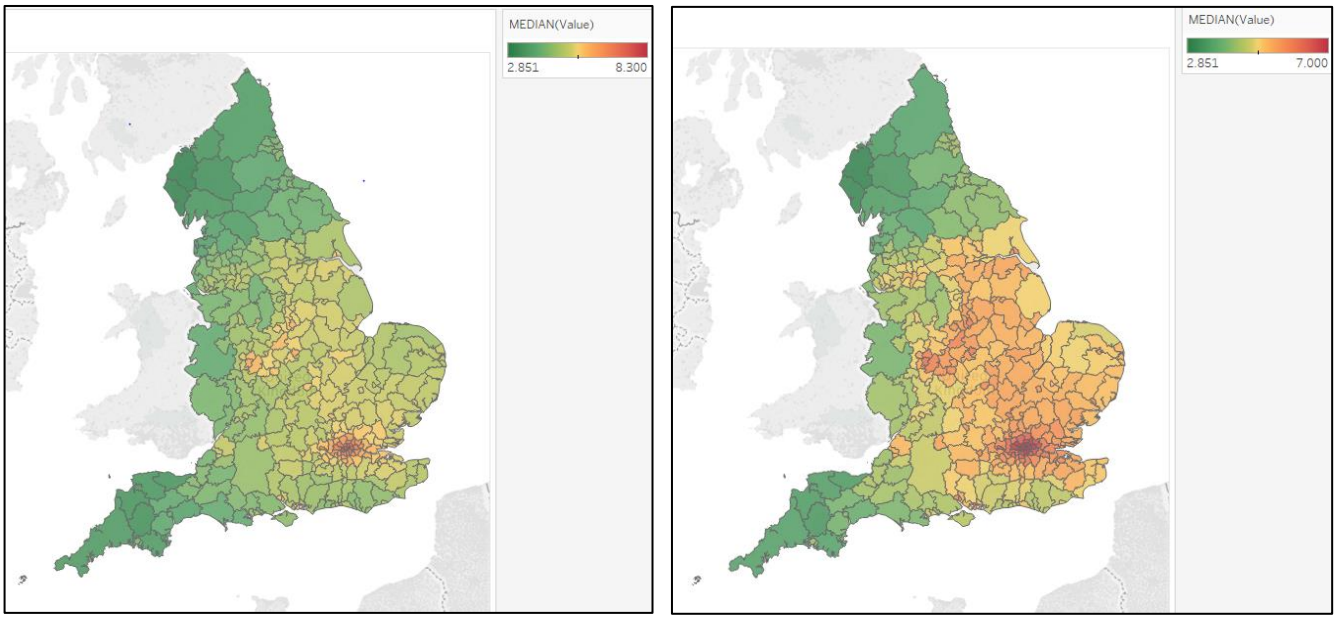
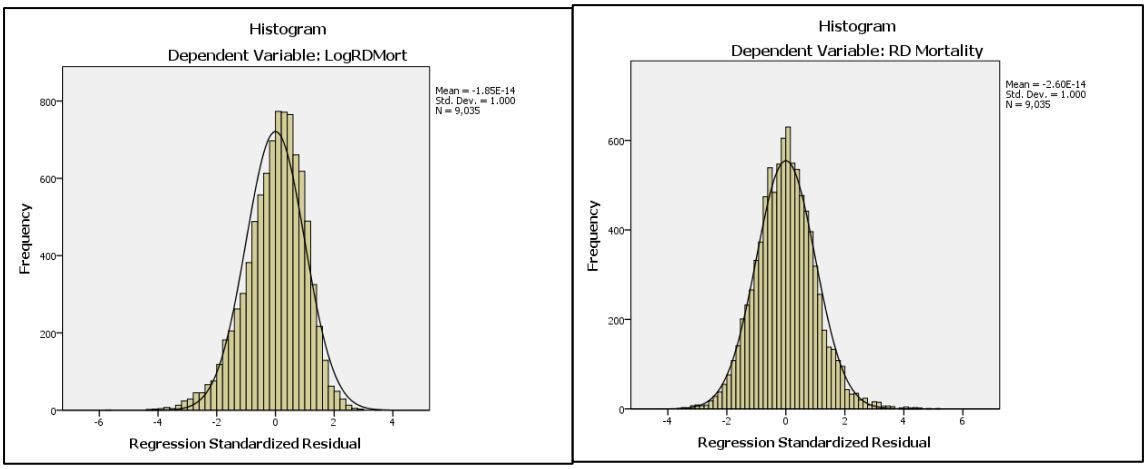


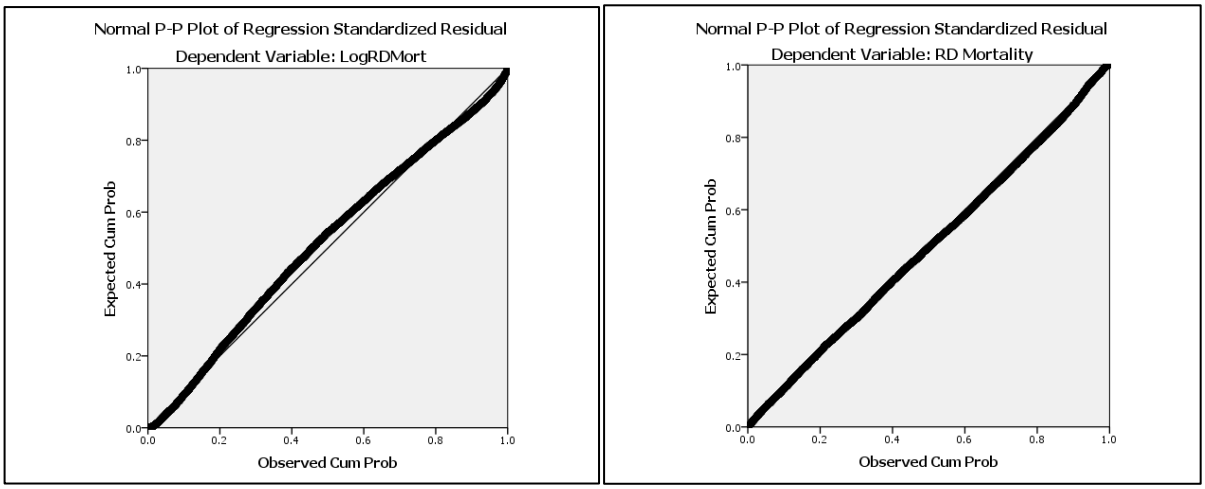
Figure 12: Fraction of mortality attributable to pollution by LAD, Median value between 2010 – 2016

An example is given in Figure 12. Because pollution levels in London are so high, London blocks out perception of variation in the rest of the country, so I reduced the scale's maximum value in the second map.

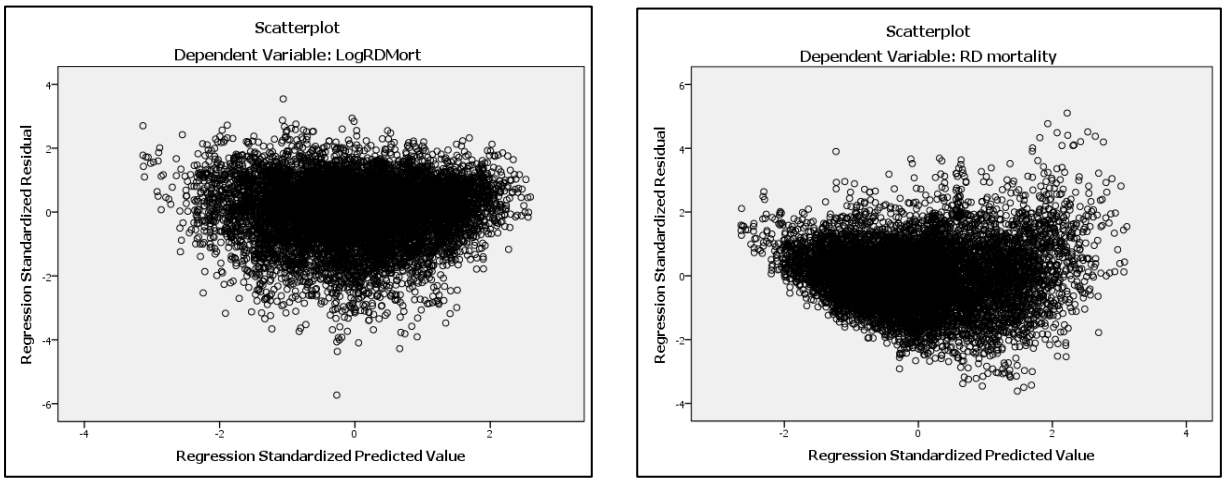
3.3.3 Regression analysis



Histogram of residuals from regression with time, gender and IMD as predictors of: Log normalised respiratory disease mortality in Figure 13 (left), and original values of RD mortality in Figure 14 (right).



Normal probability plots from regression of: log normalised respiratory disease mortality data in Figure 15 (left), and original values of RD mortality in Figure 16 (right)



Scatterplots of standardised residuals against standardised predicted values of: log normalised respiratory disease mortality data in Figure 17 (left), and original values of RD mortality in Figure 18 (right)

For the final stage of analysis, again I used visual outputs to check the assumptions underlying the regression modelling were being met. Unexpectedly, despite earlier checks showing that transformed values of mortality would be better (see Figures 1 and 2), the plots in Figures 13 to 18 indicated that the original variables should be used in modeling.

Figure 19 shows the geographic distribution of the residuals from the regressions of RD mortality v. time, gender and deprivation. The mapping software enables identification of the LADs with high values (Corby, Manchester and Salford).

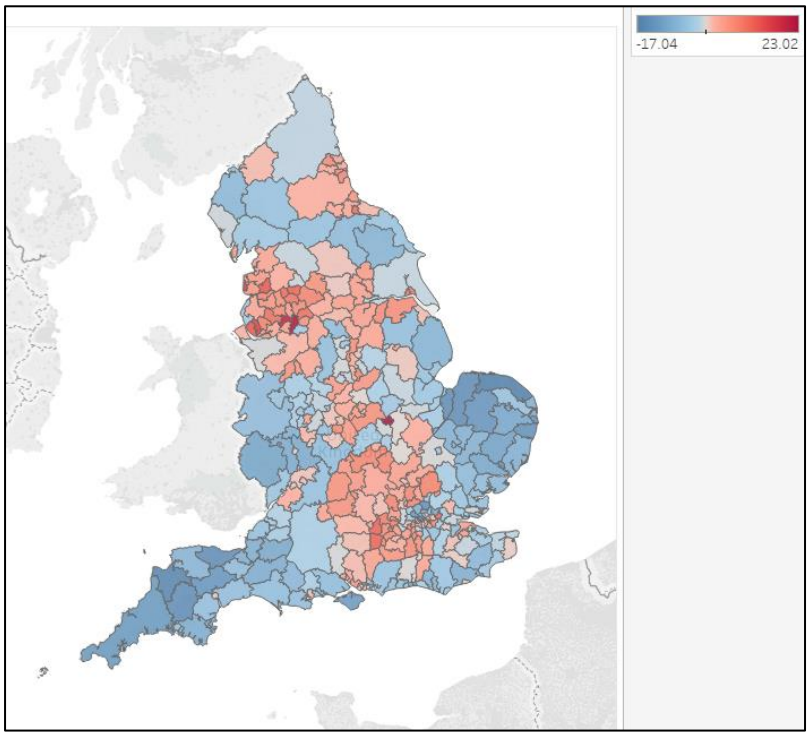


Figure 19: Geographical distribution of residuals from linear regression of RD Mortality as outcome, and time, IMD, gender as interaction terms as predictors

The last stage of analysis was using the residuals as outcomes in models with lifestyle and deprivation domain variables as predictors. Again, I used plots to check assumptions were being met.

4. FINDINGS

4.1 Patterns of variation in <75 y. mortality rates of respiratory disease in England with respect to deprivation, age, gender and time

	Females	Males
Average for England in 2016	30.0	41.0
Average decrease per year (from regression modelling)	0.4	0.7
Mean for 70 – 74 y.o in 2016	225.8	314.0
Mean in lowest deprivation decile in 2016	76.2	88.4
Mean in highest deprivation decile in 2016	11.6	16.5

Table 1: Summary statistics for RD Mortality per 100,000 population

Table 1 provides an overview of patterns of variation in < 75 y. respiratory disease mortality. Females' rate is about 25% lower than males', and rates are more than 5 times higher in the lowest deprivation decile compared to the highest. This latter difference is even greater in females. RD mortality affects mainly older people as would be expected.

Older women in the lowest deprivation decile account for the increasing inequality in RD mortality, as in this group, rates of mortality are increasing rather than decreasing (see Figure 6 and model 1 in Appendix 2). Pollution and smoking may account for these findings, however data were not available to properly test this.

4.2 Patterns for respiratory disease mortality compared to those of CVD and cancer

	< 75 y. CVD Mortality		< 75 y. cancer mortality		< 75 y. RD mortality	
	Females	Males	Females	Males	Females	Males
Average for England in 2016	45.4	101.4	122.5	149.8	30.0	41.0
Average decrease per year	3.0	5.7	1.8	2.9	0.4	0.7
Mean for 70 – 74 y.o in 2016	311.9	616.7	681.7	945.4	225.8	314.0
Mean in lowest dep decile in 2016	92.0	197.7	180.1	236.4	76.2	88.4
Mean in highest dep decile in 2016	22.6	56.9	104.9	108.6	11.6	16.5

Table 2: Summary statistics for CVD and Cancer Mortality (with RD statistics for comparison)

Table 2 provides an overview of patterns of variation in < 75 y. CVD and cancer. Both diseases have much higher rates mortality rates than RD, with cancer the highest. Male CVD mortality rates are more than twice those of females - there is a much greater difference between the genders in CVD mortality compared to respiratory disease (Table 2 and Appendix 2).

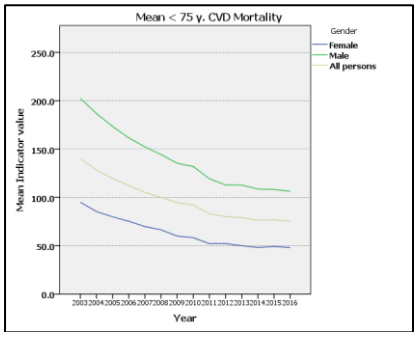
In contrast to RD mortality, there has been a big decrease in CVD mortality since 2003, but this decrease is slowing down (Figure 20). For cancer, the mortality rates are much higher, the rates have decreased and in contrast to CVD are not showing signs of slowing (Figure 21).

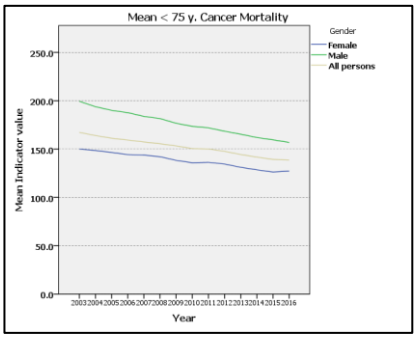
All three mortality rates have very strong associations with deprivation, with RD having the strongest (Figures 22, 23 and 4 below, and regression findings in Appendix 2[13]). CVD mortality in the lowest deprivation decile has stopped decreasing over time, in contrast to higher deciles (Figure a). Cancer mortality rates across all levels of deprivation are still decreasing in both men and women (Figure b).

Regression analysis (Appendix 2) indicated that dietary variables accounted for more unexplained variation in RD mortality than smoking or pollution, however one cannot generalise from this finding to the individual level (see section 5.1 below).

The North-west is the region of England with the highest RD median RD mortality, followed by the North-east. Kingston-Upon-Hull, Manchester, Knowsley and Liverpool are the LADs with the highest rates.

[13] The standardised coefficient for Log_IMD, and the change in R Squared is much larger for RD





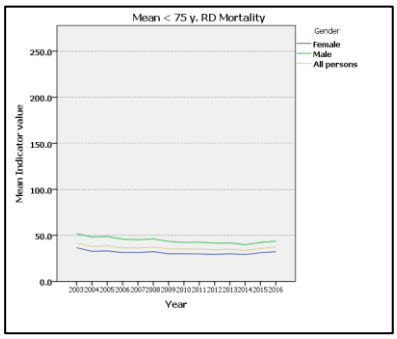
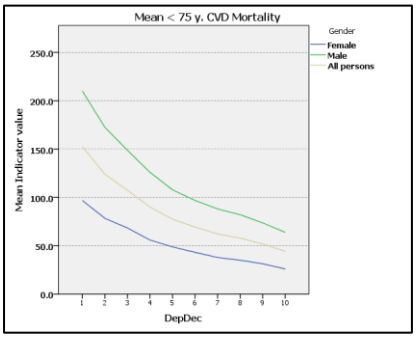
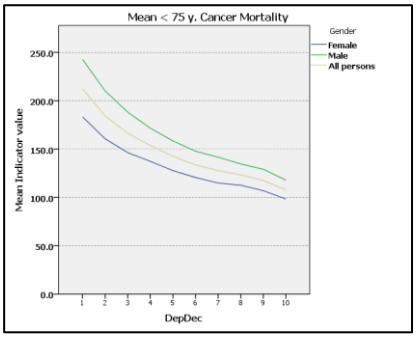


Figure 20 (left) < 75 y. CVD mortality by year (2003 – 2016) and gender, from NHOF data
Figure 21 (centre) < 75 y. Cancer mortality by year (2003 – 2016) and gender, from NHOF data
Figure 5 is reproduced on the right using the same scale, for comparison





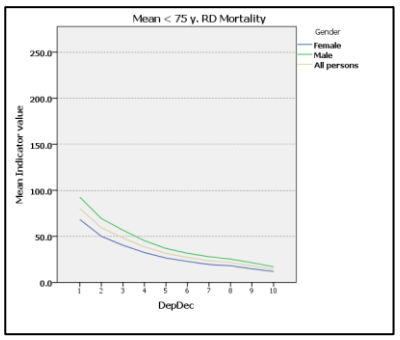


Figure 22 (right) < 75 y. CVD mortality by deprivation decile and gender, from NHOF data
Figure 23 (right) < 75 y. cancer mortality by deprivation decile and gender, from NHOF data
(Figure 4 is reproduced using the same scale, for comparison)

5. CRITICAL REFLECTION

5.1 Implications of study findings for public health policy

This study provides evidence that health inequalities in England are entrenched and for some indicators are increasing. It identifies older women in the lowest deprivation category as being more vulnerable to early mortality from respiratory disease compared to other population groups. This finding is considered reliable at the aggregate level as it is based on statistics provided by government-funded bodies which are independently quality controlled. It is not possible to generalise findings outside of England.

Limitations:
1) Findings based on aggregated data may not hold for individuals[14]. Further investigations are needed of individual level data, for example, using data from the annual cross-sectional Health Survey for England[15].

[14] This is variously knows as “Simpson’s Paradox” or the “Ecological Fallacy”
[15] Data from the 2010 survey which focussed on respiratory health would be a good dataset for this.

2) Data that would help explain the findings were lacking. Indicators on alcohol consumption, and mortality, pollution and smoking data with greater geographical granularity would have been valuable.
3) Due to time constraints, the data analysis did not include multivariate time series analysis. This would have best utilised the data on smoking prevalence over time
4) Use of more ambitious visual analytic tools, such as 3D scatterplots, may have enabled extra insights.

Lung disease is a great burden on the NHS, similar to non-respiratory cancer and heart disease, yet lung disease receives lower levels of resources compared to other diseases [8]. Since 2012, local authorities have new responsibilities and funding for public health. Local authorities have a duty to improve health inequalities in their districts, and this study indicates that women in areas of high deprivation warrant greater attention from them.

With respect to identifying priority actions, poor diet has been identified as the behavioural risk factor with highest impact on the NHS budget [9]. While the combination of unhealthy diets, physical inactivity, and high BMI is the biggest overall contributor to the indicator disability affected life years (DALYs), tobacco smoking remains another key attributable risk factor for DALYs, and is still the leading risk factor for women [3]. Smoking rates are declining overall, but it seems the peak effect of smoking on women's health is probably only now being reached in England [10]. Unfortunately, there was not sufficient time to fully test the association between smoking and mortality due to lung disease in this study.

Pollution is an obvious additional contributory factor, and Figure 24 below shows how in major cities, those living in the centre and along major transport arteries have really high exposure to pollutants, and one needs data at very high levels of geographical granularity to examine its impact[16].

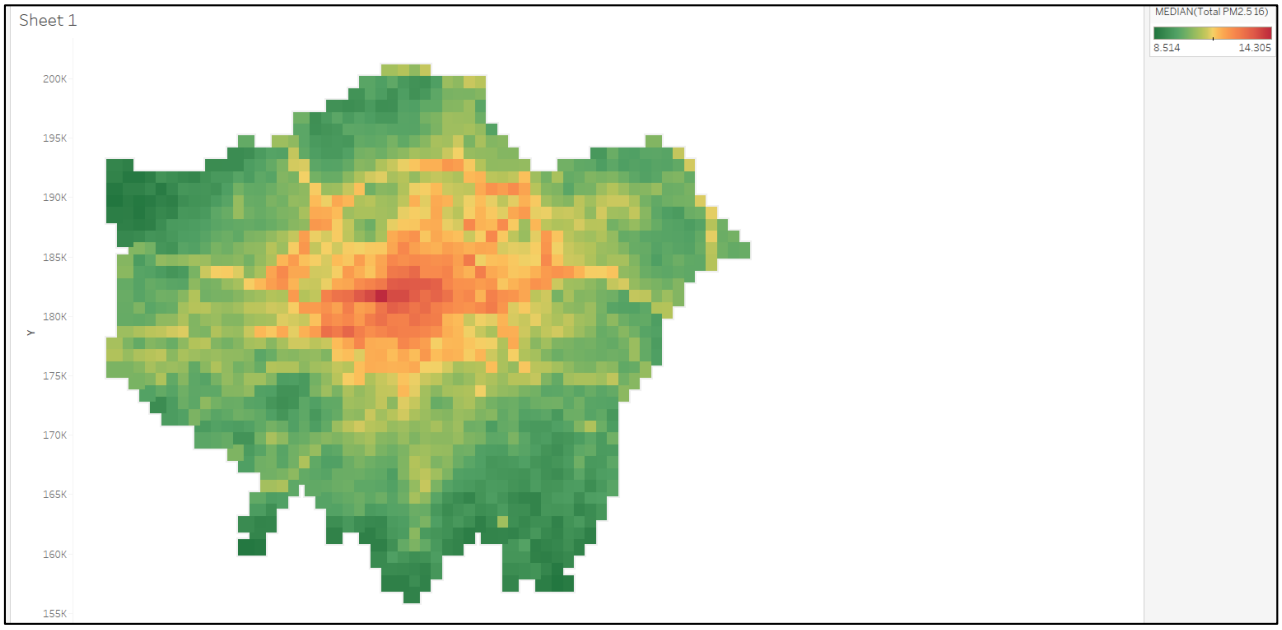


Figure 24: Annual mean PM2.5 concentration 2016 (ugm^{-3}) across London (data from https://uk-air.defra.gov.uk/data/pcm-data)

[16] Figure 24 shows the distribution in London of particulate matter less than 2.5µm in aerodynamic diameter, such fine particles tend to stay longer in the air than heavier particles, and trigger or worsen chronic disease such as asthma, heart attack, bronchitis and other respiratory problems.

5.2 Effectiveness of Visual Analytics approaches

It is important to researchers to separate out their own needs as analysts from the needs of the audience of their findings. What might be a good plot to include in a final report may not be the most effective for early unpacking of the relationships in the data. In this section I will evaluate the effectiveness of the visual analytics approaches used, with respect to an analyst's need to answer the research questions.

During data analysis, interactive visual analytic tools are helpful, so that when an unusual feature is perceived, the analyst can easily identify which datapoint or category is responsible. I used this facility in Tableau to identify LADs having unusually high or low data values, however I have not yet mastered the Plotly package in Python, and I forgot that this facility is available in SPSS. So I did not take maximum advantage of the potential functionality of the tools used.

The dataset used was large and multidimensional. The research questions necessitated examining patterns by several variables simultaneously, while the outputs also needed to be relatively simple to compare three equivalently constructed plots for the three mortality outcomes. I needed to choose for example whether to plot mortality across time, using markers or colours to show deprivation decile and gender, or to have deprivation decile as the x axis and use markers or colours for year and gender (see Figure e in Appendix). There was usually no obvious best choice, but fortunately with the statistical software it is quick and easy to produce multiple plots, from which I could build an internal picture of the variation using different viewpoints. Use of colour was maximised effectiveness. Because I am so habituated to seeing time along the horizonal axis I found other orientations harder to interpret, for example Figure 25, and so did not use them.

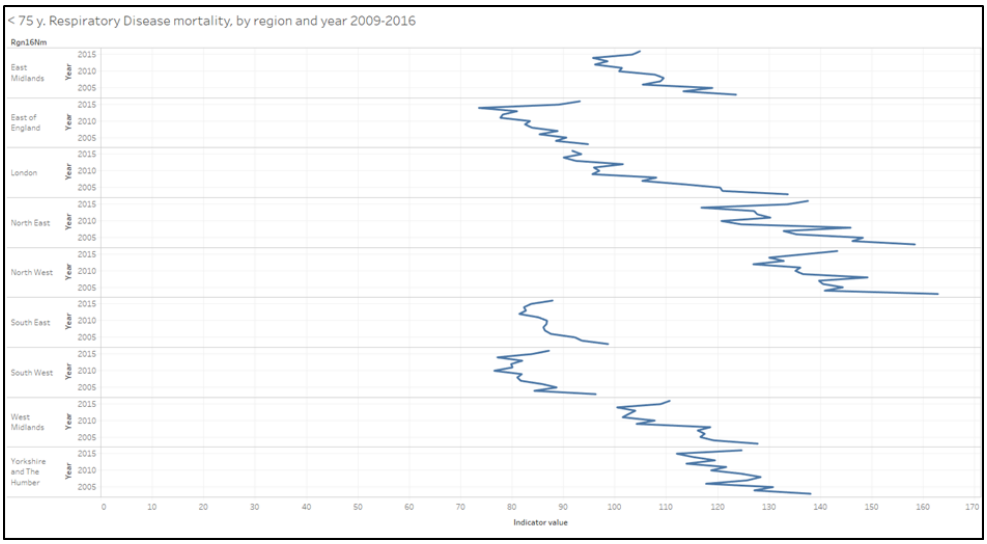


Figure 25: Plot of respiratory disease mortality by time, by region, with time on the vertical axis.

Heatmaps help analysts to gain an overview of relationships between large numbers of variables. I used them initially, for example to compare the strength of association between smoking prevalence and different deprivation domains between males and females (Figure 26), but later needed to view more than two variables simultaneously.

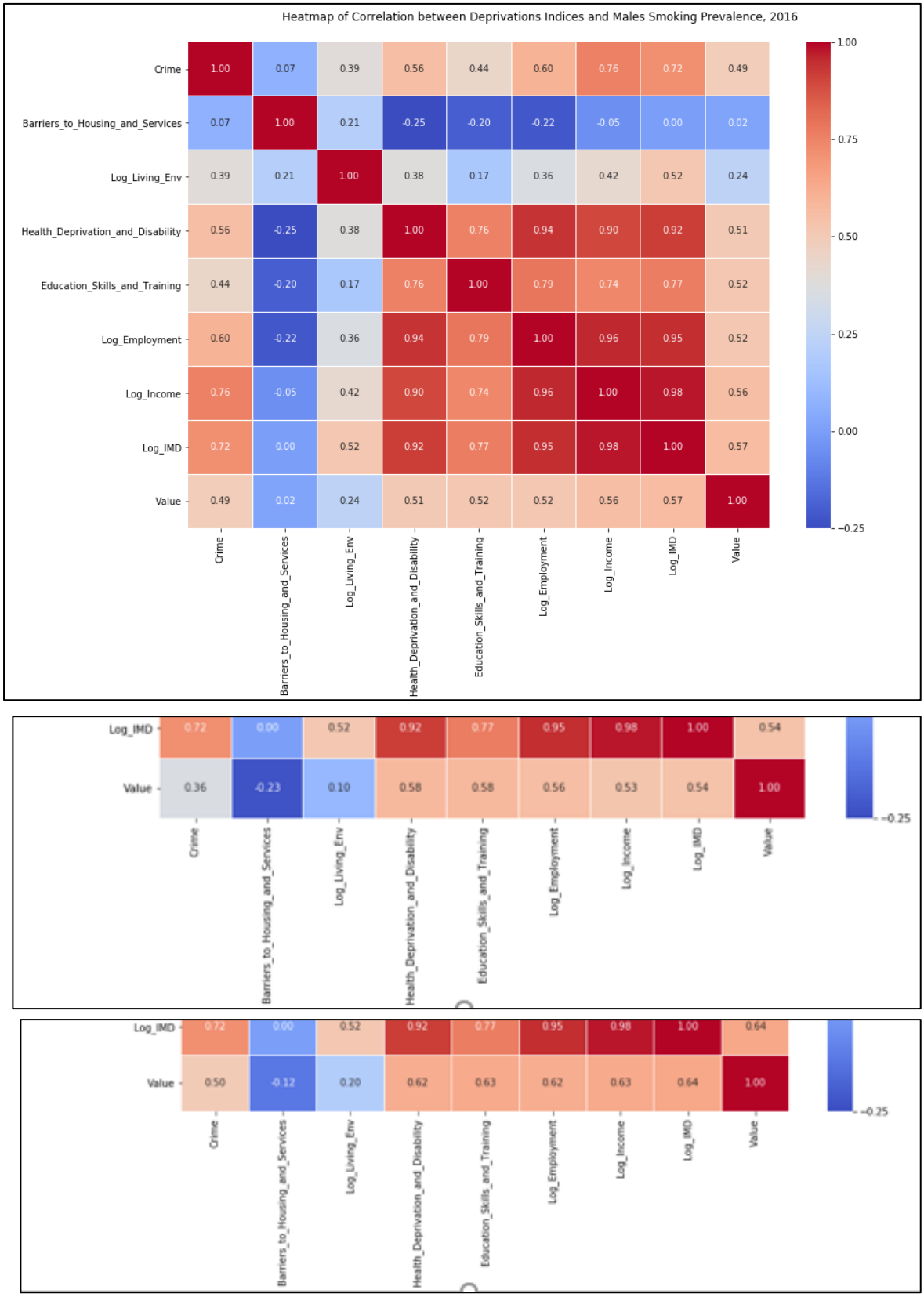


Figure 26: Heatmap of deprivation domains and smoking prevalence (labelled "Value") for males (top), and extracts from equivalent heatmaps for females (middle) and males and females combined (base)

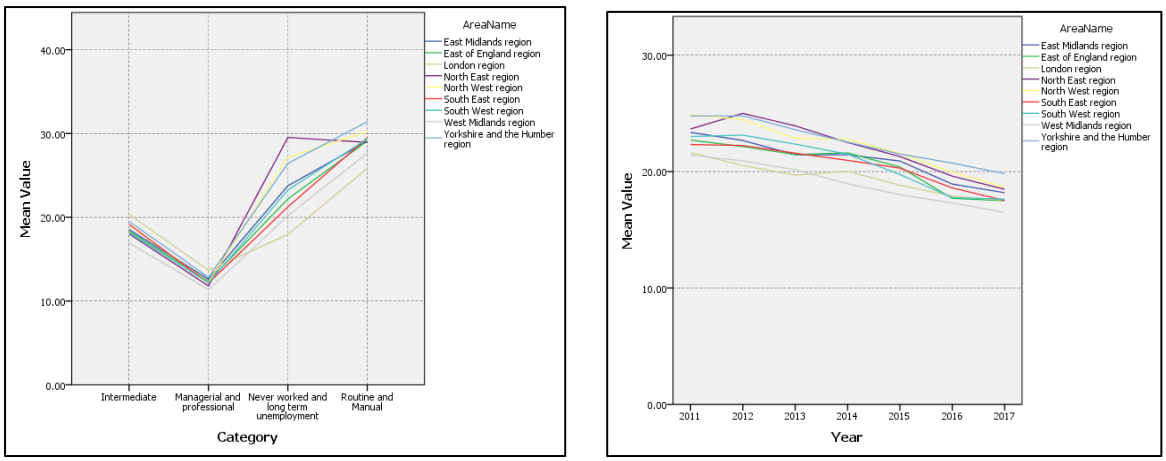


Figure 27: Smoking prevalence by socio-economic group and region (left) and by time and region (right)

When trying to decide how best to visualise patterns, choices were necessary, for example to examine variation over space and time, I chose to plot multiple minimaps rather than one large map with time plots superimposed. Minimaps allowed good perception of changes over time when working at my computer with the capacity to zoom in and query features of interest - for presentation of findings on paper they would be less effective. I found that it was not helpful to use regions as a variable rather than plotting the data geographically, except to explore the extent to which variability existed between them, since there were too many regions to internally distinguish them (see Figure 27 for example).

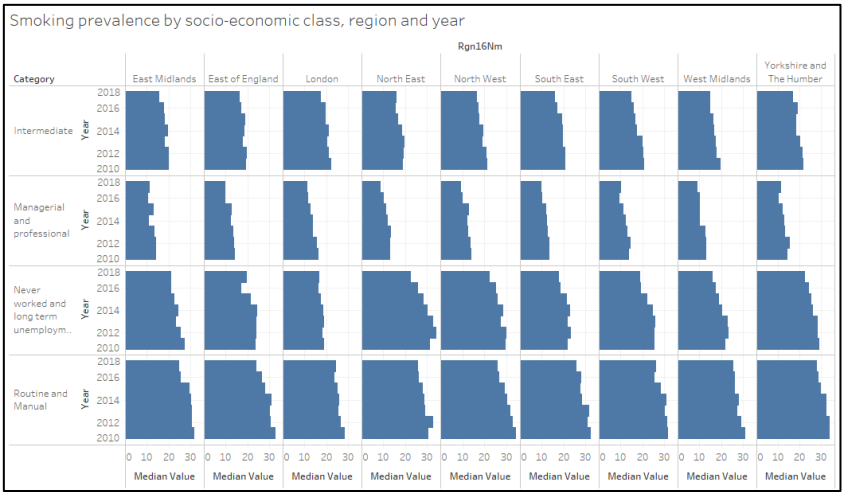


Figure 28: Display of smoking prevalence by region, time and socio-economic group using bar length

I experimented with using different colour and bars to visualise changes over time by geographical region (see Figures 28 and 29) and found myself consistently returning to chloropleth maps in preference. While they have disadvantages – such as a false impression of abrupt change at boundaries - these are outweighed by the satisfaction of viewing a map. Unfortunately, they did not directly help address the research questions since geographical variation was not mentioned.

In summary, by experimentation I gained a portfolio of plots that were effective in addressing the research questions. The most useful charts were those that plotted time along the horizontal axis, and coloured lines to distinguish categories, and the most effective maps were interactive, using two colours centred on white.

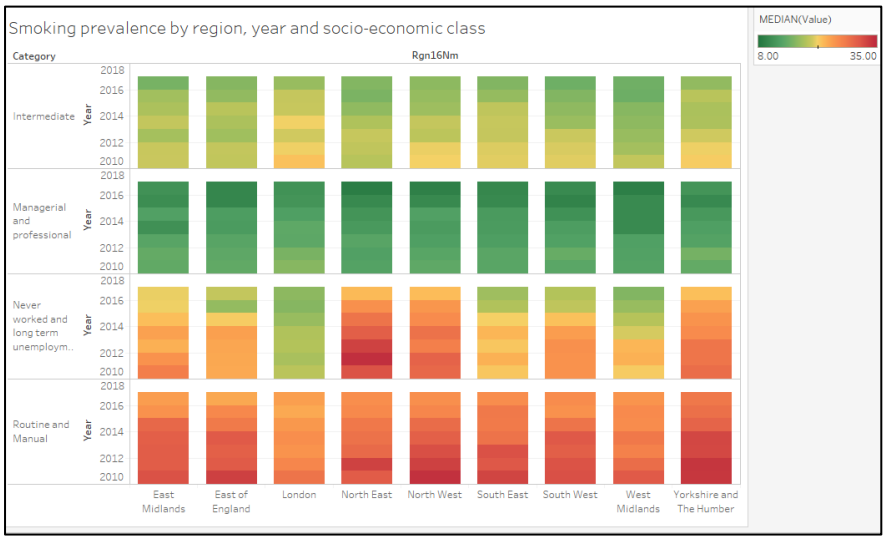


Figure 29: Display of smoking prevalence by region, time and socio-economic group, using colour

5.3 Generalizability of Visual Analytics approaches used

The approaches described above are applicable to any domain involving variables that are measured (i.e. continuous or interval data) that vary across space and time. Potential domains include environmental, educational, social, economic and political. I particularly perceive applicability to the education domain - for example, to examine variation in pupils' achievement by LAD, by gender, and by deprivation level – and that of sociology, for example to explore patterns in, and underlying determinants of, rates of teenage pregnancy and marriage.

Whenever categorical variables are converted into continuous variables by aggregating individuals and turning data into proportions (for example, the binary variable support or otherwise of a political party becomes a continuous variable when aggregated over a geographical area into % of population support for the party), then patterns of variation in the continuous variable can be explored with respect to other variables including time, location, and individual and higher-level attributes.

Aggregating individual level data using geographical units enables variables at both individual and higher level to be included in modelling, and thereby reveal potentially valuable insights for policy-making and remedial action.

6. CONCLUSION

This study illustrated how insights can be gained by using computational methods to summarise and apply statistical tests to large quantities of data, and to generate visual interfaces. By aiding interpretation of - and choice of parameters for - the statistical testing, the interfaces maximise the effectiveness of data analysis to address a question of public health significance. The approaches used are applicable to other domains with spatial and temporal references.

REFERENCES

[1] Daar AS, Singer PA, Leah Persad D, Pramming SK, Matthews DR, Beaglehole R, et al. Grand challenges in chronic non-communicable diseases. *Nature* 2007; 450**:**494.

[2] Public Health England. *Health Profile for England, 2018: Chapter 5: inequalities in health**.*** London: Public Health England, Available from: https://www.gov.uk/government/publications/health-profile-for-england-2018/chapter-5-inequalities-in-health (last accessed 21/12/18); 2018.

[3] Newton JN, Briggs AD, Murray CJ, Dicker D, Foreman KJ, Wang H, et al. Changes in health in England, with analysis by English regions and areas of deprivation, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet* 2015; 386**:**2257-2274.

[4] Dunnell K, Blakemore C, Haberman S, McPherson K, Pattison J. *Life expectancy: Is the socio-economic gap narrowing?* : Longevity Science Panel; 2018.

[5] Bennett M. *Socioeconomic inequalities in avoidable mortality, England and Wales: 2001 to 2016**.*** Office for National Statistics; 2018.

[6] Solar O, Irwin A. *A conceptual framework for action on the social determinants of health**.*** In Social Determinants of Health Discussion Paper 2 (Policy and Practice). Geneva: World Health Organization; 2010.

[7] Andrienko N, Andrienko G. *Exploratory Analysis of Spatial and Temporal Data A Systematic Approach.* Berlin Heidelberg: Springer Verlag; 2006.

[8] British Lung Foundation. *The battle for breath—the impact of lung disease in the UK**.*** London: British Lung Foundation; 2016.

[9] Scarborough P, Bhatnagar P, Wickramasinghe KK, Allender S, Foster C, Rayner M. The economic burden of ill health due to diet, physical inactivity, smoking, alcohol and obesity in the UK: an update to 2006–07 NHS costs. *Journal of public health* 2011; 33**:**527-535.

[10] Bongaarts J. Trends in causes of death in low-mortality countries: implications for mortality projections. *Population and Development Review* 2014; 40**:**189-212.

Appendix 1 – Variables

Variable name	Original Source	Full name in Source	Notes
Dependent Variables			
RD mortality	NHS and Public Health England (based on Office for National Statistics (ONS) source data)	Indicator 4.04i in PHOF and Indicator 1.2 in NHSOF: Under 75 mortality rate from all cardiovascular diseases	Age-standardised rate of mortality from all cardiovascular diseases (including heart disease and stroke) in persons less than 75 years per 100,000 population NB PHOF values are smoothed over three years
CVD mortality	NHS and Public Health England (based on ONS source data)	Indicator 4.05i in PHOF and Indicator 1.4 in NHSOF: Under 75 mortality rate from cancer	Age-standardised rate of mortality from all cancers in persons less than 75 years per 100,000 population NB PHOF values are smoothed over three years
Cancer mortality	NHS and Public Health England (based on ONS source data)	Indicator 4.07i in PHOF and Indicator 1.1 in NHSOF: Under 75 mortality rate from respiratory disease	Age-standardised rate of mortality from respiratory disease in persons less than 75 years per 100,000 population NB PHOF values are smoothed over three years
Diet and Activity Independent Variables			
PercFiveADay	Public Health England (based on Active Lives, Sport England)	2.11i - Proportion of the adult population meeting the recommended 5-a-day on a 'usual day' (adults)	Proportion of the population who, when surveyed, reported that they had eaten the recommended 5 portions of fruit and vegetables on a usual day.
PortionsFruit	Public Health England (based on Active Lives, Sport England)	2.11ii - Average number of portions of fruit consumed daily (adults)	Average (mean) number of portions reported by survey respondents aged 16+ when asked how many portions of fruit they ate on the previous day.
PortionsVeg	Public Health England (based on Active Lives, Sport England)	2.11iii - Average number of portions of vegetables consumed daily (adults)	Average (mean) number of portions reported by survey respondents aged 16+ when asked how many portions of vegetables they ate on the previous day.
Perc_OWt_or_Ob	Public Health England (based on Active Lives, Sport England)	2.12 - Percentage of adults (aged 18+) classified as overweight or obese	Percentage of adults aged 18 and over classified as overweight or obese
PercActive	Public Health England (based on Active Lives, Sport England)	2.13i - Percentage of physically active adults	Number of respondents aged 19 and over, with valid responses to questions on physical activity, doing at least 150 moderate intensity equivalent (MIE) minutes physical activity / week in bouts of 10 minutes or more in the previous 28 days, expressed as % of total number of respondents aged 19 and over.
PercInactive	Public Health England (based on Active Lives, Sport	2.13ii - Percentage of physically inactive adults	Number of respondents aged 19 and over, with valid responses to questions on physical activity, doing less than 30 moderate intensity equivalent (MIE)

	England)		minutes physical activity / week in bouts of 10 minutes or more in the previous 28 days, expressed as % of total number of respondents aged 19 and over.
Deprivation Independent Variables			
Deprivation Decile	Derived from IMD below		Deciles are calculated by ranking the 32,844 Lower Super Output Areas) in England from most deprived to least deprived using the IMD (see next row), and dividing them into 10 equal groups. '1- Most deprived' to '10 – Least deprived'
IMD	Department for Communities and Local Government	Index of Multiple Deprivation	Population weighted average of the combined scores for the LSOAs of a local authority district. Domains are combined using weights as follows: • Income Deprivation (22.5%) • Employment Deprivation (22.5%) • Education, Skills and Training Deprivation (13.5%) • Health Deprivation and Disability (13.5%) • Crime (9.3%) • Barriers to Housing and Services (9.3%) • Living Environment Deprivation (9.3%)
Income	As above	Income Deprivation	Derived from indicators: • Adults and children in Income Support families • Adults and children in income-based Jobseeker's Allowance families Adults and children in income-based Employment and Support Allowance families • Adults and children in Pension Credit (Guarantee) families • Adults and children in Child Tax Credit and Working Tax Credit families, below 60% median income not already counted • Asylum seekers in England in receipt of subsistence support, accommodation support, or both
Employment	As above	Employment Deprivation	Derived from indicators: • Claimants of Jobseeker's Allowance, aged 18-59/64 • Claimants of Employment and Support Allowance, aged 18-59/64 Claimants of Incapacity Benefit, aged 18-59/64 • Claimants of Severe Disablement Allowance, aged 18-59/64 • Claimants of Carer's Allowance, aged 18-59/64
Education_Skills_ and_Training	As above	Education, Skills and Training Deprivation	Derived from indicators: • Key stage 2 attainment: average points score • Key stage 4 attainment: average points score • Secondary school absence

			• Staying on in education post 16 • Entry to higher education • Adults with no or low qualifications, aged 25-59/64 • English language proficiency, aged 25-59/64
Health_ Deprivation_ and_Disability	As above	Health Deprivation and Disability	Derived from indicators: • Years of potential life lost • Comparative illness and disability ratio • Acute morbidity • Mood and anxiety disorders
Crime	As above	Crime	Recorded crime rates for: Violence; Burglary; Theft; Criminal damage
Barriers_to_ Housing_ and_Services	As above	Barriers to Housing and Services	Derived from indicators: • Road distance to: post office; primary school; general store / supermarket; GP surgery • Household overcrowding • Homelessness • Housing affordability
Living_ Environment	As above	Living Environment Deprivation	Derived from indicators: • Housing in poor condition • Houses without central heating • Air quality • Road traffic accidents
Other Independent Variables			
PercSmoking	Annual Population Survey	2.14 - Smoking Prevalence in adults - current smokers	Prevalence of smoking among persons 18 years and over
Pollution	DEFRA/Air Pollution and Climate Change Group Public Health England	3.01 - Fraction of mortality attributable to particulate air pollution	Fraction of annual all-cause adult mortality attributable to anthropogenic (human-made) particulate air pollution (measured as fine particulate matter, PM2.5)
Gender			Gender: Male, female and all persons (NB recoded into derived variable “MALE” where Male = 1, Female =0 and All persons = System missing)
Age-group			5-year age bands from age 0 to 74 for England and region from 2003 Deprivation: Deciles from ‘1- Most deprived’ to ’10 – Least deprived’ for males, females and all persons (by calendar year from 2009) 15 five-year categories from 0 - 4 y to 70 – 74 y.

Appendix 2 – Summary Findings from Linear Regression

Model	Dependent Variable	Independent Variables included*	In final model, coefficient for last variable of list in previous column			Adjusted R Squared	Change in R Squared
			Unstandardised	Standardised	P value		
1	RDMortality	**Constant**	-14.662		.000		
		(Constant), **Male**	-16.588	-.597	.000	.194	.194
		(Constant), Male, **YearFromBase**	-1.160	-.336	.000	.220	.026
		(Constant), Male, YearFromBase, **YFBSq**	.040	.225	.000	.222	.002
		(Constant), Male, YearFromBase, YFBSq, **Log_IMD**	17.264	.531	.000	.638	.416
		(Constant), Male, YearFromBase, YFBSq, Log_IMD, **IntMaleIMD**	11.064	1.172	.000	.651	.013
		(Constant), Male, YearFromBase, YFBSq, Log_IMD, IntMaleIMD, **IntMaleYear**	.813	.324	.000	.653	.002
		(Constant), Male, YearFromBase, YFBSq, Log_IMD, IntMaleIMD, IntMaleYear, **IntMaleYearIMD**	-.392	-.460	.000	.654	.001
2	CVDMortality	**Constant**	53.499		.000		
		(Constant), **Male**	-51.595	-.527	.000	.569	.570
		(Constant), Male, **YearFromBase**	-11.249	-.925	.000	.727	.157
		(Constant), Male, YearFromBase, Y**FBSq**	.360	.689	.000	.738	.011
		(Constant), Male, YearFromBase, YFBSq, **Log_IMD**	27.731	.242	.000	.874	.136
		(Constant), Male, YearFromBase, YFBSq, Log_IMD, **IntMaleIMD**	58.317	1.753	.000	.890	.016
		(Constant), Male, YearFromBase, YFBSq, Log_IMD, IntMaleIMD, **IntMaleYear**	3.685	.483	.000	.913	.023
		(Constant), Male, YearFromBase, YFBSq, Log_IMD, IntMaleIMD, IntMaleYear, **IntMaleYearIMD**	-2.562	-.987	.000	.917	.004
3	CancerMortality	**Constant**	75.841		.000		
		(Constant), **Male**	-17.235	-.293	.000	.344	.344
		(Constant), Male, **YearFromBase**	-2.241	-.307	.000	.451	.107
		(Constant), Male, YearFromBase, **YFBSq**	.021	.056	.056	.451	.000
		(Constant), Male, YearFromBase, YFBSq, **Log_IMD**	26.535	.386	.000	.716	.265
		(Constant), Male, YearFromBase, YFBSq, Log_IMD,	21.575	1.079	.000	.732	.017

Model	Dependent Variable	Independent Variables included*	In final model, coefficient for last variable of list in previous column			Adjusted R Squared	Change in R Squared
		IntMaleIMD					
		(Constant), Male, YearFromBase, YFBSq, Log_IMD, IntMaleIMD, **IntMaleYear**	.081	.015	.833	.738	.006
		(Constant), Male, YearFromBase, YFBSq, Log_IMD, IntMaleIMD, IntMaleYear, **IntMaleYearIMD**	-.409	-.226	.002	.738	.000
4	RD_Resids	**Constant**	5.831		.000		
	(standardised)	(Constant), **PortionsVeg**	-1.423	-.237	.000	.108	.109
		(Constant), PortionsVeg, **PortionsFruit**	-1.345	-.217	.000	.126	.017
		(Constant), PortionsVeg, PortionsFruit, **PercActive**	.024	.118	.000	.137	.011
		(Constant), PortionsVeg, PortionsFruit, PercActive, **PercSmoking**	-.014	-.064	.000	.140	.003
		(Constant), PortionsVeg, PortionsFruit, PercActive, PercSmoking, **Perc_OWt_or_Ob**	.005	.026	.023	.140	.000
5	RD_Resids	(**Constant)**	-1.892		.000		
	(standardised)	(Constant), **Barriers_to_Housing_and_Services**	-.049	-.279	.000	.104	.104
		(Constant), Barriers_to_Housing_and_Services, **Crime**	.884	.405	.000	.145	.041
		(Constant), Barriers_to_Housing_and_Services, Crime, **Log_Living_Env**	-.620	-.299	.000	.197	.052
		(Constant), Barriers_to_Housing_and_Services, Crime, Log_Living_Env**, Log_Employment**	-2.357	-.842	.000	.215	.018
		(Constant), Barriers_to_Housing_and_Services, Crime, Log_Living_Env, Log_Employment, **Health_Deprivation_and_Disability**	1.328	.858	.000	.292	.077
		(Constant), Barriers_to_Housing_and_Services, Crime, Log_Living_Env, Log_Employment, Health_Deprivation_and_Disability, **Education_Skills_and_Training**	-.016	-.133	.000	.298	.006

*"**YearFromBase**" = number of years since 2000; "**YFBSq**" = $(YearFromBase)^2$, "**Log_IMD**" = Natural logarithm of Index of Multiple Deprivation; "**IntMaleIMD**" = Interaction term Male*Log_IMD; "**IntMaleYear**" = Interaction term Male*YearFromBase; "**IntMaleYearIMD**" = Interaction term Male*YearFromBase*Log_IMD.

Appendix 3 – Additional charts

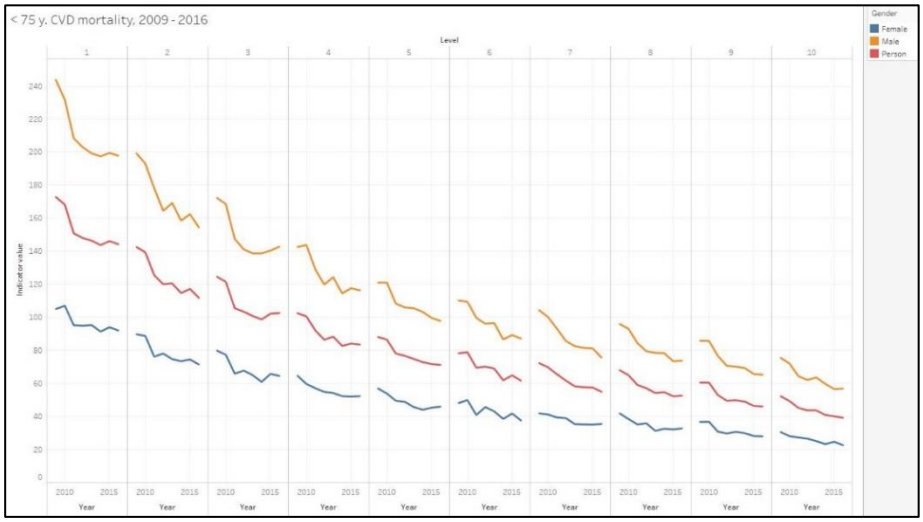


Figure a: < 75 y. CVD mortality by gender and year, by deprivation decile

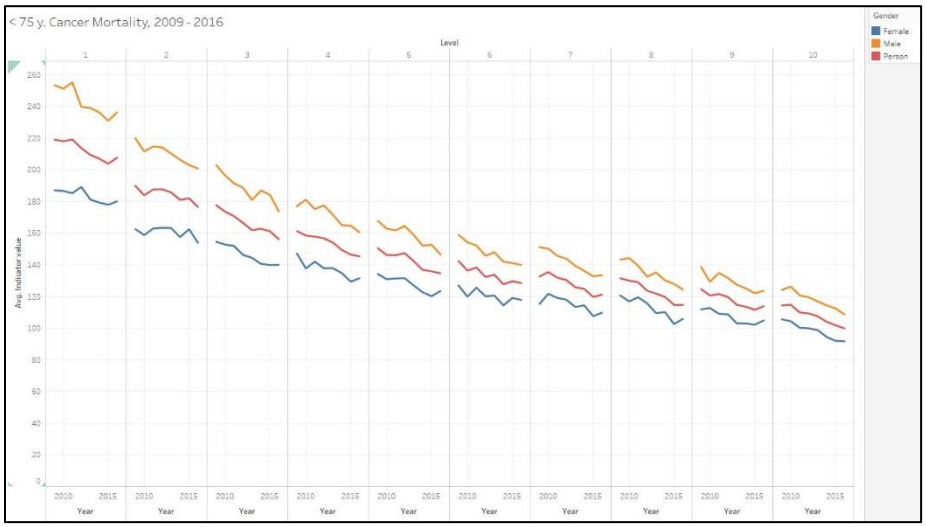


Figure b: < 75 y. cancer mortality by gender and year, by deprivation decile

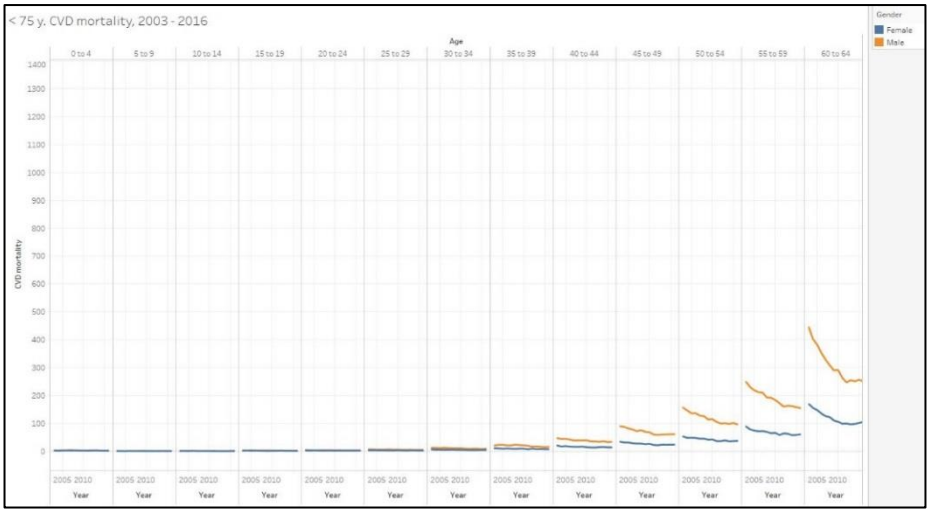


Figure c: < 75 y. CVD mortality by gender and year, by age group

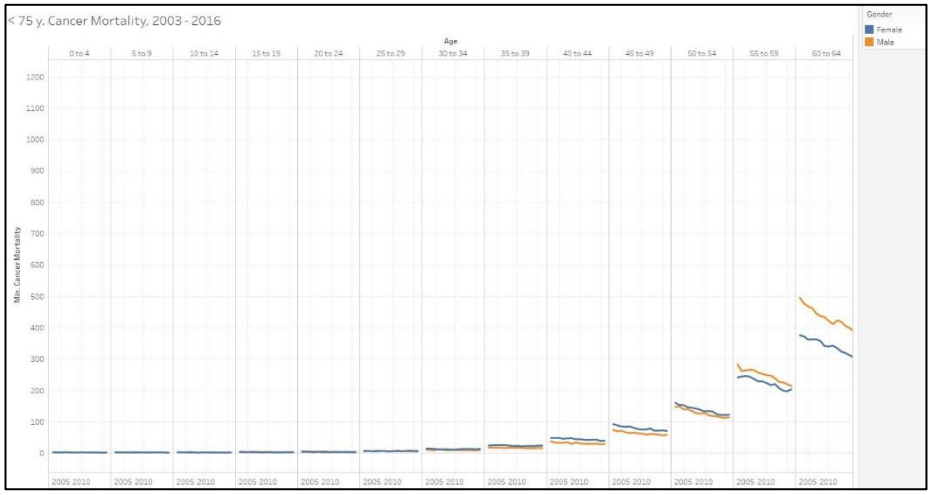
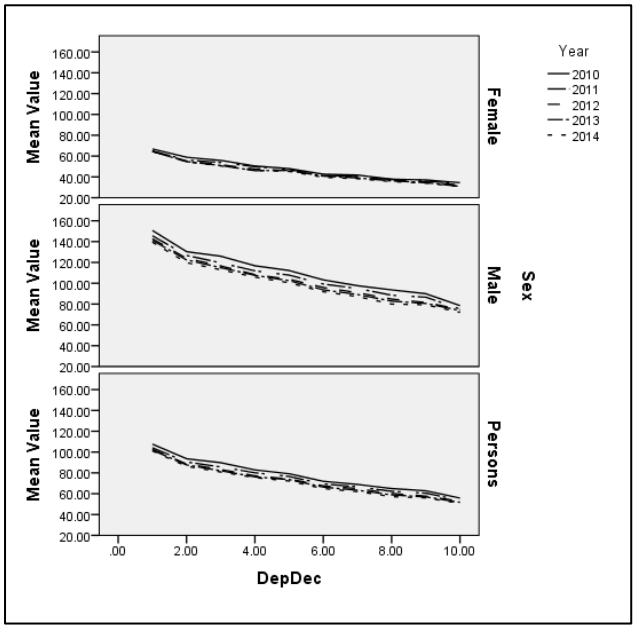


Figure d: < 75 y. cancer mortality by gender and year, by age group



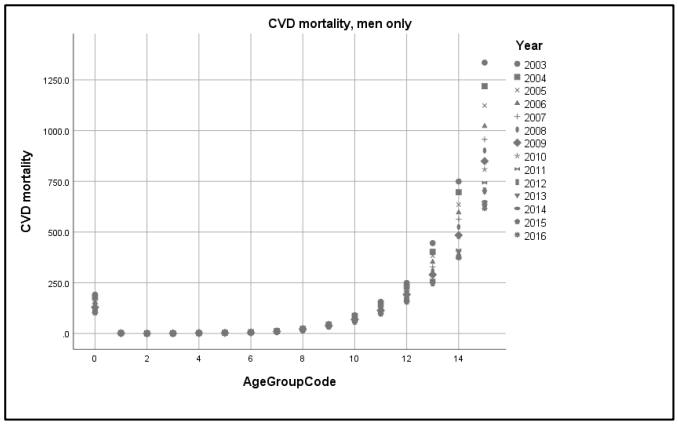


Figure e: Experiments with plotting approaches using black and white symbols and lines –
RD mortality by Deprivation decile by year (left), male CVD mortality by agegroup by year (above), and male cancer mortality by year and agegroup (below)