**Obtaining Demographic Information from Monthly GP Prescription Data**

Sarah Parks, 24/1/2014

**Summary**

This report highlights research to see if information on demographics, health or deprivation can be extracted from monthly GP prescription data aggregated by GP practice.

The GP prescription data is produced by the NHS Information Centre and is publicly available. It is available from August 2010 and is restricted to England.

The research has compared the number of prescriptions per person for sets of drugs with demographics, the proportion of the population in ill health or limited due to disability, and levels of deprivation.

The main findings are:

**There are drugs which correlate well with ill health, deprivation and the proportion of the population who are over 60**. The highest correlation is 0.9. This is for ill health, measured by whether individuals are limited, against prescriptions per person for non-opioid analgesics. For all of the variables high correlations were found. The variable with the lowest correlations is the proportion of the population over 60. The highest correlation for this is 0.76.

**Models can be built to predict ill health, deprivation, and the proportion of individuals over 60, which incorporate census data and drug data.** In these models the drug data significantly improves the models, showing that it contains information not present in any of the census data.

**There is a large amount of variation in GP prescribing patterns.** This variation is affected by a lot of factors, including, but not limited to, variation in advice given to GPs, accessibility of resources for certain treatments, variation in whether drugs are prescribed by hospital doctors of GPs, and accessibility of clinics. This variation is very hard to control against. Although additional data described in this report could help reduce some of this variability, it will not be possible to control against all of it.

This report starts with a description of the data, including previous work done on the data, and limitations of the data for our purpose. It then details the analyses which have been undertaken on the data, and finishes with conclusions and ideas for further work.

The findings in this paper show that, despite the large amount of variation in the data that is not related to the variables of interest, it is possible to find drugs which are highly correlated with demographics, the proportion of the population in ill health or limited due to disability, and levels of deprivation. What is more, these drugs significantly helped in modelling our variables of interest. This research can be used to provide evidence for obtaining more detailed data on GP prescriptions.

**Introduction**

This document describes exploratory analysis of monthly prescribing data by GP practice. Monthly GP prescription data was first made available in December 2011, initially providing information for September 2011. From 28 September 2012 more detailed data, called presentation level data, which gives descriptions of the specific drug and dose, was produced (initially for June 2012) and this data is now available from August 2010 onwards. The presentation level data is described in this document.

The aim of this project is to assess whether information on demographics, health, or deprivation can be extracted from this public GP prescription dataset. If it could this would be very useful as this data is monthly and public.

**The Data**

**What the data covers**

These datasets contain all prescriptions written in England by GPs and non-medical prescribers (nurses, pharmacists, optometrists, chiropodists and radiographers attached to GP practices) and dispensed anywhere in the UK in the community by month. This includes all prescribed medicines, dressings and appliances. For each medicine and practice, information is given for the number of items dispensed, the net ingredient cost, the actual cost and the quantity of tablets, capsules, liquid etc. Medicines are given by presentation, meaning that the exact drug and dose is provided. This is described both by the British National Formulary (BNF) name and the BNF code (a 15 digit alphanumeric code). The first 9 letters of the BNF code match chemicals listed in the Chemical file, which is also provided. All practices in England are included, as well as some "dummy practices" which represent certain environments such as specialist clinics, hospices, prisons, training units etc. Branch surgeries have been amalgamated into the main surgeries. Practices may close, open, merge or divide over time, so each months data may not contain the same practices. Each practice is listed using a code which matches the addresses in the address file. Information on Primary Care Trust (PCT) (which has recently been changed to Clinical Commissioning Group (CCG)) and Strategic Health Authority is also given for each practice. Data are released three months after the month the data is for (so in October for the July data). Prescriptions are labelled by practice they were prescribed by and date they were dispensed.

**What the data doesn't cover**

The data do not include private prescriptions, prescriptions which were not dispensed, and prescriptions prescribed anywhere other than a GP surgery or specialist GP-run clinics. Additionally 0.2% of community dispensed items were not included in the data as they could not be linked to a specific practice. The data also don't give information about the length of treatment or what the drugs were prescribed for. Information about whether prescriptions were paid for or free is also not provided.

**Accuracy of the data**

The figures provided are collected as a part of the reimbursement process of pharmacies. Pharmacists only receive payment on submission of prescriptions so the data are expected to be complete. There are however inaccuracies and the rate of them is internally audited at 2.5%. Even if errors get noticed and corrected for the pharmacies, the data are not changed so the inaccuracies remain.

**The British National Formulary**

Drugs are given by BNF code as well as by name. The BNF is the guide for GPs describing all the different drugs they can prescribe. It gives detailed information on dosages, reactions between drugs, etc. It is split up into 15 chapters which are then themselves split into numbered sections, paragraphs and sub-paragraphs. The chapters each correspond to a body system, such as the cardiovascular system, or the eye. The sections within them then correspond to a type of drug used for treatment, and then this is further split into the particular drugs and doses. Each drug has a code which corresponds to these sections; this is described below:

BNF code

Characters 1 and 2 show BNF chapter

3 and 4 BNF section

5 and 6 BNF paragraph

7 BNF sub-paragraph

8 and 9 chemical substance

10 and 11 product

12 and 13 strength and formulation

14 and 15 equivalent (same as 12 and 13 if generic or if branded but an equivalent generic exists, otherwise =A0)

Online access to the BNF is available from <https://www.medicinescomplete.com/about/> using username [kim.team@ons.gsi.gov.uk](mailto:kim.team@ons.gsi.gov.uk) and password Library1. Any problems with this should be directed to Catty Bennett.

As well as the BNF each CCG has its own formulary information online. This gives advice on cost effectiveness and suggests first choice drugs for some treatments. These fact sheets are specific to the CCG and vary over the country. They are also regularly updated. It is possible that when PCTs changed to CCGs some practices may have received new fact sheets which differed from their previous ones. GPs also get prescribing advice from their computers which now give warnings if drugs are prescribed that may interfere with each other. All of this means that care should be taken when comparing specific drugs, as advice on which drugs to give varies over the country. Broader analyses of sections or paragraphs is more appropriate.

**Other similar data**

Prescription data are also available in other forms. This data are less detailed than the monthly practice data and does not give much more information, so is probably not useful.

1) Quarterly PCT (now CCG) level prescribing data contains prescribing by GPs and all non-medical prescribers attached to practices including that which cannot be attributed to a specific practice. This is at BNF section level so is much less detailed in terms of the drug which has been provided. This is available from April 2008

2) Yearly prescription cost analysis prescribing data contains prescribing by GPs, all non-medical prescribers, dentists and all hospital prescriptions written anywhere in the UK dispensed in the community in England. This is done by BNF therapeutic class, which defines groups of medications used to treat a particular disease. This is normally BNF section level, but can be paragraph level. No location information is available for this data.

**Previous Work**

So far the data have solely been used to examine healthcare issues. The examples of this that could be found on-line are described below.

**Identifying prescription saving on statins.**

<http://theodi.org/news/prescription-savings-worth-millions-identified-odi-incubated->company

http://www.prescribinganalytics.com/

This project was carried out as a collaboration between a start-up, academics, and health care professionals. They focused on statins and used the cost data to show that prescribing branded statins instead of their generic equivalents, which are known to have exactly the same effect, wasted the NHS £27 million a month. They focussed on statins as it is well known that the branded and generic drugs have the same efficacy, however they are planning on applying the same analysis to many more drugs to try and find the NHS savings

**Looking at diabetes prevalence and spend per diabetes patient by Clinical Commissioning Group area**

http://openhealthdata.cdehub.org/

The project was carried out by the same start up as the previous project, and shows how spending per diabetes patient varies over the country. The aim of this is more to show what can be done with open data than to make any particular point. They will be adding to this website showing the use of different open data sets.

**Blog looking at the data**

http://paulbradshaw.wpengine.com/2012/04/27/exploring-gp-practice-level-prescribing-data/

This blog illustrates how you can manipulate this kind of data in Linux. The main purpose of this is to show how to explore data rather than focussing on this data in particular. He has uploaded the data to Google Fusion tables so that he can view locations on a map (some cleaning is required as some of the Post Codes are wrong as they don't all fall within England on the map, even though they should)

**Mapping English GP prescribing data: a tool for monitoring health-service inequalities**

http://bmjopen.bmj.com/content/3/1/e001363.full

They pick two drugs; metformin hydrochloride and methylphenidate, and compare prescription rates with reported incidence of relevant disease and spend per person registered at that practice. The idea is to show that patterns can be found over the country. They use demographic information for each practice from <https://indicators.ic.nhs.uk/webview> which has a breakdown by gender and age category for each practice list.

**Other Similar Work**

Prescribing data are also produced by other countries. Below are a few examples of how it has been used.

**Taiwan**

http://www.biomedcentral.com/1471-2458/11/380

Testing for heterogeneity of cardiovascular drug prescribing by township. They picked up disparity between Northern and Southern regions

**USA**

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3398217/

Geographic variation in prescribing decisions in diabetes per hospital referral region. Used regression models to investigate the relationships between prescribing rates and socio-demographic risk factors such as ethnicity, gender and income

**Limitations of the data**

There are a number of limitations of these data for obtaining population statistics. The first category of limitations concerns the details of the data provided. Currently no information is given on whether prescriptions were dispensed for the same person, either in one go or separately. There is also no diagnosis information, and even the presentation level data detailing the exact drug do not necessarily provide diagnosis information. Additionally drugs may not necessarily be given for the diagnosis they represent in the BNF. The data is also all per GP practice, meaning that the lowest geographical level which can be studied is Local Authority District (LAD) level. It also means we have to assume that individuals live near their practices. If the NHS Information Centre has the address for each individual then it could be mapped to a lower super output area (LSOA) and then the data could be provided at LSOA level.

There are also limitations of using GP prescribing data for obtaining population statistics at all, as prescribing patterns vary over the country. For example, for depression some communities may have a culture of prescribing a lot, or a culture of trying other treatment, this may also depend on the other facilities available in an area. There is also the problem that many drugs are prescribed in hospitals or specialist clinics and then continued in GP surgeries. The hand over point between hospitals and GP surgeries may vary over the country. Some drugs are also available in GP surgeries and specialist clinics and may not be handed over at all. Drug reps can also lead to variation in prescribing pattern as they may promote branded drugs, leading to these drugs being used significantly more than generic drugs in certain areas.

This second set of limitations means that a lot of care needs to be taken when interpreting any results.

**Data Analysis**

**Preliminary Analyses**

The data in this study run from August 2010 to July 2013. Over the 36 months there are a total of 12172 practices prescribing. 73% are present every month and others come and go (not necessarily in consecutive months). This is partly due to practices opening, closing, merging and splitting and also due to the “dummy practices” which represent GP run special clinics, prisons etc. which may not be prescribing monthly.

Practices have been mapped to Local Authorities. This is the lowest geographical area it is possible to go down to because at any lower level there will be many areas where there are no practices. Throughout this study it is assumed that individuals attend a GP practice within the LA they live in. This is a reasonable assumption, but is less likely to hold for regions with smaller areas, such as areas in London. We assume that the likelihood of going to a GP practice outside the area individuals live in is fairly even over the country, but this may not be the case. It is also assumed that the proportion of people using private GP services is constant over the country. This assumption is unlikely to be true, but hopefully the proportion of individuals using private services is in general low, so this will not affect the analyses too badly. Some variation in the data is likely to be seen for LADs which violate these assumptions.

The idea of the project was to explore if information on demographics, levels of health of the population, and levels of deprivation could be extracted from this data. This requires measures of these variables to compare with information from the prescription data.

Demographic figures are taken from the mid-year population estimates by LADfor the relevant year, both for the whole population, and for individual sexes where relevant. For old age the proportion of the population over 60 is used.

Deprivation is generally measured per LSOA. As the GP data can only be used at LAD level the average of the multiple indices of deprivation of the LSOAs inside the LAD, and the average rank of the LSOAs inside the LAD have been used as two measures of deprivation.

In the 2011 Census, levels of health are assessed using two questions; “how is your health in general?” which has the options “good”, “fair”, ”bad” and “very bad”, and, “Are your day-to-day activities limited because of a health problem or disability which has lasted, or is expected to last, at least 12 months?” which has the answers “yes, limited a lot”, “yes, limited a little”, and “no”. These questions have been used to calculate two measures of ill health; the proportion of people in bad or very bad health, and the proportion of people who are limited a lot.

A measure for drugs is also needed. One possibility is to use the number of prescriptions in each region; however this will be strongly affected by region size. Indeed, for the full dataset, there is a high correlation between the number of prescriptions in an LA and the population of that LA (0.948); this holds both for males (0.946) and for females (0.949). An alternative, which has been used in all of the following analyses, is to use the number of prescriptions per person in each region. This should remove effects of region size, however for very small regions estimates will not be as good as for big regions.

**Approaches for data analysis**

Two approaches of analysing the data were explored.

The first was to come up with hypotheses of drugs which may be related to one of the variables of interest, and then explore whether there were the expected correlations. Two of the hypotheses explored were:

* Assessing health by measuring the number of depression drugs prescribed
* Obtaining information about women by looking at prescriptions for contraceptives.

Unfortunately, drugs for depression pose problems when examining health, as different regions have different prescribing policies for drugs for depression, dependent on what other facilities there are available in the area. This means that patterns seen in the data may represent these differences and not health differences. Equally, contraceptives, although easily normalisable, are also a problem as they are not only prescribed by GPs, meaning that a number of the prescriptions are missing. These prescriptions are probably prescribed in sexual health clinics, which are not evenly spread over the country. This makes it difficult to discern whether variation in numbers of prescription over the country is caused by clinic access, or by demographic effects such as ethnicity, which the variation correlates with. Analysis of contraceptive data is in the R script Contraceptives.R and further details of the normalisation carried out and the analysis is presented in the file “GP prescription Data - How to use the data – practical advice and code”

All other hypotheses explored ran into similar problems so instead a data mining approach was taken. This approach involved calculating the correlation between the number of prescriptions per person in each region for all data in the BNF and then for each chapter, section and subsection in the BNF against all of the population measures described above. This splitting of the data produces sets of drugs to look at, instead of all the drugs. For example, Chapter 2 is all drugs prescribed for problems involving the cardiovascular system. The advantage of this approach is that no prior knowledge is required, and in fact that we are not limited by the knowledge we have. The disadvantage is that it is not possible to normalise the data easily, as before.

**Data mining**

The variables of interest are:

* The proportion of the population in each LAD with self reported bad or very bad health.
* The proportion of the population in each LAD measured by levels of disability.
* Average deprivation in an LAD.
* Average rank of deprivation in an LAD.
* The proportion of the population in each LAD over 60.

For each variable, the correlation of this variable with sets of drugs defined by the chapters, sections and subsections of the BNF was calculated. This gives a large number of correlations spanning the full range from -1 to 1. Taking a cut-off of correlations greater than 0.7, where the drug has been prescribed in every region, there are drugs for each variable which correlate well. The highest correlation which satisfies these conditions is 0.90 for non-opioid analgesics against ill health measured by disability. Non-opioid analgesics are prescribed pain killers such as paracetamol and asprin, but not opioid ones such as morphine and codine. It seems reasonable that individuals who are in pain, and hence on painkillers, would describe themselves as limited by their health. Tables 1-5 show all sections with a correlation of 0.7 or higher for our five variables of interest.

Table 1 Correlations between 2011 GP data split by chapter, section and subsection of the BNF against the percentage of the population in each LA with self reported bad or very bad health.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Chapter | Section | Subsection | Information | Correlation |
| 4 | 7 | 1 | Non-opioid analgesics | 0.88 |
| 4 | 7 |  | Analgesics | 0.84 |
| 3 | 1 |  | Bronchodilators | 0.82 |
| 3 | 1 | 1 | Adrenoreceptor Agonists | 0.81 |
| 3 |  |  | Respiratory System | 0.78 |
| 4 |  |  | Central Nervous system | 0.77 |
| 3 | 1 | 2 | Antimuscarinic bronchodilators | 0.77 |
| 10 | 3 |  | Drugs for the relief of soft tissue inflammation | 0.76 |
| 10 | 3 | 2 | Rubefactants and other topical antirheumatics | 0.76 |
| 2 | 6 | 1 | Nitrates | 0.76 |
| 12 | 3 |  | Drugs acting on the oropharynx | 0.75 |
| 1 | 1 |  | Dyspepsia and gastro-oesophageal reflux disease | 0.75 |
| 13 | 10 | 3 | Antiviral preparations | 0.74 |
| 1 | 1 | 2 | Compound alginates and proprietary indigestion preparations | 0.73 |
| 13 | 5 | 2 | Preparations for psoriasis | 0.72 |
| 13 | 9 |  | Shampoos and other preparations for scalp and hair conditions | 0.72 |
| 4 | 7 | 2 | Opioid analgesics | 0.72 |
| 13 | 5 |  | Preparations for eczema and psoriasis | 0.71 |
| 4 | 6 |  | Drugs used in nausea and vertigo | 0.71 |
| 1 | 2 |  | Antispasmodics and other drugs altering gut motility | 0.71 |
| 4 | 8 | 1 | Control of epilepsy | 0.71 |
| 4 | 8 |  | Antiepileptic drugs | 0.71 |

Table 2 Correlations between 2011 GP data split by chapter, section and subsection of the BNF against the percentage of the population in each LA who are limited a lot.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Chapter | Section | Subsection | Information | Correlation |
| 04 | 07 | 01 | Non-opioid analgesics | 0.90 |
| 04 | 07 | All | Analgesics | 0.88 |
| 03 | 01 | All | Bronchodilators | 0.85 |
| 03 | 01 | 01 | Adrenoreceptor Agonists | 0.83 |
| 03 | 01 | 02 | Antimuscarinic bronchodilators | 0.83 |
| 04 | All | All | Central Nervous system | 0.82 |
| 03 | All | All | Respiratory System | 0.82 |
| 02 | 06 | 01 | Nitrates | 0.81 |
| 04 | 06 | All | Drugs used in nausea and vertigo | 0.79 |
| 01 | 02 | All | Antispasmodics and other drugs altering gut motility | 0.78 |
| 04 | 07 | 02 | Opioid analgesics | 0.78 |
| 01 | All | All | Gastro Intestinal System | 0.76 |
| 01 | 01 | All | Dyspepsia and gastro-oesophageal reflux disease | 0.76 |
| 01 | 01 | 02 | Compound alginates and proprietary indigestion preparations | 0.75 |
| 10 | All | All | Musculoskeletal and joint diseases | 0.74 |
| 04 | 08 | 01 | Control of epilepsy | 0.74 |
| 04 | 08 | All | Antiepileptic drugs | 0.74 |
| 13 | 05 | 02 | Preparations for psoriasis | 0.73 |
| 03 | 02 | All | Corticosteroids | 0.73 |
| 05 | 04 | All | Antiprotozoal Drugs | 0.72 |
| 05 | 04 | 01 | Antimalarials | 0.72 |
| 10 | 03 | All | Drugs for the relief of soft tissue inflammation | 0.72 |
| 10 | 03 | 02 | Rubefactants and other topical antirheumatics | 0.72 |
| 11 | 08 | 01 | Tear deficiency, ocular lubricants and astringents | 0.70 |
| 02 | 06 | All | Nitrates, calcium-channel blockers, and other antianginal drugs | 0.70 |

Table 3 Correlations between 2011 GP data split by chapter, section and subsection of the BNF against average deprivation in an LA

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Chapter | Section | Subsection | Information | Correlation |
| 13 | 9 |  | Shampoos and other preparations for scalp and hair conditions | 0.84 |
| 1 | 7 | 1 | Soothing Haemorrhoidal preparations | 0.80 |
| 7 | 2 | 2 | Vaginal and vulval infections | 0.79 |
| 12 | 3 | 4 | Mouthwashes, gargyles, and dentrifices | 0.78 |
| 12 | 3 |  | Drugs acting on the oropharynx | 0.78 |
| 9 | 6 | 2 | Vitamin B group | 0.74 |
| 13 |  |  | Skin | 0.74 |
| 12 | 3 | 1 | Drugs for oral ulceration and inflammation | 0.74 |
| 4 | 9 | 2 | Antimuscarinic drugs used in parkinsonism | 0.73 |
| 13 | 3 |  | Topical local anaesthetics and antipuritics | 0.72 |
| 13 | 7 |  | Preparations for warts and calluses | 0.72 |
| 13 | 10 |  | Anti-infective skin preparations | 0.71 |

Table 4 Correlations between 2011 GP data split by chapter, section and subsection of the BNF against the average rank of deprivation in an LA

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Chapter | Section | Subsection | Information | Correlation |
| 1 | 7 | 1 | Soothing Haemorrhoidal preparations | 0.79 |
| 13 | 9 |  | Shampoos and other preparations for scalp and hair conditions | 0.78 |
| 7 | 2 | 2 | Vaginal and vulval infections | 0.77 |
| 12 | 3 | 4 | Mouthwashes, gargyles, and dentrifices | 0.74 |
| 12 | 3 |  | Drugs acting on the oropharynx | 0.72 |
| 13 | 3 |  | Topical local anaesthetics and antipuritics | 0.72 |
| 13 |  |  | Skin | 0.71 |

Table 5 Correlations between 2011 GP data split by chapter, section and subsection of the BNF against proportion of population who are over 60

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Chapter | Section | Subsection | Information | Correlation |
| 8 | 3 |  | Sex hormones and hormone antagonists in malignant disease | 0.76 |
| 8 | 3 | 4 | Hormone antagonists | 0.76 |
| 2 | 8 |  | Anticoagulants and protamine | 0.74 |
| 2 | 8 | 2 | Oral anticoagulants | 0.73 |
| 11 | 6 |  | Treatment of glaucoma | 0.72 |
| 6 | 3 |  | Corticosteroids | 0.71 |
| 6 | 3 | 2 | Glucocorticoid therapy | 0.71 |

This gives a number of drug sets which with high correlation to population figures. To assess reproducibility the analyses have been repeated using the GP prescription data from 2012, hypothesising that if the correlation changed significantly then drug set would not be useful for population estimates. There was very little variation in correlations between the two years, and the top hits shown in the tables above did not change.

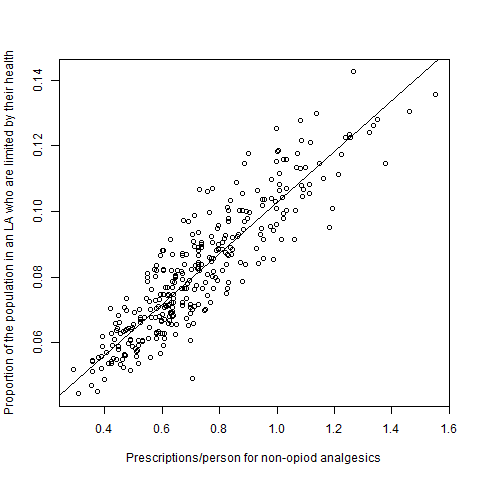
Figure 1 shows prescriptions per person for non-opioid analgesics against health measured by disability. This shows a good correlation with one obvious outlier which has a higher number of prescriptions per person than expected. This outlier is the Isles of Scilly, an LAD which is regularly an outlier. This is not surprising as this is the smallest LAD, and they will have different access to private practice, and alternative clinics, than the rest of the country, as they are an island. Some boroughs in London are also sometimes outliers. This may be because these boroughs have smaller areas, so individuals are more likely to go to GP surgeries in a different LAD to the one they live in, or it may be that people there are more likely to go to private surgeries.

**Building Models**

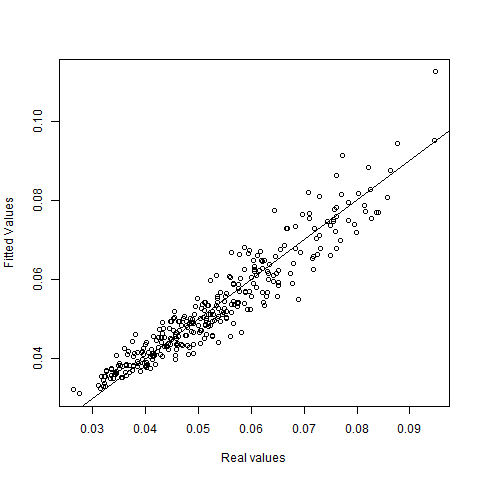
Linear models for the variables of interest can be built using the number of prescriptions per person for the top-hit drug. These can then be improved by adding other variables. Extra variables can be picked either from the census and other surveys, or from the lists of correlated drugs.

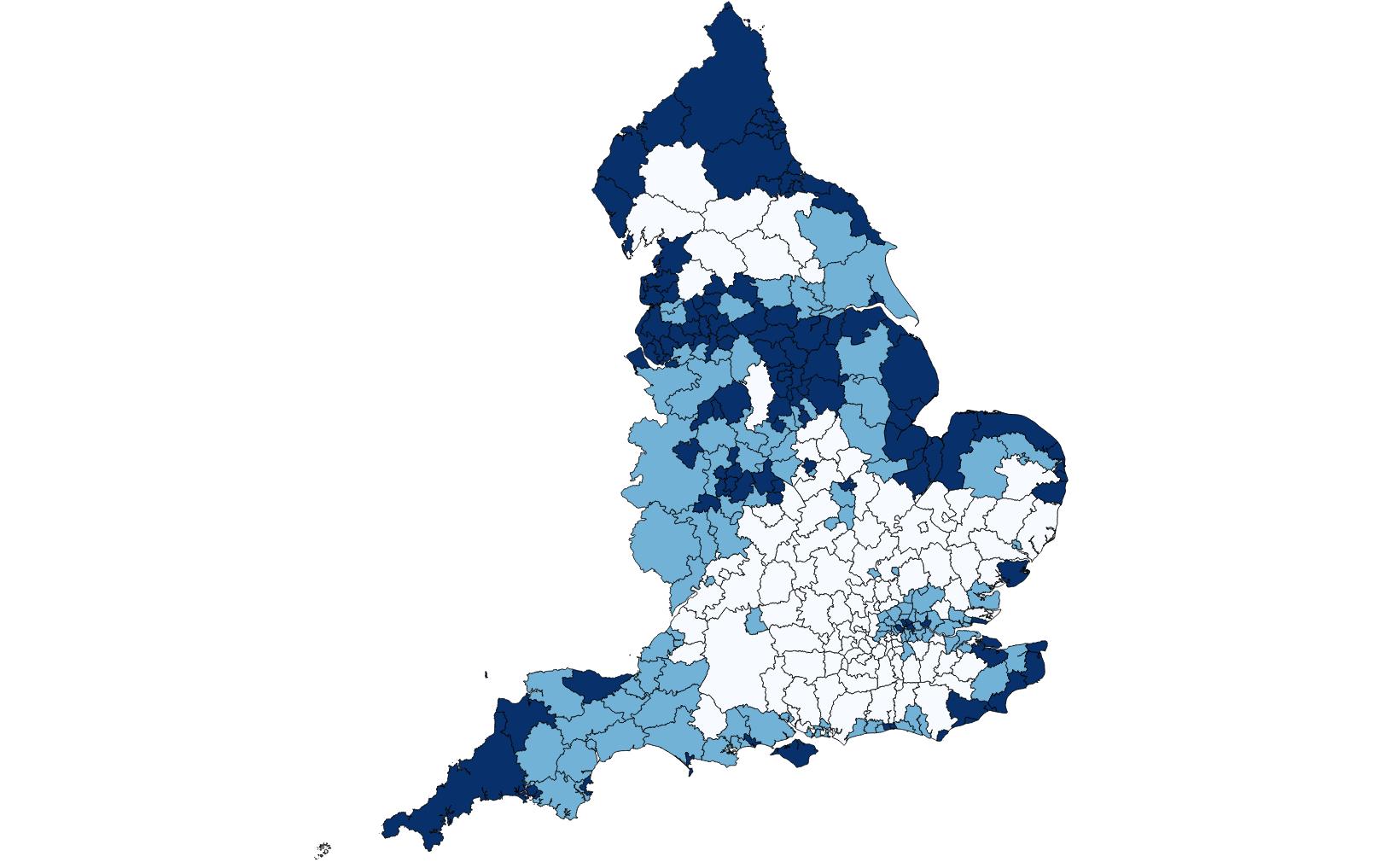
Other drugs which are also highly correlated with the variable seem ideal, as they can be easily calculated regularly. A problem with this is that as well as correlating well with the variable of interest these variables also tend to correlate well with each other. In a multiple linear regression this can lead to multicollinearity. To avoid this issue, instead of directly using the values of prescriptions/person for each drug, a principal component analysis can be performed on all the drugs with a correlation better than 0.7 to obtain orthogonal vectors which describe the variation in the data. Only principal components which correspond to at least 1% of the variation of the data have been included in the model. The cut-offs of 0.7 and 1% described above are arbitrary and

**Figure 1** Prescriptions per person for non-opioid analgesics against the proportion of the population in an LA who are limited by their health. The line shown is the line of best fit. This regression has an R2 of 0.78

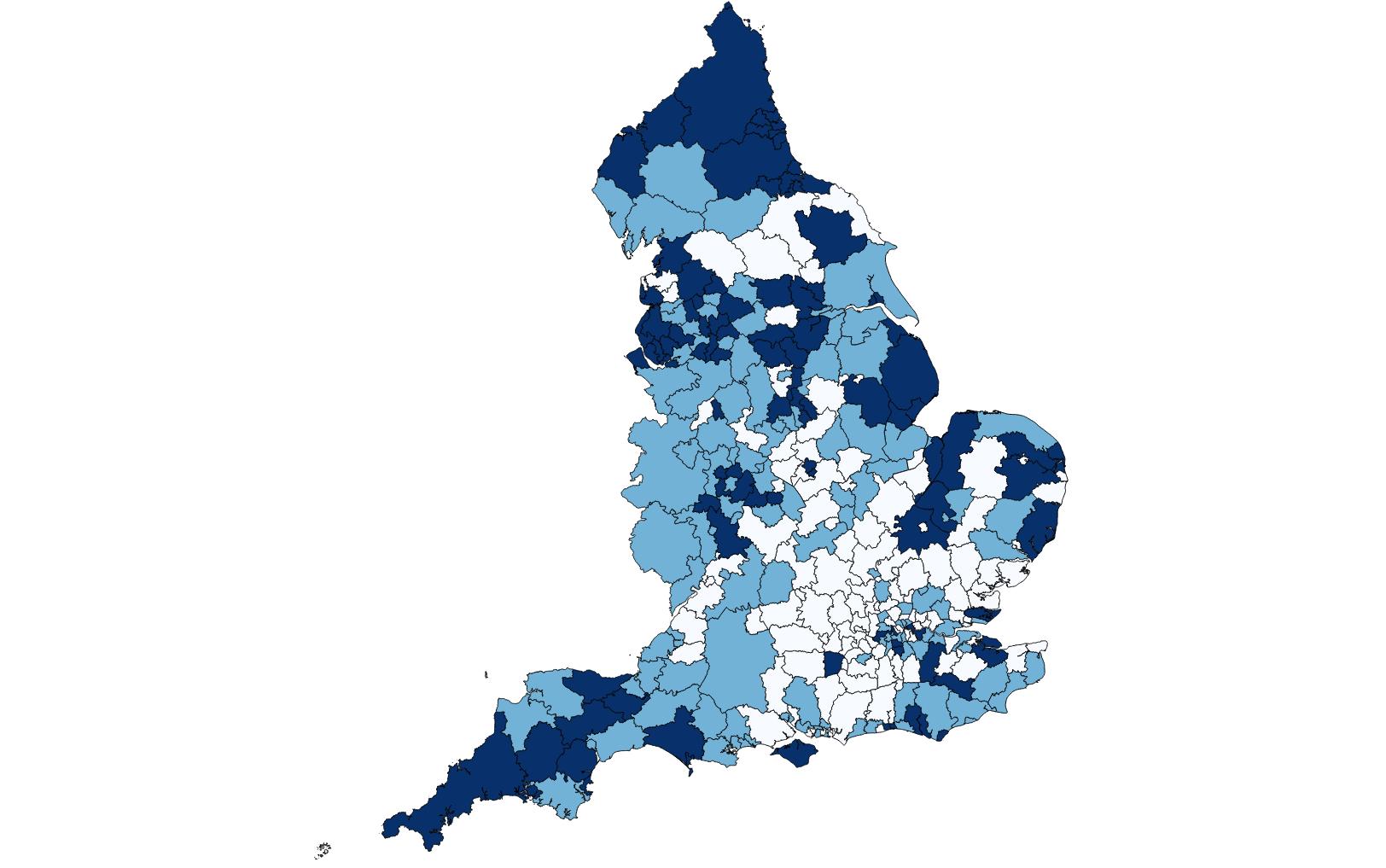


**Figure 2** Real values versus fitted values for the best model for health measured by disability. This model incorporates census data and data from the prescription dataset. This model has an adjusted R2 of 0.91.





**Figure 3** Health measured by disability mapped over England. Dark blue corresponds to the LADs in the lowest third of the Health data, light blue corresponds to LADs in the middle third of the Health data and white corresponds to LADs in the top third of the data.



**Figure 3** Fitted values of health measured by disability from the best model mapped over England. Dark blue corresponds to the LADs in the lowest third of the Health data, light blue corresponds to LADs in the middle third of the Health data and white corresponds to LADs in the top third of the data.

could be optimised. 0.7 was chosen as correlations of 0.7 or greater are often seen to be good correlations, and it gave a reasonable but not too long list of drugs. A cut-off was chosen to eliminate principal components because some of them represent only a very small amount of the variation in the data, so if the analysis was rerun with new data they would be likely to change and hence affect the model.

Other census or survey data may also be relevant to the variables of interest. For example, for levels of ill health, the proportion of individuals over 60, or the level of deprivation in areas may be relevant. There are also regularly published figures for unemployment, and in the census information was collected on the number of qualifications people have and whether people own their own houses. These things can all be incorporated into models.

The best model for each variable was built by including the principal components (as described above) and the relevant variables from other census and survey data into a linear model, and then performing stepwise backwards selection using the Akaike Information Criterion (AIC) to remove any variables which do not significantly improve the model. Model assumption checks, and outlier removal, was also performed.

For ill health, measured by whether people are in bad or very bad health, the proportion of individuals over 60, the average deprivation in the LSOAs within the LA and the top 9 principal components were included. The best model drops five of the principle components, but keeps all the other variables, and has an adjusted R2 value of 0.91. Figure 2 shows the fitted values of this model versus the real values. The adjusted R2 is a large improvement on the model which included just the top drug, which has an adjusted R2 of 0.78 (see Figure 1). It is also an improvement on a model built just using census data, and not using any of the drug data, showing that there is information in the drug data beyond what is in the census data used here. Figure 3 and 4 show the real and fitted values for health over the country with the LADs in the bottom third with respect to health in dark blue, the next third of LADs in light blue and the top third in white. It can be seen that, despite the high adjusted R2 of the model these maps look quite different.

For ill health measured by disability, and deprivation measured in both ways, models with adjusted R2 greater than 0.9 can be found. These models include both census and drug data. For the proportion of the population over 60 the best model has an adjusted R-squared of 0.86. All of these models can be built and tested using the Modeling.R script. This script will produce plots similar to the ones above, produce shape files for mapping, and carry out assumption tests producing both plots and p-values.

**Testing Models**

The models described above have high r-squared values and appear to fit well. Ideally we would like to test how good the models really are. The best way to do this would be to build the model using one year’s worth of data, and then use this model to make predictions for another year for which there is also data. Unfortunately for health and deprivation there is only data for one specific year so there is nothing to compare predictions to.

It is possible to test models based on the current data by splitting the LADs into two groups, building the models on one group and testing on the other group. For all models this shows that the models fit well, but it isn’t informative about the predictive power of the model in different years. Code using a test and a training set is also in the Modeling.R file.

**Conclusions and Future Work**

In this work both a hypothesis-driven and a data-driven approach were explored. The hypothesis driven-approach revealed limitations of the data, and issues which need to be considered in interpreting results, it did not however itself produce lots of useful results, perhaps due to difficulty in creating appropriate hypotheses. The data-driven approach did reveal a surprising number of drugs which correlated with demographic statistics. It did not however give much insight into how any of these results could be interpreted.

A limitation of the study is that although many models were produced, it is difficult to test their predictive values over time. This is because the demographic statistics predicted in the models, health and deprivation, are not calculated regularly, and so there is only one set of data for them in the three years the GP data covers. It is also difficult to know whether changes in prescriptions/person over time are related to changes in health levels in that area, or changes in policy for drug prescription. Any use of this data would need to be in close consultation with GPs to ensure results are interpreted correctly.

This work has revealed a number of limitations of GP prescription data, both in the form it is currently provided, and in the nature of the data itself. Despite this, it has been possible to find good correlations between the number of prescriptions per person in each region for certain drugs and demographic information such as the proportion of the population who are in ill health. Combining this data with other census outputs and the number of prescriptions per person for other drugs it has also been possible to build regression models covering a high proportion of the variance which could be used for prediction.

As they stand these models have a variety of uses. Firstly they could be used for quality assurance to check other outputs. Secondly the output could be used to show trends in the general health of the population over the country, and potentially to identify whether regions have particularly good or bad health. An important point to note is that including the drug data with census data in models significantly improved them. This implies that this drug data could maybe be combined with other data, such as survey data, to improve it.

I have not shown any analysis using time series, but the monthly nature of this data makes this a possibility. Given this and the variation in GP prescribing patterns which is present in this data this data is ideally situated to look at GP prescribing patterns over time, and potentially how policies affect prescribing patterns over time. An example of how different drugs can be looked at over time, which has not been discussed in this report, is given in the Contraceptives.R file.

Currently, health is measured in the census by the question “How is your health in general?” with possible answers of good, fair, bad and very bad. This question is both subjective and qualitative. This work indicates that the number of prescriptions per person of certain drugs could be used as an alternative quantitative measure of health.

The models discussed in this paper show that useful outputs could be produced from GP data. These outputs could be greatly improved if more detailed data was available. One big limitation of the data as it stands is that it is only possible to perform analyses at LA level due to the distribution of GP practices over the country. If the prescriptions were mapped to the address of the individual instead of the address of the GP practice then this would allow much more fine scale analyses. It is possible that this data is collected at the same time as the item and GP practice data. Prescriptions also usually contain information on the date of birth and sex of an individual, as well as information on whether they paid for their prescription or not, and why. This information would allow for much wider spread analyses to be run, and may allow for further demographic information to be obtained. This information may also allow for prescriptions to be aggregated per person, meaning that it would be possible to work out the distributions of prescription per person in different regions, potentially allowing for improvement of the measure of number of prescriptions per person.

Further improvements could be made if hospital and private prescriptions were also present in the data. It should however be noted that even if all prescriptions were present in the data there would still be prescription pattern variation present in the data. This variation may also vary over time, making it hard to control for. This is an inherent limitation of prescription data.

In conclusion this work has opened up possible avenues of research for calculating measures of ill health or deprivation, which could be greatly improved if more data was made available. The results of this work could be used to provide evidence for the potential usefulness of GP data in an attempt to prove that more detailed data is required.