ULSTER UNIVERSITY

DOCTORAL THESIS

# Analytics, Visualisation and Machine Learning of General Practitioner Prescribing using Open Health Data

*Author:*
Frederick G BOOTH

*Supervisors:*
Prof. M D MULVENNA
Prof. R R BOND

*A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy*

September 2022

I confirm that the word count of this thesis is less than 100,000 words.

*"In my life I have found two things of priceless worth – learning and loving. Nothing else, not fame, not power, not achievement for its own sake – can possibly have the same lasting value. For when your life is over, if you can say "I have learned" and "I have loved" you will also be able to say "I have been happy."*

Arthur C Clarke, Rama II

# Contents

# List of Figures

# List of Tables

# Acknowledgements

From the world of academia, I would like to acknowledge and thank Professor Maurice Mulvenna and Dr Raymond Bond for providing expert supervision, guidance, and expertise without which this project would not have been possible. I also want to thank Dr Kieran McGlade for his contributions to this research providing his unique perspective as a working GP. Finally, I would like to thank my co-authors for their contributions, and messages of support over the past three years.

To the people who supported me outside the university, I would like to thank my family for their patience; Sharon, Rachel for her technical contributions to my python and latex code and Grace for the constant supply of drugs (diabetic and blood pressure) and for her proof-reading skills. Special mention goes to Rick for constantly reminding me how good my supervisory team were and Peter for his brutally honest comments which, if nothing else, kept me smiling.

Lastly, I must acknowledge the people who provided daily companionship and kept me sane. Till Lindemann, Richard Kruspe, Paul Landers, Oliver Riedel, Christoph Schneider, Christian "Flake” Lorenz, Francis Rossi, Rick Parfitt, David Draiman and Leo Moracchioli to name but a few.

# Abstract

This thesis examined open General Practitioner (GP) prescribing data from 2015 onward to investigate the nature of GP practices in Northern Ireland (NI). Contribution to knowledge is embodied in the linking of multiple open data sources to create a novel data set, the use of data analytics techniques and machine learning to develop a method for the categorisation of GP practices based on location and prescribing behaviours, the comparison of these categories to discover differences in prescribing behaviours and possible contributing factors. One unexpected factor, COVID-19, and the national lockdown changed prescribing behaviours, and this was also examined. Finally, people's attitudes to the concept of citizen science via a GP prescribing dashboard was surveyed as a possible next step in making open data more accessible for anyone to analyse. It was found that whilst registered patients in NI had risen in line with population, the number of GP practices had fallen by 3.6% and comparing levels to that of other UK nations, NI had the highest prescribing in 6 of the 20 British National Formulary (BNF) chapters. The new method of categorisation employing machine learning clustering techniques found that two types of GP practice exist in NI, Metropolitan and Non-Metropolitan. Whilst they had similar prescribing patterns, prescribing levels were higher in half of the BNF chapters for Metropolitan practices with the largest variation being in the prescribing of Antidepressants and Analgesics. Possible factors contributing to the variations observed found a possible link to deprivation as a larger proportion of Metropolitan practices were in areas of high deprivation. The effects on prescribing due to the national lockdown showed a pattern of peak, trough, recovery. Antibiotic prescribing however did not recover to pre lockdown levels. Attitudes to citizen science were positive with 15.1% of participants contributing comments on resulting graphical output.

# List of Abbreviations

| | |
|---|---|
| **ADE** | **A**dverse **D**rug **E**vent |
| **API** | **A**pplication **P**rogramming **I**nterface |
| **ATC** | **A**natomical **T**heraputic **C**hemical |
| **BERT** | **B**idirectional **E**ncoder **R**epresentations from **T**ransformers |
| **BNF** | **B**ritish **N**ational **F**ormulary |
| **BSO** | **B**usiness **S**ervices **O**rganisation |
| **CSV** | **C**omma **S**eparated **V**alues |
| **CT** | **C**omputer **T**omography |
| **EPD** | **E**nglish **P**rescribing **D**ataset |
| **FHIR** | **F**ast **H**ealthcare **I**nteroperability **R**esources |
| **FSIP** | **F**ood **S**afety **I**nformation **P**latform |
| **GP** | **G**eneral **P**ractitioner |
| **HTML** | **H**yper**T**ext **M**arkup **L**anguage |
| **JSON** | **J**avascript **O**bject **N**otation |
| **kNN** | **k** **N**earest **N**eighbor |
| **LAD** | **L**ocal **A**uthority **D**istrict |
| **LCG** | **L**ocal **C**ommissioning **G**roup |
| **LGD** | **L**ocal **G**overnment **D**istrict |
| **LOF** | **L**ocal **O**utlier **F**actor |
| **LSOA** | **L**ower Layer **S**uper **O**utput **A**rea |
| **MRI** | **M**agnetic **R**esonance **I**maging |
| **MSOA** | **M**iddle Layer **S**uper **O**utput **A**rea |
| **NHS** | **N**ational **H**ealth **S**ervice |
| **NISRA** | **N**orthern **I**reland **S**tatistics **R**esearch **A**gency |
| **NI** | **N**orthern **I**reland |
| **NLP** | **N**atural **L**anguage **P**rocessing |
| **NN** | **N**eural **N**etwork |
| **OA** | **O**utput **A**rea |
| **OBJ** | **Obj**ective |
| **ONS** | **O**ffice for **N**ational **S**tatistics |
| **PCA** | Principal Component Analysis |
| **PET** | **P**ositron **E**mission **T**omography |
| **PRISMA** | **P**referred **R**eporting **I**tems for **S**ystematic Reviews and **M**eta-**A**nalysis |
| **RF** | **R**andom **F**orest |
| **RMSE** | **R**oot **M**ean **S**quare **E**rror |

| | |
|---|---|
| **RQ** | **R**esearch **Q**uestion |
| **SOA** | **S**uper **O**utput **A**rea |
| **SVM** | **S**upport **V**ector **N**etwork |
| **UI** | **U**ser **I**nterface |

# Chapter 1

# Introduction

> "In the beginning the universe was created. This has made a lot of people very angry and been widely regarded as a bad move."
>
> Douglas Adams,
> Hitchhiker's Guide to the Galaxy

Medical research is quite often the result of a collaboration between medical professionals and academic researchers in order to investigate a particular problem or illness. These studies are generally conducted using medical data which, in other circumstances, would not be available outside the National Health Service (NHS). As the question is already known, researchers formulate a hypothesis and then gather data to test whether the hypothesis is true or not. As this research is focused on a particular problem or outcome, a limited number of variables need to be included and these are identified at the start of the study. With the adoption of information technology (Prokosch and Ganslandt, 2009) large volumes of data are now being collected on a regular basis. The analysis of these data using machine learning and artificial intelligence algorithms provides a basis for informing both medical and policy decisions (Simpao et al., 2014).

Medical informatics, a sub-discipline of health informatics, directly impacts the patient – physician relationship. It focuses on using digital systems to collect data, develop medical knowledge and assist in the delivery of patient medical care. The goal of medical informatics is to ensure that patient medical information is available at the precise time and place it is needed to make medical decisions. Medical informatics is also used for the management of medical data for research and education (University of Illinois, 2021).

Health data analytics, on the other hand, is the analysis of healthcare data using quantitative and qualitative techniques in order to identify trends and patterns in the data (Grandview Research, 2021).

## 1.1 Big data

As technology has advanced, the ability to capture and store large amounts of data has resulted in the phenomenon we now refer to as *big data*. Analysis of these large data sets, often using machine learning or artificial intelligence, allow researchers to discover patterns or anomalies which in turn lead to the formulation of previously unknown research questions. As this is the case, we can see that big data relates not only to the volume and variability of the data but to the possibility of analysing these data in new ways to gain new knowledge (Krumholz M., 2014).

The analysis of big data presents a number of challenges which can be summarised in the '6 Vs of big data' (Andreu-Perez et al., 2015), - Value, Volume, Velocity, Variety, Veracity and Variability.

Value - As the cost of gathering and storing large volumes of data is not negligible, the data must provide value. This is not always evident at the time of collection and is only realised in full once the data has been analysed and the outcomes determined.

Volume - The volume of the data refers to the size of the data being generated. This may include sources such as text, audio, video, social networking, research studies, medical data, space images, crime reports, weather forecasting etc. These data are normally unorganised and not suitable for storage in conventional relational databases.

Velocity - Since, by its nature, big data are generated dynamically, the velocity relates directly to the volume of data being generated. Systems such as Stock Market applications must be able to capture and process large volumes of data at the same velocity as it is being generated in order to produce results that can be acted on immediately.

Variety - The data generated are in many different formats: audio, video, images, text etc., and is generated by humanity itself. Given the different formats, this constitutes variety in itself but adding the inevitable errors generated by the human element complicates things further.

Veracity - How trustworthy is the data? Since it is generated by various sources, we need to be able to stand over any results from the analysis of these data. For this reason, it needs to be checked for accuracy and duplication. Taking the Volume and Velocity of the data and the inevitable errors being generated, Veracity is of high import.

Variability - This may be an important factor when analysing large volumes of data. Trends may be seasonal or may change due to outside influences and these must be taken into account during analysis.

## 1.2 Open data

The advent of the *open data* movement in the 1990s has seen the increase of big data sets being made available for analysis. Unlike their big data counterparts, the open data files are often anonymised, aggregated and published on a regular basis rather than streaming. Many of the medical data sets contain geographical references which have not been adequately explored (Bohm et al., 2011), opening the possibilities for discovering patterns within geographical areas not previously defined.

## 1.3 Open health data

The term open data appeared for the first time in 1995 relating to geophysical and environmental data with the concept of open public data being defined in 2007, twelve years later. The following year the newly elected President of the United States, Barack Obama, signed two presidential orders relating to the concept of open government which encompass the concept of open data. Originally open data was meant to show transparency in government affairs but this quickly changed to the concept of improving government systems through research (Chignard, 2013). The concept of open data did not stay confined to government and quickly spread to other public bodies with anonymised medical data being included. The prescribing of medications by GPs, seen as an indicator of the health of the population, was one of the open data sets published in the United Kingdom with prescription data being first published in Northern Ireland from April 2013. The other three nations followed suit with England publishing prescription data from January 2014, Scotland from October 2015 and Wales from April 2018.

Analysis of open data must also consider the '6 Vs' although not all of the V's will necessarily apply to a single data set.

In choosing the data sets to be analysed, careful consideration must be made to ensure that it actually provides **Value** to the study, that it is relevant and provides insight into medical knowledge. The analysis mechanism must be able to cope with the **Volume** of data being fed into it with monthly updates being added. The **Velocity** of the data is not a concern although the system must be able to produce results in a reasonable time. It is likely that a **Variety** of different data sources will need to be linked in order to consider different aspects of the subject matter and these must be examined to ensure compatibility. The **Veracity** of the data must be considered to ensure that the data quality reliable. **Variability** may play a factor, especially when considering medical data as trends may change over time or be affected on a seasonal basis. In analysing geo-referenced data, a seventh V, **Visualisation**, will also be utilised in order to visualise results geographically in order to clarify trends.

## 1.4 Aim and objectives

As there is a perceived lack of analysis of geo-referenced medical data (Bohm et al., 2011), this thesis will use open prescribing data issued by Northern Ireland General Practitioner (GP) practices and dispensed at various pharmacies to provide new knowledge on prescribing practices and trends in Northern Ireland. The broad aim of the study is to provide new analysis on open health data by combining open data sources with geographical data to identify patterns or trends that can be used by healthcare professionals to inform policy and clinical decisions. The methodology for categorising GP practices should also prove valuable to other researchers, and concerned citizens who are interested in exploring open health prescription data. As the healthcare system works towards personalised health treatment, it is imperative that all variables are considered. As far as I am aware there has never been an analysis of healthcare patterns which has focused on matching the geographical location of the trends within Northern Ireland. It is therefore important to identify what, if any, influence geographical location has. In identifying these trends, it should then be possible to spot anomalous behaviour that could be indications of overloading of services, lack of services or possibly fraudulent behaviour (Carvalho et al., 2017).

The objectives of this study are:

- OBJ1 - to identify open source health data containing geographical references.
- OBJ2 - to identify and link relevant supplementary published data to create a novel data set.
- OBJ3 - to identify the types of GP practice using geographical location and their relationship with dispensing pharmacies using prescription data.
- OBJ4 - to investigate differences in prescribing behaviours between identified types of GP practice.
- OBJ5 - to understand what possible factors contributing to differences in prescribing behaviours of identified types of GP practice.
- OBJ6 - to understand the effect the COVID-19 pandemic, and in particular, the first national lockdown had on prescribing behaviours.
- OBJ7 - to develop and assess the usefulness of a GP prescribing dashboard in relation to being used as a citizen science tool.

## 1.5 Research Questions

- RQ1 - Can GP practices be classified in terms of their location with respect to the location of the pharmacies dispensing their prescriptions and does this validate the traditional classifications of Urban, Rural and Semi-Rural?

- RQ2 - If GP practices can be classified in terms of their location, what are the differences in prescription behaviour between classes?
- RQ3 - Is it possible to identify additional contributing factors to differences in classes?
- RQ4 - What effect has COVID-19 and the national lockdown had on prescribing behaviours?
- RQ5 - What interest do citizens have in prescribing behaviours?

Having potentially identified a gap in current knowledge regarding types of GP practice and their prescribing behaviours in Northern Ireland, an investigation into relevant literature and technical considerations will be undertaken in the next chapter.

## 1.6 Validation

Data validation, the process of ensuring the accuracy and quality of data was implemented in this project by building several checks to ensure the logical consistency of input and stored data. In the creation of the Local Data Store, all data was subjected to: data type checks to verify that each variable contains the correct type of data with no special characters which could lead to data rejection, range checks to verify that variable ranges are within expected parameters and format checks to check that variables with specific formats are entered correctly (e.g. Dates DD-MM-YYYY). In the machine learning elements of the project the Silhouette method was be used to validate the optimum number of clusters (k) with the Calisnki-Harabasz co-efficient used to validate both the choice of clustering algorithm and the final clustering solutions. Furthermore, the results of the analysis within the project was submitted as academic papers for peer review.

## 1.7 Research Publications

These are the articles that have been published as a result of the work in this thesis.

Examining the Effect of Deprivation on Prescribing Behaviours in Northern Ireland - 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) DOI:10.1109/BIBM49941.2020.9313132. (Booth et al., 2020a).

Examining the Effect of General Practitioner Practice Size on Prescribing Behaviours in Northern Ireland - 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) DOI:10.1109/BIBM49941.2020.9313570. (Booth et al., 2020b)

COVID-19 and lockdown: The highs and lows of general practitioner prescribing - 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI) DOI:10.1109/BHI50953.2021.9508575. (Booth et al., 2021a)

Discovering and Comparing Types of General Practitioner Practices Using Geolocational Features and Prescribing Behaviours by Means of K-Means Clustering - Scientific Reports Volume 11 DOI:10.1038/s41598- 021-97716-3. (Booth et al., 2021b)

Local Data Store developed for this project. DOI:10.5281/zenodo.6409927 (Booth, 2022)

## 1.8 Contribution to knowledge

The following chapters will provide detail on the contributions to knowledge resulting from this study. The contributions are as follows:

- Creation of a novel set of data within a Local Data Store for the analysis of GP practices and their relationship with the pharmacies dispensing prescriptions.
- First study to attempt to scientifically validate the Urban, Rural and Semi-Rural classifications traditionally attributed to GP practices.
- Creation of a novel technique to discover the types of GP practice based on location and relationship with dispensing pharmacies using prescription data.
- Identification of types of practice based on location and relationship with dispensing pharmacies using scientific analysis.
- Discovery of how deprivation affects prescribing behaviours and whether this is different each practice type.
- Discovery of how GP practice size affects prescribing behaviours within practice types.
- Discovery that practice size is not a factor in the identification of different types of GP practice.
- First study to examine the effect COVID-19 and the first national lockdown had on prescribing behaviours.
- Discovery of public attitudes to GP prescribing when presented as a citizen science data analytics dashboard.

## 1.9 Structure of thesis

This thesis presents a study primarily on the investigation of how geographical location of GP practices in combination with their associated pharmacies, in the form of the prescriptions issued and dispensed by both entities, define the type of GP practice and seeks to explore the differences in prescribing behaviours between archetypes. In doing so, it will demonstrate that value can be gained from the analysis of open data. Figure 1.1 illustrates the structure of this document and the relationship between chapters, which are summarised below.

- **Chapter 2**, provides a review of previous work that has been conducted combining GP prescription data, machine learning and open data.
- **Chapter 3**, details the data sources chosen for this study, the wrangling and how they were linked to create a novel data set.
- **Chapter 4**, provides a baseline for the following chapters by providing a study of GP practices and pharmacies at national level.
- **Chapter 5**, seeks to identify different types of GP practice using the geographical profile of the practices and that of the pharmacies that dispense prescriptions issued by them.
- **Chapter 6**, examines the differences in prescribing profiles of the identified archetypes.
- **Chapter 7**, explores two possible factors (deprivation and practice size) influencing the differences observed between archetypes in Chapter 6.
- **Chapter 8**, examines how the COVID-19 pandemic and in particular the first national lockdown affected prescribing behaviours in NI and compares these behaviours to those seen in England.
- **Chapter 9**, details the development and subsequent results of a study using a prototype data science dashboard based on the open data sets used in this thesis to explore attitudes of ordinary people to the concept of Citizen Science.
- **Chapter 10**, provides a summary of the main conclusions of the work, discusses the implications and suggests opportunities for further work in this field.

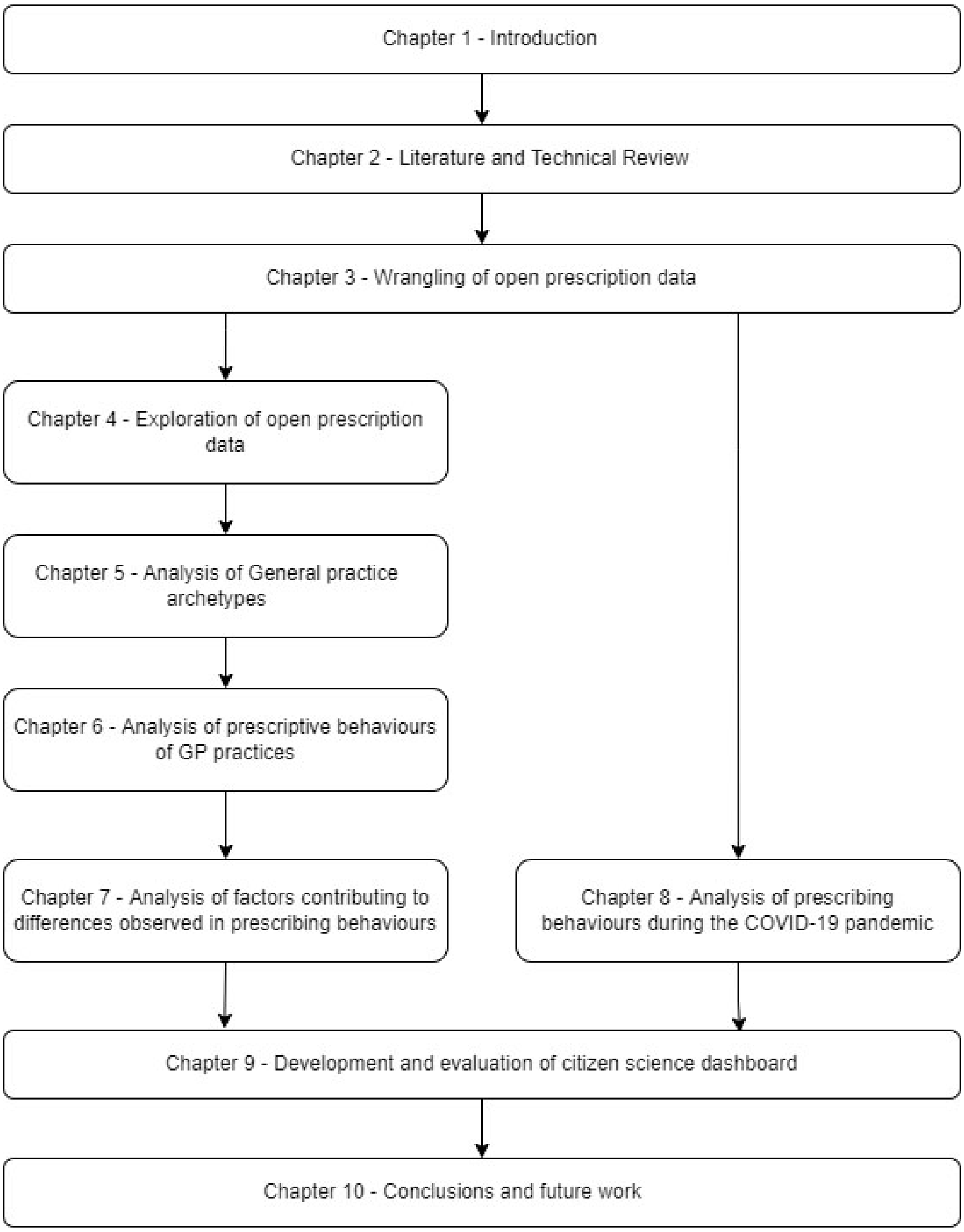

Chapter 1 - Introduction
Chapter 2 - Literature and Technical Review
Chapter 3 - Wrangling of open prescription data
Chapter 4 - Exploration of open prescription data
Chapter 5 - Analysis of General practice archetypes
Chapter 6 - Analysis of prescriptive behaviours of GP practices
Chapter 7 - Analysis of factors contributing to differences observed in prescribing behaviours
Chapter 8 - Analysis of prescribing behaviours during the COVID-19 pandemic
Chapter 9 - Development and evaluation of citizen science dashboard
Chapter 10 - Conclusions and future work

FIGURE 1.1: Structure of thesis document and relationship between chapters

# Chapter 2

# Literature review

> "Insufficient facts always invite danger."
>
> Spock
> Star Trek

The previous chapter outlined the perceived gap in current knowledge regarding the analysis of GP practices and their prescribing behaviours in Northern Ireland. In order to explore this further, previously published articles relating to the analysis of open prescription data using machine learning techniques with the different techniques available for analysis have been reviewed in order to inform this study.

## 2.1 Literature review strategy

Using the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) (Moher et al., 2009) as a guide, a review of relevant literature was performed to assess research undertaken in the area of study (Figure 2.1). The key words initially identified to search under were "General Practice", “Prescribing”, “Prescription”, “Data Science”, “Machine Learning”, “Artificial Intelligence”, “Data Analytics” and "Open Data". These were then combined to produce the following search string which was used to interrogate the IEEE Explore, Google Scholar, Scopus and PubMed databases:

- (“General Practice” OR “Prescribing” OR “Prescription”) AND (“Data Science” OR “Machine Learning” OR “Artificial Intelligence” OR “Data Analytics”) AND “Open Data”

The original search produced a total of 2,357 results (IEEE Explore = 135, Google Scholar = 1,930, Scopus = 183, PubMed = 109). Screening these on the basis of the title reduced the overall results by 2,296 with the majority of articles being rejected on the grounds of not relating to healthcare. The remaining 61 articles were then screened by reading their abstracts and a further 28 were rejected. Finally the full text of the remaining 33 was reviewed resulting in the 25 articles chosen below.

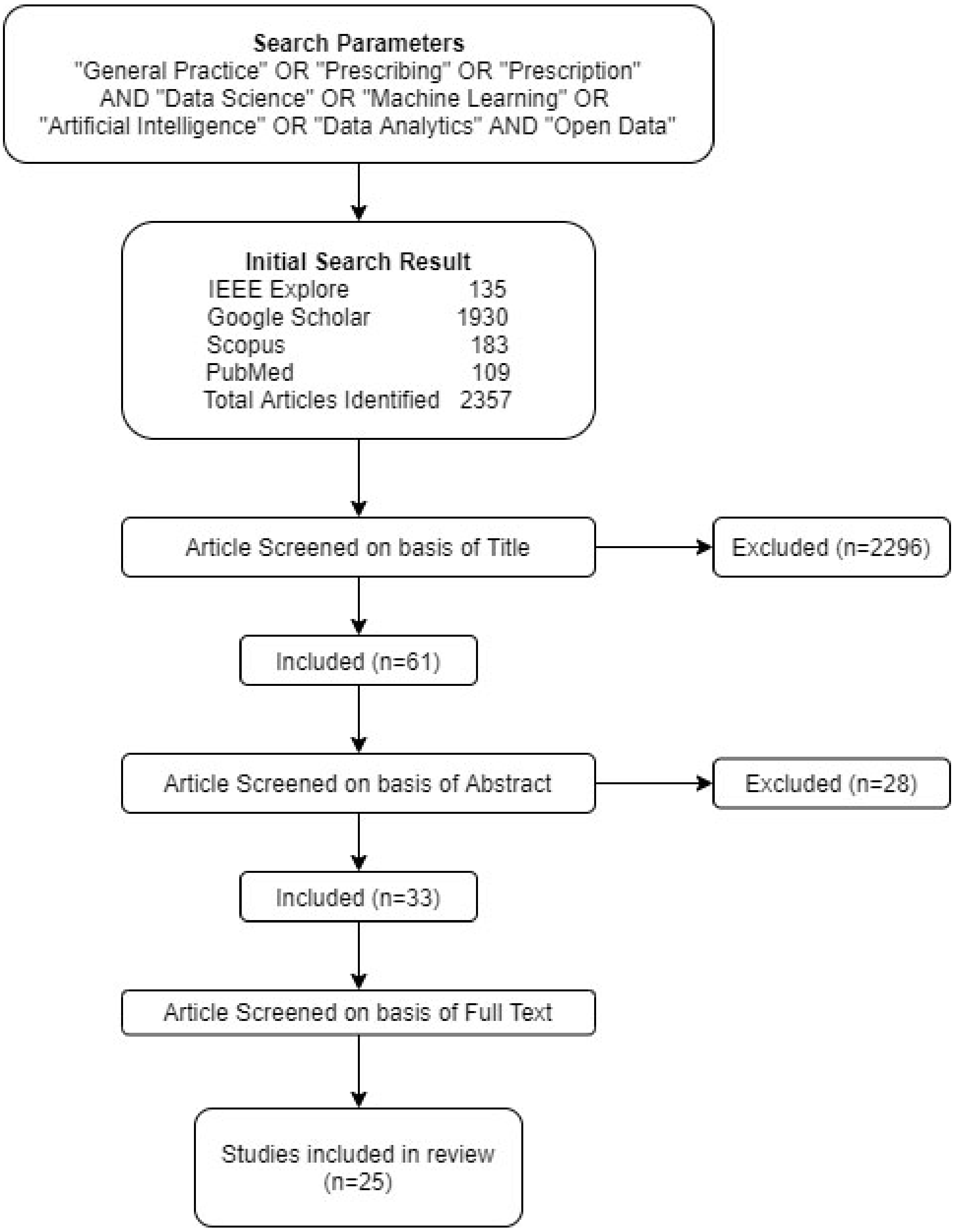

Search Parameters
"General Practice" OR "Prescribing" OR "Prescription"
AND "Data Science" OR "Machine Learning" OR
"Artificial Intelligence" OR "Data Analytics" AND "Open Data"
Initial Search Result
IEEE Explore 135
Google Scholar 1930
Scopus 183
PubMed 109
Total Articles Identified 2357
Article Screened on basis of Title
Excluded (n=2296)
Included (n=61)
Article Screened on basis of Abstract
Excluded (n=28)
Included (n=33)
Article Screened on basis of Full Text
Studies included in review
(n=25)

FIGURE 2.1: Literature Review process

## 2.2 Literature review summary

As the review failed to produce papers specific to the classification of GP practices using open source administrative data, literature dealing with aspects of the proposed research were reviewed in order to help frame the research. The results of the literature review produced 25 papers for review of which 13 related to studies undertaken involving open data or prescriptions (Fig 2.2). Three papers were identified providing literature reviews on machine learning and artificial intelligence in healthcare with a further four papers relating directly to artificial intelligence in healthcare. A further six papers were identified discussing issues around the provision and use of open data in healthcare with two papers detailing the development of applications to assist healthcare researchers analyse open data. Of the studies, five used prescription data from other sources with eight using open source data. The purpose of these studies were to identify general prescribing trends (4), identify trends in the prescribing of specific medications (2), estimate disease prevalence (1), identify adverse drug events (1), investigate the equality of prescribing between different areas and identify outliers in prescribing data (2). To achieve this, these studies used several machine learning / artificial intelligence techniques including Natural Language Processing, K-means clustering, Deep Learning, K Nearest Neighbour and Bayesian Regression.

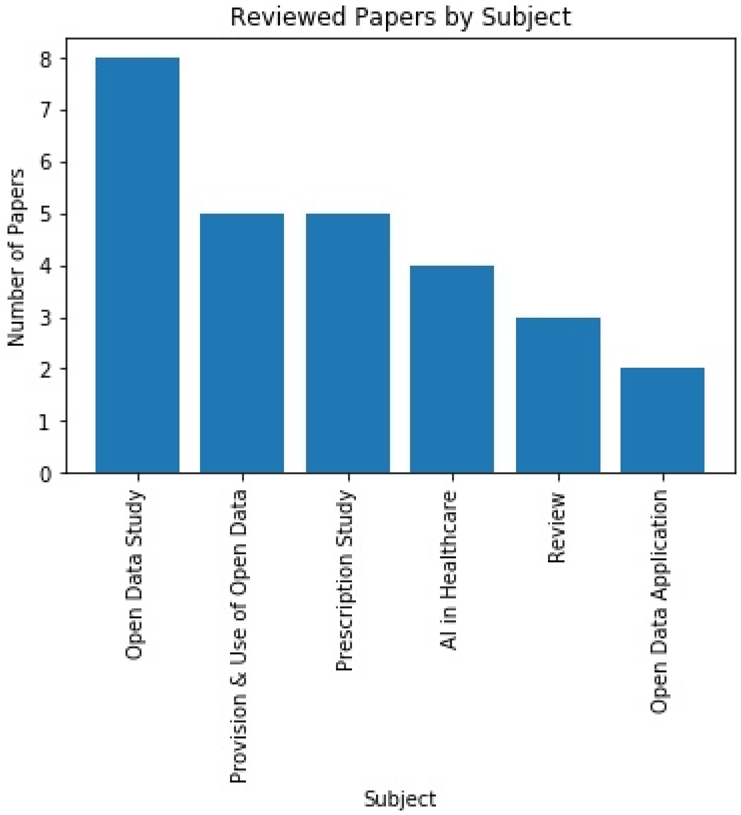


FIGURE 2.2: Reviewed papers by subject

### 2.2.1 Review papers

Three papers were identified providing a review of healthcare and healthcare analytics using big data.

Imran et al. (2021) comment on the growth of healthcare data and the opportunities presented by the analysis of such data. Ranging from the causation of diseases to the diagnosis of the same and increased efficiency and opportunities for financial savings. They do however acknowledge that the analysis of big data is often complicated and resource intensive with a high probability of failure. In order to present strategies for the analysis of big data, the authors provide a systematic literature review of publications dealing with research on big data analysis. They identify five challenges in big data analysis as being Confidentiality and Data Security, Granular Access Control, Interoperability, Data and Analytics reliability and Data Provenance. Centering their review around applications using NoSQL four types of databases, Key-value stores, Columnar stores, document stores and graph stores, they identify the limitations of these applications. The authors highlight a number of success strategies for the analysis of big data and propose a new architecture which reportedly solves the limitations previously identified in other applications. In comparing their work with other literature reviews they highlight the novelty of their work and present it as a road map for clinical administrators.

Bahri et al. (2019) seek to provide a survey of the issues surrounding big data in the healthcare sector. Defining the characteristics of big data, they look at how big data is generated, how it is collected and stored and what pre-processing may be involved before analysis can begin. Recognising that analysis is the most important step in the process, the authors look at the technologies available in each of the steps listing them and summarising their main attributes. Focusing on big data applications in the healthcare sector, the authors explore the use of machine learning and artificial intelligence systems within the sector. Looking at healthcare monitoring, the authors explore the use of wearable sensor devices and the potential benefits if these data could be streamed to healthcare professionals. The use of social networks for the collection of data on individuals allowing for a system of health prediction is also explored. Looking at the possible benefits to organisations, the use of machine learning and artificial intelligence systems to enhance the performance of departments by streamlining workflow or providing recommendation systems to allow healthcare professionals to make better decisions would be very beneficial. These systems could be built in tandem with healthcare knowledge and management systems providing all the data needed in one place for the treatment of patients. Finally the authors acknowledge that along with the opportunities big data analysis provides, it also comes with its own set of challenges. Given that the data comes from multiple sources it is often in differing formats and requires cleaning and pre-processing before any analysis can be done. The large volumes of data being generated also provides challenges in the provision of data storage and the processing power needed for analysis.

Reviewing the subject of big data analytics in healthcare, Yousef (2021) begins by defining big data as being too big to be analysed by traditional database protocols and having three main characteristics encapsulated in the three Vs – Volume, Velocity and Variety. Volume being how large the data is, Velocity referring to the speed with which the data is produced and Variety being the differing types of data. The authors see big data analysis as a means to manage performance of organisations or departments, assist in diagnosis of illness and give a more personalised patient experience. Detailing their search criteria for papers to be reviewed the authors then focus on the sources of big data and categorise these as being Electronic Health Records, Electronic Medical Records, Patient-Reported Outcomes, data collected from wearable sensors and data extracted from social media. Analysis is described as being Descriptive, Diagnostic, Predictive or Prescriptive with six techniques being Cluster Analysis, Data Mining, Graph Analysis, Natural Language Processing, Neural Networks and Machine Learning. Finally, looking at the challenges faced in analysing big data in healthcare, the authors list theses as being Privacy and Security of data, the issues around the storage and processing of large volumes of data, the ownership of the data and the level of skill needed in cleaning analysing and visualising the data.

A summary of the scope and number of papers reviewed (Table 2.1) shows that within the three studies listed, over 200 peer reviewed papers have been published during the past 16 years on the subject of artificial intelligence in healthcare. This review produced a total of 25 papers published between 2011 and 2021 (Fig 2.3) with 80% of these being published since 2017.

TABLE 2.1: Summary of review papers

| **Reference** | **Search Criteria** | **Scope of Review** | **Number of Papers** |
|---|---|---|---|
| Bahri et al. (2019) | N/A | 2013-2018 | 50 |
| Yousef (2021) | "big data" or "big data analytics" and "healthcare" or "medicine" or "biomedicine" | 2013-2020 | 58 |
| Imran et al. (2021) | Six basic search queries "big data", "NoSQL", "NewSQL", "big data tools", "big data techniques" and "big data analytics" were combined with "healthcare" then "healthcare analytics" resulting in a total of 18 search queries. | 2005-2021 | 99 |

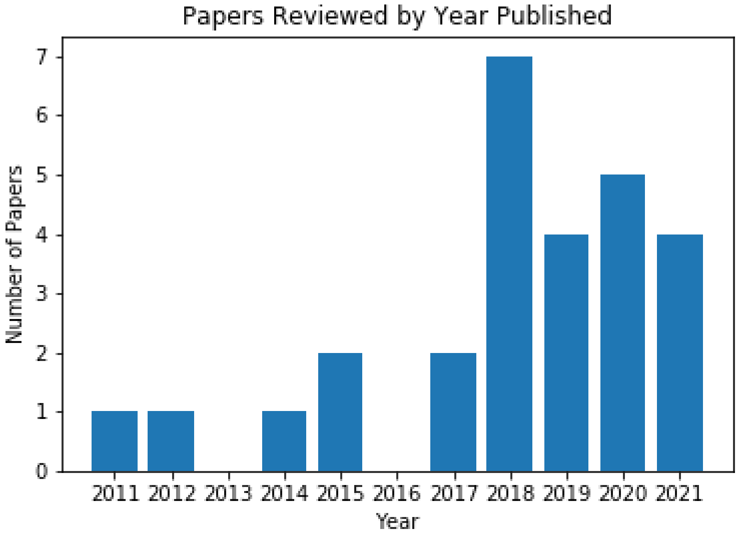


FIGURE 2.3: Reviewed papers by year published

### 2.2.2 Artificial intelligence in healthcare

Four papers were identified discussing the role of artificial intelligence in healthcare. Table 2.2 provides a summary of these papers highlighting the specific topic investigated in each.

Examining the role of artificial intelligence in data science, Gujral (2020) categorise its application into three categories – Descriptive, Predictive and Prescriptive. Unlike a human doctor, Machine Learning can look at the high volumes of healthcare data with artificial intelligence making recommendations based on that data. For artificial intelligence algorithms to perform well they rely on high quality training sets and the authors stress the need for research databases to be set up for this purpose. These databases should have access to arrange of data sources including mobile health application data, hospital data, insurance data along with data from social services and government reports. To identify specific diseases and match individuals to the correct treatment, personal health data, genetic profiles and life experiences must be added to theses databases to create a complete picture. Recognising that individual health information is scattered across various platforms the authors suggest that this information will eventually become linked to be used by predictive health models. This new linked data will need new data governance strategies to be out in place and Artificial Intelligence will only be successful if this happens. Relating this specifically to India, the authors list the current initiative to promote artificial intelligence in the country. They recognise the potential for artificial intelligence in the country but state that access to data and a multi-stakeholder plan are needed along with the promotion of research and development in this area.

In their paper Jiang et al. (2017) survey the current state of artificial intelligence applications in the healthcare sector taking into account the popular techniques and the major disease areas in which they are used. In particular they focus on machine learning algorithms for structured data such as Support Vector Machine (SVM), Neural Network (NN) and modern deep learning. For the analysis of unstructured data they focus on natural language processing (NLP). In considering the disease areas in which artificial intelligence is used, they identify cancer, neurology and cardiology as the three main areas and focus on early detection and diagnosis, treatment and outcome prediction and prognosis evaluation applications currently used for stroke victims.

Reviewing artificial intelligence in healthcare, Iliashenko et al. (2019) focus their attention on existing projects, AI company startups and developed applications. Two popular telemedicine applications using artificial intelligence are singled out for mention. These are Ada, a German application developed by a team of doctors, scientists and engineers and launched in 2016, and Your.MD, a powerful health information service founded in Norway, eventually moving their headquarters to the UK. Mapping the top artificial intelligence start up companies in order of their funding, the authors found that the highest number of startups were in the USA with 49 startups. This was followed by Israel with 7 startups then the UK with 6 startups making these three countries the world leaders using AI in healthcare. The top project is identified as BenevolentAI, a British AI company based in London which uses large quantities of mined and inferred biomedical data to advance the development of new drugs and medications. Taking a sample list of artificial intelligence applications, the authors then classify these by their purpose, their means of collecting data, their types of users and by the types of processed data. Finally the authors discuss the opportunities and challenges facing artificial intelligence in healthcare concluding that artificial intelligence is most commonly used for diagnosis assistance, management of healthcare enterprises and assistance in keeping a healthy lifestyle. The challenges being the necessity for specific architecture to be available, general prejudice against artificial intelligence by the public, concerns regarding privacy and information safety and the necessity for high quality, reliable services.

In reviewing recent breakthroughs in artificial intelligence technologies and their biomedical applications, Yu et al. (2018) focus on the development of the Aravind eye care system in India. This system, developed by a collaboration between ophthalmologists and computer scientists seeks to automatically classify diabetic retinopathy from retinal photographs of diabetic patients. As diabetic retinopathy affects more than 90 million people worldwide and is a leading cause of blindness in adults it is vital to be able to monitor the condition, identifying patients who could benefit from early treatment. Given that there are not enough ophthalmologists with the skills needed to read and interpret these photographs and follow up with each patient, the development of an AI system would greatly ease the burden. A team of researchers in collaboration with Google Inc successfully trained an AI system on thousands of images resulting in a physician level sensitivity and specificity in diagnosing diabetic retinopathy with the added bonus of being able to recognise cardiovascular risk factors from the same images. The technology has since been integrated into a chain of hospitals in India with the US Food and Drug Administration (FDA) approving its use in the United States.

TABLE 2.2: Summary of artificial intelligence in healthcare papers

| **Reference** | **Topic Investigated** |
|---|---|
| Jiang et al. (2017) | Early detection and diagnosis, treatment and outcome prediction and prognosis evaluation applications currently used for stroke victims. |
| Yu et al. (2018) | Automatic classification of Diabetic retinopathy from retinal photographs. |
| Iliashenko et al. (2019) | Existing healthcare AI projects, AI company startups and developed applications. |
| Gujral (2020) | AI Healthcare initiatives in India. |

### 2.2.3 Provision and use of open data

Five papers were identified on the subject of the provision and use of open data. Table 2.3 provides a summary of these papers highlighting the topic investigated in each.

In their thesis (Temiz, 2018) explores the concept of open data, how it occurs and attempts to understand the challenges surrounding the provision of open data and its use. Analysing the current provision of open data, the author seeks to identify the challenges presented to organisations who supply the data. The author believes that there are three main issues influencing organisations providing open data for use, these being the readiness of the organisation to provide it, the perceived effort in providing it and the perceived benefits. They also identify three issues which they believe do not influence the provision of open data, namely the perceived usefulness of the data, the perceived risk in providing the data and any external pressure to make the data available for use. The readiness of the organisation to provide open data was found to be the greatest influencing factor. Analysing the current state of open data in Stockholm, the author found that the provision took the form of closed websites providing an application programming interface (API) to supply the data and whilst this does provide an opportunity for the public to explore the data, it does not provide transparency or accountability for the organisation providing it. Furthermore, the author concluded that the lack of a centralised repository for open data in Sweden and the attitude that the analysis of these data may be important but vital for decision making contributed to the lack of knowledge in Sweden of what open data was.

Examining hospital medicines data, Goldacre and Mackenna (2020) highlight the contrast between primary care data that is open source and freely available and that of secondary care i.e. hospital data to which access is restricted. The authors point out that the data is kept by hospital pharmacies and that the barriers to access to these data is not technological but political, cultural, and contractual. The restricted nature of secondary prescription data limits any analysis to that of primary prescribing. The authors argue that the inclusion of secondary prescription data into any analysis would improve the overall analytics for the National Health Service and in turn generate insights to improve patient care.

Recognising the growing problem with negative rumors being circulated on social media Hsu and Cheng (2020) sought to develop an architecture to enable users to verify the veracity of information being posted. To do this they integrated six modules to create a Food Safety Information Platform (FSIP) that was integrated into Facebook to identify whether information on food safety being posted was credible or not. The six modules comprised of open data, machine learning, cloud computing, chatbot and an intelligent social application module. The application was tested using several cloud computing environments along with three machine learning algorithms – decision tree, logistic regression and support vector. Support vector proved to give the best results where training articles were less than 1000 but once the training articles exceeded 90,000 there was to difference observed in the algorithms. Testing the application on Facebook resulted in an accuracy of 0.769 proving that the application worked.

Examining the challenges and opportunities in sharing open data, Olsson (2020) hosted five focus groups comprising of both private company representatives and public officials and researchers. The concept of open data was compared to that of open-source software, a concept familiar with several the participants. It was acknowledged that whilst some had shared their software as open source, none had made their data open. Each focus group shared and discussed what data their organisations produced and could share, their attitude to sharing these data and the challenges and opportunities afforded by sharing the data. The concept of open data collaboration (ODC) between organisations was discussed leading to open innovation. Whilst participants found the concept interesting they felt that opening their data and processes to other was potentially giving away assets with the concept of losing a competitive advantage in the short-term. The participants acknowledged that change does not come easily and issues around business, technical, organisational and legal considerations must be explored to enable open data collaboration between entities.

Gebka et al. (2019) propose a methodology for generating value with open data that is not confined to the programmer. The methodology requires the participation of non-technical citizens to gather their priorities and understanding of various issues. The methodology comprises of six steps: 1. Formulate the challenge, 2. Invite participation of public servants and citizens with relevant experience of the situation, 3. Begin the workshop setting the challenge in context with the views and experience of participants and introducing the open government data available, 4.

Foster ideas for possible solutions, 5. Presentation of solutions with the best being voted for. and 6. Ideas published for others to implement. This methodology captures the needs of citizens allowing the publishers of open government data to identify which data sets are most useful, where gaps exist or which data is weak. It also promotes the knowledge of what open government data exists and involves citizens in the process of identifying solutions to problems.

TABLE 2.3: Summary of provision and use of open data papers

| **Reference** | **Topic Investigated** |
|---|---|
| Temiz (2018) | Identification of the challenges presented to organisations who supply the data, the lack of a centralised repository for open data in Sweden and the lack of knowledge on what open data is. |
| Gebka et al. (2019) | Proposal for a methodology for generating value with open data that is not confined to the programmer. |
| Goldacre and Mackenna (2020) | A case for the inclusion of secondary prescription data (Hospital prescribing) to be included in open data repositories. |
| Hsu and Cheng (2020) | Development of an application to verify the veracity and filter fake news from social media. |
| Olsson (2020) | Focus groups examining the challenges and opportunities in sharing open data. |

### 2.2.4 Development of open data applications

Two papers were identified describing the development of applications for use in analysing open health data. Table 2.4 provides a summary of these papers highlighting the application developed in each case.

Curtis and Goldacre (2018) chronicled the development of a web-based tool for research into prescription data in England. They complied open-source data for monthly prescriptions over an eighteen-year period. Each medication was mapped to the British National Formulary (BNF) model with financial data adjusted to account for inflation over the years. This tool allows the researcher to graph trends in English prescribing over time drilling down into the BNF structure down to chemical name for cost per item and quantity per item although it does not take into account the differing strengths and formulations of medication.

Recognising that not all medical researchers have a background in coding Hao et al. (2017) propose a system of creating user friendly applications based on Jupyter Notebooks with Jupyter Declarative Widgets and Jupyter Dashboards providing the user interaction with the underlying code. Jupyter Declarative Widgets allow the developer to link front end elements such as HTML to variables in the backend python code. Jupyter Dashboard provides the mechanism for the presentation of results generated by the back end to be displayed as HTML layouts. This practice would allow non-technical researchers to concentrate on the analysis of the data rather than the mechanisms in producing results. Users could adjust the analysis settings with a click of a button on the HTML front end without needing to know what the code below is doing. An advantage of this system is that since the HTML front end is bound to the underlying code, any settings adjustments input by the user will be reflected in the results in real-time. Whilst this method has the potential of making data analysis simpler for health researchers, the authors acknowledge that different data sets contain differing variables or order of variables (schema) limiting the developed application to the type of data source. Different applications would have to be developed for different schema.

TABLE 2.4: Summary of development of open data applications papers

| **Reference** | **Application Developed** |
|---|---|
| Curtis and Goldacre (2018) | Open Prescription dashboard for English prescribing data. |
| Hao et al. (2017) | User Friendly applications using Jupyter notebooks to allow non technical researchers to explore open data sources. |

### 2.2.5 Prescription studies

Five papers were identified relating to studies undertaken using prescription data. Table 2.5 provides a summary of these papers highlighting the subject of the study, the source of the data and methodology used along with the result.

Tackling the problem of identifying the prescribing of opioids in the United States of America, Wang et al. (2020) state that the problem lies in the fact that the information is generally held within the free text area of a prescription. Since this information is a valuable resource for medical research, they propose using machine learning in the form of natural language processing to extract this information. They propose to do this in four steps: 1. Identification of prescriptions from medical notes, 2. Identification of medications along with dosage and frequency from prescriptions, 3. Normalisation of medicine names and the filtering of opioid medications and, 4. Mapping these attributes to the Fast Healthcare Interoperability Resources (FHIR), a clinical standards mechanism. Testing the tool on 505 discharge summaries and 2000 opioid prescription records the results were promising but a few errors were encountered. These errors were categorised as being information of multiple medications being parsed to a single group of medication, missing information in the original prescriptions and ambiguous patterns forming in the data due to low frequency or not present in the training data. In conclusion the authors feel that the system could be adapted to other types of medication becoming a fundamental tool for the extraction of medication data from clinical notes.

In their study, Lee et al. (2014) hypothesised that there was a unique prescription pattern attributable to each hospital and that representing each hospital as a node on a network based on the similarity of prescription patterns, insight could be gained into the interaction between hospitals. This was done for 2010 and 2011 (before and after a new Healthcare policy was implemented). Calculating standard network statistics relating to the network in each instance, the authors compared these to ascertain whether the policy was effective or not. In most cases the authors saw a positive effect of the new healthcare policy. The authors also performed some clustering on the models to gain insight into clusters with similar prescribing patterns and to identify hospitals with a more central position within the network. The authors anticipated that using this method of analysis could provide valuable insights into public healthcare management.

Shin et al. (2015) performed an analysis of prescription patterns taking into account the patient's symptoms. Using a text mining approach, the authors treated prescriptions as documents and analysed the words on the prescriptions. In doing this, the authors were able to analyse the relationship between the patient's diagnosis and the prescribed medicines. The aim was to identify different prescription patterns for different practitioners allowing the authors to identify anomalous behaviour. Analysis of the data allowed the authors to examine the medications prescribed for specific symptoms and identify where costly medicines were being prescribed over the cheaper generic alternatives.

In this study, the authors (Bucholc et al., 2018) examined the prescription rates and prescription costs for General Practitioner surgeries in the Western Health and Social Care Trust in Northern Ireland with the aim to gauge the variations between surgeries and gain insights into the reasons behind them. Data for a consecutive three year period from April 2013 to March 2016 were analysed with Surgeries being categorised as Urban or Rural based on the number of residents in the same settlement as the surgery. Deprivation statistics were also used to categorise surgeries as being situated in either 'more deprived' or 'less deprived' areas. The results of the study indicated that there was higher prescribing rates for surgeries in more deprived areas and that there was a growing trend in prescribing rates. The authors acknowledged that these trends and the variations between surgeries did not necessarily indicate bad practice for surgeries with higher rates. The data examined in the survey allowed the authors to compare prescribing practices relating to branded drugs versus generic, and cheaper alternatives and showed evidence of inefficient prescribing by General Practitioners although surgeries prescribing more items per head seemed to prescribe cheaper drugs.

Pezzotti et al. (2011) posed the question of whether pharmacy data could be used to estimate the prevalence of chronic conditions, the authors compared data from the Italian Drug Prescription database to that collected from the Italian Board of Census National Health Survey (ISTAT) 2004-2005 and other sources. The authors found that 20 chronic conditions could be identified via the pharmacy data and that prevalence rates were generally higher from the ISTAT survey. Diseases such as Cardiovascular, Diabetes and thyroid disorder matched the prevalence rates in the ISTAT survey. It was found that estimates based on small geographical areas were not constrained by sampling problems but this was not possible at national level. Considering the limitations of the analysis, the authors concluded that for certain chronic conditions such as diabetes, pharmacy data could be used to approximate prevalence.

TABLE 2.5: Summary of prescription studies papers

| Reference | Subject | Data Source | Methodology | Result |
|---|---|---|---|---|
| Pezzotti et al. (2011) | Prevalence of Chronic conditions | Italian Health Administration databases | Descriptive statistics | By identifying the drugs used to treat chronic conditions the authors were able to estimate the prevalence of these conditions within the Italian population. |
| Lee et al. (2014) | Hospital prescribing patterns | Two data sources comprising 9,589 and 9,715 prescriptions | Network Analysis | The method was applied before and after policy changes within the network and verified that the changes had a positive effect on prescribing. |
| Shin et al. (2015) | Analysis of prescribing patterns to identify anomalous behaviour | Sample of outpatient prescription data | Text Mining using Naive Bayesian | Identified costly medications prescribed for specific symptoms over the cheaper generic alternatives. |
| (Bucholc et al., 2018) | General Practitioner prescribing patterns | prescription data from 55 GP surgeries | Descriptive statistics | Comparison of prescribing practices showed evidence of inefficient prescribing by General Practitioners. |
| Wang et al. (2020) | Opioid Prescribing in USA | 505 discharge summaries and 2000 opioid prescription records | Natural Language Processing (NLP) | Proof of concept for a tool to extract medication data from clinical notes. |

### 2.2.6 Open data studies

Eight papers were identified describing studies undertaken using open health data. Table 2.6 provides a summary of these studies highlighting the subject of the study, the source of the data and methodology used along with the result.

Rezaei-darzi et al. (2021) utilising the open prescription data published by the National Health Service (NHS) and combining this with demographic data provided by the Index of Multiple Deprivation provided by the Ministry of Housing, Communities and Local Government, the authors propose to provide an analysis of the equality of prescribing Novel Oral Anticoagulants in England. These Novel Oral Anticoagulants are used in the treatment of a heart condition called Atrial fibrillation. Using a Bayesian model the authors examine prescription patterns for these medications by NHS Clinical Commissioning Group in England allowing the investigation of their geographical distribution over time. The authors cite this model as being a new approach to modeling prescription data which can be used for medications prescribed for chronic illnesses with the medication being taken over long periods.

Fan et al. (2020) highlight the problem of side effects from taking prescription drugs with approximately a third of hospital admissions being a result of adverse drug events (ADEs). The aim of the study was to create a method to identify unreported events from open data sources. Two web-based sources, WebMD and Drugs.com were used with ten thousand reviews gathered for analysis. These were manually labeled, and a deep learning algorithm was then applied to detect ADE's. Bidirectional Encoder Representations from Transformers (BERT) based models were used and compared to standard deep learning models. The authors found that the use of BERT models improved both the extraction and detection results over those of the standard models and could be used for other healthcare information extraction tasks.

Recognising that little had been done in relation to analysing open health data with a view to inform both policy makers and decision makers within the health service, Cleland et al. (2018) set out to explore the relationships between antidepressant prescribing, depression prevalence and economic deprivation. Data was gathered and cleaned with prescribing data converted from the British National Foundry (BNF) system used in the UK to the Anatomical Therapeutic Chemical (ATC) classification system to allow international comparisons. GP practices were also assigned to Super Output Areas (SOA) to link these to deprivation data. The authors found that there was a strong correlation between antidepressant prescribing practices and the multiple deprivation figures for each practice noting a couple of anomalies in the results. On the other hand, weak correlations were recorded between Prevalence and Depression and Prevalence and Prescribing, respectively. Focusing on the anomalies identified, the authors felt that identified practices with unusually high prescribing practices warranted further investigation and possible intervention. The authors also concluded that considering the weak correlation between depression prevalence and antidepressant prescribing, further investigation would be required as to why prescription rates were rising in Northern Ireland.

Using publicly available data relating to emergency calls of suspected heroin overdoses, Richard et al. (2019) used a Bayesian regression model to examine the relationship between different demographics and environmental characteristics attributed to the areas in which the calls originated. Environmental variables included the proportion of parks, commercial premises, manufacturing premises and downtown development zones along with the distance from pharmacies. Demographic variables included the age and sex of patients, level of education, household income, the number of fast-food restaurants, distance to local hospitals and to opioid treatment centres. As a result, the authors were able to build a statistical model of the environmental and demographic features associated with higher numbers of emergency calls relating to heroin overdoses in the city of Cincinnati with the aim to inform policy makers.

Tackling the problem of missing data in patient records Khan et al. (2012) propose a hybrid solution combining an artificial intelligence knowledge-based system with a logic-based inference system to provide a robust decision support system for clinicians. The system was built to tackle the problem of insomnia and was based on three real world data sets – Patient records from a telephone survey conducted by the Centre for Disease control, prescription protocols relating to the prescription of sleeping medication and a drug interaction registry to identify negative interactions between drugs. Mapping the patient record data to named variables (e.g. sex = 2 maps to sex = Female) the authors created a knowledge based system with a logic-based reasoning machine learning system to impute missing data in the data set. This was then linked to known insomnia prescriptions to identify the conditions under which various medications were prescribed. Using complete records from the patient data set, the authors tested their system by inputting the missing data for the incomplete records and subjecting the complete data set to the knowledge-based system to decide what medications should be prescribed in each case. These results were then compared to the prescription protocols to establish the accuracy of the system. The hybrid model was also tested against a simple knowledge-based model. It was found that both models suffered from a degradation in performance as the amount of missing data increases although the hybrid model only suffered a 1% degradation as opposed to a 4% degradation in the knowledge-based system. The hybrid system was found to provide effective decision making in recommending insomnia medication.

In this study Rich et al. (2015) investigated the use of Open health data in understanding health trends. Acknowledging that clinical data on its own was not particularly a good indicator, the authors looked to enhance their data set by linking it to other external factors such as possible causes and population trends in order to attempt to predict future trends based on this historical data. Analysing diabetes related hospitalisations in the New York Area, the authors developed a platform for analysing and visualising the data using geoplots. In developing this platform, the authors acknowledged a number of issues. The frequency of updates to the data could be sporadic with some data being released at different time periods than others. Often the released data was not in a suitable format and required a good deal of reformatting before it was usable. Also, the use of anonymised data did not lend itself to analysing trends at neighbourhood level.

Rao et al. (2018) proposed an iterative k-means method for identifying outliers in open health data. In essence, the authors proposed that a k-means algorithm be applied to the data to identify outliers. This was done for a fixed number of times, or until no further outliers were identified. In this way they could drill down into anomalies to show that, for instance, an anomaly existed in 'Suicide', 'Age Group 30-49' or that Hospitals in a particular geographic area were under performing. These outliers could then be examined by a human to ascertain the reason for them.

Working on unsupervised data sets, Zhang and Wang (2018) investigated a possible method for detecting outliers from prescription data. Having considered the correlations in the data such as drug and age, the authors identified these as criterion for identifying outliers. Combining calculations for Global Outliers, Local Outliers and Feature weight the authors derived an Outlier figure representing the largest distance from their k nearest neighbour. The authors claimed that their method minimised the number of false positives identified and, having tried their method on real world data, felt that the method showed promise.

TABLE 2.6: Summary of open data studies papers

| **Reference** | **Subject of Study** | **Data Source** | **Methodology** | **Result** |
|---|---|---|---|---|
| Khan et al. (2012) | Missing Patient Data | 450,000 patient records | Knowledge bases | Proved that missing data could be imputed in order to support clinical decision making. |
| Rich et al. (2015) | Diabetes control and prevention | Hospital discharge information | Time based simulation using geoplots | The simulation was able to highlight neighbourhood hotspots for diabetes. |
| Zhang and Wang (2018) | Outlier detection | Prescription data on 7,979 individuals | Global and Local outlier detection | Method promising for detecting outliers in real healthcare data sets. |
| Rao et al. (2018) | Outlier detection | New York State open healthcare data | k-means clustering | Method identified several anomalous trends in the data. |
| Cleland et al. (2018) | Antidepressant prescribing | Northern Ireland open prescription data | k-means clustering | No link between antidepressant prescribing and social deprivation found. |
| Richard et al. (2019) | Heroin related overdose incidents | Call data for 6,246 Emergency calls. | Bayesian space-time Poisson model | Identification of environmental and demographic characteristics associated with higher levels of heroin overdose calls. |
| Fan et al. (2020) | Adverse Drug Event (ADE) Detection | 10,000 social media health related reviews, | Deep Learning | Study shows the viability of using deep learning for ADE detection. |
| Rezaei-darzi et al. (2021) | Prescribing Novel Oral Anticoagulants in England | English open prescription data | Bayesian small area analysis | New approach to modeling prescription data for medications for chronic illnesses over long periods. |

## 2.3 Discussion

Almost half of the articles found (48%) considered issues around the use of open data, machine learning and artificial intelligence in the healthcare sector. Three provided reviews of previous published literature, four considered the role of artificial intelligence in healthcare specifically and two described the development of applications to enable non-technical researchers to analyse open data. Of the studies identified, six investigated prescribing patterns although four of these were relating to specific medications. Four of the studies compared prescribing between different geographical areas although none attempted to define the prescribing in terms of geographical location and prescribing behaviours. Table 2.7 provides an overview of the subject matter of the identified studies with general prescribing and outlier detection being the most popular. The amount of data collected and analysed in the studies (Fig 2.4) ranged from one to eighteen years although many of these comprised of samples, not complete data sets. Analysis methods varied depending on the purpose of the study (Fig 2.5) with the majority of studies using k-Means clustering.

The review papers provide an interesting overview of the issues to be aware of when analysing big data, the technologies available for analysis and previous work that has been undertaken. As such, they provide valuable background reading and will help to inform decisions during the course of this project and provide an overarching synopsis of the available literature. Papers dealing with Artificial Intelligence in healthcare all document AI systems developed to facilitate the screening of patients for various diseases and use the train/test model to validate their results. As such the processes developed in these papers are likely to have a global impact with their use being adopted by various healthcare organisations. Papers under the heading "Provision and Use of Open Data" provide a valuable overview of the issues involved in open data. They provide insight into the cost of providing open data, the challenges an opportunities associated with sharing open data and proposals for more data to be added to open data repositories. On the subject of the development of open data applications, these papers document the development of applications to enable citizens with no programming/data science skills to analyse data. In both cases the developed applications are applied to specific domains but could be reengineered for other types of data. The authors validate their solutions by gathering feedback from users of the applications. The solutions differ in that one uses web technologies which can be accessed by anyone with an internet connection and a browser, the other relies on a combination of Jupyter notebooks and Jupyter Declarative Widgets to provide functionality. This approach would require some technical knowledge to install onto a personal device and would be unlikely to be used by the public. Finally, papers detailing studies involving prescription or open data use a variety of methods to achieve results. Whilst most use some form of electronic records as a data source, only one uses open source prescription data from an administrative source. Each paper provides a unique analysis specific to the area in which they are interested in, they propose that the methods used could be beneficial to other researchers with their results applicable to their local healthcare communities.

TABLE 2.7: Summary of study subjects

| **Subject** | **Number of studies** |
|---|---|
| General Prescribing Patterns | 3 |
| Opioid Prescribing | 1 |
| Antidepressant Prescribing | 1 |
| Discovering Prevalence of Chronic Conditions | 1 |
| Detection of Atrial Fibrilation (AF) | 1 |
| Identification of Heroin related overdose incidents | 1 |
| Diabetes Control and Prevention | 1 |
| Detection of Adverse Drug Events | 1 |
| Estimation of Missing Patient Data | 1 |
| Outlier Detection | 2 |

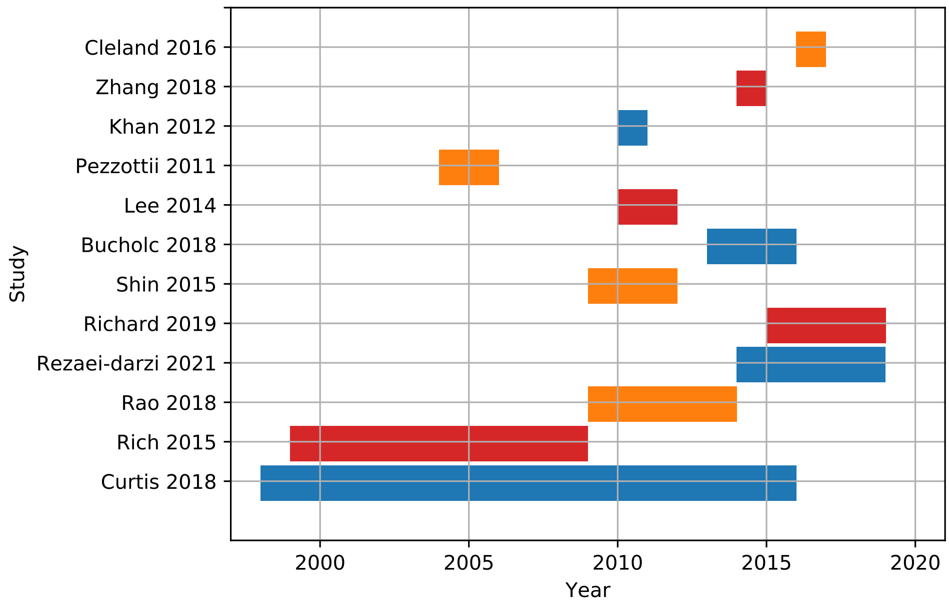


FIGURE 2.4: Scope of data used in studies

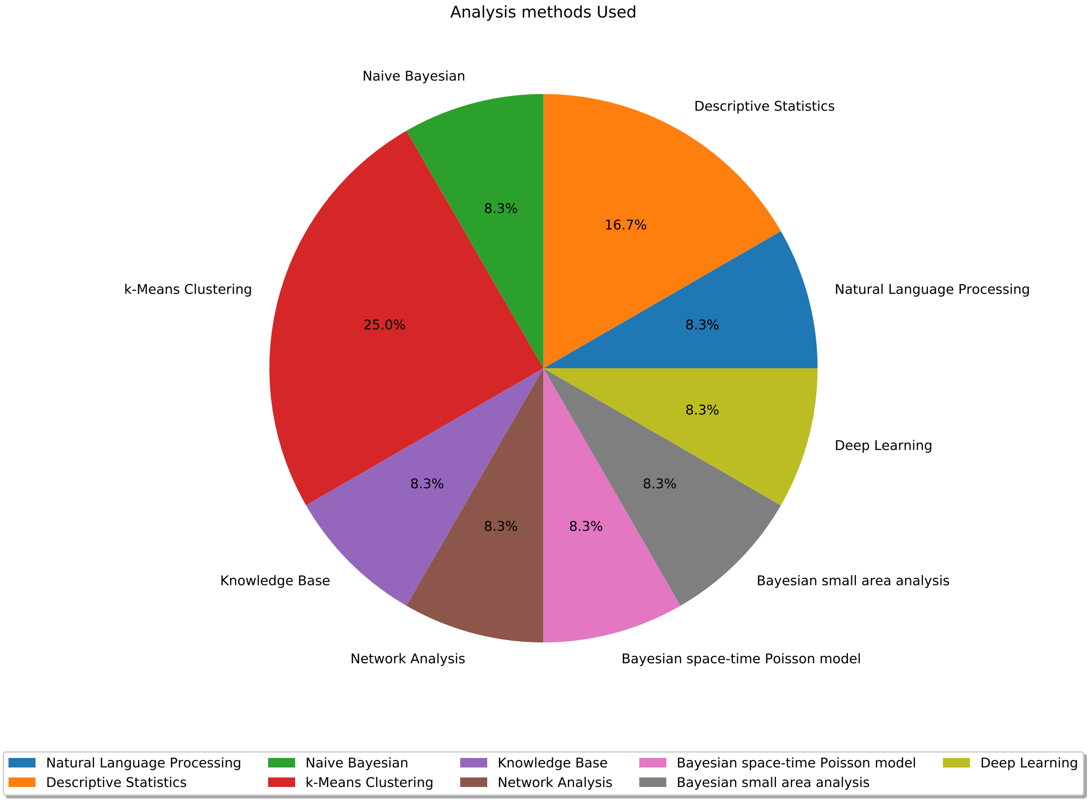


FIGURE 2.5: Analysis methods used in studies

## 2.4 Limitations

The literature review was conducted using the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) (Moher et al., 2009) as a guide. This review was carried out independently with no peer review being performed due to the nature of the PhD project. This means that the choice of papers to be reviewed and the analysis of their contents could be subject to an unconscious bias.

## 2.5 Conclusion

In this chapter, it has been established that there are a growing number of peer reviewed papers being published since 2017 providing analysis on open health data. Analysis of prescription data has already been used to analyse trends in the prescribing of antidepressants, opioids and the prevalence of chronic conditions with this being broken down in some cases to the standard Rural / Urban split based on population numbers attributed to the location in which the GP surgery or hospital was located. Classification in all studies was limited to the classification of different types of drugs, not to the type of practices with the most popular method being k-Means clustering. In all cases, trends were examined using descriptive statistics. No literature was found investigating links between prescribing behaviours of GP practices, their locations and the categorisation of surgeries based on both these criteria. Chapter 3 outlines the selection of data used in this study and linkages created to provide an integrated data set for analysis. Chapter 4 will provide an initial analysis of Northern Ireland GP surgeries and prescribing trends before seeking to identify different types of GP practice using the geographical profile of the practices and that of the pharmacies that dispense prescriptions issued by them in Chapter 5.

# Chapter 3

# Wrangling of open prescription data

> "The most important step a person can take is the next one ."
>
> Brandon Sanderson,
> The Stormlight Archive

This chapter provides a background to the choice of data sources, the wrangling involved in order to make them fit for purpose and the linkages made in order to create a novel data store providing data on GP practices in Northern Ireland. A technical review is included for information on the techniques available for the analysis of these data. All the data used is publicly available for download and a list of these sources is included at Appendix C. One of the challenges of this study was identifying, cleaning and integrating multiple data sets from diverse open data sources. A data pipeline that could be tracked and audited was developed to integrate these data sets into a local data store allowing them to be compared and analysed. The resulting process is outlined in Figure 3.1.

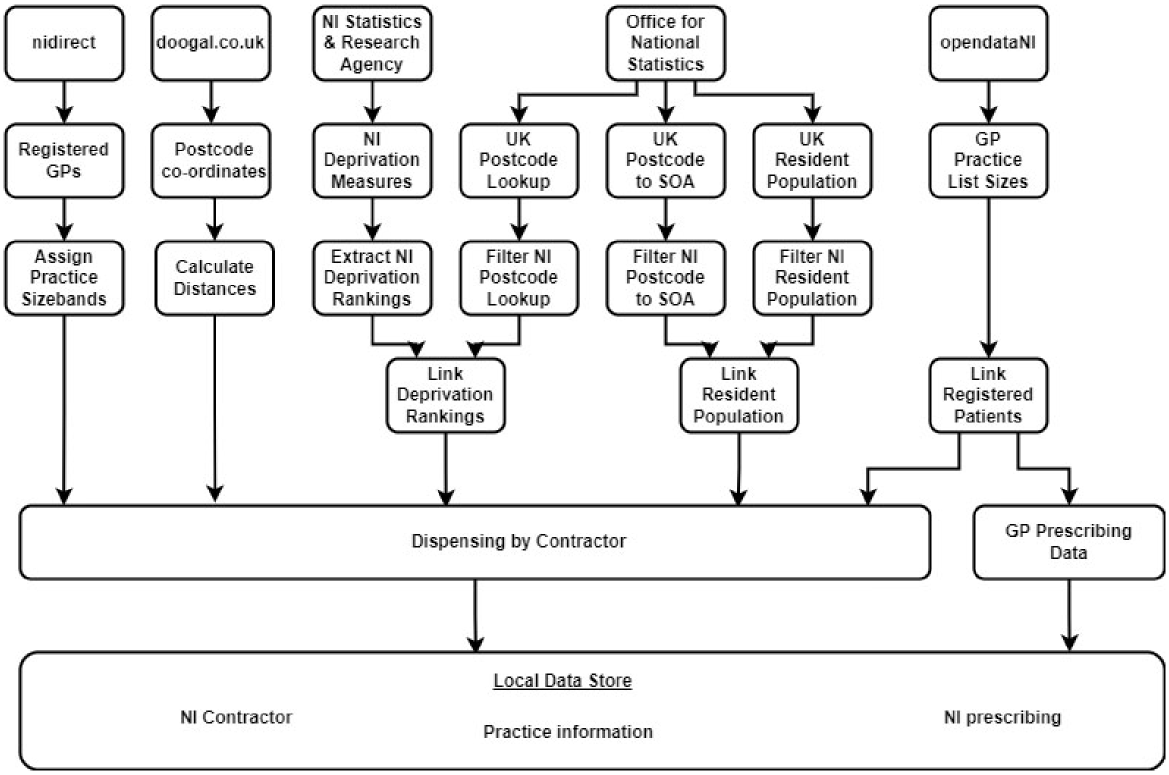


FIGURE 3.1: Data flow diagram illustrating the main data sources and data processing activities.

## 3.1 Data wrangling

Data wrangling is the term used for the acquisition, cleaning, structuring and enriching of raw and often data messy data sets in order to make them useful for analysis. As the number of data sources increase and the data becomes more diverse and unstructured it is essential for data wrangling to take place in advance of data mining (Sharma, 2021; Stobierski, 2021; Boopathy, 2021). The data wrangling process was achieved using the following steps:

- Data discovery - This is the process of searching for suitable data sources and examining the variables within each to ascertain their usefulness and compatibility with other data sets.

- Data structuring - Once appropriate raw data has been obtained, it is necessary to standardise the structure of each data set to make them compatible with each other. This can often involve setting variable types or transforming variables.

- Data cleaning - Raw data must be cleaned to eliminate errors or missing data. In some cases, cleaning could also include the removal of outliers from the data.

- Data enriching - This is the process whereby additional data sets are linked to provide additional variables thus enriching the original data set.

- Data validation - This process involves the assessment of the quality of the enriched data set, checking that linked variables are correct and relevant to the original data.

## 3.2 Software

The Python programming language and Jupyter Notebooks from Anaconda[1] were used throughout, utilising Pandas DataFrames for data wrangling and Scikit-Learn (Version 0.21.2)[2] for clustering. Matplotlib[3] and IpyLeaflet[4] were used for data visualisation.

## 3.3 Data sources

### 3.3.1 Dispensing by contractor

Dispensing by contractor data, provided by the Health and Social Care Business Services Organisation (BSO) was downloaded from the OpenDataNI website[5]. Each file relates to prescriptions issued and dispensed within a specific month. Table 3.1 gives a summary of the variables in each file.

These data originates from the BSO's prescribing and dispensing information systems and covers prescriptions that are prescribed in Northern Ireland by GPs or nurses (within a GP practice), pharmacists, optometrists, chiropodists and (potentially) radiographers that are subsequently dispensed by a community pharmacist, dispensing doctor or appliance supplier and are finally submitted to the BSO for payment. Prescriptions issued but not presented to a pharmacy for dispensing and private prescriptions issued by GPs are not included in the data set.

All registered practices in Northern Ireland are included in these data with claim data being recorded where (in the relevant month) a prescription has been dispensed and submitted to the BSO by the dispensing contractor (pharmacist) as a claim for payment.

Data have been published from March 2018 and is published each month covering the dispensing information from the 2 months previous (e.g. June 2021 data was published in August 2021).

**GP practices** - Each practice is identified by a unique practice number with the Practice details, including names and addresses, included for ease of reference. These figures only cover practices in Northern Ireland with no comparable data sets being published for England, Scotland or Wales.

[1] Jupyter Notebooks by Anaconda, Available at: https://www.anaconda.com/products/individual

[2] Scikit-Learn, Available at: https://scikit-learn.org/stable/install.html

[3] Matplotlib, Available at: https://anaconda.org/conda-forge/matplotlib

[4] Ipyleaflet, Available at: https://pypi.org/project/ipyleaflet/

[5] HSC Business Services Organisation (2020) Dispensing by Contractor, Available at: https://www.opendatani.gov.uk/dataset/dispensing-by-contractor

**Number of items** - A prescription item is a single supply of a medicine, and does not take into account the quantity of medicine prescribed. Patients with a long term condition usually get regular prescriptions. While many prescriptions are for one month (28 or 30 days supply), items will be for varying length of treatment and quantity.

TABLE 3.1: List of variables in dispensing by contractor data set

| Variable | Description | Data Type |
|---|---|---|
| Practice | A unique code attributed to each practice | int64 |
| Practice Name | Registered name of practice | object |
| Address1 | Registered address of practice line 1 | object |
| Address2 | Registered address of practice line 2 | object |
| Address3 | Registered address of practice line 3 | object |
| Postcode | Postcode of practice registered address | object |
| Chemist | A unique code attributed to each pharmacy | int64 |
| Contractor Name | Registered name of pharmacy | object |
| Contractor Address Line 1 | Registered address of pharmacy line 1 | object |
| Contractor Address Line 2 | Registered address of pharmacy line 2 | object |
| Contractor Address Line 3 | Registered address of pharmacy line 3 | object |
| Contractor Address Line 4 | Registered address of pharmacy line 4 | object |
| Contractor Postcode | Postcode of pharmacy registered address | object |
| Year | Year (YYYY) | int64 |
| Month | Month (1-12) | int64 |
| Number of Items | Number of Items prescribed | int64 |

### 3.3.2 GP prescribing data

GP prescribing data are also provided by the Health and Social Care Business Services Organisation (BSO) and downloaded from the OpenDataNI website[6]. Each file relates to prescriptions issued and dispensed within a specific month. Table 3.2 gives a summary of the variables in each file.

These data also originate from the BSO's prescribing and dispensing information systems and covers prescriptions that are prescribed in Northern Ireland by GPs or nurses (within a GP practice), pharmacists, optometrists, chiropodists and (potentially) radiographers that are subsequently dispensed by a community pharmacist, dispensing doctor or appliance supplier and are finally submitted to the BSO for payment. Prescriptions issued but not presented to a pharmacy for dispensing and private prescriptions issued by GPs are not included in the data set.

[6] HSC Business Services Organisation (2020) GP Prescribing Data, Available at: https://www.opendatani.gov.uk/dataset/gp-prescribing-data

Within each file, data is available for prescriptions issued / dispensed for each GP practice in Northern Ireland, and for each medication. Each medication is identified by it's British National Formulary (BNF) code along with the following information:

- The number of prescribed items that are dispensed
- The quantity of tablets, capsules, liquid etc. dispensed
- The gross cost (£), and
- The actual cost (£)

NI data has been published from April 2013 and is published each month covering the dispensing information from the 2 months previous (e.g. June 2021 data was published in August 2021).

Northern Ireland GP prescribing data differs from the GP prescribing data published for England and Wales. English prescribing data has been published by the NHS Business Services Authority[7] since December 2011, NHS Wales[8] providing similar data for Wales since April 2013 and Public Health Scotland[9] since October 2015. All three bodies publish GP prescribing data by medication on a monthly basis with data covering the same variables as the data published in the NI data set with some minor differences.

**Differences in data sets** - There are a number of differences between the NI data and that of England and Wales. For example, in the calculation of actual cost, the English figures include container fees whereas these are not part of the NI calculation. As container fees are generally less than £300 per month, this does not present a major problem when comparing costs. The names of medications may not be directly comparable between the English and NI data sets as the English names are based on their Business Services Authority Drug and Appliance database, and the NI names are based on the UK wide standardised Dictionary of Medicines and Devices.

**Number of items** - A prescription item is a single supply of a medicine, and does not take into account the quantity of medicine prescribed. Patients with a long term condition usually get regular prescriptions. While many prescriptions are for one month (28 or 30 days supply), items will be for varying length of treatment and quantity.

**Gross cost v actual cost** - The gross cost is the basic price of a drug, i.e. the price listed in the Drug Tariff or price lists, the actual cost is the estimated cost to the NHS. Actual cost is calculated by subtracting the discount per item from the gross cost.

[7] NHS Business Services Authority English Prescribing Dataset (EPD) https://digital.nhs.uk/data-and-information/publications/statistical/practice-level-prescribing-data

[8] NHS Wales General Practice Prescribing Data https://nwssp.nhs.wales/ourservices/primary-care-services/general-information/data-and-publications/general-practice-prescribing-data-extract/

[9] Public Health Scotland Prescriptions in the Community https://www.opendata.nhs.scot/dataset/prescriptions-in-the-community

**British National Formulary (BNF)** - The BNF is a joint publication by the British Medical Association and the Royal Pharmaceutical Society, providing information on the selection, prescribing, dispensing and administration of medicines available in the UK. In the BNF, medicines are classified by therapeutic group, for example: cardiovascular, respiratory, gastro-intestinal, etc. with sub divisions drilling down to specific medications. These divisions are known as Section, Paragraph and Sub-Paragraph with each medication having its own unique BNF code. BNF Chapters 1-15 cover the main medications with other items being classified in 'pseudo BNF chapters' 18-23. Table 3.3 provides a list of all BNF chapters. BNF structure at Chapter, Section and Paragraph level can be found in Appendix D.

**Quantity** - The quantity of a drug dispensed is measured in units depending on the type of product, which is given in the drug name. Quantities cannot be added together because of the different strengths and types. For example, where the type is tablet, capsule, ampoule, vial etc. the quantity will be the number of tablets, capsules, ampoules, vials etc. Where it is a liquid, the quantity will be the number of millilitres and where it is a solid form (e.g. Cream, gel, ointment), the quantity will be the number of grams.

TABLE 3.2: List of variables in GP prescribing data set

| **Variable** | **Description** | **Data Type** |
|---|---|---|
| Practice | A unique code attributed to each practice | int64 |
| Year | Year (YYYY) | int64 |
| Month | Month (1-12) | int64 |
| VTM_NM | Substance / Product name | object |
| VMP_NM | Generic Name | object |
| AMP_NM | Branded / Generic Name | object |
| Presentation | Type of medication (e.g. capsule, cream etc) | object |
| Strength | Medication strength (e.g. 5mg) | object |
| Total Items | Number of Items prescribed | int64 |
| Total Quantity | Quantity in each item (e.g. 30 capsules, 5ml) | int64 |
| Gross Cost (£) | The basic price of a drug (£) | float64 |
| Actual Cost (£) | The estimated cost to the NHS (£) | float64 |
| BNF Code | This is a unique code attributed to each medication by the British National Formulary (BNF), the UK pharmaceutical reference guide. | object |
| BNF Chapter | Top level split using first 2 characters of the BNF code | int64 |
| BNF Section | Second level split subdividing BNF chapter using second set of 2 characters of the BNF code | int64 |
| BNF Paragraph | Third level split subdividing BNF section using third set of 2 characters of the BNF code | int64 |
| BNF Sub-Paragraph | Fourth level split subdividing BNF paragraph using fourth set of 2 characters of the BNF code | int64 |

TABLE 3.3: British National Formulary (BNF) chapters

| **Chapter** | **Description** |
|---|---|
| 1 | Gastro-Intestinal System |
| 2 | Cardiovascular System |
| 3 | Respiratory System |
| 4 | Central Nervous System |
| 5 | Infections |
| 6 | Endocrine System |
| 7 | Obstetrics |
| 8 | Malignant Disease & Immunosuppression |
| 9 | Nutrition And Blood |
| 10 | Musculoskeletal & Joint Diseases |
| 11 | Eye |
| 13 | Skin |
| 14 | Immunological Products & Vaccines |
| 15 | Anesthesia |
| 18 | Preparations used in Diagnosis |
| 19 | Other Drugs And Preparations |
| 20 | Dressings |
| 21 | Appliances |
| 22 | Incontinence Appliances |
| 23 | Stoma Appliances |

### 3.3.3 GP practice list sizes

GP practice list sizes are published on a quarterly basis for all registered NI GP practices. These files are also available from the OpenDataNI website[10] and provided the Health and Social Care Business Services Organisation (BSO). Table 3.4 gives a summary of the variables in each file.

In addition to the number of registered patients belonging to each GP surgery at the beginning of each quarter, this file also provides data in which Local Commissioning Group (LCG) each practice is located. Local Commissioning Groups are better known in Northern Ireland as Health Trusts of which there are five - Belfast, Northern, South Eastern, Southern and Western.

[10] HSC Business Services Organisation (2020) GP Practice List Sizes, Available at: https://www.opendatani.gov.uk/dataset/gp-practice-list-sizes

TABLE 3.4: List of variables in GP practice size lists data set

| Variable | Description | Data Type |
|---|---|---|
| PracNo | A unique code attributed to each practice | int64 |
| PracticeName | Registered name of practice | object |
| Address1 | Registered address of practice line 1 | object |
| Address2 | Registered address of practice line 2 | object |
| Address3 | Registered address of practice line 3 | object |
| Postcode | Postcode of practice registered address | object |
| LCG | Local Commissioning Group (Health Trust) | object |
| Registered Patients | Number of Registered patients in the practice. | int64 |

### 3.3.4 Postcode to Output Area to Lower Layer Super Output Area to Middle Layer Super Output Area to Local Authority District (February 2019)

The Office for National Statistics (2019) Postcode to Output Area to Lower Layer Super Output Area to Middle Layer Super Output Area to Local Authority District (February 2019)[11] provides a best-fit lookup between postcodes, frozen 2011 Census Output Areas (OA), Lower Layer Super Output Areas (LSOA), Middle Layer Super Output Areas (MSOA) and current local authority districts (LAD) as of February 2019 in the UK. Table 3.5 gives a summary of the variables in this file.

[11] The Office for National Statistics (2019) Postcode to Output Area to Lower Layer Super Output Area to Middle Layer Super Output Area to Local Authority District (February 2019) Available at: http://geoportal.statistics.gov.uk/datasets/c479d770cba14845a0e43db4e3eb5afa

TABLE 3.5: List of variables in Postcode to Output Area to Lower Layer Super Output Area to Middle Layer Super Output Area to Local Authority District (February 2019) data set

| Variable | Description | Data Type |
|---|---|---|
| pcd7 | Unit postcode – 7 character version | object |
| pcd8 | Unit postcode – 8 character version | object |
| pcds | Unit postcode - variable length | object |
| dointr | Date of introduction | object |
| doterm | Date of termination | object |
| usertype | Postcode user type (Small User or Large User) | object |
| oa11cd | 2011 Census Output Area (OA)/ Small Area (SA) | object |
| lsoa11cd | 2011 Census Lower Layer Super Output Area (LSOA)/ Data Zone (DZ)/ SOA | object |
| msoa11cd | 2011 Census Middle Layer Super Output Area (MSOA)/ Intermediate Zone (IZ) | object |
| ladcd | Local Authority District | object |
| lsoa11nm | 2011 Census Lower Layer Super Output Area (LSOA)/ Data Zone (DZ)/ SOA Name | object |
| msoa11nm | 2011 Census Middle Layer Super Output Area (MSOA)/ Intermediate Zone (IZ) Name | object |
| ladnm | Local Authority District Name | object |
| ladnmw | Local Authority District Name (Welsh) | object |

### 3.3.5 National Statistics Postcode Lookup (February 2020)

The Office for National Statistics (2020) National Statistics Postcode Lookup (February 2020)[12], provides a 'best-fit' between postcodes to a range of current statutory geographies using the 2011 Census Output Areas including Northern Ireland Wards. Table 3.6 gives a summary of the variables in this file.

TABLE 3.6: List of variables in National Statistics Postcode Lookup (February 2020) data set

| Variable | Description | Data Type |
|---|---|---|
| pcd | Unit postcode – 7 character version | object |
| pcd2 | Unit postcode – 8 character version | object |
| pcds | Unit postcode - variable length | object |
| dointr | Date of introduction | int64 |

Continued on next page

[12] The Office for National Statistics (2020) National Statistics Postcode Lookup (February 2020) Available at: https://geoportal.statistics.gov.uk/datasets/national-statistics-postcode-lookup-february-2020

**Table 3.6 – continued from previous page**

| Variable | Description | Data Type |
|---|---|---|
| doterm | Date of termination | float64 |
| usertype | Postcode user type | int64 |
| oseast1m | National grid reference - Easting | object |
| osnrth1m | National grid reference - Northing | object |
| osgrdind | Grid reference positional quality indicator 1 | object |
| oa11 | 2011 Census Output Area (OA)/ Small Area (SA) | object |
| cty | County | object |
| ced | County Electoral Division | object |
| laua | Local authority district (LAD)/unitary authority (UA)/ metropolitan district (MD)/ London borough (LB)/ council area (CA)/district council area (DCA) | object |
| ward | (Electoral) ward/division | object |
| hlthau | Health Authority | object |
| nhser | NHS England (Region) (NHS ER) | object |
| ctry | Country | object |
| rgn | Region | object |
| pcon | Westminster parliamentary constituency | object |
| eer | European Electoral Region | object |
| teclec | Local Learning and Skills Council (LLSC)/ Dept. of Children, Education, Lifelong Learning and Skills (DCELLS)/ Enterprise Region (ER) | object |
| ttwa | Travel to Work Area (TTWA) | object |
| pct | Primary Care Trust (PCT)/ Care Trust/ Care Trust Plus (CT)/ Local Health Board (LHB)/ Community Health Partnership (CHP)/ Local Commissioning Group (LCG)/ Primary Healthcare Directorate (PHD) | object |
| nuts | Nomenclature of territorial units for statistics | object |
| park | National park | object |
| lsoa11 | 2011 Census Lower Layer Super Output Area (LSOA)/ Data Zone (DZ)/ SOA | object |
| msoa11 | 2011 Census Middle Layer Super Output Area (MSOA)/ Intermediate Zone (IZ) | object |
| wz11 | 2011 Census Workplace Zone (WZ) | object |

Continued on next page

**Table 3.6 – continued from previous page**

| Variable | Description | Data Type |
|---|---|---|
| ccg | Clinical Commissioning Group (CCG)/ Local Health Board (LHB)/ Community Health Partnership (CHP)/ Local Commissioning Group (LCG)/ Primary Healthcare Directorate (PHD) | object |
| bua11 | Built-up Area | object |
| buasd11 | Built-up Area Sub-division | object |
| ru11ind | 2011 Census rural-urban classification | object |
| oac11 | 2011 Census Output Area classification (OAC) | object |
| lat | Decimal degrees latitude | float64 |
| long | Decimal degrees latitude | float64 |
| lep1 | Local Enterprise Partnership (LEP) - first instance | object |
| lep2 | Local Enterprise Partnership (LEP) - first instance | object |
| pfa | Police Force Area (PFA) | object |
| imd | Index of Multiple Deprivation (IMD) | object |
| calncv | Cancer Alliance | object |
| stp | Sustainability and Transformation Partnership | object |

### 3.3.6 UK usual resident population

The Office for National Statistics (2014) usual resident population[13], provides data on the usual resident population in the UK by Super Output Area. The usual resident population is a count of the number of residents living in households and communal establishments and includes the number of students and schoolchildren who would reside in each area if they were not living away from their family home during term-time. Additional information provided includes the area and population density for each area. Table 3.7 gives a summary of the variables in this file.

TABLE 3.7: List of variables in UK usual resident population data set

| Variable | Description | Data Type |
|---|---|---|
| SOA | Super Output Area Name | object |
| SOA Code | Super Output Area Code | object |
| All usual residents | Total Number of Usual Residents | int64 |
| Area (hectares) | SOA area in hectares | int64 |
| Population Density (number of usual residents per hectare) | Total number of Usual Residents divided by Area | float64 |

[13] Office for National Statistics (2014) Usual resident population Available at: https://www.nomisweb.co.uk/census/2011/ks101uk

### 3.3.7 Northern Ireland Multiple Deprivation Measure 2017

A deprivation ranking is provided for each ward in Northern Ireland by the Northern Ireland Statistics Research Agency (NISRA). The Northern Ireland Multiple Deprivation Measure 2017 (NIMDM2017)[14], is the measure of deprivation using multiple distinct domains of deprivation which can be recognised and measured separately. These are experienced by individuals living in an area. People may be counted as deprived in one or more of the domains depending on the number of types of deprivation that they experience. The overall measure is expressed as a weighted aggregation of the measurements of each domain.

The seven 'domains' which make up the Multiple Deprivation Measure are:

- Income deprivation
- Employment deprivation
- Health Deprivation & Disability
- Education, Skills & Training
- Access to Services
- Living Environment, and
- Crime & Disorder

The overall multiple deprivation measure is a weighted average of the seven domains with weights of 25%, 25%, 15%, 15%, 10%, 5% and 5% applied respectively.

Deprivation measures are a ranking from most deprived to least deprived areas with the most deprived area being ranked '1'. NISRA does not stipulate which areas are classed as deprived or non-deprived leaving this determination up to the researcher. Table 3.8 gives a summary of the variables in this file.

TABLE 3.8: List of variables in Northern Ireland Multiple Deprivation Measure 2017 data set

| **Variable** | **Description** | **Data Type** |
|---|---|---|
| LGD2014NAME | Local Government District Name | object |
| 2015 Default Urban/Rural | Ward Type | object |
| Ward2014 code | Ward Code | object |
| Ward2014 name | Ward Name | object |
| Multiple Deprivation Measure Rank | Deprivation Rank (1 being the highest) | int64 |

[14] Northern Ireland Statistics Research Agency (NISRA) Northern Ireland Multiple Deprivation Measure 2017 (NIMDM2017) Available at: https://www.nisra.gov.uk/statistics/deprivation/northern-ireland-multiple-deprivation-measure-2017-nimdm2017

### 3.3.8 Northern Ireland GP practice size bands

As no source for this measure was published, the number of registered doctors in each practice was calculated using the nidirect Find a GP practice[15] to manually look up each practice. Practices were then categorised as: Single-handed practices (1 Doctor), Small practices (2 doctors), Medium practices (3-4 doctors) and Large practices (5+ doctors). No information was available on Locums so this was ignored. As no data was available for 5 practices (1.5%), these were assigned as a Medium practice as the mean for all known practices was calculated as approximately 4 registered doctors.

### 3.3.9 Mid year population estimates

In order to compare prescribing in the four regions of the UK, Numbers of Items prescribed per head of population was used. These figures were obtained using mid year population estimates[16] published by the Office for National Statistics.

## 3.4 Preparation of additional data sources

### 3.4.1 GP practice list sizes

GP practice list sizes are published on a quarterly basis providing details of each GP practice by assigned identification number, name, address, postcode, Local Commissioning Group (Health Trust in NI) and the number of patients registered at the start of the quarter. These data sets have been published since July 2015 with each file being less than 40 kilobytes and the number of rows equaling the number of active GP practices. In order to standardise variable names, the practice identification field (PracNo) was renamed Practice in all files.

### 3.4.2 Postcode to Output Area to Lower Layer Super Output Area to Middle Layer Super Output Area to Local Authority District (February2019)

The Postcode to Output Area to Lower Layer Super Output Area to Middle Layer Super Output Area to Local Authority District provides a conversion from postcode to super output area for every postcode in the United Kingdom. This file is 392 megabytes in size with over 2.6 million rows. In order to minimise the resources needed to link postcodes from other files this original file was filtered creating a new conversion file with Northern Ireland postcodes, i.e. those beginning with BT. This new conversion file was approximately 8 megabytes in size with 61,403 records. Algorithm 1 details this process.

[15] nidirect Find a GP practice Available at: https://www.nidirect.gov.uk/services/gp-practices

[16] Office for National Statistics Mid Year Population Estimates Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates

**Algorithm 1** Algorithm for creating NI file of Postcodes to Super Output areas

```
local = blank dataframe
batch = 100
while rows exist do
    Read batch from UK data file
    NI = batch where postcode starts with 'BT'
    local = local append NI
end while
Write local dataframe to local hard disk
```

### 3.4.3 National Statistics Postcode Lookup (February 2020)

The National Statistics Postcode Lookup provides a conversion from postcode to ward for every postcode in the United Kingdom. This file is similar to the Postcode to Output Area to Lower Layer Super Output Area to Middle Layer Super Output Area to Local Authority District but contains different variables. This file was also filtered to create a postcode to ward conversion file for Northern Ireland. This file does also provide lower super output codes for Northern Ireland but does not contain the matching names for them hence the previous file was more suitable. Algorithm 2 details this process.

**Algorithm 2** Algorithm for creating NI file of postcodes to ward conversion file

```
local = blank dataframe
batch = 100
while rows exist do
    Read batch from UK data file
    NI = batch where postcode starts with 'BT'
    local = local append NI
end while
Write local dataframe to local hard disk
```

### 3.4.4 UK usual resident population

Taken from the 2011 Census, this publication provides a separate download file specifically covering Northern Ireland in csv format. The file is 32 kilobytes in size providing data for all 890 super output areas in Northern Ireland. No wrangling was required for this file.

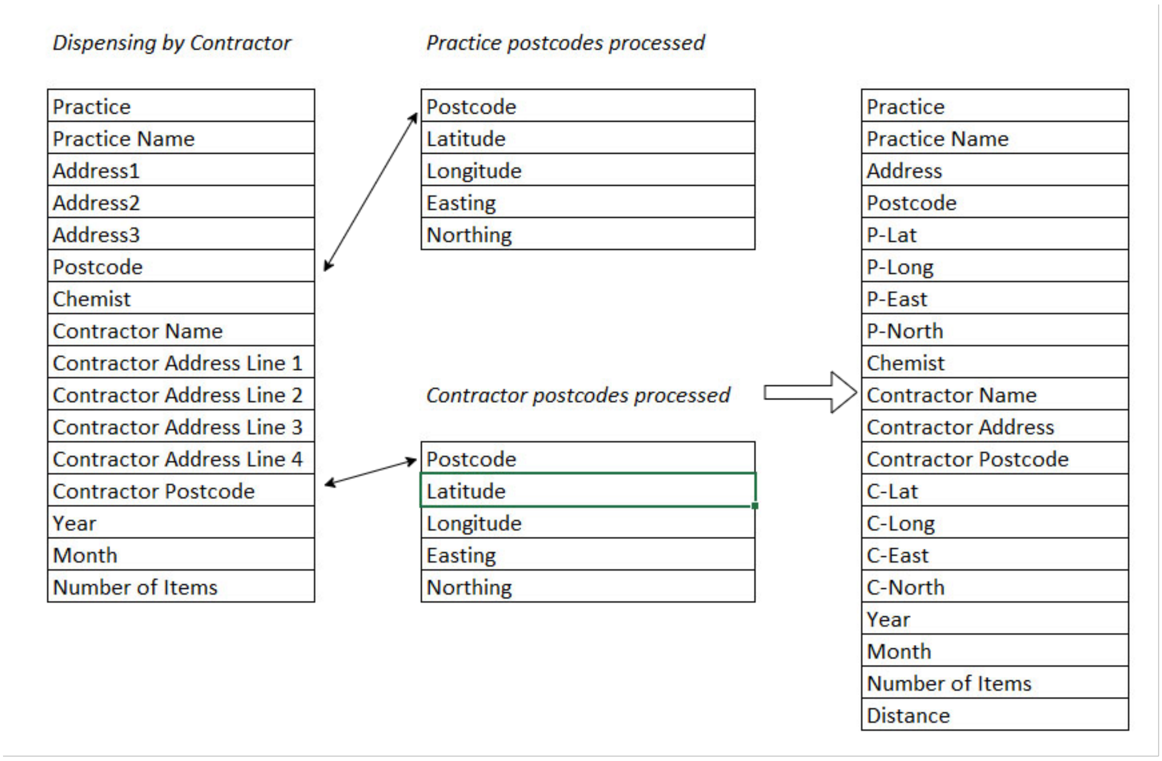
### 3.4.5 Northern Ireland Multiple Deprivation Measure 2017

The Northern Ireland Multiple Deprivation Measure 2017, published at ward level is available as an Excel file (NIMDM17_Ward2014.xls). Within this file there are two worksheets, the second worksheet (NIMDM2017) providing data on each ward within Northern Ireland along with the Multiple Deprivation Rank for each ward. Wards are ranked from 1 to 462 with 1 representing the most deprived ward and 462 the least deprived. The relevant columns of this worksheet were copied into a new file and saved as a comma separated values (csv) file.

## 3.5 Creation of Local Data Store

### 3.5.1 NI Contractor data

The Dispensing by contractor data sets from April 2018 to June 2021 were used in this study. Each file was downloaded as a comma separated values (csv) file with file ranging from 3 megabytes to approximately 7 megabytes (18,400 - 19,700 rows of data). Each row details a request for payment by a contractor (pharmacy) for medication dispense in that month for a specific pharmacy although only number of items is specified. Figure 3.2 shows the process used to wrangle these data each month with the relevant code described in Algorithm 3.



*Dispensing by Contractor*

| |
|---|
| Practice |
| Practice Name |
| Address1 |
| Address2 |
| Address3 |
| Postcode |
| Chemist |
| Contractor Name |
| Contractor Address Line 1 |
| Contractor Address Line 2 |
| Contractor Address Line 3 |
| Contractor Address Line 4 |
| Contractor Postcode |
| Year |
| Month |
| Number of Items |

*Practice postcodes processed*

| |
|---|
| Postcode |
| Latitude |
| Longitude |
| Easting |
| Northing |

*Contractor postcodes processed*

| |
|---|
| Postcode |
| Latitude |
| Longitude |
| Easting |
| Northing |

| |
|---|
| Practice |
| Practice Name |
| Address |
| Postcode |
| P-Lat |
| P-Long |
| P-East |
| P-North |
| Chemist |
| Contractor Name |
| Contractor Address |
| Contractor Postcode |
| C-Lat |
| C-Long |
| C-East |
| C-North |
| Year |
| Month |
| Number of Items |
| Distance |

FIGURE 3.2: NI contractor data workflow

**Practice/Contractor locations** - In order to calculate the distance between each prescribing practice and the corresponding dispensing contractor, it was necessary to obtain northings and eastings for both the practice and contractor. Initially, a list of unique postcodes for practices and contractors was obtained and batches of 50 were submitted to the Batch Geocoding service[17]. In addition to northings and eastings, the latitude and longitude was also captured. A postcode conversion file was created with this information. Any postcodes that did not convert using this method

[17] Batch geocoding service, Available at: https://www.doogal.co.uk/BatchGeocoding.php

**Algorithm 3** Algorithm for data wrangling of dispensing by contractor data files

```
Read monthly data file
Read Practice Postcode errors file
Read Pharmacy Postcode errors file
Read Practice northings and eastings file
Read Pharmacy northings and eastings file
Match and replace incorrect practice postcodes identified in Practice Postcode errors file
Match and replace incorrect pharmacy postcodes identified in Pharmacy Postcode errors file
Match practices postcodes to Practice northings and eastings file to assign northings, eastings, atitude and longitude
Match pharmacy postcodes to Pharmacy northings and eastings file to assign northings, eastings, atitude and longitude
Calculate distance between each practice and pharmacy using northings and eastings
filter dataframe for missing distance data        ▷ Any missing data must be investigated and fixed
if missing distance data is null then
    Drop unwanted columns
    Write cleaned dataframe to local drive
end if
Create unique list of practice for current month
Read previous months cleaned data file
Create unique list of practice for previous month
Compare lists to identify new practices (gains) or lost practices (losses)
Print lists of gains and losses.
```

were checked manually using internet searches for the company website and google maps in order to obtain the correct postcode. One practice postcode and 15 contractor postcodes were found to be in error and a file was created listing these with the new postcodes and relevant northings, eastings, latitude and longitude being added to the post the postcode conversion file. A script was then run to replace the erroneous postcodes with the correct ones before matching the postcode conversion file to the original download file creating eight new variables - practice northing, practice easting, practice latitude, practice longitude, contractor northing, contractor easting, contractor latitude and contractor latitude. These conversion files were used for all subsequent monthly data files with checks to ensure that all postcodes were converted.

**Calculating distance** - Having obtained the northings and eastings for each of the prescribing practices and corresponding dispensing contractors it was possible to calculate the distance between each. Equation 3.1 presents the formula for calculating distance using northings and eastings, the quotient being the square root of the difference in northings squared plus the difference in eastings squared divided by one thousand.

$$Distance = \frac{\sqrt{(Northing_{Pharmacy} - Northing_{Practice})^2 + (Easting_{Pharmacy} - Easting_{Practice})^2}}{1000} \tag{3.1}$$

**Validation** - Whilst no numerical data was changed, checks were made prior to and after processing the data to ensure the quality of the enriched data. The mean 'Number of Items' for the month, along with the range (minimum and maximum) were checked before and after the process to ensure they had not changed. The number of rows of data for each variable was also checked to ensure that it contained no null value. A summary of the results of the validation on each data set is included in Appendix B.

**Practice gains and losses** - Establishing that there were 333 unique GP practices listed in the original data file (April 2018), it was important to formulate a strategy of how to deal with gains to this list (new practices opening) or to a lesser extent losses (practices closing or being amalgamated into another practice). Each data file was compared with the previous month to check whether any gains or loses had occurred. It was found that over the 39 months (April 2018 - June 2021), no new practices were added to the list with 12 practices ceasing to exist. It is unknown what happened these practices, they may have been merged into larger practices or simply ceased to operate. It was decided that no procedures had to be put in place as no gains were identified and any losses would be reflected in the reported data.

### NI prescribing

Having established a cohort of 333 GP practices which could be tracked, it was important to match these practices to the corresponding GP prescribing files. GP prescribing data files are available from April 2013 on the Opendatani website although the unique list of 333 GP practices established from the April 2018 Dispensing by Contractor file could only be tracked from July 2015 onward. Each file (July 2015 - June 2021) was downloaded as a comma separated values (csv) file with file ranging between 61 megabytes and 69 megabytes (450,000 - 477,000 rows of data). Figure 3.3 shows the process used to wrangle the monthly data with Algorithm 4 providing a listing of the code used.

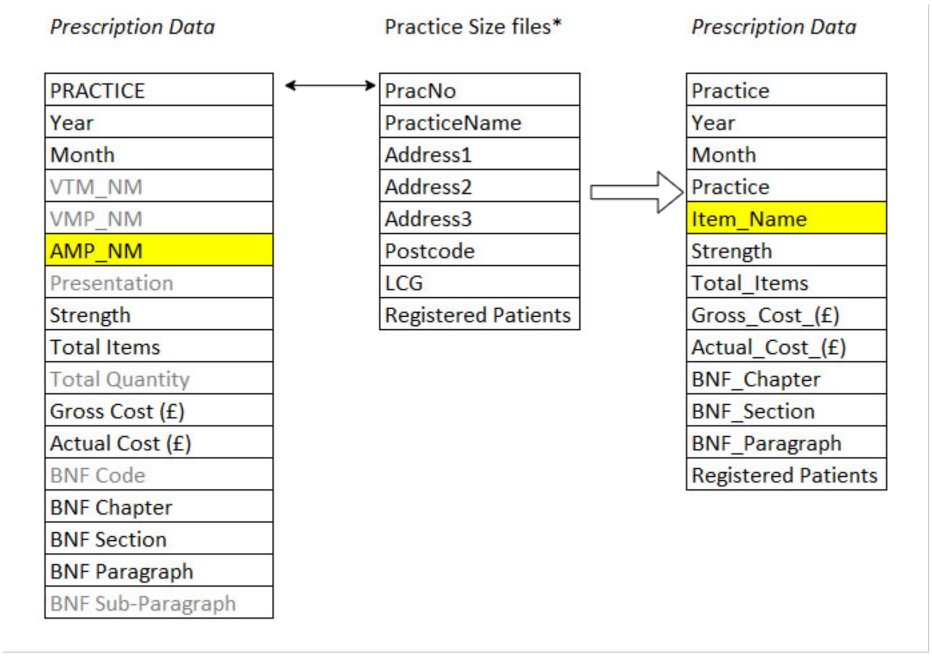


FIGURE 3.3: NI Prescribing data workflow (Coloured background indicates variables which were renamed)

**Variable names** - Comparing the downloaded files, it was quickly established that some variable names had changed slightly over the period and did not match those used in the Dispensing by contractor data files (e.g. Practice, PRACTICE). In order to standardise these, key variables were renamed. Table 3.9 provides a summary of these name changes.

TABLE 3.9: List of GP prescribing data variables renamed in order to standardise them.

| Original Name | Standardised Name |
|---|---|
| PRACTICE | Practice |
| Total Items | Total_Items |
| Total Quantity | Total_Quantity |
| AMP_NM | Item_Name |
| Gross Cost (£) | Gross_Cost_(£) |
| Actual Cost (£) | Actual_Cost_(£) |
| BNF Chapter | BNF_Chapter |
| BNF Section | BNF_Section |
| BNF Paragraph | BNF_Paragraph |

**Missing data** - The key fields in each data file were checked for missing data. With the exception of BNF_Chapter, BNF_Section and BNF_Paragraph, no missing fields were identified. Approximately 0.24% of the BNF variables had missing data and these were assigned to 99 - Unclassified.

**Enhancing with Practice size data** - As practices are not the same size in terms of the number of patients they have registered, it is important to include the number of registered patients to allow the normalisation of numbers of items prescribed thus facilitating a like-for-like comparison. For this reason, each prescribing data file was matched with the relevant monthly GP practice size file.

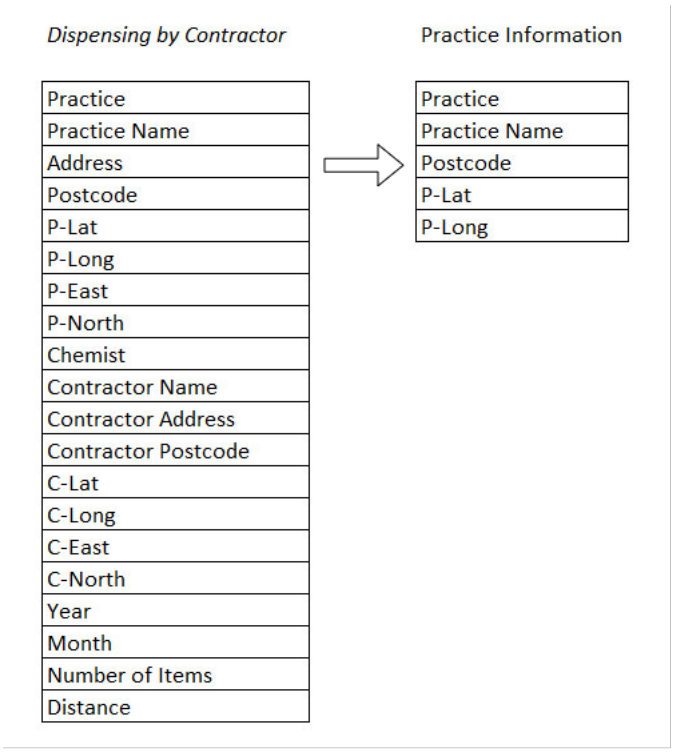
**Validation** - Whilst no numerical data was changed, checks were made prior to and after processing the data to ensure the quality of the enriched data. The mean for all numerical data variables for the month, along with the range (minimum and maximum) were checked before and after the process to ensure they had not changed. The number of rows of data for each variable was also checked to ensure that it contained no null value. A summary of the results of the validation on each data set is included in Appendix A.

### 3.5.2 Practice information file

The practice information file is a master file of all 333 practices tracked in this project. It contains data which describes each practice but which does not change each month. This file was initially created using the cleaned NI Contractor file from April 2018 located in the Local Data Store. From this file, a list of the 333 practices was created (Figure 3.4) containing the practice identification number, practice name, postcode and longitude and latitude coordinates.

**Algorithm 4** Algorithm for data wrangling GP prescribing data

```
Read monthly data file
Standardise column names by renaming them.
Replace missing BNF Chapter data with 99
Replace missing BNF Section data with 99
Replace missing BNF Paragraph data with 99
Replace BNF Chapter data = "-" with 99
Replace BNF Section data = "-" with 99
Replace BNF Paragraph data = "-" with 99
Convert BNF Chapter datatype to integer
Convert BNF Section datatype to integer
Convert BNF Paragraph datatype to integer
Drop unwanted columns
Read quarterly practice size file
Merge prescribing data with practice size file to add registered patients data to all rows.
Write cleaned dataframe to local disk.
```



*Dispensing by Contractor*

Practice
Practice Name
Address
Postcode
P-Lat
P-Long
P-East
P-North
Chemist
Contractor Name
Contractor Address
Contractor Postcode
C-Lat
C-Long
C-East
C-North
Year
Month
Number of Items
Distance

Practice Information

Practice
Practice Name
Postcode
P-Lat
P-Long

FIGURE 3.4: Workflow for creating practice information file

**Adding Population Density** - In order to add the population density for the super output area (SOA) in which each practice is located, the practice information file was linked to the National Statistics Postcode Lookup by postcode allowing the SOA name (lsoa11nm) and SOA Code (lsoa11cd) to be associated with each practice.This was then linked by SOA code to the UK usual resident population giving the Number of usual residents, the area in Hectares and the population density by hectare for the SOA in which each practice was located. Population density per square kilometre was calculated by multiplying population density by hectare by one hundred.

This metric was then added as a new column in the practice information file ensuring that each practice had a population density assigned to it.

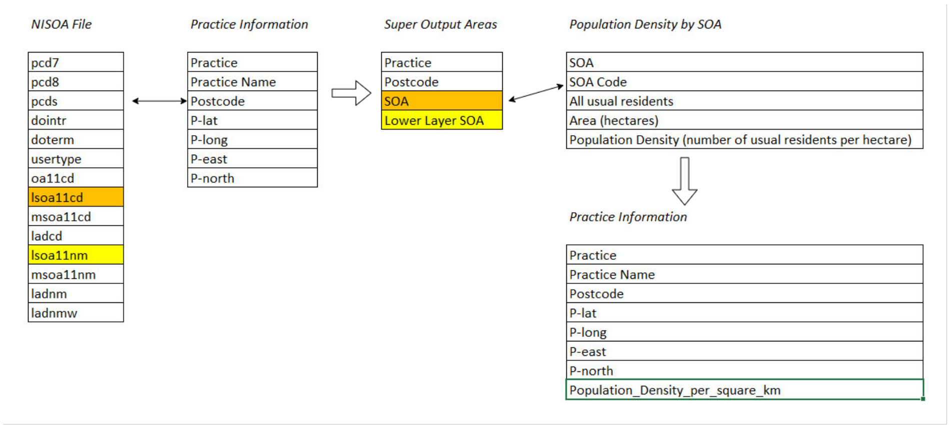


FIGURE 3.5: Workflow for adding population density to practice information file

**Adding Deprivation Level** - In order to associate the deprivation level for the area in which each practice was located, it was necessary to first associate each practice's postcode with the ward in which the practice was located using the National Statistics Postcode Lookup. Having linked the ward data, it was then possible to link each ward to the deprivation rank assigned to that ward in the Northern Ireland Multiple Deprivation Measure 2017. The python rank function was then used to rank the deprivation rankings and summary statistics were obtained using python's describe function. Using these statistics deprivation levels were assigned to each practice location with Quartile 1 (Low deprivation) being those ranked below the lower quartile figure, Quartile 2 (Low/Medium deprivation) as those ranked from the lower quartile figure but below the median, Quartile 3 (Medium/High deprivation) as those ranked from the median but lower than the upper quartile figure and Quartile 4 (High deprivation) as those ranked from the upper quartile and higher. These Deprivation levels were then mapped back onto the practice information file ensuring that each practice had a deprivation level assigned to it.

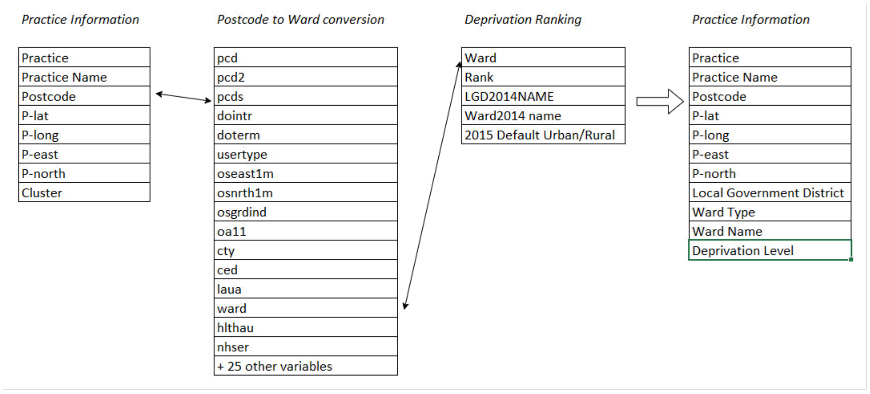
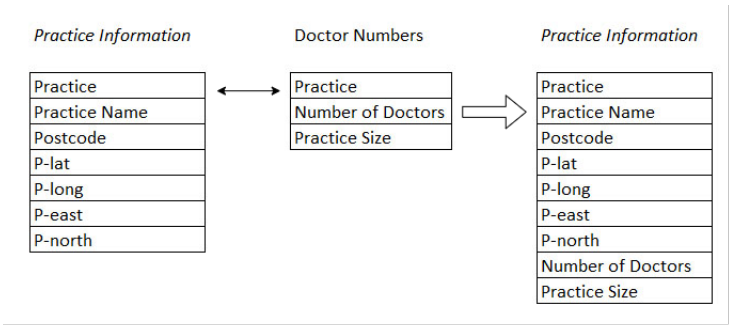


FIGURE 3.6: Workflow for adding deprivation level to practice information file

**Adding Practice Size (based on number of registered doctors)** - The number of registered doctors in each practice was manually calculated using the 'Find a GP practice' service on the nidirect website. This file was matched using the practice variable in order to add 'Number of Doctors' and 'Practice Size' as additional columns to the Practice information file.



*Practice Information*

Practice
Practice Name
Postcode
P-lat
P-long
P-east
P-north

Doctor Numbers

Practice
Number of Doctors
Practice Size

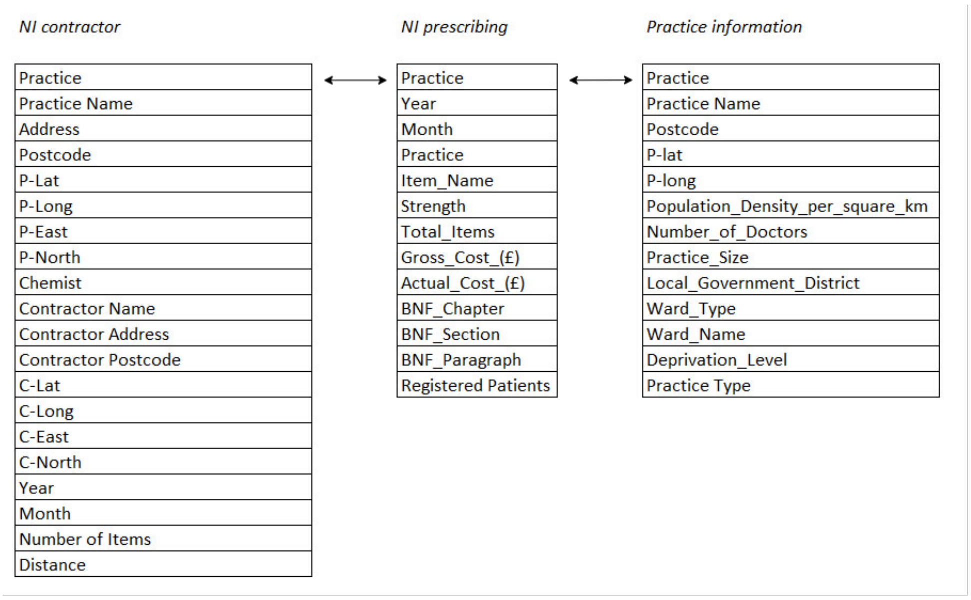
*Practice Information*

Practice
Practice Name
Postcode
P-lat
P-long
P-east
P-north
Number of Doctors
Practice Size

FIGURE 3.7: Workflow for adding practice size to practice information file (Coloured background indicates variable which were renamed)

### 3.5.3 Novel data set in Local Data Store

The completed data set for NI practices comprised of 41 NI contractor files from 2018-04 to 2021-06 taking up 229Mb, 1 Practice information file (108Kb) and 73 NI prescribing files from 2015-07 to 2021-06 taking up 2.61Gb. Figure 3.8 provides a variable list for each file type all of which can be linked by Practice. The combined NI contractor files consist of 20 variables and 662,925 rows (a breakdown by file can be seen in Appendix B), the NI Practice information file consists of 13 variables and 333 rows (one for each practice) and the combined NI prescribing files contain 13 variables with 33,151,722 rows (a breakdown by file can be seen in Appendix A).



*NI contractor*

Practice
Practice Name
Address
Postcode
P-Lat
P-Long
P-East
P-North
Chemist
Contractor Name
Contractor Address
Contractor Postcode
C-Lat
C-Long
C-East
C-North
Year
Month
Number of Items
Distance

*NI prescribing*

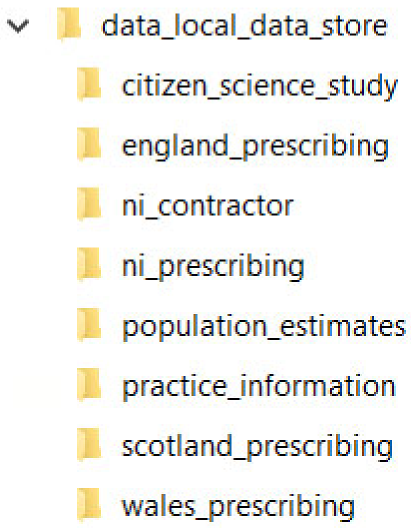
Practice
Year
Month
Practice
Item_Name
Strength
Total_Items
Gross_Cost_(£)
Actual_Cost_(£)
BNF_Chapter
BNF_Section
BNF_Paragraph
Registered Patients

*Practice information*

Practice
Practice Name
Postcode
P-lat
P-long
Population_Density_per_square_km
Number_of_Doctors
Practice_Size
Local_Government_District
Ward_Type
Ward_Name
Deprivation_Level
Practice Type

FIGURE 3.8: File types and variable lists in Local Data Store

In addition to the data held on NI practices within the LDS, supplementary data was added to enable comparisons with other UK nations (England, Scotland and Wales) along with data from the citizen science study (Figure 3.9). In total the zipped size of the LDS is 757.5Mb and can be downloaded from Zenodo (Booth, 2022).



data_local_data_store
- citizen_science_study
- england_prescribing
- ni_contractor
- ni_prescribing
- population_estimates
- practice_information
- scotland_prescribing
- wales_prescribing

FIGURE 3.9: Structure of Local Data Store

## 3.6 Discussion

The creation of this novel data store used the monthly Dispensing by contractor and GP prescribing data files as a basis for analysis. These were subsequently linked with other sources to enhance them in order to gain a deeper understanding of GP prescribing behaviours.

The earliest data available for Dispensing by contractor was April 2018, reporting data on 333 GP practices. Examining the number of practices reporting over time, it was established that no new practices had been added to this list and a total of 12 practices had been removed from it. A general exploration of the novel data set was undertaken in order to gain an understanding of what factors may affect prescribing behaviours and how NI prescribing behaviors compare with those of the other UK nations. Details of the results of this initial exploration are discussed in Chapter 4.

Traditionally, researchers have categorised GP surgeries as serving a rural, urban or semi-rural community (Eccles et al., 2019). These categories reflect the population of the area in which the surgery is located but does not reflect the prescribing behaviours of the practice or its relationship with the pharmacies dispensing prescriptions. In order to investigate whether these categories truly reflects the structure of GP practices, the novel data set was be used to explore whether these categories could be supported by use of a scientific method of identifying GP practice types and, if not, what alternative categories would better describe the resulting data types. The results of this analysis are discussed in Chapter 5.

Having discovered the types of GP practice that exist in Northern Ireland it was important to examine the differences in prescribing behaviours and explore some of the possible factors contributing to any differences observed. Two factors not initially considered as being contributing features in identifying GP practice types, Deprivation and practice size (by number of registered doctors) were then examined to gain insight into the their effect on prescribing of different types of GP practice. The results of this analysis are discussed in Chapter 6.

## 3.7 Limitations

The resultant NI contractor, NI prescribing and Practice information files can be linked using the 'Practice' variable which is an identification number unique to each practice. Each row in the NI contractor files details the total number of items dispensed by a contractor (pharmacy) for the month, issued by each a specific practice. The NI dispensing data is solely a monthly breakdown of items prescribed by each practice. As there is no breakdown of items in the NI contractor data, it is impossible to infer which medications are dispensed by each pharmacy.

## 3.8 Technical review

In considering how we are going to analyse a data set, we must first understand the types of data available, the types of analytics and techniques available and the areas these methods can be applied to. In addition to theoretical studies, there are three types of analytics (descriptive, predictive and prescriptive), four types of data (Web/Social Media, Sensor, Biometric, Administrative), with machine learning techniques generally being classified as supervised or unsupervised. The application of these methods fall into six areas (Healthcare administration, privacy and fraud detection, Public Health, Mental Health, Pharmacovigilience and Clinical decision support) (Fig. 3.10).

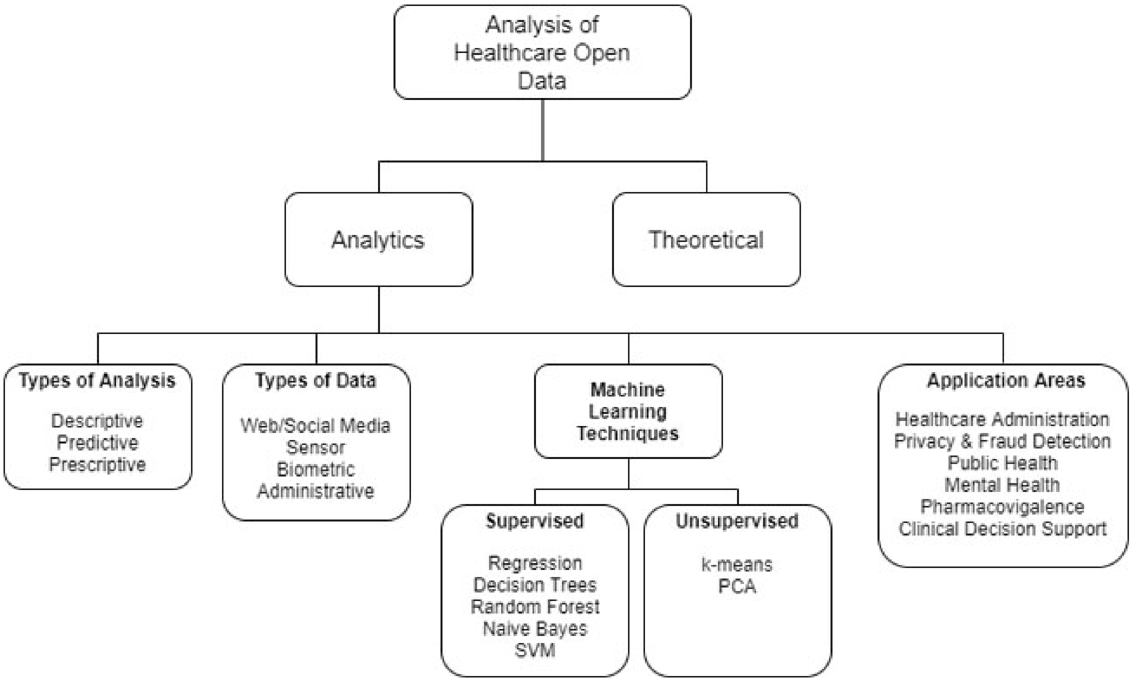


FIGURE 3.10: Healthcare analysis approaches

### 3.8.1 Types of analytics

In general, data analytics falls into one of three categories - Descriptive, Predictive or Prescriptive.

**Descriptive** analytics are those which seek to explore the data in order to discover new information from the data.

**Predictive** analytics are those which seek to predict upcoming events based on historical data.

**Prescriptive** analytics utilise scenarios in order to provide support for decision making.

### 3.8.2 Types of data

Healthcare data is collected from multiple sources which generally fall into one of four categories - Web/Social media, Sensor, Biometric or administrative.

**Web/Social media** - Data extracted from websites, blogs and social media like Twitter and Facebook are becoming more prevalent in medical research. With the development of Natural Language Processing (NLP) these data is becoming easier to analyse (Baldwin et al., 2013).

**Sensor** - This relates to data gathered from medical sensors or devices that can range from temperature sensors, pressure detectors, flow sensors, acoustic sensors and gas sensors to cameras, image sensors and magnetic field sensors. Image sensors and cameras used in the medical sector can be optical, X-ray or ultrasonic (Baldwin et al., 2013).

**Biometric** - these data relates to data such as genetics, fingerprints, retinal scans handwriting and blood pressure readings. Also included are medical imaging data such as test results from Ultrasound-Mammography, magnetic resonance imaging (MRI), computer tomography (CT), positron emission tomography (PET), radiography and X-rays. In some cases these data can be gained using medical sensors (Baldwin et al., 2013).

**Administrative** - This relates to data collected through administrative processes such as insurance claims, requests for payment in regard to prescriptions dispensed by pharmacists and electronic medical records relating to patient admissions, number of beds available or usage of specific medical equipment (Baldwin et al., 2013).

### 3.8.3 Data analytics

Data analytics encompasses statistics and computer programming with some elements of machine learning and involves raw data being analysed in order to identify trends or patterns, predict future events or provide support for decision making. Machine learning algorithms are procedures that are run on a data set performing pattern recognition so that they can learn from the data or be fit to the data. Different algorithms have different purposes. There are four categories of machine learning, supervised, unsupervised, semi-supervised and reinforcement. Supervised machine learning is commonly used when data is labeled while unsupervised is used when no labels exist. Semi-supervised or Deep Learning is used when there is a combination of labeled and unlabeled data available and reinforcement learning requires feedback in order for the algorithm to learn.

### 3.8.4 Supervised learning

The input data in supervised machine learning algorithms is called training data and has known labels. The data is trained using an algorithm resulting in a model on which predictions can be based. These models are usually used for regression or classification purposes.

**Regression** attempts to estimate the relationship between variables. The simplest form is linear regression where the output is the sum of the weighted attribute values. This is easiest to visualise in two dimensions where the line of best fit is drawn through points on a scatter graph. Another form is a Decision Tree where an instance's attributes are compared at each node in the tree determining their path through the tree. When a leaf is encountered then the instance is classified according to the class assigned to that leaf. Similar to decision trees, rules are a popular alternative where attributes are passed through a set of rules in order to determine which class the instance belongs to. Popular algorithms used for regression are Linear or Polynomial regression, Decision Trees and Random Forest. These are all used where the data is continuous, that is, data that can take any values. Examples include time, height and weight. Because continuous data can take any value, there are an infinite number of possible outcomes.

**Classification** is used to map attributes to predefined classes based on shared attributes. Popular supervised classification algorithms include K Nearest Neighbor, Decision Trees, Logistic Regression, Naive-Bayes and Support Vector Machine (SVM). These algorithms are used where the data is categorical, i.e. the data fits into a finite number of groups or categories. Categorical data can take on numerical values (such as “1” indicating Yes and “2” indicating No), but the numbers don't have mathematical meaning.

**Linear Regression** - Regression is generally used to predict a target value based on previous variables and the relationship between them. Simple linear regression assumes that there is a direct relationship between two variables and that changes in one variable results in a predictable change in the other. This is demonstrated by plotting each point on a scatter graph and calculating the line of best fit (i.e. the line which models the data best). The linear equation for this line can then be calculated.

**Decision Tree** - This algorithm is used on supervised data an although it can be used for both classification and regression it is most commonly used for classification. A decision tree is basically a series of 'if else' statements in the form of "if this is true pass to node A else pass to node B". This procedure is repeated until all possibilities have been exhausted and a final classification has been determined.

**Random Forest (RF)** - This classification algorithm, as implied by its name, consists of a number of decision trees working together. Each decision tree within the forest calculates a class prediction and the class with the most votes becomes the model's prediction. This algorithm works on the premise that a committee of decisions will invariably come to the right conclusion with those producing erroneous decisions being in the minority. In essence, the collective decision made by the forest is more likely to be correct than that of an individual decision tree.

**Nearest Neighbor** - The Nearest Neighbor, more commonly known as k Nearest Neighbor (kNN) works on the principal that an individual data point will be influenced by its nearest neighbors. As with the k-means algorithm used for classification, the choice of k is important. If k=1 then the algorithm is very specific and will classify the data point in the same class as the closest data point (overfitting), whereas if k=110 this is too generalised with the data point being classified to the class that the majority of the nearest data points belong to (underfitting).

**Logistic Regression** - This algorithm is used for classification purposes and is similar to Linear regression. As this algorithm is based on probability, the function that separates the classes is not a straight line but an 'S' curve known as a 'Sigmoid function'. Inputted numbers are weighted and compared to the Sigmoid function in order to find a relationship between the features and a probable outcome. In doing so, it established a relationship between one or more features and a binary outcome hence enabling it to classify new data or forecast trends in the data.

**Naive Bayes** - This algorithm is a supervised learning algorithm used for classification. Its name comes from the fact that it is naive (it assumes that all features are independent of each other) and it is based on Bayes Theorem. Bayes theorem or Bayes Law calculates the probability of an outcome based on previous knowledge. For example, if it rains on two sunny days and stays dry on five sunny days, there is a high probability that it will be dry on the next sunny day.

**Support Vector Machine (SVM)** - This is a supervised machine learning algorithm which can be used for both classification or regression. The algorithm creates a hyperplane (or line) which separates the data into two classes. The closest data points from each class has to be equidistant from the hyperplane.

### 3.8.5 Unsupervised learning

Unsupervised learning is used where the data is not labeled. The unsupervised machine learning algorithms are used to discover general structures within the data. These algorithms are generally used to cluster the data into groups with similar characteristics but can also be used for dimensionality reduction, anomaly detection or creation of association rules. Popular algorithms for clustering are k-means and Hierarchical and Principal Component Analysis (PCA) being used for dimensionality reduction. Discriminant Analysis can be used for both dimension reduction and classification.

**Clustering** is used in the identification of groups or categories in data where there is no predefined class to be predicted but instead the instances can be divided into natural groups. There are different ways in which the outcome of clustering can be expressed, the instance may belong to only one group (Exclusive), may belong to several groups (Overlapping), belongs to each group with a degree of probability (Probabilistic) or be hierarchical with groups being refined from a top level down to individual instances.

Clustering is an unsupervised learning data analysis technique used for discovering interesting patterns within data. There are many different clustering algorithms available for a data scientist to choose from and no single technique provides a best fit for all cases. Most algorithms use a measure of similarity or distance between data points to discover groups of data which can be grouped together. Some algorithms require the data scientist to guess the number of groups, or clusters of data, to be discovered, others require the data scientist to specify what distance data points are to be regarded as being close. All clustering algorithms are iterative whereby identified clusters are fed back into the algorithm for re-evaluation until a desired result is achieved. Some of the algorithms available include:

- Affinity Propogation - this algorithm measures similarities between pairs of data points exchanging real-valued messages between data points until a high quality set of examples and associated clusters emerge (Frey and D., 2007).

- Agglomerative Clustering - this algorithm is one of several hierarchical clustering methods and involves merging data points until the required number of clusters is achieved (Müllner, 2011).

- BIRCH - short for Balanced Iterative Reducing and Clustering using Hierarchies, this algorithm involves constructing a tree structure from which cluster centroids are identified (Zhang et al., 1996).

- DBSCAN - short for Density-Based Spatial Clustering of Applications with Noise, this algorithm involves finding high-density areas in the data and expanding those areas around them as clusters (Ester et al., 1996).

- K-Means - probably the most widely known and a popular choice of clustering algorithm as the only parameter is k (the number of clusters) and is less expensive in terms of processing power and memory than other methods. k points are chosen at random as centroids and all other points are assigned to a cluster based on their Elucidean distance from these centroids in an effort to minimize the distance (Dash and Liu, 2000) between data points within each cluster (MacQueen, 1967). Table 3.10 summarises the advantages and disadvantages of k-means clustering.

TABLE 3.10: Summary of the advantages and disadvantages of k-means clustering (Google Developers, 2022)

| **Advantage** | **Disadvantage** |
|---|---|
| Relatively simple to implement. | k must be chosen manually. |
| Scales to large data sets. | Dependent on initial values |
| Guarantees convergence. | Trouble clustering data of varying sizes and density |
| Can warm-start the positions of centroids. | Clusters outliers |
| Easily adapts to new examples. | Requires scaling when using high numbers of features |
| Generalizes to clusters of different shapes and sizes, such as elliptical clusters. | |

- Mini-Batch K-Means - this algorithm is a modified version of the k-means algorithm which updates clusters during each iteration by sampling mini-batches rather than the entire data set (Sculley, 2010).

- Mean Shift - this algorithm's clustering method involves finding and adapting centroids based on the density of data points shifting the mean during each iteration in order to increase the density of each cluster (Jones and Bhanu, 1999).

- OPTICS - is short for Ordering Points To Identify the Clustering Structure and is a modified version of DBSCAN (Ankerst et al., 1999).
- Spectral Clustering - This algorithm utilises linear algebra in order to cluster data points to a specified number of clusters (CFFR, 2016).
- Gaussian Mixture Model - is a probabilistic model which summarizes a multivariate probability density function with a mixture of Gaussian probability distributions in order to identify clusters within a data set (Biernacki et al., 2000).

**Anomaly detection** - Anomalies are data points that do not fit into the usual structure of the data being analysed. These anomalies are often labeled as outliers in the data and can be detected using machine learning algorithms such as k-means clustering. Another well known method of detecting anomalies is Local Outlier Factor (LOF) which is carried out in most nearest neighbor based algorithms.

**Association** is used to find relations between variables. Association rules are similar to classification rules except that they can be used to predict any attribute, not just the final class if the instance.

**Discriminant analysis** - there are two types of Discriminant analysis; Linear and Quadratic both of which are used for classification. Linear Discriminant Analysis works on the basis that any random variable comes from one of a number of classes. A discriminant rule tries to divide the data set into the number of disjoint regions that represent all the classes. In trying to classify a data point it is then allocated to the region that it is closest to using two rules to guide this decision. The Maximum Likelihood Rule assumes that each class could occur with equal probability and the Bayesian Rule which calculates the classes prior probabilities for allocation of the data point. Quadratic Discriminant Analysis works on the same basis although the discriminant function in this instance is quadratic, not linear.

**Principal Component Analysis (PCA)** - this method reduces a data set with multiple variables down to its principal components. The principal components are the underlying structure of the data or the components that provide the most variance. This is often used to reduce the dimensions of the data set in order to graphically display the results of clustering.

### 3.8.6 Correlation

Correlation is the measure of the extent to which two variables are related. For example, two variables with a strong positive correlation would exhibit the same trends, if one increased, the other would increase also. A popular method of testing for correlation is the use of Pearson's correlation coefficient which produces a result based on the covariance of two continuous variables which is in the range -1 to +1. In interpreting the results, a coefficient near +1 represents a near perfect positive correlation whist a coefficient near -1 represents a near perfect negative correlation (in other words when one variable increases a similar decrease is seen in the second variable). Coefficient values between 0.50 and 1 represent a strong positive correlation, values between 0.30 and 0.49 a moderate correlation, values below 0.29 a low correlation and a value of 0 being no correlation. Conversely, coefficient values between -0.50 and -1 represent a strong negative correlation, values between -0.30 and -0.49 a moderate negative correlation and values below -0.29 a low negative correlation (Statistics Solutions, 2021).

### 3.8.7 Statistical Significance tests

Statistical significance tests are used to gauge the probability that a relationship between variables is due only to random chance. In other words, they tell us what the probability is that we would be making an error if we were to assume that we have found relationship between these variables (California State University, 2022).

The Wilcoxon signed-rank test can be used for comparing the results from supervised machine learning models. Using k-fold cross-validation this test calculates k accuracy scores in order to test whether the two samples differ significantly from each other. If they do, then one is more accurate than the other. (Grootendorst, 2019)

Another significance test used to compare supervised learning modes is the McNemar's test and is used to check the extent to which the predictions between one model and another match. If the resulting p-value is lower than 0.05, we can reject the null hypothesis and see that one model is significantly better than the other (Grootendorst, 2019).

As supervised learning models will not be appropriate for the data sets in this project, Independent t-tests, appropriate for use with unsupervised machine learning techniques will be used. A t-test is a statistical test that is used to compare the means of two groups. It is often used in hypothesis testing to determine whether two groups are different from one another commonly known as an independent two-sample t-test that compares the means for two groups. Common assumptions made when doing a t-test include the scale of measurement being similar, the randomness of the sampling of data, normality of data distribution, adequacy of sample size, and equality of variance in standard deviation (Maverick, 2021). The test produces two metrics known as the t-value and the p-value. The t-value indicates the difference between the two groups with a large t-value indicating a larger difference between the groups and a low score indicating that the groups are similar. The larger the t-value the more likely that the results are repeatable. The p-value is the probability that the results from your data occurred by chance. Written as a decimal, a p-value of 0.01 means that there is only a 1% chance that the data happened by chance. It is generally accepted that a p-value of 0.05 is acceptable to mean that the data is valid while a p-value less than 0.01 represents very strong evidence that the two groups are different (Statistics How To, 2021; Fernandez, 2020; Whitley and Ball, 2002).

### 3.8.8 Multiple Testing Correction Techniques

In the statistical testing of a hypothesis it is generally accepted that a probability of less than 0.05 (5%) is sufficient to control the maximum experimentwise error rate (MEER) and thus the probability of rejecting falsely at least one true individual null hypothesis. This probability rate is referred to as the alpha value ($\alpha$). The probability of committing false statistical inferences increases considerably when more than one hypothesis is simultaneously tested (i.e. multiple comparisons). There are two ways of adjusting for this, calculate adjusted probabilities for each set of data or adjust the alpha value to which these probabilities (p-values) are compared. It is this latter type of adjustment that will be used in this project (Chen et al., 2017; Bender and Lange, 2001). These are some of the methods for adjusting alpha to minimise the possibility of false positives creeping into the results:

**Bonferroni adjustment** - One of the most commonly used, this method adjusts the alpha value down by the number of simultaneously tested hypotheses (m) where adjusted alpha equals alpha divided by the number of tested hypotheses (Equation 3.2) (Bland and Altman, 1995).

$$\alpha' = \frac{\alpha}{m} \tag{3.2}$$

**Holm adjustment** - Based on the Bonferroni method, the Holm adjustment computes significance levels based on the p value based on the rank of the p value in that for the i$^{\mathrm{th}}$ ordered hypothesis $\mathrm{H}_{(\mathrm{i})}$, $\mathrm{p}_{(\mathrm{i})}$ is compared to alpha$_{(\mathrm{i})}$ ordered from the smallest p value first. This is continued until $\mathrm{p}_{(\mathrm{i})}$ is greater than or equal to adjusted alpha$_{(\mathrm{i})}$ (Holm, 1979). Equation 3.3 is used to calculate the adjusted alpha value.

$$\alpha'_{(i)} = \frac{\alpha}{m - i + 1} \tag{3.3}$$

**Hochberg adjustment** - This method is similar to the Holm method and uses the same equation (3.3) to calculate the adjusted alpha value. Where Hochberg differs is that this method starts with the largest p value and iterates downward until $\mathrm{p}_{(\mathrm{i})}$ is less than adjusted alpha$_{(\mathrm{i})}$ (Hochberg, 1988)

### 3.8.9 Validation Techniques

Validation of any machine learning algorithm is important in order to ensure the model is providing good results. In supervised learning it is important to validate the model not only to ensure good results on the training data but on the live test data as well. With the absence of labels in unsupervised data it is harder to test whether the model is producing good results therefore metrics such as cohesion within the cluster (a measure of how close objects are within the cluster) and separation between different clusters (how well each cluster is separated from the others) are used (Grootendorst, 2019; Jain, 2020). Table 3.11 provides an overview of validation techniques commonly used.

TABLE 3.11: Summary of Supervised and Unsupervised Machine learning validation techniques

| **Technique** | **Type of Model** | **Description** |
|---|---|---|
| Train/test split | Supervised | Data is split randomly into 70%/30% where 70% is used to train the model with 30% used to test it. This allows the researcher to see how the model reacts to previously unseen data. (Grootendorst, 2019) |
| k-Fold Cross-Validation | Supervised | Data is split into k folds, then trained on k-1 folds and test on the one fold that was left out. It does this for all combinations and averages the result on each instance. (Grootendorst, 2019) |
| Leave-one-out Cross-Validation | Supervised | Leave-one-out Cross-Validation uses each sample in the data as a separate test set while all remaining samples form the training set. (Grootendorst, 2019) |
| Nested Cross-Validation | Supervised | This technique involves nesting two k-fold cross-validation loops |
| Time-series Cross-Validation | Supervised | In this technique information from the future is reserved for testing ensuring that all training data happens before the test data. (Grootendorst, 2019) |
| Silhouette coefficient | Unsupervised | Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1 with 1 representing well defined clusters, 0 representing clusters where the distance between them is indifferent an -1 representing wrongly aligned clusters. (Bhardwaj, 2020) |
| Calisnki-Harabasz coefficient | Unsupervised | The Calinski-Harabasz index meaures variance between clusters with well-defined clusters having a large between-cluster variance and a small within-cluster variance. The optimal number of clusters corresponds to the solution with the highest Calinski-Harabasz index value. (MathWorks, 2022) |

Continued on next page

**Table 3.11 – continued from previous page**

| Technique | Type of Model | Description |
|---|---|---|
| Dunn index | Unsupervised | The Dunn index is calculated as a ratio of the smallest inter-cluster distance to the largest intra-cluster distance. A high score means better clustering since observations in each cluster are closer together, while clusters themselves are further away from each other. (PyShark, 2022) |
| Hartigan index | Unsupervised | The Hartigan index is based on the ratio of the logarithm between the within and between sum-of-squares. (Scherl, 2010) |

### 3.8.10 Areas of application in healthcare

**Healthcare administration** - Data mining techniques can be used on medical administrative data held in data warehouses (Post et al., 2013) in order to improve quality and reduce costs (Phillips-Wren et al., 2008), optimise the utilisation of resources and assist with patient management. In addition to the clinical uses, data warehouses can be used for research, quality control and training purposes.

**Privacy and fraud detection** - The privacy of patients and the data collected on them is of the utmost importance. For this reason machine learning techniques have been developed to provide anonymisation to the data in order for it to be used for research purposes (E. Youssef, 2014), often outside the healthcare system. Techniques have also been developed to detect abuse of medical systems and fraud. These systems have he ability to highlight suspicious care activity, misrepresentation of information and financial irregularities (Yang and Hwang, 2006).

**Public health** - Whilst most research focuses on a specific problem, person or disease, machine learning is also used to investigate general public health problems. These can range from designing preventative healthcare systems (Nimmagadda and Dreher, 2014), predicting hospitalisation numbers due to influenza outbreaks (Buczak et al., 2016) and creating data mining systems to extract knowledge for non-expert users (Santos et al., 2013).

**Mental health** - With the mental health of each population becoming a global concern (Chong et al., 2012), clinicians are looking to artificial intelligence to aid them in the early diagnosis and treatment of mental health issues. Studies have explored the use of AI to identify and intervene in the developmental delay of children (Chang, 2007), provide personalised treatment for anxiety disorder (Panagiotakopoulos et al., 2010) and predict and provide early diagnosis for disorders such as Insomnia and dementia (Carús Candás et al., 2014).

**Pharmacovigilience** is the monitoring of medications in order to identify adverse drug reactions caused by combination of drugs or specific patient biology (Harpaz et al., 2013). Whilst most studies focus on adverse reactions due to multiple drugs (Harpaz et al., 2010; Eriksson et al., 2014; Jin et al., 2008), there have been studies on adverse reactions caused by anticancer agents (Sakaeda et al., 2011b) and Statins used in Cardiovascular disease and muscular and renal failure treatment (Sakaeda et al., 2011a; Maguire et al., 2007).

**Clinical decision support** refers to the analysis of large volumes of medical data using machine learning algorithms in order to filter information specific to an individual or situation. These systems are used to improve care quality, avoid errors or adverse events, and medical staff to be more efficient and provide descriptive and/or predictive analysis to inform clinical decisions. These systems are already being used in hospitals to inform decisions on treatment for cardiovascular disease (Karaolis et al., 2010; Tsipouras et al., 2008; Bandyopadhyay et al., 2015), cancer (Delen, 2009; Agrawal et al., 2012) and diabetes (Huang et al., 2007; Razavian et al., 2015; Barakat et al., 2010).

## 3.9 Conclusion

The combination of several data sources, linked by GP practice and location by way of postcode has resulted in the creation of a data store containing a novel data set facilitating the investigation into the types of GP practice in Northern Ireland, the exploration of their prescribing behaviours and the investigation of what impact factors such as deprivation and size of practice have on those behaviours. A review of the techniques available has also been undertaken to inform the following analysis decisions. The discussion of the results of these investigations are detailed in Chapter 4, 5 and 6.

# Chapter 4

# Exploration of open prescription data

> "Keep going, and don't worry about your speed. You're making progress, even if it doesn't seem like it. Forward is forward, no matter how slow."
>
> Lori Deschene

Chapter 3 discussed the creation of a local data store consisting of a novel data set relating to 333 NI GP practices and their prescribing behaviours from July 2015 onward. This chapter provides an initial analysis of elements within the LDS to assess whether they contribute to the prescribing profile of GP practices and should be considered as features to be used in the scientific categorisation of GP practices (RQ1). To facilitate like for like comparisons with the other UK nations, prescribing per head of population was used.

The aims of this chapter is to:

- Use basic analytics to explore and describe GP practices within the novel data set constructed for this study.
- Discover the relationship between population and registered patients.
- Explore the distance travelled from GP practice to dispensing pharmacy over time.
- Explore the relationship between deprivation and the location of GP practices.
- Explore the relationship between GP practice size and their location.
- Explore the location of GP practices using the traditional rural / urban categorisations.
- Compare prescribing trends and levels with England, Scotland and Wales.

## 4.1 GP practices

The Dispensing by contractor data set was first published for April 2018 reporting data on prescriptions issued by the 333 GP practices actively operating at that time. Figure 4.1 shows a map of the locations of these practices. In tracking prescribing data it was necessary to ensure that all practices were included in the study. This meant that any new practices being added over time would need to be accounted for. Practices that ceased reporting prescription data were assumed to have either closed or have been amalgamated into another practice. As each practice is assigned a unique identification number, it was possible to check each month to check on the numbers of practices reporting prescribing data and identify those no longer reporting. Over the period April 2018 - June 2021, no new GP practices were recorded issuing prescriptions while 12 practices ceased reporting over the same period (Figure 4.3).

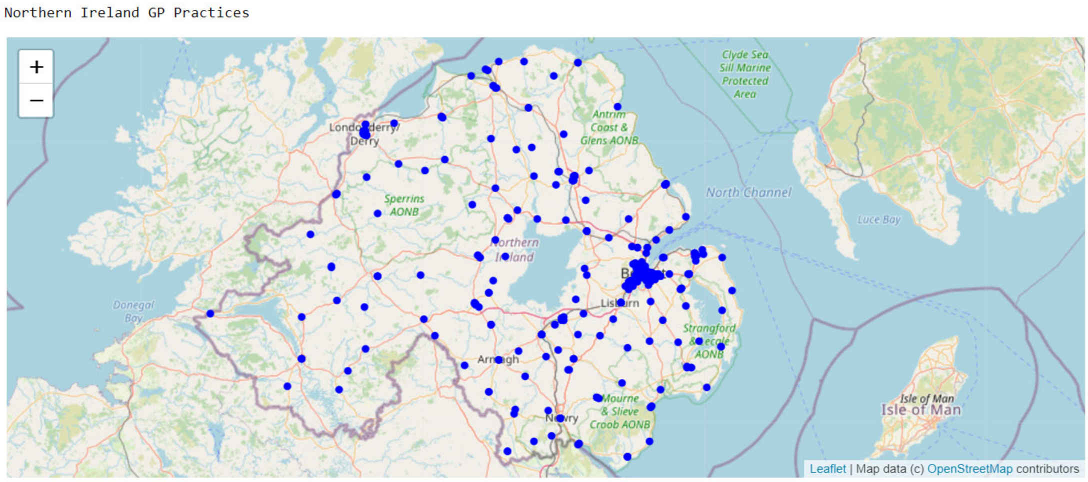


FIGURE 4.1: Map of NI GP practice locations

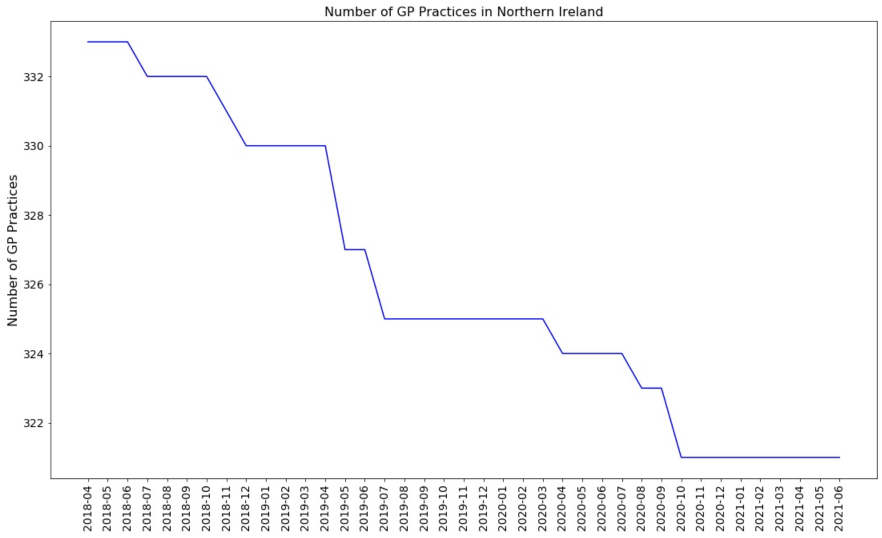


FIGURE 4.2: Number of GP practices reporting prescription data (April 2018 - June 2021)

## 4.2 Registered patients

The total number of registered patients in Northern Ireland has risen from 1,943,085 in April 2018 to 2,006,937 in June 2021, a rise of 3.29%. This is in line with population figures estimated by the Office for National Statistics although the number of registered patients is consistently higher than population estimates (Figure 4.3). The discrepancy between the two figures can be attributed to two causes:

- **Estimation** - The UK population is counted during the decennial census with mid-year estimates being produced every six months based on the number of registered births and deaths and net migration.
- **Ghost patients** - Ghost patients are created due to issues with records management within surgeries (Royal College of General Practitioners, 2019). Ghost patients can occur when the registered patient has died but the surgery has not been informed or on occasion when the patient has moved to another country without deregistering from the practice. As practices are paid based on the number of registered patients they have, they perform occasional list cleansing exercises to remove ghost patients.

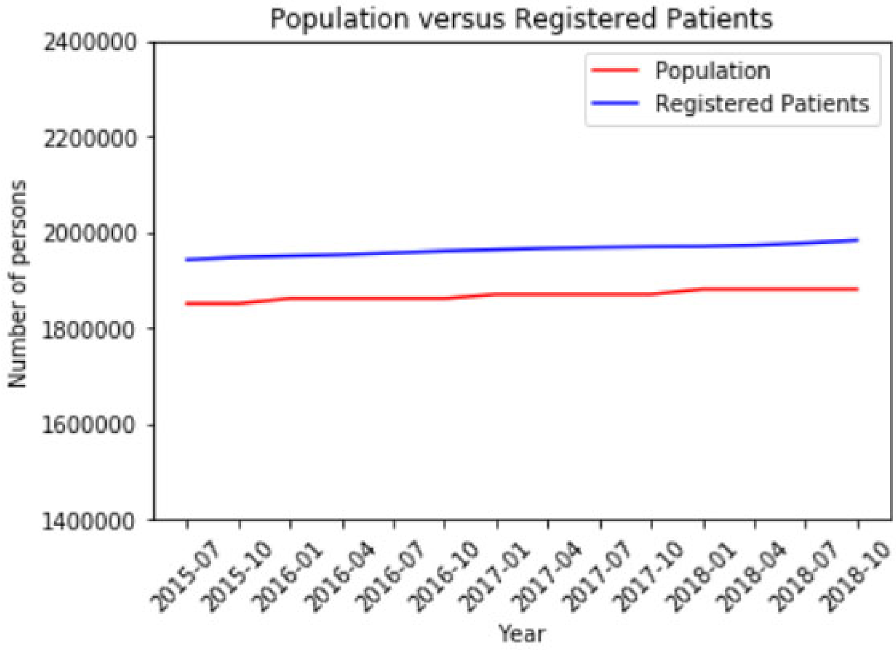


FIGURE 4.3: Comparison of Number of registered patients with population estimates in Northern Ireland

## 4.3 Distance traveled to dispense prescriptions

A brisk walking speed is typically a minimum of 3 miles per hour or 5 kilometres per hour, which means brisk walkers should be able to complete 5 kilometres in 1 hour (The Pacer Blog, 2021). The Department of Infrastructure Travel Survey for Northern

Ireland 2014-2016 (NI Department for Infrastructure, 2021) estimated that 71% of all journeys made were by car, 5% by public transport and 18% on foot. This being the case, an arbitrary distance of 5 kilometres was chosen as a reasonable distance to travel to dispense prescriptions.

Analysis of the distance traveled by patients to dispense prescriptions shows that less than 43% of the items prescribed are dispensed within 5 kilometres of the issuing practice with half of items prescribed being dispensed at pharmacies over 20 kilometres from the issuing practice (Figure 4.4). This dispels the assumption that most prescriptions are dispensed close to the issuing practice and supports the assumption that the majority of patients do not live in the same super output area as the practice they are registered with. Admittedly, some prescriptions may be dispensed close to patient's work addresses but there is no way to verify if this is the case.

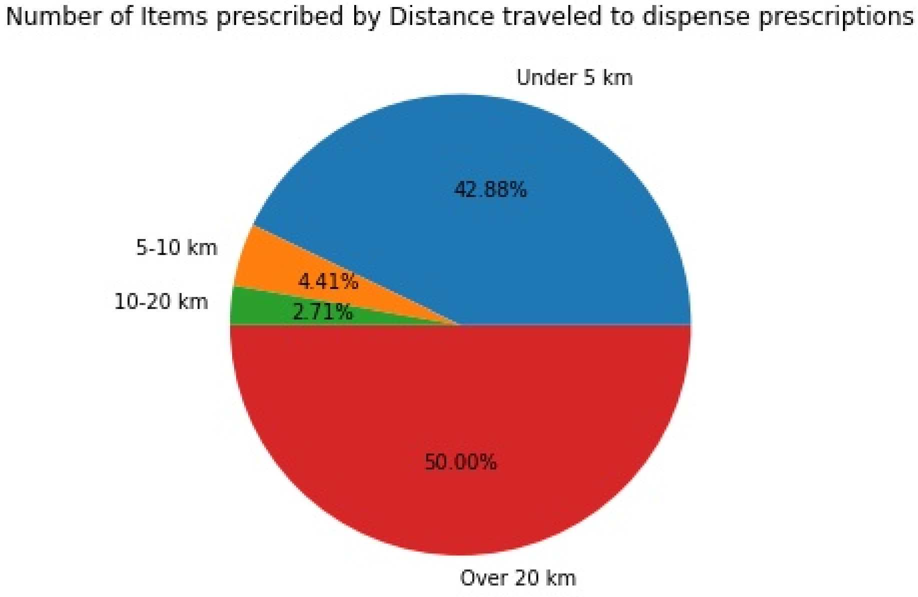


FIGURE 4.4: Number of items prescribed by distance traveled to dispense prescriptions

As the distance travelled between the prescribing GP surgery and the dispensing pharmacy was already calculated during the construction of the local data store detailed in Chapter 3, a monthly average was calculated. Analysis of these averages showed that the distance travelled had fallen from 14.8 km to 13.1 km over the period April 2018 - June 2021 (Figure 4.5). Whilst the COVID-19 pandemic and subsequent lockdowns are a likely contributor to this reduction since March 2020, the trend had already been established before this. Interestingly, travel distances seem to spike seasonally around July each year. This may be due to patients being on holiday.

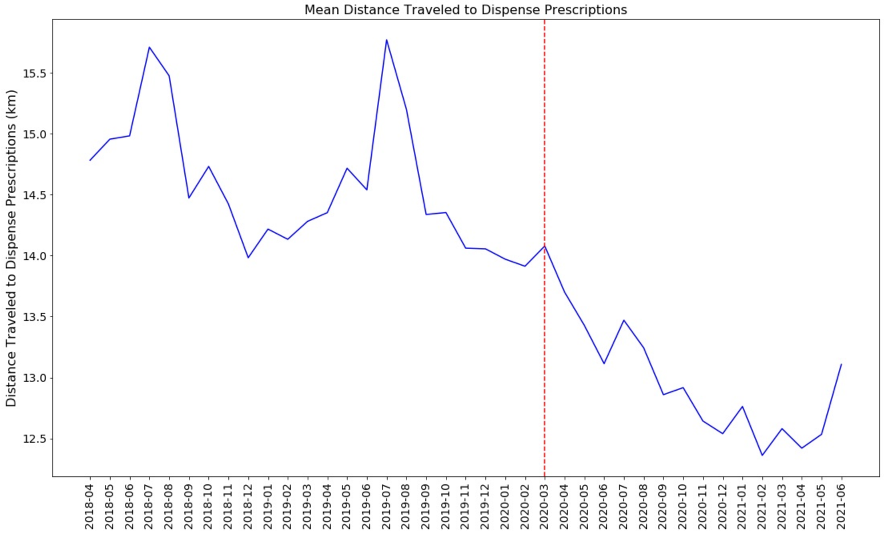


FIGURE 4.5: Mean distance traveled to dispense prescriptions in Northern Ireland with red line indicating the first lockdown imposed as a result of the COVID-19 pandemic.

## 4.4 GP practices by deprivation area

Examining the level of deprivation attributed to the area in which each GP surgery was located showed that there was an even distribution of surgeries with 84 (25.2%) being located in areas with low deprivation (quartile 1), 83 (24.9%) in areas with low/medium deprivation (quartile 2), 85 (25.5%) in areas with medium/high deprivation (quartile 3) and 81 (24.3%) in areas with high deprivation. Looking at these surgeries on a map (Figure 4.6) showed that the majority of practices in high deprivation areas were located in major towns and cities.

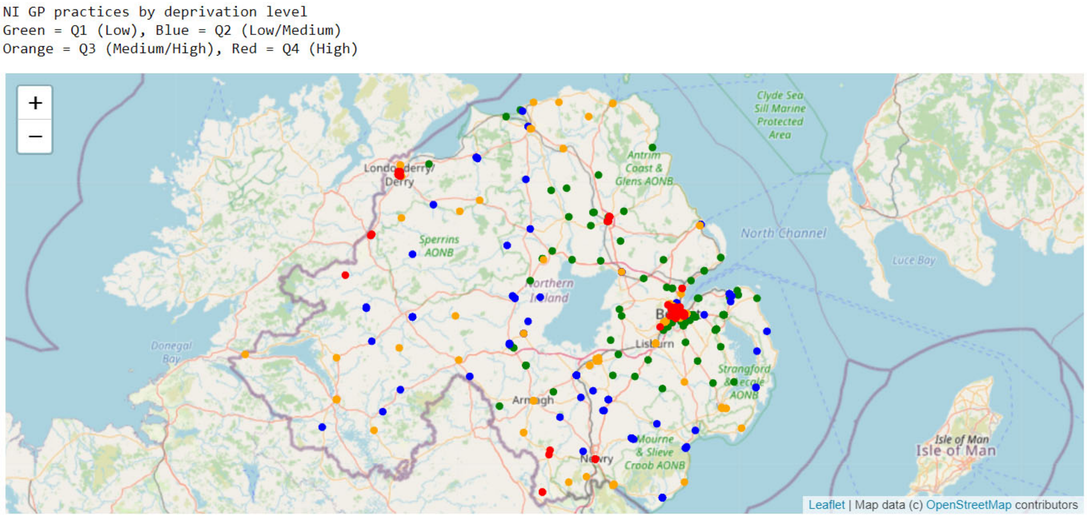

NI GP practices by deprivation level
Green = Q1 (Low), Blue = Q2 (Low/Medium)
Orange = Q3 (Medium/High), Red = Q4 (High)
+
−
Clyde Sea Sill Marine Protected Area
Londonderry/Derry
Antrim Coast & Glens AONB
Sperrins AONB
North Channel
Luce Bay
Northern Ireland
Belfast
Donegal Bay
Lisburn
Strangford & Lecale AONB
Armagh
Mourne & Slieve Croob AONB
Newry
Isle of Man
Leaflet | Map data (c) OpenStreetMap contributors

FIGURE 4.6: NI GP practices by deprivation level attributed to the location in which they are located. Green = Q1 (Low), Blue = Q2 (Low/Medium), Orange = Q3 (Medium/High), Red = Q4 (High).

## 4.5 GP practices by practice size

Categorising GP practices by their size based on the number of GP's in the practice showed that there were 29 (8.7%) Single-Handed practices with only one registered doctor, 61 (18.3%) Small practices with two registered doctors, 122 (36.6%) Medium practices with three to four doctors and 121 (36.3%) Large practices with over five doctors. Looking at the locations of these surgeries on a map (Figure 4.7), no clear pattern can be seen dictating any geographical reasons for the respective sizes of surgeries.

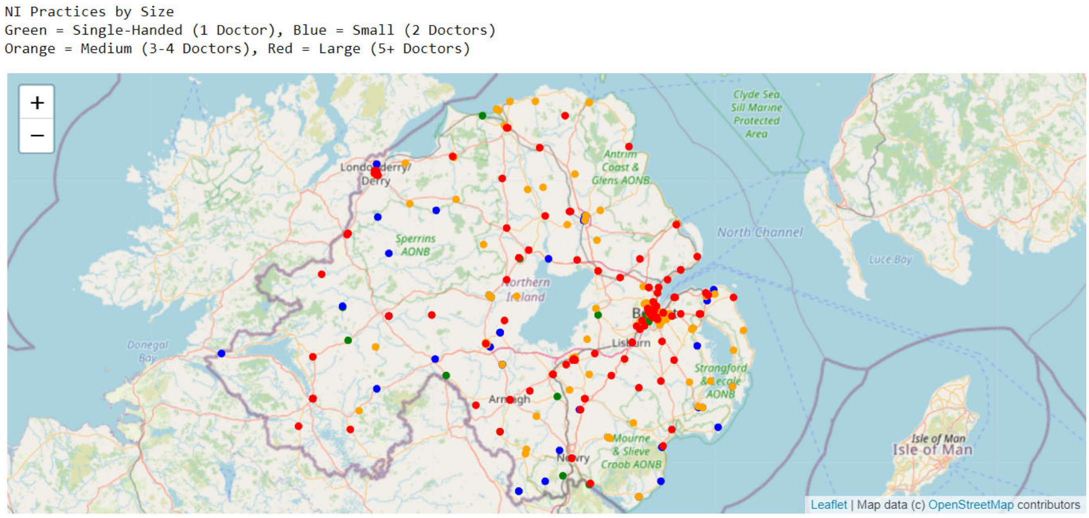

NI Practices by Size
Green = Single-Handed (1 Doctor), Blue = Small (2 Doctors)
Orange = Medium (3-4 Doctors), Red = Large (5+ Doctors)
+
−
Londonderry/Derry
Sperrins AONB
Northern Ireland
Antrim Coast & Glens AONB
Clyde Sea Sill Marine Protected Area
North Channel
Luce Bay
Belfast
Lisburn
Donegal Bay
Strangford & Lecale AONB
Armagh
Mourne & Slieve Croob AONB
Newry
Isle of Man
Leaflet | Map data (c) OpenStreetMap contributors

FIGURE 4.7: Location of NI GP practices by practice size. Green = Single-Handed (1 Doctor), Blue = Small (2 Doctors), Orange = Medium (3-4 Doctors), Red = Large (5+ Doctors).

## 4.6 Geographical breakdown of GP practices

Traditionally geographical areas have been designated depending on the population in those areas. Classifications are made at Output area level, which is the smallest geographical area considered by the Office for National Statistics[1]. Output areas are classified as Urban if they have a population of 10,000 people or more whilst all other areas are designated Rural. In Northern Ireland, the NI Statistics Research Agency performed a review of the Statistical Classification and Delineation of Settlements in 2015 that considered settlements and their proximity to areas designated as Urban. This review suggested that a new categorization, mixed - rural/urban[2]. Using this revised classification scheme, 225 (67.6%) practices are designated to be located in urban areas, 85 (25.5%) in rural areas and 23 (6.9%) in mixed - rural/urban areas. Figure 4.8 shows a map of NI practices using these classifications.

[1] 2011 rural/urban classification, Office for National Statistics, Available at: https://www.ons.gov.uk/methodology/geography/geographicalproducts/ruralurbanclassifications/2011ruralurbanclass

[2] Review of the Statistical Classification and Delineation of Settlements, NISRA, Available at: https://www.nisra.gov.uk/publications/settlement-2015-documentation

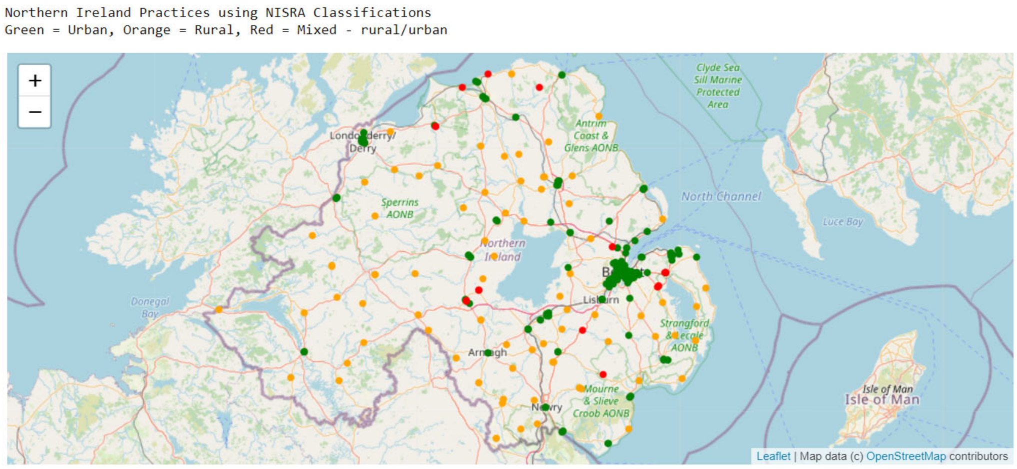

Northern Ireland Practices using NISRA Classifications
Green = Urban, Orange = Rural, Red = Mixed - rural/urban
+
−
Londonderry/Derry
Sperrins AONB
Antrim Coast & Glens AONB
Clyde Sea Sill Marine Protected Area
North Channel
Luce Bay
Northern Ireland
Belfast
Lisburn
Donegal Bay
Strangford & Lecale AONB
Armagh
Newry
Mourne & Slieve Croob AONB
Isle of Man
Leaflet | Map data (c) OpenStreetMap contributors

FIGURE 4.8: Locations of NI GP practices using NISRA classifications of Urban, Rural and Mixed-rural/urban. Green = Urban, Orange = Rural, Red = Mixed - rural/urban.

## 4.7 Comparison with UK nations

Comparable GP prescription data is published for all UK nations although for differing periods. The NHS Business Services Authority have published English prescribing data from January 2014[3], NHS Wales from April 2018[4] and Public Health Scotland from October 2015[5]. The total Number of items prescribed each month was extracted from these data sets and normalised using the mid-year estimates for the population of each nation[6]. The resulting graph (Figure 4.9) shows that overall prescribing per head of population is highest in Wales, Northern Ireland second highest and England and Scotland, having similar levels, being the lowest prescribers. Pearson's correlation co-efficient, used to measure the relationship or association between two continuous variables, was calculated to gauge the correlation between NI and the other nations. It was found that there was a high correlation between NI and England (r=.79), NI and Scotland (r=.77) and between NI and Wales (r=.88).

[3] NHS Business Services Authority English Prescribing Dataset (EPD) https://digital.nhs.uk/data-and-information/publications/statistical/practice-level-prescribing-data

[4] NHS Wales General Practice Prescribing Data https://nwssp.nhs.wales/ourservices/primary-care-services/general-information/data-and-publications/general-practice-prescribing-data-extract/

[5] Public Health Scotland Prescriptions in the Community https://www.opendata.nhs.scot/dataset/prescriptions-in-the-community

[6] Office for National Statistics Population Estimares, Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates

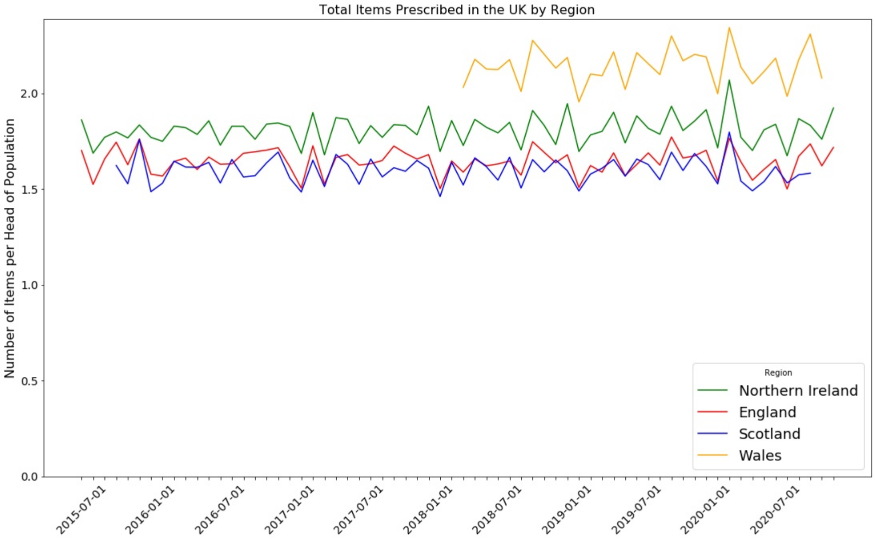


FIGURE 4.9: Comparison of number of items prescribed per head of Population in England, Scotland, Wales and Northern Ireland

### 4.7.1 Comparison by British National Formulary chapter

Examining the prescribing relationship between NI and the other UK nations at BNF chapter level provides more insight into the types of disease more prevalent in the UK nations. To account for seasonality, each years data was averaged and compared with the previous giving an indication of the trend over time. No comparable comparisons were possible with Wales as data were only available from 2018 onward. BNF chapter 18 (Preparations used in Diagnosis) refers mainly to preparations used for X-rays and, as the majority of these are performed in hospitals, the figures recorded in this category are negligible. For this reason, chapter 18 has been removed from this analysis.

A detailed analysis of each BNF chapter was carried out (Appendix E) with the average number of items prescribed per head of population being compared for each nation where data was available. Where data was only available for a specific period (e.g. Wales 2018-04 onward), only this period was compared with that of NI. The results of this analysis (Table 4.1) shows that prescribing by NI GP practices is higher than the rest of the UK nations in 5 of the BNF chapters: 4 (Central Nervous System), 5 (Infections), 10 (Musculoskeletal & Joint Diseases), 13 (Skin) and 20 (Dressings). Correlation of prescribing trends between NI prescribing and that of other UK nations was also examined (Table 4.2). This showed that there was a high correlation in the majority of BNF chapters. Notable exceptions were chapter 14 (14 - Immunological Products & Vaccines) which showed a weak negative correlation with prescribing in both England and Wales, Chapter 15 (15 - Anaesthesia) showing a weak negative correlation with prescribing in England and Chapter 19 (19 - Other Drugs And Preparations) showing strong negative correlations with prescribing in England and Scotland and a weak negative correlation with prescribing in Wales.

TABLE 4.1: Average number of items prescribed per head of population (Chapters in bold indicate where prescribing in NI is highest in the UK.)

| **BNF Chapter** | **NI** | **England** | **Scotland** | **Wales** |
|---|---|---|---|---|
| 1 - Gastro-Intestinal System | 0.177 | 0.150 | 0.158 | 0.196 |
| 2 - Cardiovascular System | 0.416 | 0.485 | 0.375 | 0.621 |
| 3 - Respiratory System | 0.137 | 0.108 | 0.119 | 0.159 |
| **4 - Central Nervous System** | **0.456** | 0.312 | 0.360 | 0.444 |
| **5 - Infections** | **0.075** | 0.057 | 0.071 | 0.063 |
| 6 - Endocrine System | 0.151 | 0.165 | 0.131 | 0.212 |
| 7 - Obstetrics | 0.046 | 0.045 | 0.043 | 0.055 |
| 8 - Malignant Disease & Immunosuppression | 0.008 | 0.007 | 0.008 | 0.011 |
| 9 - Nutrition And Blood | 0.088 | 0.087 | 0.068 | 0.109 |
| **10 - Musculoskeletal & Joint Diseases** | **0.066** | 0.047 | 0.059 | 0.060 |
| 11 - Eye | 0.025 | 0.027 | 0.028 | 0.034 |
| 12 - Ear, Nose And Oropharynx | 0.025 | 0.017 | 0.026 | 0.025 |
| **13 - Skin** | **0.081** | 0.047 | 0.072 | 0.054 |
| 14 - Immunological Products & Vaccines | 0.002 | 0.021 | 0.001 | 0.022 |
| 15 - Anaesthesia | 0.003 | 0.002 | 0.003 | 0.003 |
| 19 - Other Drugs And Preparations | 0.002 | 0.002 | 0.002 | 0.002 |
| **20 - Dressings** | **0.015** | 0.011 | 0.014 | 0.011 |
| 21 - Appliances | 0.025 | 0.043 | 0.044 | 0.051 |
| 22 - Incontinence Appliances | 0.002 | 0.003 | 0.004 | 0.002 |
| 23 - Stoma Appliances | 0.006 | 0.009 | 0.007 | 0.009 |

TABLE 4.2: Correlation of UK nations Average number of items prescribed per head of population with NI (Chapters in bold indicate where a negative correlation is observed.)

| **BNF Chapter** | **England** | **Scotland** | **Wales** |
|---|---|---|---|
| 1 - Gastro-Intestinal System | 0.875 | 0.751 | 0.908 |
| 2 - Cardiovascular System | 0.834 | 0.634 | 0.904 |
| 3 - Respiratory System | 0.873 | 0.798 | 0.878 |
| 4 - Central Nervous System | 0.895 | 0.798 | 0.939 |
| 5 - Infections | 0.971 | 0.970 | 0.958 |
| 6 - Endocrine System | 0.827 | 0.746 | 0.852 |
| 7 - Obstetrics | 0.786 | 0.630 | 0.889 |
| 8 - Malignant Disease & Immunosuppression | 0.747 | 0.874 | 0.839 |
| 9 - Nutrition And Blood | 0.558 | 0.744 | 0.873 |
| 10 - Musculoskeletal & Joint Diseases | 0.925 | 0.835 | 0.930 |
| 11 - Eye | 0.884 | 0.853 | 0.834 |
| 12 - Ear, Nose And Oropharynx | 0.907 | 0.876 | 0.837 |
| 13 - Skin | 0.937 | 0.929 | 0.931 |
| 14 - Immunological Products & Vaccines | **-0.212** | 0.910 | **-0.176** |
| 15 - Anaesthesia | **-0.006** | 0.765 | 0.290 |
| 19 - Other Drugs And Preparations | **-0.738** | **-0.759** | **-0.282** |
| 20 - Dressings | 0.915 | 0.819 | 0.757 |
| 21 - Appliances | 0.721 | 0.876 | 0.750 |
| 22 - Incontinence Appliances | 0.495 | 0.282 | 0.372 |
| 23 - Stoma Appliances | 0.723 | 0.597 | 0.380 |

## 4.8 Discussion

Dispensing by Contractor data for NI is only available from April 2018. For this reason, the 333 GP practices in operation at this date were examined over time. Examining the geographical spread of practices (Figure 4.1) it is evident that there is higher concentrations of practices in the cities of Belfast and Derry/Londonderry. Whilst practices are more spread out in counties Tyrone and Fermanagh this reflects the rural nature of these counties. For this reason, it is likely that population density is a key feature influencing the location of a GP practice. In this regard, the traditional method of classification agrees with these findings. It was found that the number of GP practices has declined over the period with no new practices being established (Figure 4.2). This is not as a result of a declining population or that of declining registered patients as figures show that the population has grown steadily over the period with registered patients matching this growth (Figure 4.3). With this in mind, it is reasonable to assume that the patients registered with practices which cease to operate will be transferred to another practice. On this assumption, the number of registered patients will have an effect on the prescribing behaviours of a GP practice and should be considered as a contributing factor to the classification of the practice. In considering the distance travelled by patients to dispense prescriptions, it was found that 50% travelled over 20 kilometres while only 43% dispensed prescriptions within 5 kilometres of their GP practice (Figure 4.4). On the assumption that patients are most likely to dispense prescriptions at a location convenient to them, it is likely that the location of the pharmacy used is an indication of the location in which the patient resides or works. On this basis it is reasonable to assume that the distance patients are willing to travel to see their doctor and the number of pharmacies used to dispense prescriptions issued by a particular practice are likely features to be considered when categorising it. As expected, GP practices located in areas of high deprivation are generally found within large towns or cities (Figure 4.6). Whilst this is the case, it has already been established that patients do not necessarily live in the same location as their GP practice meaning that the level of deprivation is not likely to have a major influence on the categorisation of a practice and will therefore not be included a a describing feature. Similarly, the size of the GP practice in regards to the number of registered doctors working in it shows no distinct pattern with almost three quarters having 3 or more registered doctors (Figure 4.7). Again it was not felt that the size of the practice would be a major contributor to its prescribing behaviours and was not considered as a feature for classification purposes. Comparing the overall number of items prescribed per head of population in NI against that of the other UK nations revealed that NI practices had the second highest prescribing levels, Wales being consistently higher (Figure 4.9). Analysing the prescribing levels by BNF chapter showed that NI practices had the highest prescribing levels in 5 chapters: 4 (Central Nervous System), 5 (Infections), 10 (Musculoskeletal & Joint Diseases), 13 (Skin) and 20 (Dressings) (Table 4.1). Whilst this is the case, the prescribing trends in these chapters were highly correlated with that of the other UK nations (Table 4.2) and do not stand out as being noteworthy. Negative correlations were found in 3 chapters: 14 (Immunological Products & Vaccines), 15 (Anesthesia) and 19 (Other Drugs and Preparations). This indicates that prescribing in the respective nations fell as it was rising in NI. These negative correlations are generally weak and seem to reflect differing reporting structures in each nation. Chapter 14 (Immunological Products & Vaccines) (Figure E.14) shows similar (highly positively correlated) prescribing trends in NI and Scotland whereas prescribing of vaccines in England and Wales is negligible. This suggests that whist practices in NI and Scotland prescribe and administer vaccines locally this is possibly done at regional level in England and Wales by the NHS. Prescribing in Chapter 15 (Anesthesia) (E.15) is minimal in all nations indicating that Anesthesia is not routinely prescribed by GP surgeries. The large fall in prescribing in England and Wales at the end of 2019 can be attributed to the COVID-19 pandemic. Trends for NI and Scotland do not fall in the same way because the data provided by NHS England and Wales includes prescribing by dental practices who routinely use anesthetic. Due to a change in reporting in Chapter 19 (Other Drugs and Preparations) (Figure E.16) it is not clear if there is a parity of reporting between all nations. This is the only BNF chapter where strong negative correlations are evident and it is likely that this can be attributed to a disparity of reporting indicated by the almost non existent reporting in NI before July 2018. Given the generally strong correlations and similar prescribing levels observed, there is no indication that factors outside of NI contribute to the classification of GP practices within NI.

## 4.9 Limitations

The number of patients registered with a practice is reported each quarter. For the purpose of this analysis it was assumed that this figure would remain the same for the following two months which in reality is unlikely. Also, an accurate count of the population is only taken every 10 years during the census. The figures used for comparison in this study are mid year estimates produced by the Office for National Statistics taking into account births, deaths and net migration since the previous year.

## 4.10 Conclusion

Initial analysis of individual elements within the LDS indicate that in addition to location (traditionally used for categorising GP practices) other features should be considered when attempting to classify GP practices. As the number of practices has fallen over time without a corresponding fall in the overall number of registered patients, it is assumed that registered patients within individual GP practices fluctuates making this a likely feature to be considered. Distance travelled to dispense prescriptions and the number of pharmacies associated with an individual GP practice were also considered to be features contributing to the prescribing profile of practices. Examination of Deprivation and Practice size showed no signs of being important to the classification of GP practices and was not considered as features. Similarly, comparison on NI prescribing levels and patterns with that of the other UK nations showed no indications that NI GP practices are substantially different than those in England, Scotland or Wales. Chapter 5 details the process used to scientifically evaluate the types of GP practice with the aim to test the validity of the traditional categorisations of Urban, Rural and Semi-rural.

# Chapter 5

# Analysis of General Practice archetypes

> "A wise man once said that you should never believe a thing simply because you want to believe it."
>
> Tyrion Lannister
> Game of Thrones

The previous chapter provides a baseline for comparing GP prescribing trends in Northern Ireland establishing that the number of registered patients has increased in line with population and that NI has the highest prescribing levels per head of population in six of the twenty BNF chapters including chapter 4 which covers the prescribing of antidepressants and analgesics. The work presented in this chapter was published in the Scientific Reports journal as a paper entitled "Discovering and Comparing Types of General Practitioner Practices Using Geolocational Features and Prescribing Behaviours by Means of K-Means Clustering" (Booth et al., 2021b) and modified here to fit within the framework and context of this thesis. This chapter discusses the work conducted in the discovery of of General Practitioner practices using Geolocational Features and Prescribing Behaviours by Means of K-Means Clustering in greater depth. Comparison of identified archetypes will be discussed in chapter 6.

## 5.1 Background

Traditionally, General Practices have been categorised using the density of the local population as a benchmark resulting in the urban, rural and semi-rural categorisations (Eccles et al., 2019). Formal mechanisms exist for the classification of geographical areas based on population [1] [2], but are based solely on population density.

[1] Northern Ireland Statistics Research Agency (2018) Urban - Rural Classification, Available at: https://www.nisra.gov.uk/support/geography/urban-rural-classification

[2] Department for Environment, Food & Rural Affairs (2017) The 2011 Rural-Urban Classification for Output Areas in England, Available at:

In the previous chapter, research suggests that GP practice have a wider influence due to their prescribing profiles and the distance patients travel to dispense prescriptions. Working on the hypothesis that patients dispense their prescriptions at a pharmacy most convenient to them, the distance travelled is likely to be more representative of the location in which each patient resides. As such, practices should not be categorised solely on their location but account should be taken of their prescribing profile. This is vital as accurate categorisation will enable like-for-like comparisons between practices within a given archetype leading to the identification of incorrect prescribing practices.

The aim of this chapter is to propose a new mechanism for categorising GP practices using machine learning clustering techniques to identify clusters of practices which differ in their prescribing patterns with the view to establishing whether traditional categorisations are valid. If not, these new clusters will be examined and categorised accordingly.

## 5.2 Methods

### 5.2.1 Data sources

This analysis was performed using a subset of variables from the local data store which was created for this project. Chapter 3 details the data sources, wrangling and linking of these data sets.

As unsupervised clustering was to be performed on the data with the aim of identifying patterns or clusters within the data, it was important to choose the most relevant features to cluster on. Important features help to create clusters and unimportant features may hinder the formation of clusters. In considering what metrics define the attributes of a GP practice in relation to categorising it using not only its geographical location but its relationship to its patients in the form of the prescriptions prescribed and subsequently dispensed, six features were chosen. From initial investigations (Chapter 4) the distance travelled to dispense prescriptions may be significant in the categorising of GP practices. The assumption was made that the majority of patients would take their prescriptions to their local pharmacy for dispensing, the distance being an indication of the distance patients live from the GP practice and the overall influence of the practice.

In order to eliminate seasonal variations, one years data (April 2018 - March 2019) was used to create the clustering data set. Figure 5.1 shows the combination of variables from the LDS to create the six features used. This resulted in an initial data set of 7 variables with 231,777 rows of data taking up 10.6Mb of memory. These data were then aggregated to provide data points for each practice resulting in 333 rows of data to be submitted to the clustering algorithm.

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file /591462/RUCOA_leaflet_Jan2017.pdf

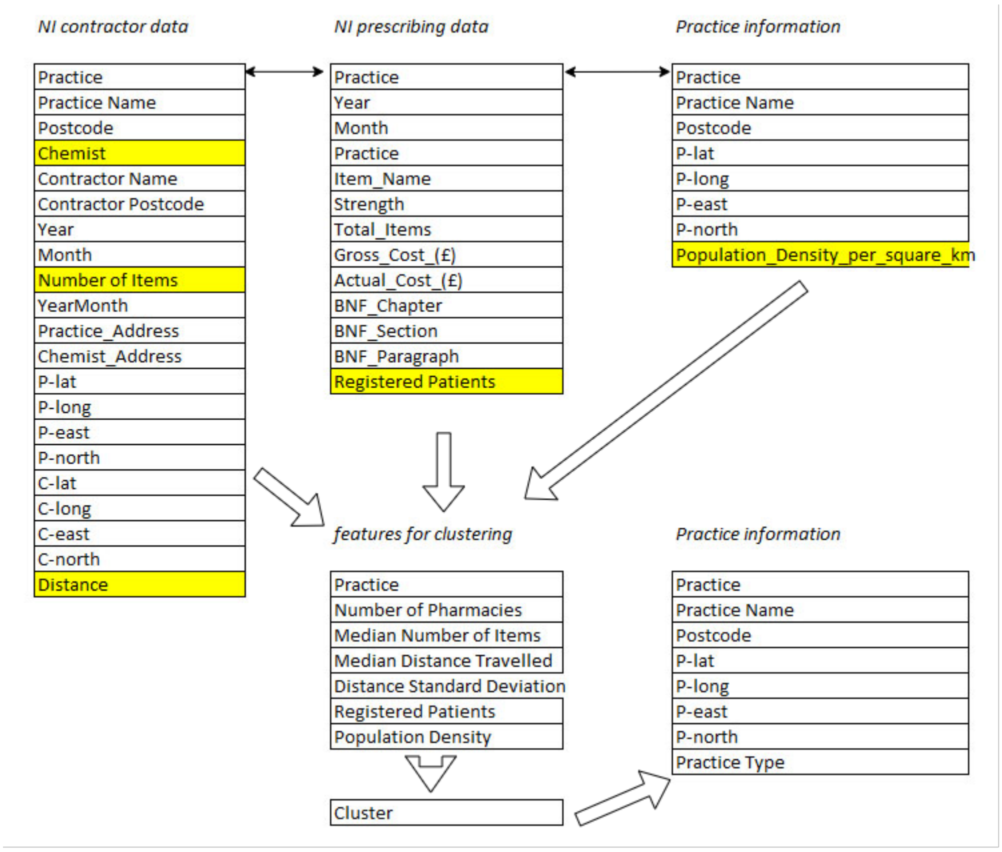


FIGURE 5.1: Workflow for identification of practice type using clustering

### 5.2.2 Feature selection

Six features were chosen that best represented the relationship of a practice with its geolocation and prescribing profile. As geographical coordinates are not linear and pose a challenge for machine learning, distance was used as a proxy for geolocation resulting in the following features.

- **Number of pharmacies** - This is the total number of pharmacies that has dispensed prescriptions for each GP practice during the year. While the majority of pharmacies will dispense numerous prescriptions for a GP practice, it is only counted once during period giving an indication of the influence each GP practice has on the surrounding area.

- **Number of items per registered patient** - The number of items prescribed by a GP practice gives an indication of the health of the patients registered with the practice. This figure in itself is not comparable as there are differing numbers of registered patients within each practice. In order to normalise the number of items prescribed to make it comparable, it has been divided by the number of registered patients in the practice.

- **Median distance to pharmacy (km)** - Using the postcodes of both GP practice and pharmacy, the distance traveled by a patient to dispense their prescription could be calculated. The calculated distances were then examined and, due to the presence of some extreme distances, it was decided that the median distance would be more representative of a typical journey than the average distance traveled. The assumption was that if the patient did not dispense their prescription at a pharmacy close to the GP practice, they were more likely to dispense it close to where they lived. This figure also gives an indication of the influence each practice has on the surrounding area.

- **Distance standard deviation (km)** - the standard deviation associated with the median distance to pharmacy was also calculated in order to gauge the variation in distance traveled and shows the extent to which the median distance may vary.
- **Population density per square km** - This figure is the population density of the super output area in which each practice is located. This is the figure that would be used by statistical agencies to decide whether the area should be categorised as urban, rural or semi-rural. This feature is important to link the new methodology with the previous.
- **Registered patients** - This is the number of patients registered with each practice and gives an indication of the size of the practice in terms of patients.

In preparation for clustering, data for each practice was collated to provide a single data point for each feature. The distribution graph for distance traveled to dispense prescriptions showed that this variable was skewed to the left and therefore not Gaussian (Figure 5.2). For this reason, the median for this feature was used for clustering as opposed to the mean which would be skewed also. Over the twelve-month period, the pharmacies where a particular GP practice prescription was dispensed were counted uniquely.

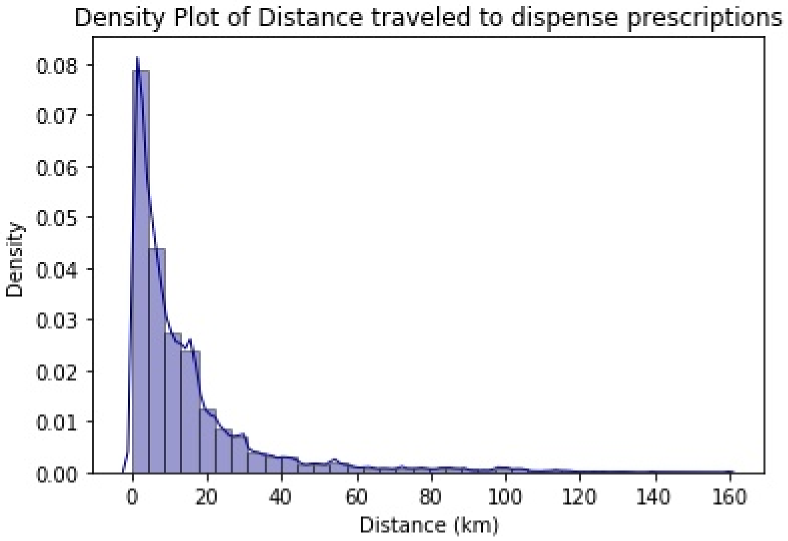


FIGURE 5.2: Density plot of distance traveled to dispense prescriptions

### 5.2.3 Normalisation

As the six selected features did not share the same range of data, it was important to normalise the data set. This was done using the Whitening function in pythons scipy library. The whitening function transforms the original data set with known co-variances into one where the co-variances form the identity matrix (Wikipedia, 2021) i.e.all variables range between 0 and 1. This is achieved by dividing the difference between the variable to be normalised and the minimum value by the difference between the maximum value and the minimum value (Equation 5.1).

$$X_{Normalised} = \frac{(X - X_{Minimum})}{(X_{Maximum} - X_{Minimum})} \tag{5.1}$$

### 5.2.4 Selection of clustering algorithm

It was important to choose the best clustering algorithm suitable for the data. For this reason a selection of 9 algorithms were tested with the resulting clustering solutions evaluated using the Calinski-Harabasz Index, which measures of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The cohesion is estimated based on the distances from the data points in a cluster to its cluster centroid and separation is based on the distance of the cluster centroids from the global centroid (Dey, 2021). The higher value of resultant index means the clusters are dense and well separated.

### 5.2.5 Choosing optimal parameters

Three types of clustering algorithm were tested: Partitioning clustering, Hierarchical clustering and Density-Based Spatial Clustering.

**Partition clustering** subdivides the data into a set of k groups where k is specified by the analyst. It is important to find the optimum value for k to ensure that the correct number of clusters are identified. Two methods were used for this task in order to verify that the correct k value was chosen (DataNovia, 2018a).

- The Elbow method - This method runs k-means clustering on the data set for a range of values (e.g. 1 - 10) calculating the average distances to the centroid from all data points. These distances are then plotted on a graph which will show where the distances fall suddenly creating an elbow in the graph. This elbow represents the optimum value for k. This method relies on the data scientist visually choosing the correct value and is not always accurate.
- Silhouette analysis - This method calculates a silhouette coefficient for each data point. This coefficient is a measure of how similar a data point is within-cluster (cohesion) compared to other clusters (separation). Similar to the Elbow

method, silhouette coefficients are calculated for a range of clusters with the average of the resultant coefficients for each value of k being reported. This was calculated using the silhouette_score module within the sklearn library.

**Hierarchical clustering** does not need the number of clusters to be specified. The result of hierarchical clustering is a tree-based representation of the objects known as dendrogram. Observations can be subdivided into groups by cutting the dendrogram at a desired similarity level (DataNovia, 2018a).

**Density-Based Spatial Clustering** identifies dense regions, which can be measured by the number of objects close to a given point. The parameter, epsilon, defines the radius in which other data points can be considered as belonging to the same cluster. The optimal value of epsilon can be calculated using Knee point detection. The average of the distances of every point to its k nearest neighbours is calculated with the value of k being specified by the user and corresponds to MinPts. The larger the data set, the larger the value of minPts should be chosen although minPts must be at least 3. In the resulting graph, a knee corresponds to a threshold where a sharp change occurs along the k-distance curve and corresponds to the optimal value for the epsilon parameter (DataNovia, 2018b).

### 5.2.6 Visualising clusters

In order to visually verify that the output from the k-means clustering algorithm is correct, the number of dimensions (features) must be reduced so that the data points can be plotted in two dimensions. This study used Principal Component Analysis (PCA) to convert the six identified features of each practice into a two-dimensional array of uncorrelated variables for visualisation. Once the number of clusters was verified using this method, the cluster labels were mapped back onto the original data set to allow analysis of the clusters.

### 5.2.7 Dispersion of pharmacies

The dispersion of pharmacies (distance between pharmacies dispensing prescriptions for a given practice) for each cluster type was also examined to see if this information would support the cluster labelling established by the k-means clustering algorithm. To do this the average distance between each pharmacy dispensing prescriptions for a particular practice was calculated and presented as a box plot for analysis.

### 5.2.8 Hypothesis Testing

As multiple hypothesis tests (7 in all) have been employed in this chapter there is an increased possibility of false discoveries. To address this a corrected Bonferroni alpha value has been calculated as 0.05/7 = 0.007.

## 5.3 Results

### 5.3.1 Choosing the optimum value for k

In choosing the optimum number of clusters (k), both the Elbow and Silhouette methods were used with the resulting graph from the elbow method (Figure 5.3) open to interpretation and the Silhouette method (Figure 5.4) clearly showing that the optimum number of clusters was in fact two (k=2).

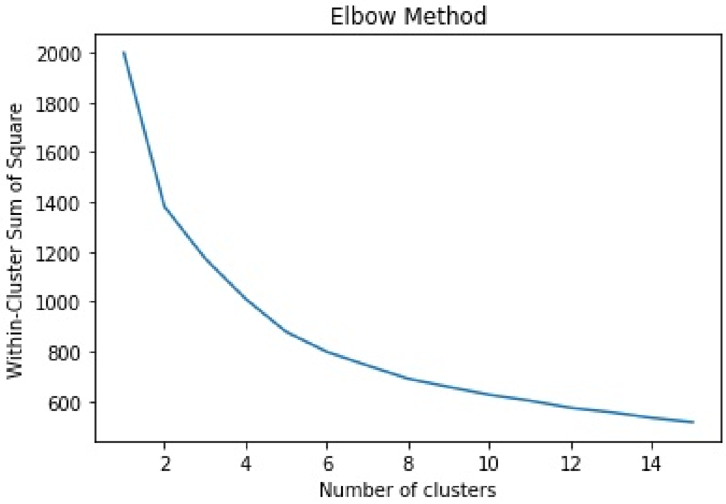


FIGURE 5.3: Plot of results using the Elbow method to determine the optimum number of clusters in the data set.

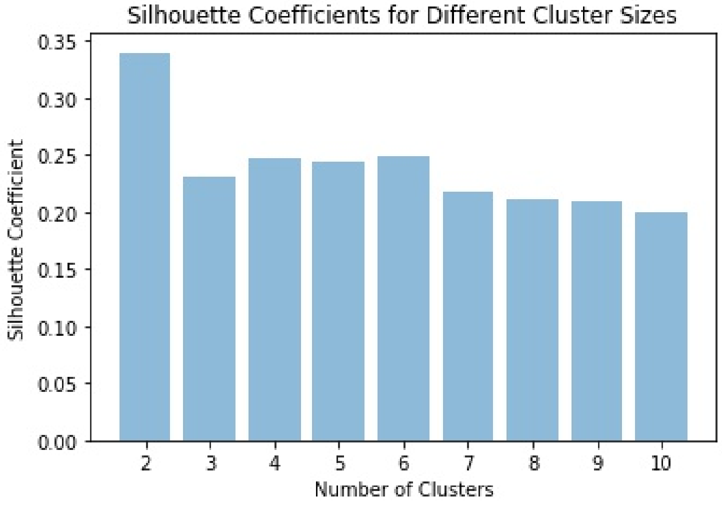


FIGURE 5.4: Plot of results using the Silhouette method to determine the optimum number of clusters in the data set.

### 5.3.2 Choosing the optimum value for epsilon

The Knee point method was used to calculate the optimum value of epsilon (Figure 5.5) with the resulting knee indicating that epsilon equals 1.88.

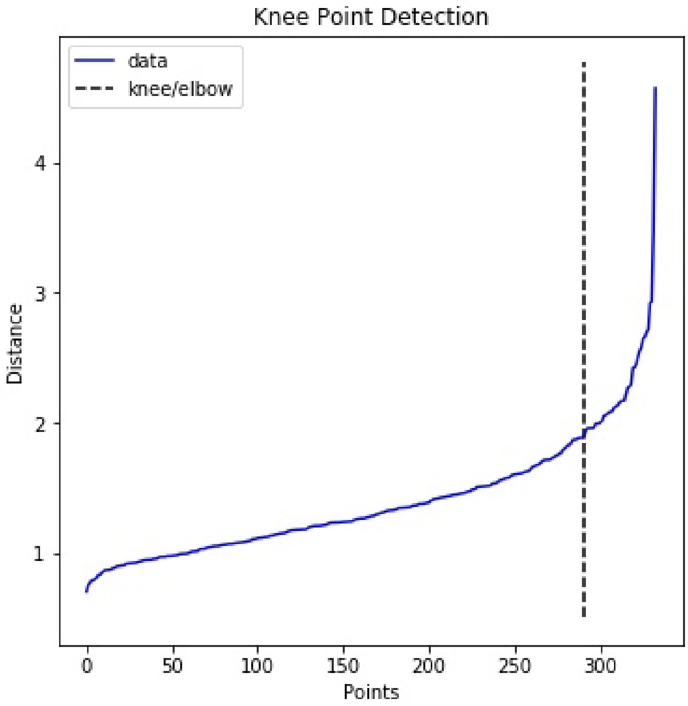


FIGURE 5.5: Knee Point detection method to find the optimum value of epsilon.

### 5.3.3 Hierarchical clustering Dendrogram

The resulting dendrogram (Figure 5.6) produced when submitting the data to hierarchical clustering clearly indicated the presence of two main clusters with sub clusters also evident.

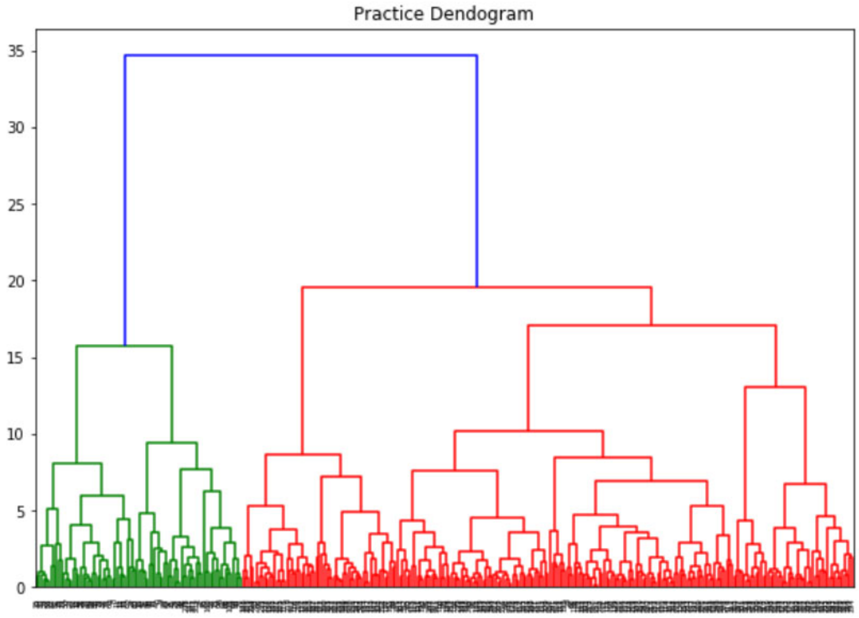


FIGURE 5.6: Dendrogram produced as the result of hierarchical clustering.

### Analysis of clustering algorithms

Using the optimum values for k and epsilon previously calculated, 9 different algorithms were used to provide clustering solutions on the data set (Figure 5.7). The algorithms evaluated were Affinity Propagation, BIRCH, Dbscan, Hierarchical, k-means, Mean Shift, Mini-batch k-means, Optics and Spectral. The resulting clusters were scored using the Calinski-Harabasz Index with the results ordered highest to lowest (Table 5.1). Based on these scores k-means and Mini-batch k-means were identified as having the highest identical scores and subsequently k-means was chosen. In choosing the value for the maximum number of iterations to be used in the k-means algorithm, research has shown that K-means converges after 20-50 iterations in all practical situations (Broder et al., 2014). For this reason it was felt that using the python script default value of 100 was sufficient.

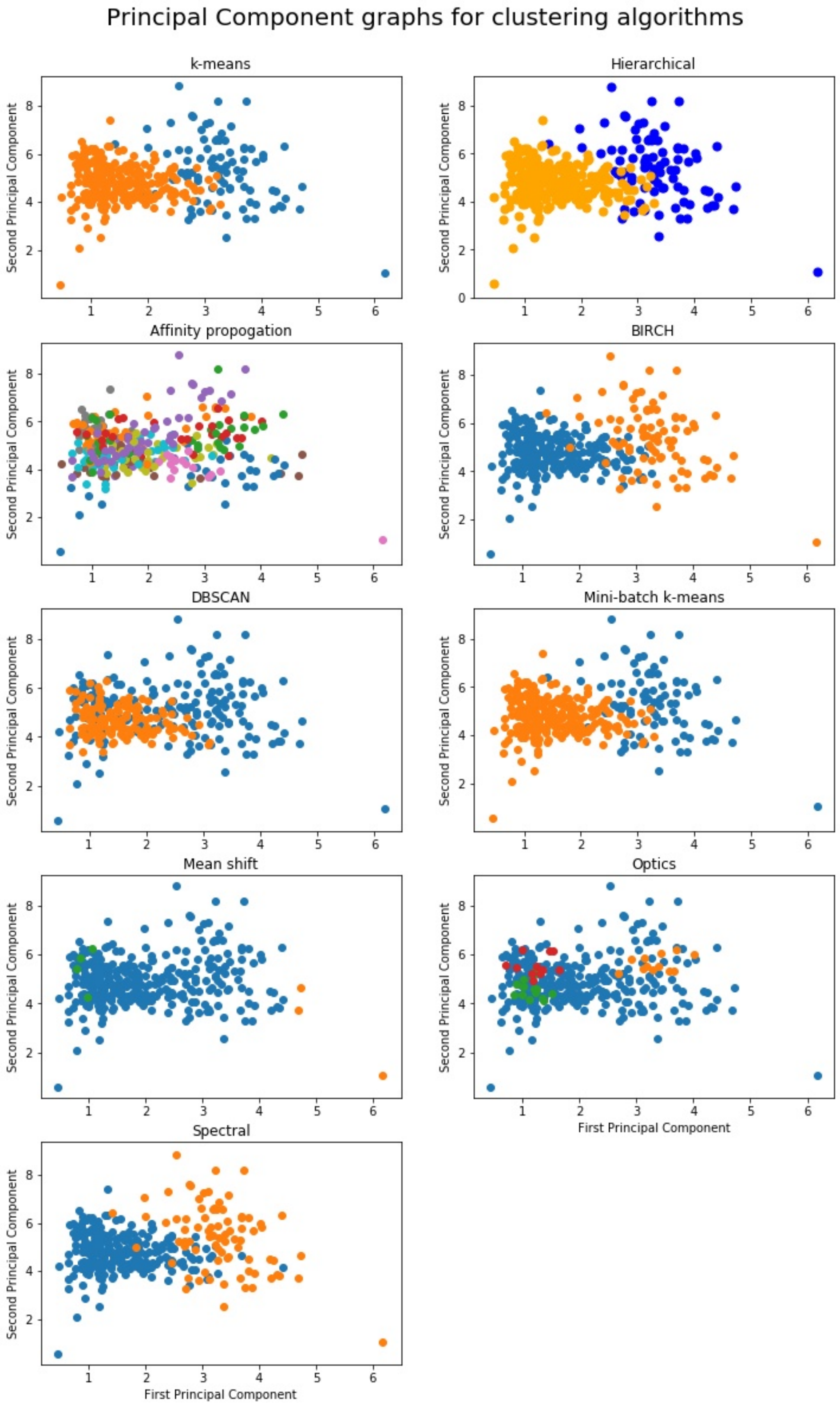


FIGURE 5.7: Comparison of clustering algorithms

TABLE 5.1: Analysis of algorithm performance using the Calinski-Harabasz Index

| Algorithm | Score |
|---|---|
| k-means | 147.754016 |
| Mini-batch k-means | 147.754016 |
| Hierarchical | 142.856024 |
| BIRCH | 142.856024 |
| Spectral | 142.077419 |
| Affinity Propagation | 53.620322 |
| Dbscan | 32.155019 |
| Optics | 11.335246 |
| Mean shift | 10.335918 |

### 5.3.4 Exploration of the urban/rural/semi-rural categories

Whilst the all indicators (Elbow, Silhouette and Dendrogram) indicated that k=2 is the optimum number of clusters, out of curiosity, it was decided to see if the methodology would support the hypothesis that there exists 3 categories (ie Urban, Rural and Semi-Rural). To this end, k-means clustering was performed on the data set with k=3. The results were then reduced to two dimensions using PCA and visualised (Figure 5.8). The resultant clusters were scored using the Calinski-Harabasz Index giving a score of 116.10. Visual inspection of the resulting PCA plot demonstrated that three clusters were not a good fit for the data and this was upheld by the lower Calinski-Harabasz Index score than that of the two cluster solution (147.75). This result disproved the hypothesis that GP practices in Northern Ireland can be categorised as Urban, Rural and Semi-rural in the same way that geographical areas are basing the categorisation solely on the population in that area.

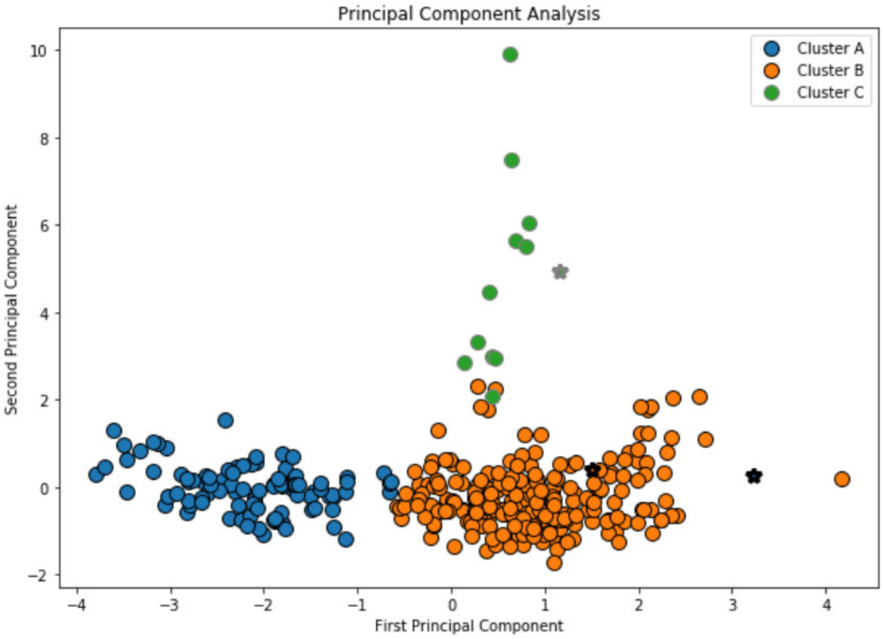


FIGURE 5.8: Principal Component Plot based on hypothesis that three types of GP practice exist in Northern Ireland

### 5.3.5 Categorisation of two clusters

Having subjected the data set to k-means clustering where k=2, PCA was used to convert the six identified features for each practice into a two-dimensional array of uncorrelated variables for visualisation. The resulting graph (Figure 5.9) showing two distinct clusters verified that this was the optimum setting.

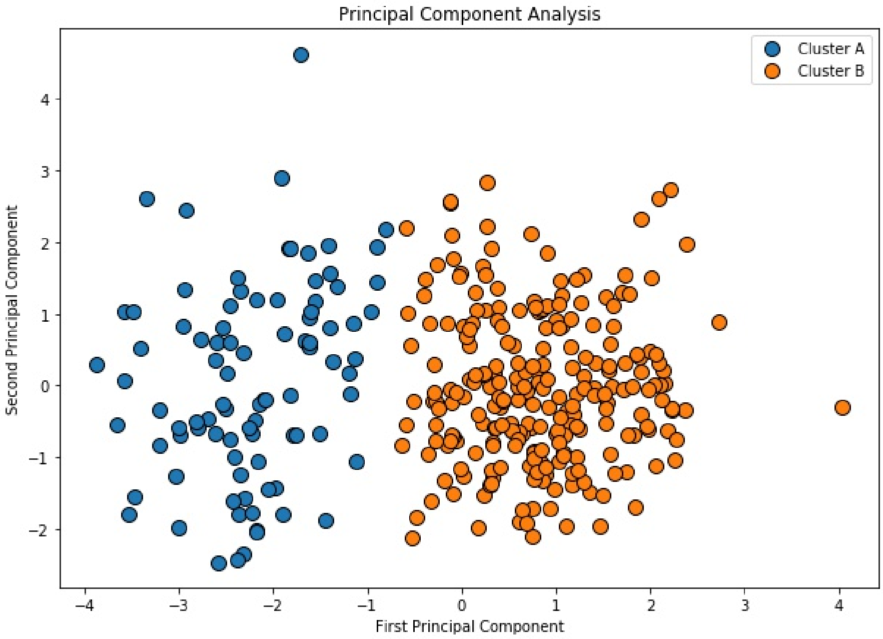


FIGURE 5.9: Principal Component Plot based on two clusters

The identified clusters (provisionally named Cluster A and Cluster B) were mapped back to the original data set and presented in the form of a map of Northern Ireland with the clusters identified (Figure 5.10).

Northern Ireland Practices by Cluster
Green = Cluster A (Metropolitan), Orange = Cluster B (Non-Metropolitan)

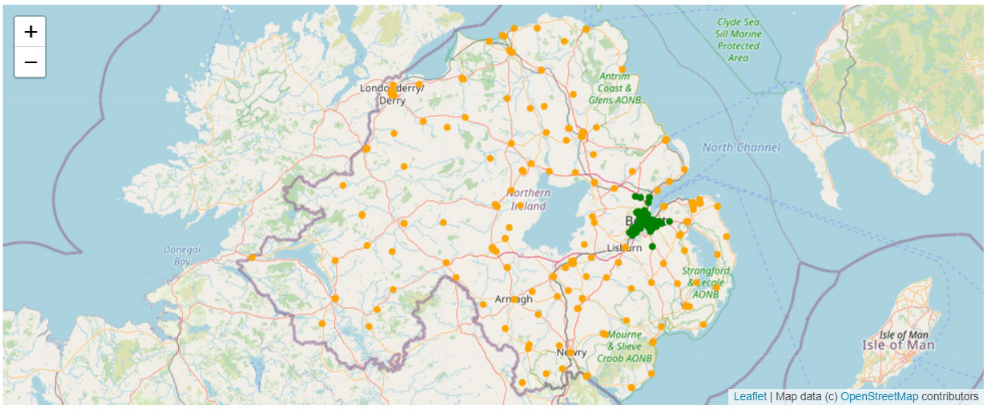

+
−
Londonderry/Derry
Antrim Coast & Glens AONB
Clyde Sea Sill Marine Protected Area
North Channel
Northern Ireland
Donegal Bay
Lisburn
Armagh
Strangford & Lecale AONB
Mourne & Slieve Croob AONB
Newry
Isle of Man
Isle of Man
Leaflet | Map data (c) OpenStreetMap contributors

FIGURE 5.10: Map of Northern Ireland showing locations of GP practices and colour coded to represent identified clusters.

### 5.3.6 Feature Statistics

**Number of Pharmacies** - Examining the statistics associated with the number of pharmacies dispensing prescriptions for practices in each cluster (Figure 5.11) it was found that there were an average of 212 pharmacies (with a standard deviation of 46.8) for practices in cluster A and an average of 98 pharmacies (with a standard deviation of 38.4) for practices in cluster B. The higher number of pharmacies associated with practices in cluster A support the categorisation of these practices as being Metropolitan whilst the lower number of pharmacies associated with practices in cluster B supports the categorisation of these practices as Non-Metropolitan. Performing T-tests on this feature revealed that the two categories were significantly different statistically ($p < 0.007$).

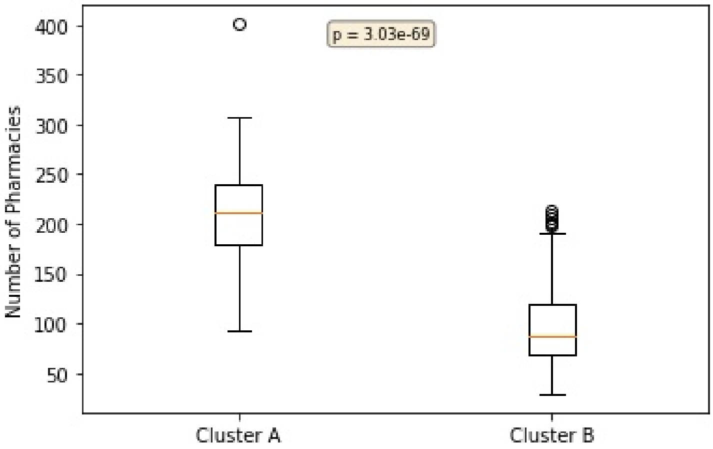


FIGURE 5.11: Number of pharmacies associated with practices by cluster.

**Number of items per registered patient** - Examining the statistics associated with the number of items prescribed per registered patient by practices in each cluster (Figure 5.12) it was found that an average of 268.8 items were prescribed per registered patient (with a standard deviation of 5.8) for practices in cluster A and an average of 246 items prescribed per registered patient (with a standard deviation of 3.4) for practices in cluster B per year. The higher number of items prescribed per registered patient associated with practices in cluster A suggests a higher prevalence of disease in cluster A and supports the categorisation of these practices as being Metropolitan whilst the lower number of items prescribed per registered patient associated with practices in cluster B suggests a lower prevalence of disease in cluster B supporting the categorisation of these practices as Non-Metropolitan. Performing T-tests on this feature revealed that the two categories were significantly different statistically (p < 0.007).

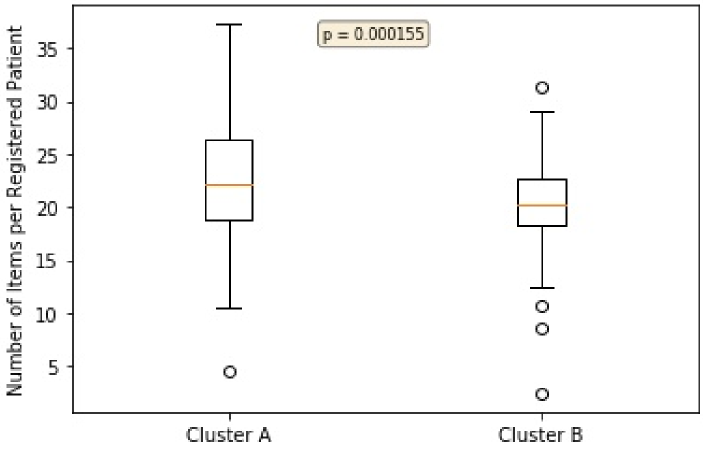


FIGURE 5.12: Number of items prescribed per registered patient associated with practices by cluster.

**Distance traveled** - Examining the statistics associated with the distance traveled by patients to dispense prescriptions for practices in each cluster (Figure 5.13) it was found that patients traveled an average of 4.6 kilometres (with a standard deviation of 1.5 kilometres) for practices in cluster A and an average of 14.8 kilometres (with a standard deviation of 5.9 kilometres) for practices in cluster B. The lower travelling distance associated with practices in cluster A support the categorisation of these practices as being Metropolitan whilst the higher travelling distance associated with practices in cluster B supports the categorisation of these practices as Non-Metropolitan. Performing T-tests on this feature revealed that the two categories were significantly different statistically (p < 0.007).

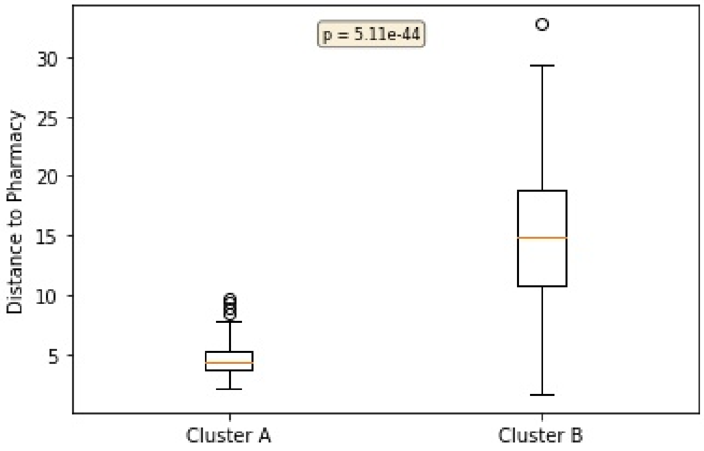


FIGURE 5.13: Distance traveled to dispense prescriptions associated with practices by cluster.

**Standard deviation of distance traveled** - Examining the statistics associated with the standard deviation of distance traveled by patients, a measure of the variation associated with the distance traveled metric, to dispense prescriptions for practices in each cluster (Figure 5.14) it was found that the average variation in distance traveled was 13.1 kilometres (with a standard deviation of 3.4 kilometres) for practices in cluster A and an average of 19.8 kilometres (with a standard deviation of 5.9 kilometres) for practices in cluster B. The lower variation in travelling distance associated with practices in cluster A support the categorisation of these practices as being Metropolitan whilst the higher variation in travelling distance associated with practices in cluster B supports the categorisation of these practices as Non-Metropolitan. Performing T-tests on this feature revealed that the two categories were significantly different statistically (p < 0.007).

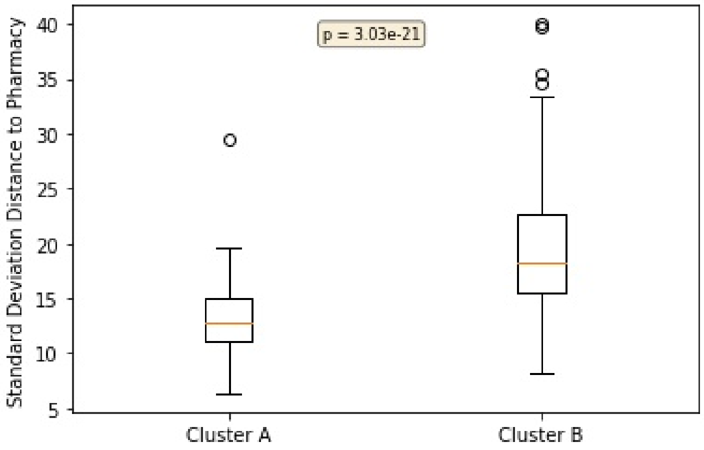


FIGURE 5.14: Standard deviation of distance traveled to dispense prescriptions associated with practices by cluster.

**Population per square kilometer** - Examining the statistics associated with the population per square kilometer of the super output area in which each practice is located (Figure 5.15), it was found that there was an average of 5,180 individuals resident per square kilometre (with a standard deviation of 2,578) for practices in cluster A and an average of 1,272 individuals resident per square kilometre (with a standard deviation of 1,230) for practices in cluster B. The higher number resident population associated with practices in cluster A support the categorisation of these practices as being Metropolitan whilst the lower number of resident population associated with practices in cluster B supports the categorisation of these practices as Non-Metropolitan. Performing T-tests on this feature revealed that the two categories were significantly different statistically (p < 0.007).

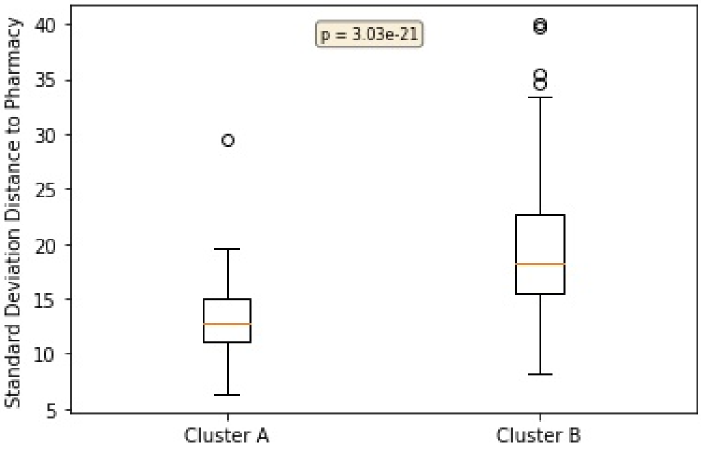


FIGURE 5.15: Population per square kilometre associated with practices by cluster.

**Registered patients** - Examining the statistics associated with the numbers of registered patients associated with practices in each cluster (Figure 5.16), it was found that there was an average of 5,645 individuals registered (with a standard deviation of 2,724) for practices in cluster A and an average of 6,030 individuals registered (with a standard deviation of 2,859) for practices in cluster B. Performing T-tests on this feature revealed that the two categories were not significantly different statistically (p=0.27) and this feature did not contribute to the overall classification of GP practices.

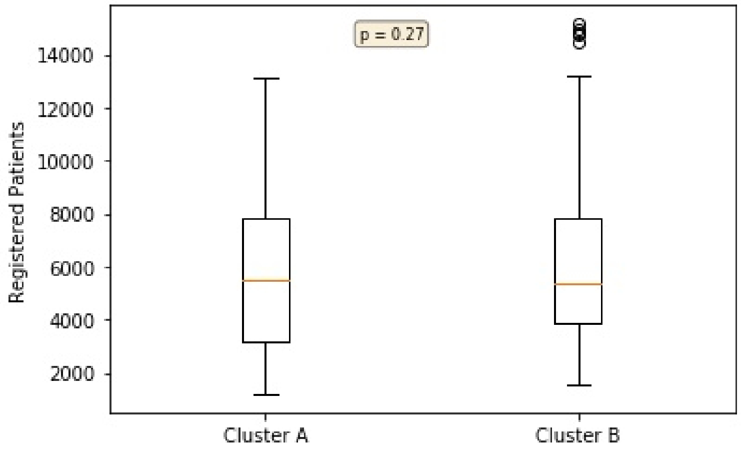


FIGURE 5.16: Number of registered patients associated with practices by cluster.

### 5.3.7 Summary of feature statistics

It is notable that the number of registered patients does not contribute significantly to the difference in the two clusters and could be ignored in future calculations. Centroid data calculated for each identified cluster (Table 5.2) was used as a basis to characterise the practice archetypes and to name each cluster. From the observations and the geographical locations of practices in cluster A, we classified this cluster with the label 'Metropolitan' given that it had a high number of pharmacies serving an area with a high population density located in the largest city in Northern Ireland. The lower number of items prescribed, and the shorter distances traveled also support this. Similarly, from the observations and the geographical locations of practices in cluster B, we originally surmised that with the longer distances being traveled and the lowest population density this cluster should be classified as Non-Metropolitan as these practices are located in both areas commonly regarded as rural and the other cities within Northern Ireland. Table 5.3 provides a breakdown of the numbers of practices in each category.

TABLE 5.2: Archetypical characteristics each cluster (i.e. Centroid Feature Values) for the period April 2018 – March 2019

| | **Archetypical Metropolitan practice (Cluster A)** | **Archetypical Non-Metropolitan practice (Cluster B)** |
|---|---|---|
| Number of Pharmacies | 212 (+-46.8) | 98 (+-38.4) |
| Number of Items per Registered Patient | 268.8 (+-5.8) | 246 (+- 3.4) |
| Distance to Pharmacy (km) | 4.6 (+-1.5) | 14.8 (+-5.9) |
| Distance Standard Deviation (km) | 13.1 (+-3.4) | 19.8 (+-5.9) |
| Population Density per Square km | 5180 (+-2578) | 1272 (+-1230) |
| Registered Patients | 5645 (+-2724) | 6030 (+-2859) |

TABLE 5.3: Breakdown of number of practices in each cluster

| | **Metropolitan practices (Cluster A)** | **Non-Metropolitan practices (Cluster B)** |
|---|---|---|
| Number of practices | 90 (27%) | 243 (73%) |

As can be seen in Table 5.2, a typical Metropolitan practice is one that normally has over 200 pharmacies associated with it, typically prescribing around 269 items per patient per year. These patients usually travel almost 5km to collect their medication but could travel up to 18km. These practices are typically located in areas of high population density with over 5,000 people per square kilometre and have around 5,600 registered patients. A typical Non-Metropolitan practice is one that normally has under 100 pharmacies associated with it, typically prescribing around 246 items per patient per year. These patients usually travel almost 15km to collect their medication but could travel up to 35km. These practices are typically located in areas of lower population density with around 1,300 people per square kilometre and have around 6,000 registered patients.

### 5.3.8 Principal Component Analysis Explained Variance Ratio

Analysing each of the features used for clustering using Principal Component Analysis, explained variance ratios were generated for each feature (Table 5.4). The results show that the feature which contributes most to the variance observed between the clusters is Number of pharmacies (0.402) with Number of Items per Registered Patient providing the second highest contribution (0.216).

TABLE 5.4: Principal Component Explained Variance Ratios

| **Feature** | **PCA Explained variance ratio** |
|---|---|
| Number of Pharmacies | 0.402 |
| Number of Items per Registered Patient | 0.216 |
| Distance to Pharmacy (km) | 0.160 |
| Distance Standard Deviation (km) | 0.118 |
| Population Density per square km | 0.056 |
| Registered Patients | 0.048 |

### 5.3.9 Outliers

**Metropolitan** - Analysis of the Metropolitan cluster shows 6 practices which are considered as outliers. Examining their locations geographically (Figure 5.17), these can be sub divided into two groups, those in the centre of the Belfast and those on the periphery of the city. Taking into account the reasons why these practices are considered to be outliers (Table 5.5), these practices can be explained.

Practice 157, located in the Queen's University area of Belfast is considered an outlier

due to three metrics - a high number of pharmacies serving it, a high variation in the distance travelled to dispense prescriptions and a low number of items per patient prescribed. Considering the location of this practice, it is likely to serve the student population living in the surrounding area. This would explain the low number of items per patient as a younger demographic are less likely to suffer multiple ailments. As they most likely reside in the area on a temporary basis, travelling home at weekends and holidays, this could also account for the variation in the distance travelled to dispense prescriptions and the high number of pharmacies serving this practice.
Practice 144, located on the Ravenhill Road in Belfast is identified as being an outlier due to high population density. This area is predominately residential and, as such, this metric is not surprising.
The remaining four practices are identified as outliers due to the high distance travelled to dispense prescriptions. As these practices are all located on the outskirts of Belfast, it is likely that patients work in the city, dispensing prescriptions during business hours accounting for the higher distance travelled.

Analysis of these outlier practices by the level of deprivation (Figure 5.18) shows no pattern with only one practice being located in an are of high deprivation (quartile 4), one in an area of Medium/High deprivation (quartile 3) and the remaining four in areas of low deprivation (quartile 1).

Analysis of these outlier practices by the size of the practice (Figure 5.19) shows that four of the six practices are considered to be large practices with 5 or more registered doctors, one as a medium sized practice with three to four registered doctors and only one practice, located near Queens University being considered as small with two registered doctors.

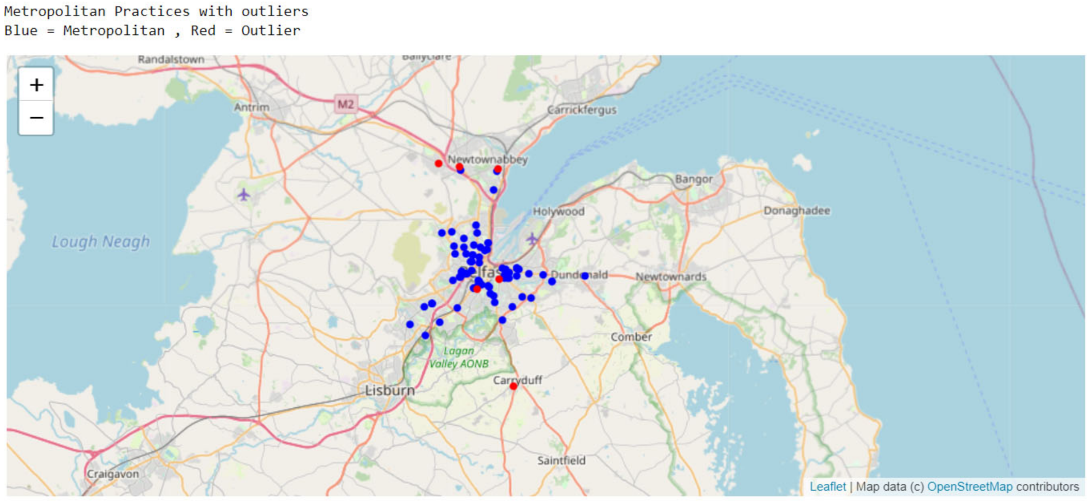


FIGURE 5.17: Map of practices considered to be outliers within the Metropolitan archetype.

TABLE 5.5: Outlier practices in the Metropolitan cluster

| Practice | Location | Pharmacies | Items | Distance | Dist. SD | Pop. | Patients |
|---|---|---|---|---|---|---|---|
| 157 | Queens, Belfast | High | Low |  | High |  |  |
| 6 | Carryduff |  |  | High |  |  |  |
| 336 | Newtownabbey |  |  | High |  |  |  |
| 337 | Newtownabbey |  |  | High |  |  |  |
| 440 | Newtownabbey |  |  | High |  |  |  |
| 144 | Revenhill Rd, Belfast |  |  |  |  | High |  |

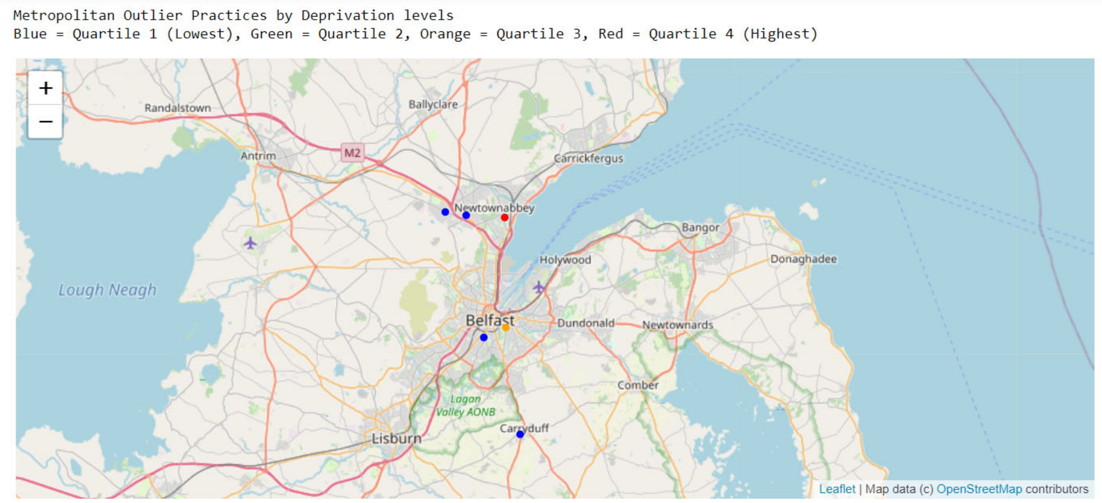

Metropolitan Outlier Practices by Deprivation levels
Blue = Quartile 1 (Lowest), Green = Quartile 2, Orange = Quartile 3, Red = Quartile 4 (Highest)
+
−
Randalstown
Ballyclare
Antrim
M2
Carrickfergus
Newtownabbey
Bangor
Holywood
Donaghadee
Lough Neagh
Belfast
Dundonald
Newtownards
Comber
Lagan Valley AONB
Carryduff
Lisburn
Leaflet | Map data (c) OpenStreetMap contributors

FIGURE 5.18: Map of Metropolitan outliers by deprivation level of the super output area in which they are located.

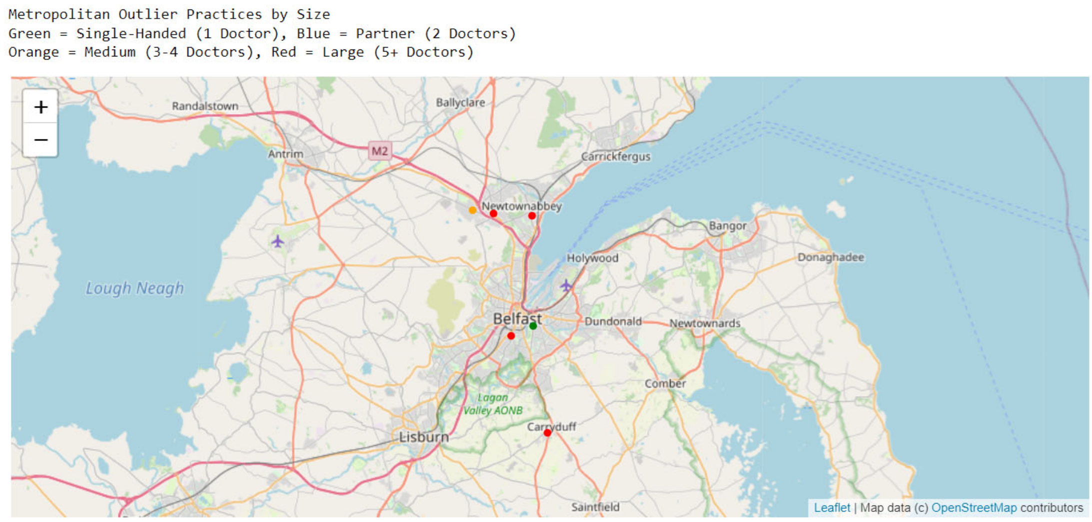

Metropolitan Outlier Practices by Size
Green = Single-Handed (1 Doctor), Blue = Partner (2 Doctors)
Orange = Medium (3-4 Doctors), Red = Large (5+ Doctors)
+
−
Randalstown
Ballyclare
Antrim
M2
Carrickfergus
Newtownabbey
Bangor
Holywood
Donaghadee
Lough Neagh
Belfast
Dundonald
Newtownards
Comber
Lagan Valley AONB
Carryduff
Lisburn
Saintfield
Leaflet | Map data (c) OpenStreetMap contributors

FIGURE 5.19: Map of Metropolitan outliers by practice size.

**Non-Metropolitan** - Analysis of the Non-Metropolitan cluster shows 23 practices which are considered as outliers. Examining their locations geographically (Figure 5.20), it can be seen that a large proportion of these practices are located close to the Metropolitan area. Three practices (267, 517 and 541), identified as outliers due to the low number of items dispensed can be discounted as they ceased operating during the period. The remaining practices are considered outliers due to one of the feature metrics being considered as high with only one practice located in Enniskillen having both high distance travelled to dispense prescriptions and a high variation in this distance.

Analysis of these outlier practices by the level of deprivation (Figure 5.21) shows no distinct pattern with the majority being located in an area of low deprivation (quartile 1). It is notable that outlier practices in the border areas of counties Londonderry and Fermanagh are generally located in areas of medium/High (quartile 3) or High (quarile 4) areas of deprivation.

Analysis of these outlier practices by the size of the practice (Figure 5.22) shows that the majority of these practices are considered to be large practices with 5 or more registered doctors.

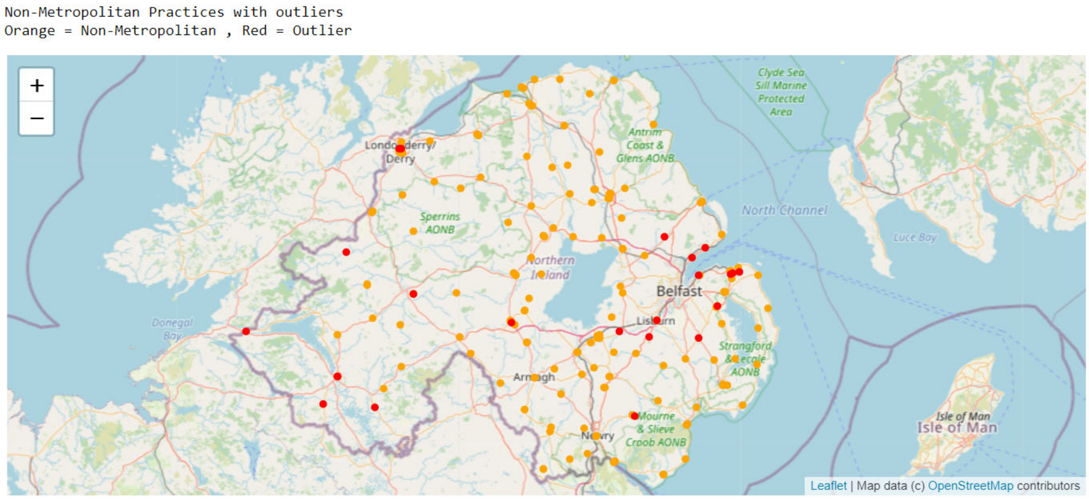


FIGURE 5.20: Map of practices considered to be outliers within the Non-Metropolitan archetype.

TABLE 5.6: Outlier practices in the Non-Metropolitan cluster

| Practice | Location | Pharmacies | Items | Distance | Dist. SD | Pop. | Patients |
|---|---|---|---|---|---|---|---|
| 198 | Saintfield | High | | | | | |
| 221 | Ballynahinch | High | | | | | |
| 227 | Lisburn | High | | | | | |
| 274 | Lisburn | High | | | | | |
| 385 | Greenisland | High | | | | | |
| 473 | Craigavon | High | | | | | |
| 267 | Comber | | Low | | | | |
| 517 | Dungannon | | Low | | | | |
| 541 | Rathfrisland | | Low | | | | |
| 663 | Omagh | | High | | | | |
| 576 | Enniskillen | | | High | High | | |
| 563 | Enniskillen | | | | High | | |
| 564 | Enniskillen | | | | High | | |
| 585 | Florencecourt | | | | High | | |
| 264 | Bangor | | | | | High | |
| 281 | Bangor | | | | | High | |
| 604 | Derry/Londonderry | | | | | High | |
| 615 | Derry/Londonderry | | | | | High | |
| 252 | Holywood | | | | | | High |
| 390 | Carrickfergus | | | | | | High |
| 433 | Ballyclare | | | | | | High |
| 574 | Enniskillen | | | | | | High |
| 616 | Castlederg | | | | | | High |

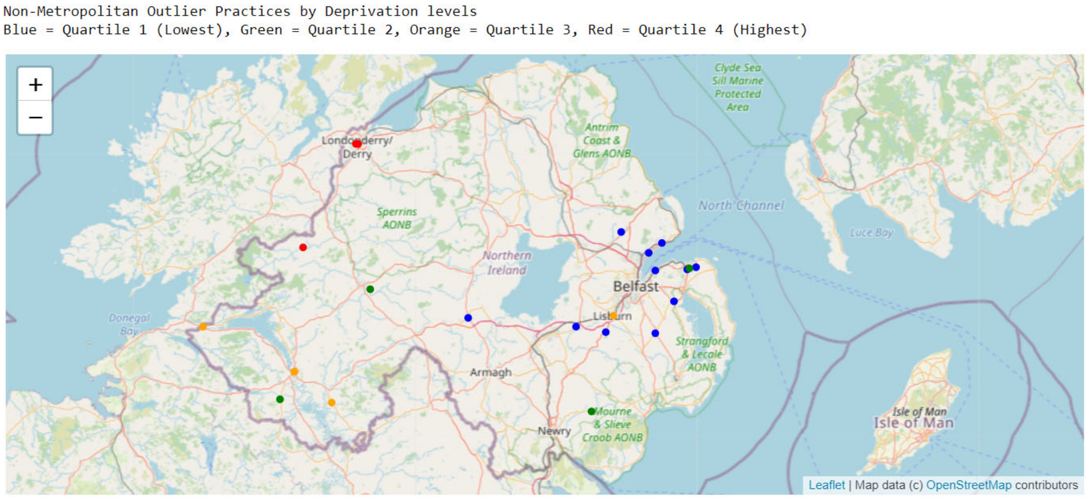
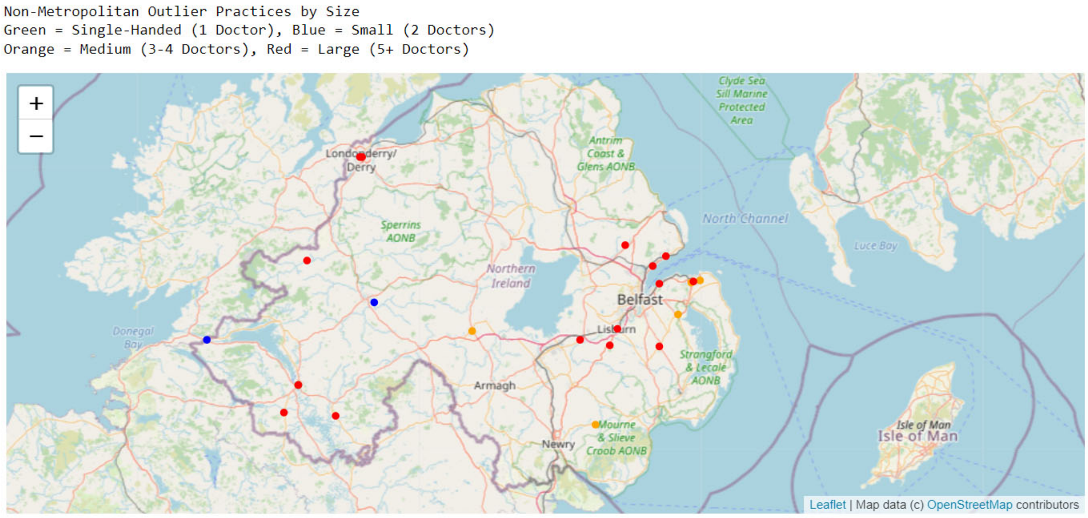

Non-Metropolitan Outlier Practices by Deprivation levels
Blue = Quartile 1 (Lowest), Green = Quartile 2, Orange = Quartile 3, Red = Quartile 4 (Highest)

FIGURE 5.21: Map of Non-Metropolitan outliers by deprivation level of the super output area in which they are located.



Non-Metropolitan Outlier Practices by Size
Green = Single-Handed (1 Doctor), Blue = Small (2 Doctors)
Orange = Medium (3-4 Doctors), Red = Large (5+ Doctors)

FIGURE 5.22: Map of Non-Metropolitan outliers by practice size.

### 5.3.10 Dispersion of pharmacies

Calculating the average distance between pharmacies dispensing prescriptions for each practice produced an average dispersion distance measured in kilometres (Figure 5.23). These figures supported the cluster labelling with Metropolitan pharmacies on average 26.2km apart with a standard deviation of 6.3km and Non-Metropolitan pharmacies on average 40.4km apart with a standard deviation of 9.6km. Performing T-tests revealed that the two categories were significantly different statistically (p < 0.007).

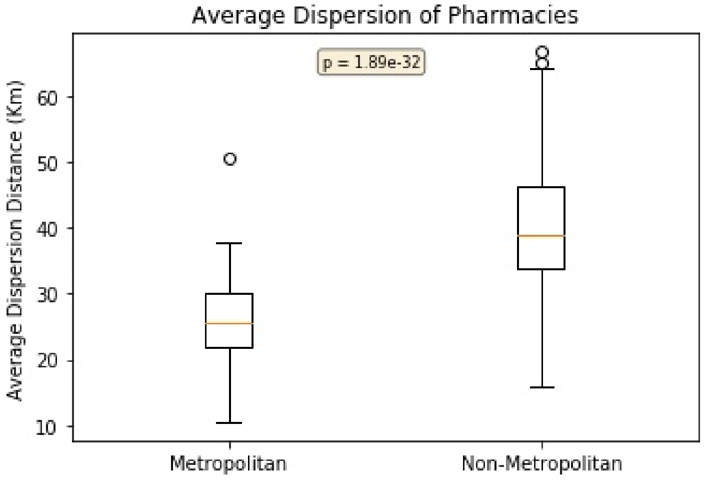


FIGURE 5.23: Dispersion of pharmacies associated with practices by cluster.

## 5.4 Discussion

The aim of this study was to discover what types of GP practice exist in Northern Ireland based not only on their location but on the influence they have on their surrounding areas based on their prescribing behaviours. Traditionally, the categories Urban and Rural have been used in medical circles but a formal means of identifying which practices belonged to which category did not exist previously other than their geographical location. Rural and Urban have been used as binary indicators in previous research (Zhang and Wang, 2018) with Semi-Rural being loosely defined as "a small town, and the surrounding Rural area" (Hogg et al., 2013) but no other justification for this classification is given. Whilst these categories may hold true, initial investigations in Chapter 4 suggest that GP practices have a wider influence in the community through their prescribing and subsequent relationship with the pharmacies dispensing these prescriptions.

In identifying the features most representative of what a GP practice is and its influence in the wider community it was felt that location of both the GP practice itself and that of the patients receiving prescriptions should be considered. As open data is not available on patients or their home addresses, the distance travelled to dispense prescriptions was used as a proxy for the distance patients live from their surgery. This is almost certainly not true in all cases as patients may attend the pharmacy closest to their surgery or their work location instead. That being said, figures show that almost 60% of prescriptions are dispensed at pharmacies located over five kilometres from the issuing GP practice. Consideration was made for the distribution of distances travelled which showed that the distribution was not Gaussian but instead skewed towards the lower distances travelled. Taking the average distance would therefore be biased towards these lower distances so the decision was taken to incorporate two distance features: Median Distance and Standard Deviation of Distance. The median distance would provide an indication of sphere of influence each GP practice had on the surrounding community whilst the standard deviation of this distance gives an indication of how this distance may vary. Having theorised that distance is an important feature in categorising GP practices, it was important to remember that GP practices are generally located in areas to serve the local community.

By applying k-means clustering to six identifying features of a GP practice, it has been discovered that two main types of GP practice exist in Northern Ireland. These have been labelled as Metropolitan (located around the city of Belfast) and Non-Metropolitan (all other areas). Interestingly, practices located in other cites in NI (e.g. Derry/Londonderry, Armagh etc) which would traditionally be classed as Urban do not fall within the Metropolitan archetype. Whilst no evidence has been found to support the semi-rural classification traditionally used, it must be noted that it is most likely that sub clusters exist within these two main archetypes as indicated in the dendrogram produced (Figure 5.6). Further evidence of sub clusters can

be seen in the analysis of outliers with four practices exhibiting high distances travelled in the Metropolitan archetype and 20 practices within the Non-Metropolitan archetype showing high metrics in at least one of the clustering features.
The level of deprivation attributed to the location of each practice and the size of the practice in terms of the number of registered doctors were not originally considered as features of the practice. The subsequent analysis of outliers has shown that the majority of outliers in both archetypes are large practices with five or more registered doctors which may indicate that practice size has an effect on the prescribing behaviours of a practice. Outlier practices in the border regions of counties Londonderry and Fermanagh have been identified as being located in areas of high deprivation, and whilst there is no indication in either set of outliers that the level of deprivation is a major factor, further analysis of the effect of derivation levels should be carried out.

Analysis of the contribution each feature provided to the variance observed between each cluster showed that Number of Pharmacies and Number of Items per Registered Patient provided the highest contributions. Distance travelled provided the third highest contribution with Population Density per square kilometre being one from last. This reinforced the hypothesis that categorising GP surgeries based solely on their location is not an appropriate method of categorisation. Interestingly, the Distance travelled by patients to dispense their prescriptions only has a 16% contribution to the variance observed although this will contribute to the number of pharmacies associated with the GP practice.

The findings are significant in that they provide clear indications of how GP surgeries should be categorised and what the contributing features are. Refining this method of categorisation with further research into any sub-clusters which may exist within each archetype will provide a basis for like-for-like comparisons of individual GP practices with the potential to identify anomalous prescribing activities.

## 5.5 Limitations

This study has sought to categorise and track GP practices operating in Northern Ireland over time. The data set on which categorisation was performed only became available from April 2018 limiting the available data to a 1-year period (April 2018 - March 2019). In addition, only GP practices that operated during the whole period were included (333 practices) with no provision made for practices which closed or those opening during the period. The analysis is based on tracking the location of the GP Practice issuing a prescription to the location of the pharmacy dispensing it. Whilst this study uses number of items per registered patient as a proxy for the levels of sickness experienced, this may not be accurate as some GPs may be over prescribing or prescribing where anther GP would ask the patient to buy over the counter (e.g. paracetamol). It is likely that the majority of registered patients do not reside in the same super output area as the practice they attend. As this does not necessarily reflect the actual residential location of the patient receiving the prescription, it is assumed that patients will dispense their prescriptions at their local pharmacy meaning that distance traveled can be used as a proxy. Similarly, the population density used as a feature in clustering practices is the population density of the super output area in which the Practice is located.

## 5.6 Conclusion

This chapter has set out to investigate the relationship between the geographic location of GP practices (to categorise practices based on their geographic location) and their prescription profile. To do this we have linked each practice with their associated dispensing pharmacies in Northern Ireland to ascertain whether this has any effect on prescribing patterns. In doing so, we have presented a methodology to compute archetypes based on areas of interest (in this case geolocation attributes) for subsequent comparisons to determine if the archetypes differ in behaviour. It was found that it was possible to classify GP practices based on geolocation attributes and two different archetypes of GP practice were identified: Metropolitan and Non-Metropolitan (the labels urban and rural were not appropriate and no evidence was found for the semi-rural category commonly used by healthcare researchers in Northern Ireland). Evidence shows that there are likely sub clusters within both archetypes. Average dispersion distances were calculated for each set of pharmacies dispensing prescriptions for each practice, the results supporting the two categorisations. Analysis of outliers within both archetypes indicate that practice size is likely to effect the prescribing behaviours of a practice, and, although deprivation level does not seem to contribute to the identification of outliers, it may also contribute to prescribing behaviours. In Chapter 6 both these factors will be explored using time series analysis of the prescribing trends of similar types of practice within each archetype.

# Chapter 6

# Analysis of the prescriptive behaviours of GP practices

> "Courage is knowing it might hurt and doing it anyway. Stupidity is the same. And that's why life is hard."
>
> Jeremy Goldberg

In Chapter 5 a methodology for categorising GP practices based on their geographical location and relationship with dispensing pharmacies was proposed. This method identified two distinct archetypes of GP practice operating in Northern Ireland - Metropolitan and Non-Metropolitan. This chapter will compare the prescriptive behaviours of these archetypes.

The work presented in this chapter was published in the Scientific Reports journal as a paper entitled "Discovering and Comparing Types of General Practitioner Practices Using Geolocational Features and Prescribing Behaviours by Means of K-Means Clustering" (Booth et al., 2021b) and modified here to fit within the framework and context of this thesis.

The objectives of this chapter are to:

- Investigate prescribing trends associated with Metropolitan and Non-Metropolitan practices over time to identify differences in prescribing trends.
- Investigate levels of prescribing associated with Metropolitan and Non-Metropolitan practices over time to identify differences in prescribing levels.

## 6.1 Methods

The total number of items prescribed per patient were analysed by cluster for the period July 2015 to December 2019 at both NI and BNF chapter levels. BNF chapters not conforming to the overall trend observed at national level were investigated in order to identify the types of medication contributing to any differences. In considering the differences in prescribing levels between the two clusters, Root Mean

Square Error (RMSE) was calculated for each BNF chapter to establish the variation in prescribing between the two clusters. Those chapters showing the highest variation were investigated further to establish which types of medication contributed to the variations.

## 6.2 Results

Comparing prescribing trends and levels of Metropolitan and Non-Metropolitan practices over this period showed that both archetypes displayed similar trends although Metropolitan practices had consistently higher levels of prescribing (Figure 6.1).

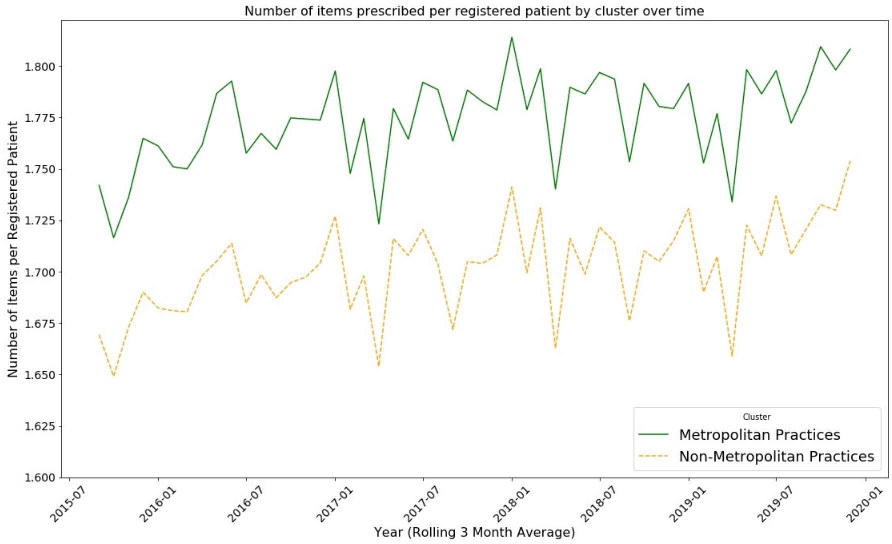


FIGURE 6.1: Total items prescribed per patient by archetype

Further investigation at BNF chapter level showed that prescribing trends at this level were also similar with no chapters showing any significant deviation between archetypes (Appendix F).

Examining prescribing levels, it was found that although Metropolitan prescribing was higher nationally, this only held true in approximately half of the BNF chapters. Prescribing levels for Non-Metropolitan practices were found to be marginally higher in BNF chapters 2 (Cardiovascular System), 6 (Endochrine System), 7 (Obstetrics), 8 (Malignant Disease & Immunosuppression), 11 (Eye), 12 (Ear, Nose and Oropharynx), 20 (Dressings), 22 (Incontinence Appliances) and 23 (Stoma Appliances)

In order to gain insight into which types of medication contributed to the differing prescribing levels observed nationally, Root Mean Square Error (RMSE) was calculated between the clusters (Figure 6.2). This showed that the largest variation between clusters occurred in chapter 4 (Central Nervous System) with smaller variations evident in chapters 3 (Respiratory System), 2 (Cardiovascular System), 6 (Endocrine System) and 13 (Skin). Notably, the chapter identified as contributing most to the variation observed between archetypes has higher prescribing levels in Metropolitan practices.

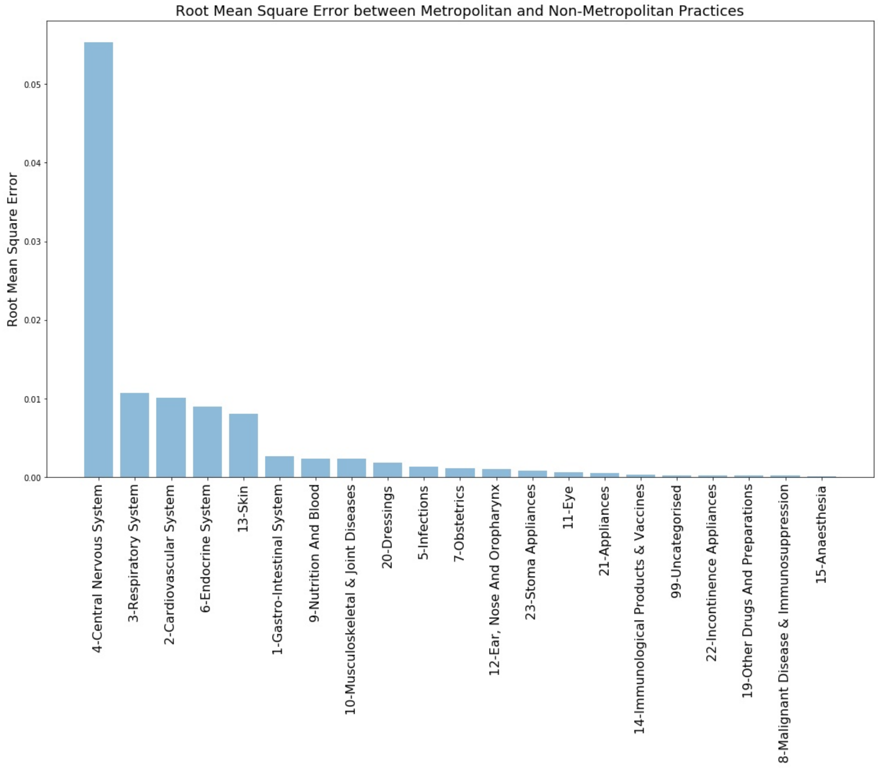
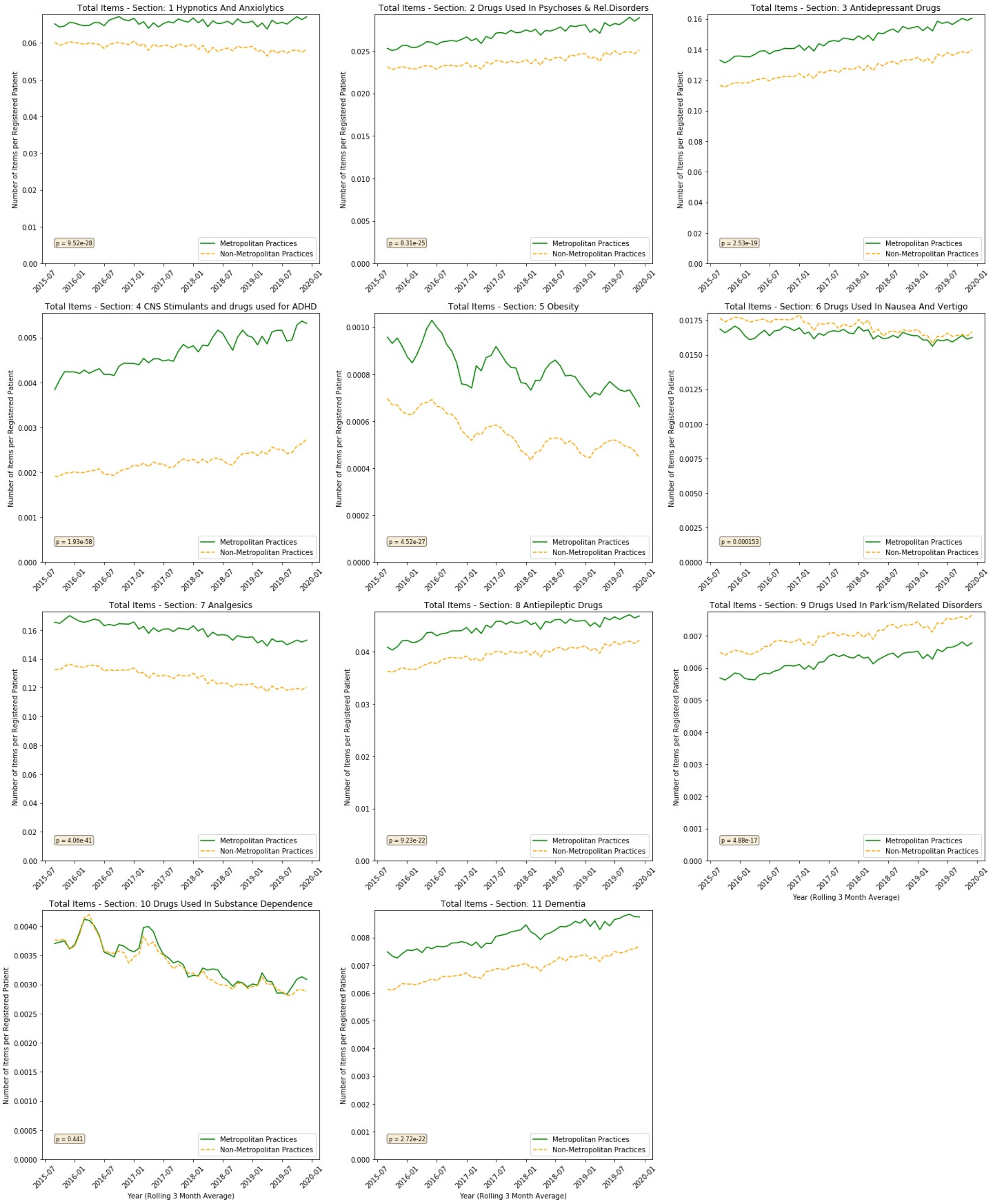


FIGURE 6.2: Root mean square error between Metropolitan and Non-Metropolitan practices

The following sections drilling deeper into the each of the BNF chapters and their contributions to variations between archetypes.

### 6.2.1 Chapter 4 - Central Nervous System

The prescribing trends for sections in BNF chapter 4 (Central Nervous System) were graphed for comparison (Figure 6.3) with independent t-tests being applied to indicate whether the archetypes were significantly different statistically (Table 6.1). In order to account for type 1 errors which could occur when performing multiple comparisons a revised Bonferroni alpha value was calculated where Bonferroni alpha equals standard alpha divided by the number of comparisons (0.05/11 = 0.005).



Total Items - Section: 1 Hypnotics And Anxiolytics

Total Items - Section: 2 Drugs Used In Psychoses & Rel.Disorders

Total Items - Section: 3 Antidepressant Drugs

Total Items - Section: 4 CNS Stimulants and drugs used for ADHD

Total Items - Section: 5 Obesity

Total Items - Section: 6 Drugs Used In Nausea And Vertigo

Total Items - Section: 7 Analgesics

Total Items - Section: 8 Antiepileptic Drugs

Total Items - Section: 9 Drugs Used In Park'ism/Related Disorders

Total Items - Section: 10 Drugs Used In Substance Dependence

Total Items - Section: 11 Dementia

FIGURE 6.3: Items per registered patient for BNF Chapter 4 (Central Nervous System) sections

TABLE 6.1: Summary of independent t-tests indicating statistical significance of differences observed in Metropolitan and Non-Metropolitan practices for BNF chapter 4 (Central Nervous System)

| **Section** | **p-value** | **Statistically significant ($p<0.005$)** |
|---|---|---|
| 1 - Hypnotics and Anxiolytics | $9.52x10^{-28}$ | Yes |
| 2 - Drugs used in Psychoses & Rel.Disorders | $8.31x10^{-25}$ | Yes |
| 3 - Antidepressant drugs | $2.53x10^{-19}$ | Yes |
| 4 - CNS Stimulants and drugs used for ADHD | $1.93x10^{-58}$ | Yes |
| 5 - Obesity | $4.52x10^{-27}$ | Yes |
| 6 - Drugs used in Nausea and Vertigo | 0.000153 | Yes |
| 7 - Analgesics | $4.06x10^{-41}$ | Yes |
| 8 - Antiepileptic drugs | $9.23x10^{-22}$ | Yes |
| 9 - Drugs used in Parkinsonism / related disorders | $4.88x10^{-17}$ | Yes |
| 10 - Drugs used in substance dependence | 0.441225 | No |
| 11 - Dementia | $2.72x10^{-22}$ | Yes |

Having the highest contribution to the overall variations seen at BNF chapter level, chapter 4 sections show higher prescribing levels in nine of the eleven sections for practices in Metropolitan areas. Only in sections 6 (Drugs used in Nausea and Vertigo) and 9 (Drugs used in Parkinsonism/Related Disorders) were prescribing levels higher in Non-Metropolitan practices.

The main contributors to the variation seen in this chapter come primarily from section 7 (Analgesics) and secondly from 3 (Antidepressant Drugs) (Figure 6.4) showing that prescribing in these two sections are considerably higher in Metropolitan practices than in Non-Metropolitan practices.

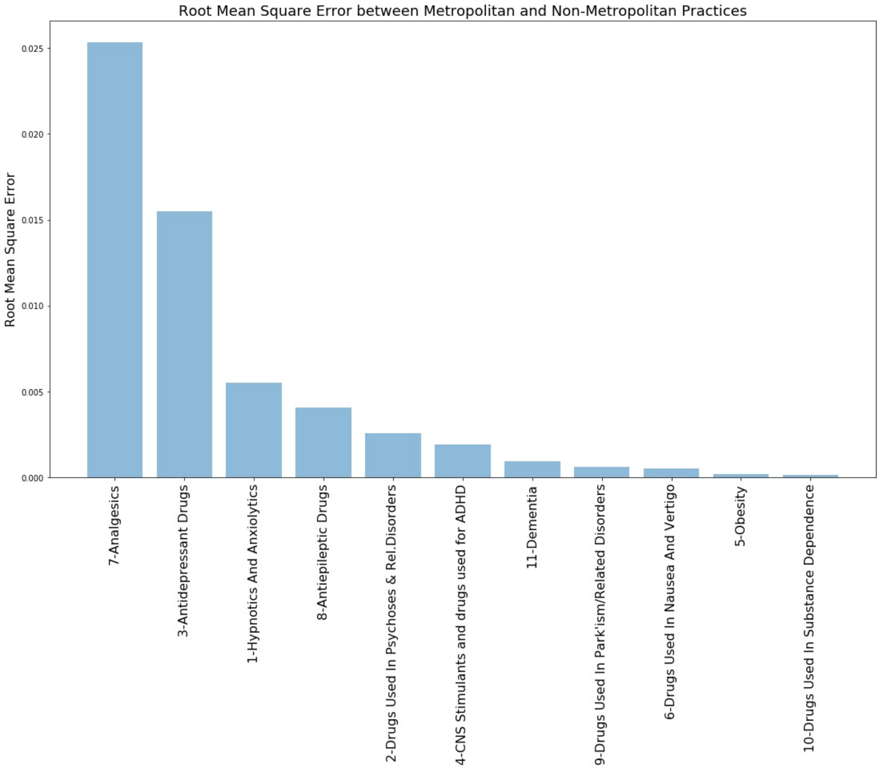


FIGURE 6.4: Comparison of RMSE by cluster at BNF section level (Chapter 4)

### 6.2.2 Chapter 3 - Respiratory System

Prescribing trends for sections in BNF chapter 3 (Respiratory System) were graphed for comparison (Figure 6.5) with independent t-tests also being applied (Table 6.2) using a revised Bonferroni alpha value of 0.005 (0.05/9).

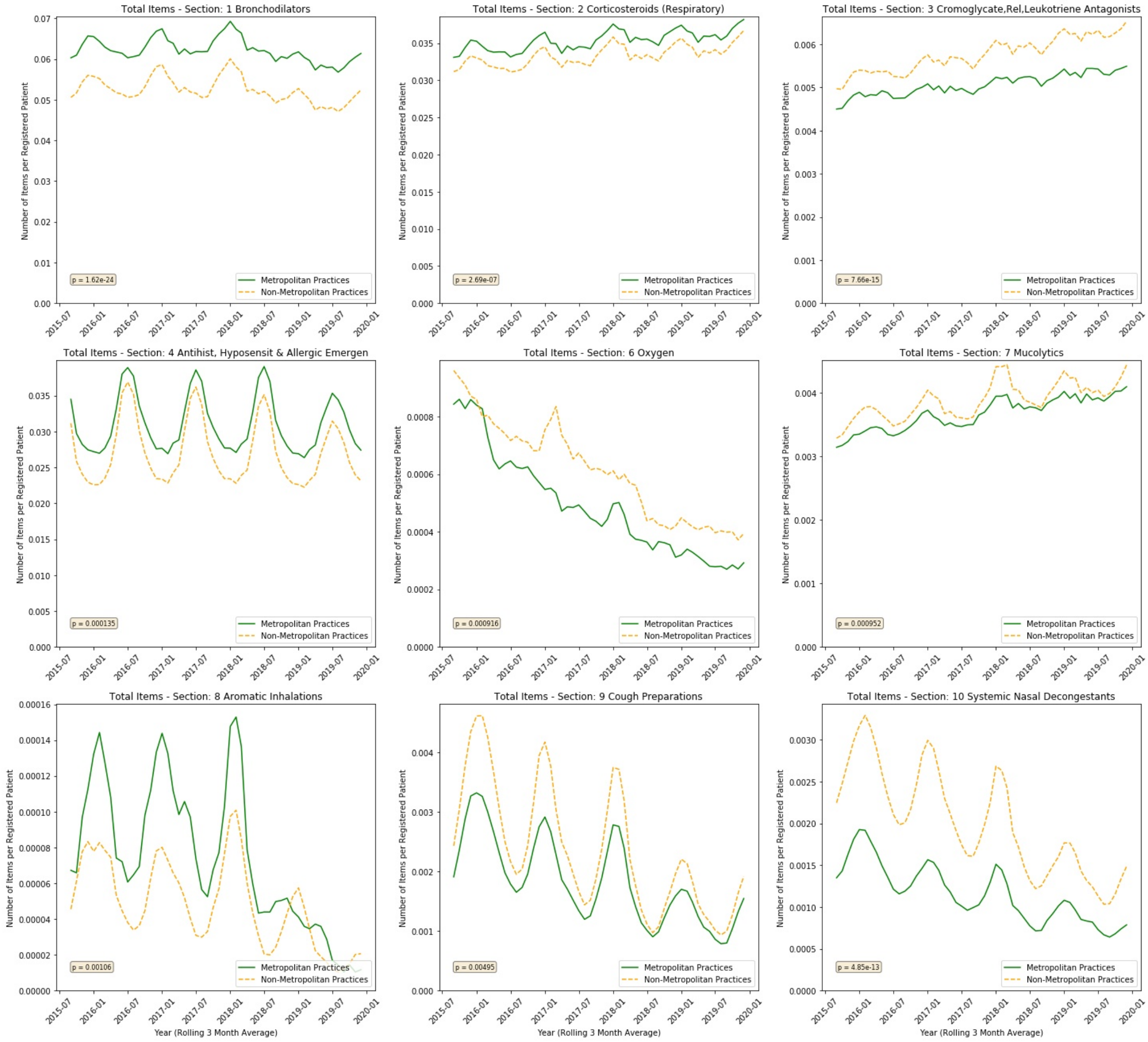


FIGURE 6.5: Items per registered patient for BNF Chapter 3 Respiratory System sections

TABLE 6.2: Summary of independent t-tests indicating statistical significance of differences observed in Metropolitan and Non-Metropolitan practices for BNF chapter 3 (Respiratory System)

| Section | p-value | Statistically significant ($p<0.005$) |
|---|---|---|
| 1 - Bronchodilators | $1.62 \times 10^{-24}$ | Yes |
| 2 - Corticosteroids (Respiratory) | $2.69 \times 10^{-07}$ | Yes |
| 3 - Cromoglycate, leukotriene and phosphodesterase type-4 inhibitors | $7.66 \times 10^{-15}$ | Yes |
| 4 - Antihistamines, hyposensitisation and allergic emergencies | 0.000135 | Yes |
| 6 - Oxygen | 0.000916 | Yes |
| 7 - Mucolytics | 0.000952 | Yes |
| 8 - Aromatic inhalations | 0.001059 | Yes |
| 9 - Cough preparations | 0.004953 | Yes |
| 10 - Systemic nasal decongestants | $4.85 \times 10^{-13}$ | Yes |

The overall variations seen at BNF chapter level, chapter 3 sections show higher prescribing levels in only four of the nine sections (where figures were reported) for practices in Metropolitan areas: 1 (Bronchodilators), 2 (Corticosteroids (Respiratory)), 4 (Antihistamines, hyposensitisation and allergic emergencies) and 8 (Aromatic Inhalations). Performing independent t-tests on the sections (Table 6.2) revealed that statistically significant differences existed in all BNF sections between prescribing levels observed for Metropolitan and Non-Metropolitan practices.

The main contributor to the variation seen in this chapter is section 1 (Bronchodilators) and with lesser contributions made by section 4 (Antihistamines, hyposensitisation and allergic emergencies) and 2 (Corticosteroids (Respiratory)) (Figure 6.6) showing that prescribing in these sections are considerably higher in Metropolitan practices than in Non-Metropolitan practices.

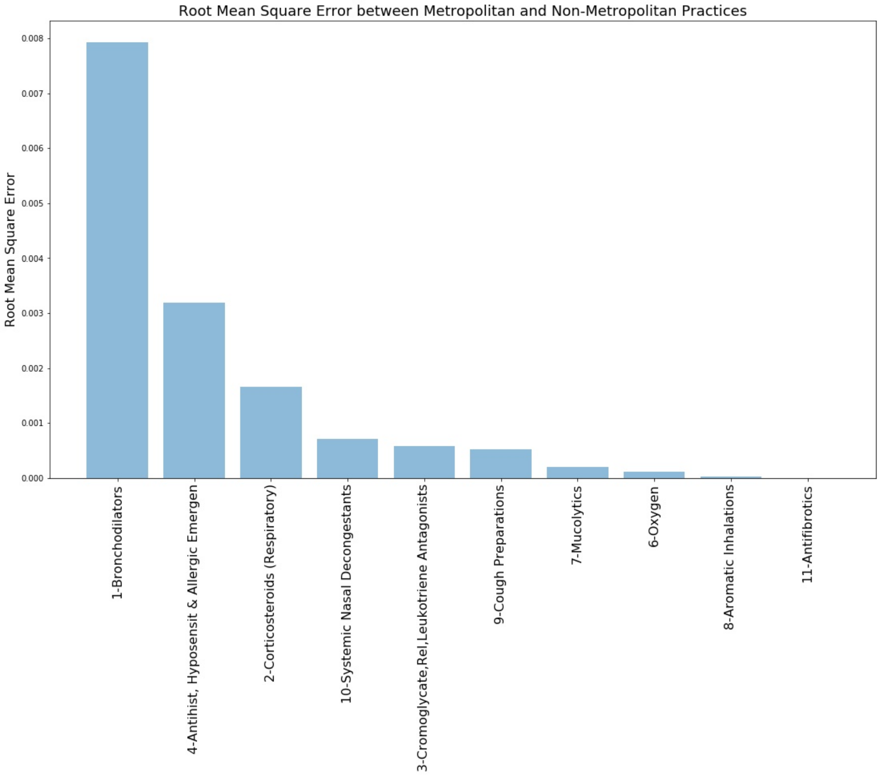


FIGURE 6.6: Comparison of RMSE by cluster at BNF section level (Chapter 3)

### 6.2.3 Chapter 2 - Cardiovascular System

Prescribing trends for sections in BNF chapter 2 (Cardiovascular System) were graphed for comparison (Figure 6.7) with independent t-tests also being applied (Table 6.3) using a revised Bonferroni alpha value of 0.005 (0.05/11).

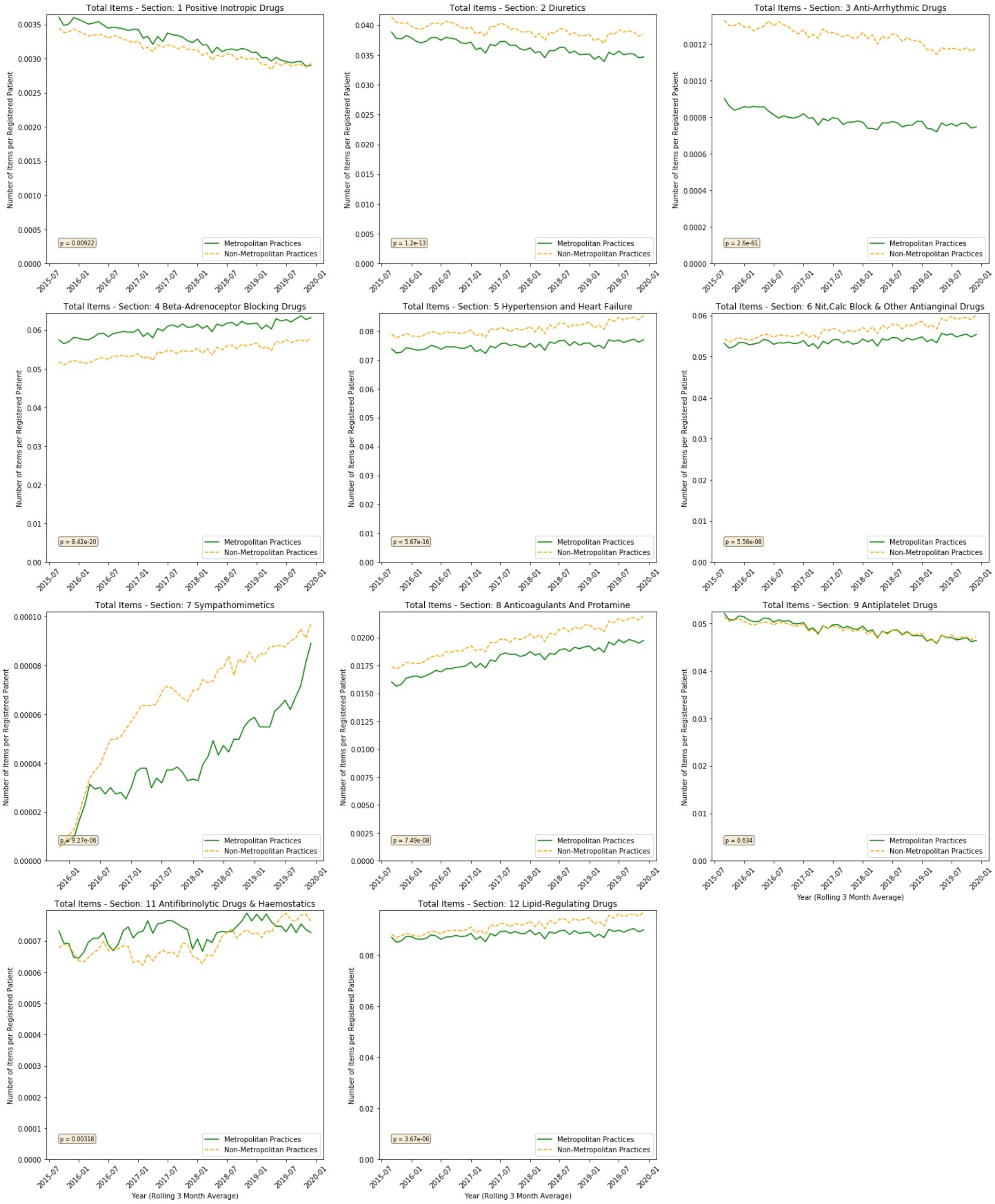


FIGURE 6.7: Items per registered patient for BNF Chapter 2 Cardiovascular System sections

TABLE 6.3: Summary of independent t-tests indicating statistical significance of differences observed in Metropolitan and Non-Metropolitan practices for BNF chapter 2 (Cardiovascular System)

| **Section** | **p-value** | **Statistically significant (p<0.005)** |
|---|---|---|
| 1 - Positive inotropic drugs | 0.00922 | No |
| 2 - Diuretics | $1.20\text{x}10^{-13}$ | Yes |
| 3 - Anti-arrhythmic drugs | $2.60\text{x}10^{-61}$ | Yes |
| 4 - Beta-adrenoceptor blocking drugs | $8.42\text{x}10^{-20}$ | Yes |
| 5 - Hypertension and heart failure | $5.67\text{x}10^{-16}$ | Yes |
| 6 - Nitrates, calcium-channel blockers & other antianginal drugs | $5.56\text{x}10^{-08}$ | Yes |
| 7 - Sympathomimetics | $9.27\text{x}10^{-06}$ | Yes |
| 8 - Anticoagulants and Protamine | $7.49\text{x}10^{-08}$ | Yes |
| 9 - Antiplatelet drugs | 0.634309 | No |
| 11 - Antifibrinolytic drugs & Haemostatics | 0.003177 | Yes |
| 12 - Lipid-regulating drugs | $3.67\text{x}10^{-06}$ | Yes |

The overall variations seen at BNF chapter level, chapter 2 sections show higher prescribing levels in only three of the eleven sections (where figures were reported) for practices in Metropolitan areas: 1 (Positive inotropic drugs), 4 (Beta-adrenoceptor blocking drugs) and 11 (Antifibrinolytic drugs & Haemostatics). Performing independent t-tests on the sections (Table 6.3) revealed that statistically significant differences existed in all BNF sections between prescribing levels observed for Metropolitan and Non-Metropolitan practices with the exception of Section 9 (Antiplatelet drugs).

Although the differences observed were small compared with those seen in chapter 4, the main contributors to the variation seen in this chapter were sections 5 (Hypertension and heart failure) and 4 (Beta-adrenoceptor blocking drugs) with lesser contributions made by sections 12 (Lipid-regulating drugs), 2 (Diuretics) and 6 (Nitrates, calcium-channel blockers & other antianginal drugs) (Figure 6.8).

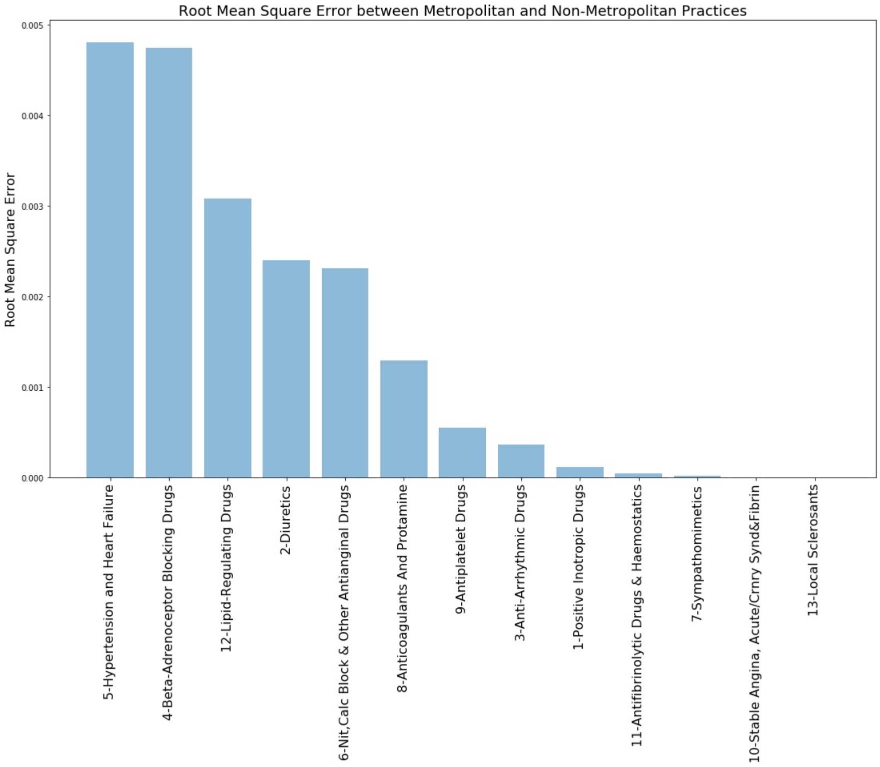


FIGURE 6.8: Comparison of RMSE by cluster at BNF section level (Chapter 2)

### 6.2.4 Chapter 6 - Endocrine System

Prescribing trends for sections in BNF chapter 6 (Endocrine System) were graphed for comparison (Figure 6.9) with independent t-tests also being applied (Table 6.4) using a revised Bonferroni alpha value of 0.007 (0.05/7).

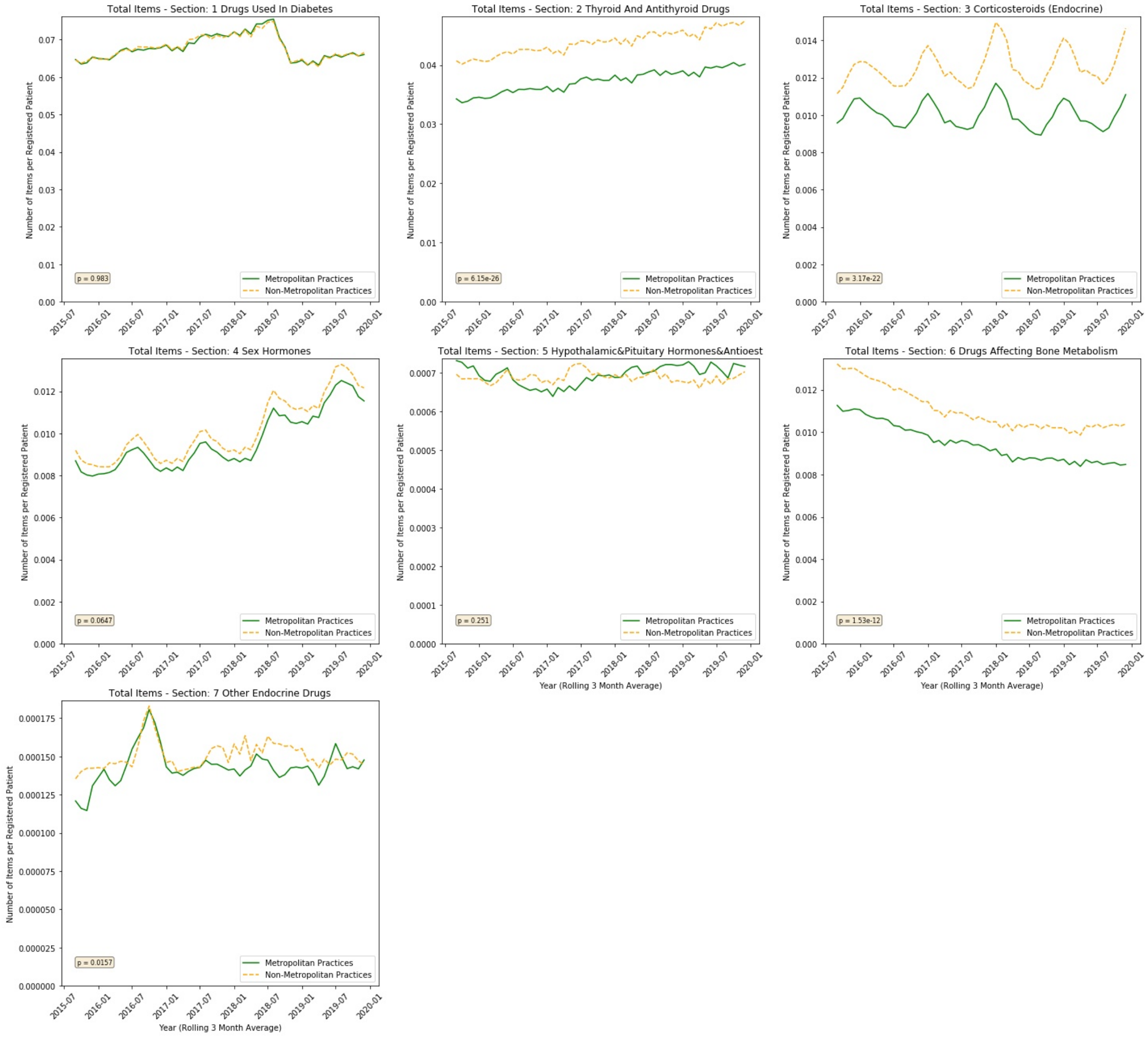


FIGURE 6.9: Items per registered patient for BNF Chapter 6 Endocrine System sections

TABLE 6.4: Summary of independent t-tests indicating statistical significance of differences observed in Metropolitan and Non-Metropolitan practices for BNF chapter 6 (Endocrine System)

| Section | p-value | Statistically significant (p<0.007) |
|---|---|---|
| 1 - Drugs used in Diabetes | 0.982554 | No |
| 2 - Thyroid and Antithyroid drugs | 6.15x10$^{-26}$ | Yes |
| 3 - Corticosteroids (Endocrine) | 3.17x10$^{-22}$ | Yes |
| 4 - Sex hormones | 0.064655 | No |
| 5 - Hypothalamic and pituitary hormones and anti oestrogens | 0.251081 | No |
| 6 - Drugs affecting bone metabolism | 1.53x10$^{-12}$ | Yes |
| 7 - Other endocrine drugs | 0.01567 | No |

The overall variations seen at BNF chapter level, chapter 6 sections show higher prescribing levels in three of the seven sections for practices in Non-Metropolitan areas: 2 (Thyroid and Antithyroid drugs), 3 (Corticosteroids (Endocrine)) and 6 (Drugs affecting bone metabolism), all of which proved to have significantly different prescribing levels from those of Metropolitan practices (Table 6.4). Prescribing levels and trends were observed to be similar for both archetypes in the remaining four sections.

The main contributor to the variation seen in this chapter was sections 2 (Thyroid and Antithyroid drugs) (Figure 6.10).

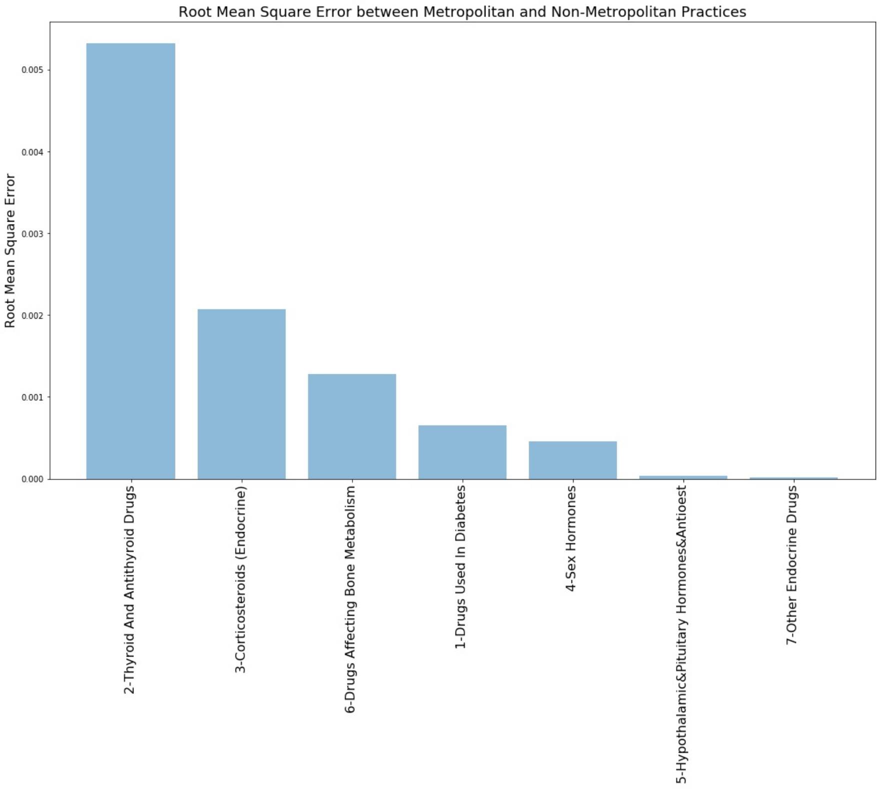


FIGURE 6.10: Comparison of RMSE by cluster at BNF section level (Chapter 6)

### 6.2.5 Chapter 13 - Skin

Prescribing trends for sections in BNF chapter 13 (Skin) were graphed for comparison (Figure 6.11) with independent t-tests also being applied (Table 6.5) using a revised Bonferroni alpha value of 0.003 (0.05/15).

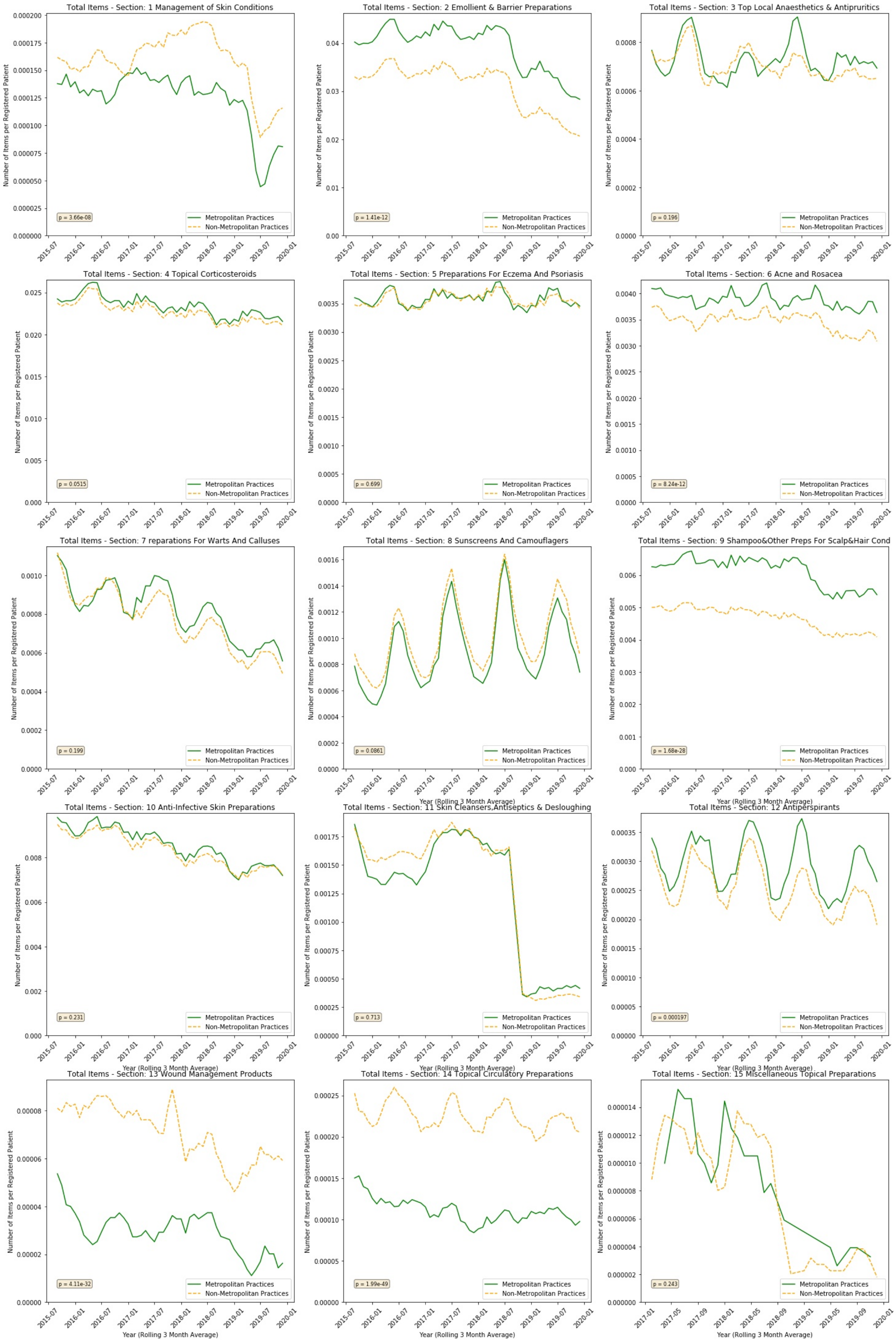


FIGURE 6.11: Items per registered patient for BNF Chapter 13 Skin sections

TABLE 6.5: Summary of independent t-tests indicating statistical significance of differences observed in Metropolitan and Non-Metropolitan practices for BNF chapter 13 (Skin)

| **Section** | **p-value** | **Statistically significant ($p<0.003$)** |
|---|---|---|
| 1 - Management of skin conditions | $3.66 \times 10^{-08}$ | Yes |
| 2 - Emollient & barrier preparations | $1.41 \times 10^{-12}$ | Yes |
| 3 - Topical anaesthetics & Antipruritics | 0.195825 | No |
| 4 - Topical corticosteroids | 0.051468 | No |
| 5 - Preparations for Eczema and Psoriasis | 0.69858 | No |
| 6 - Acne and Rosacea | $8.24 \times 10^{-12}$ | Yes |
| 7 - Preparations for warts and calluses | 0.19859 | No |
| 8 - Sunscreens and camouflagers | 0.086087 | No |
| 9 - Shampoos and other preparations for scalp and hair conditions | $1.68 \times 10^{-28}$ | Yes |
| 10 - Anti-infective skin preparations | 0.230657 | No |
| 11 - Skin cleansers, Antiseptics & Desloughing | 0.71288 | No |
| 12 - Antiperspirants | 0.000197 | Yes |
| 13 - Wound management products | $4.11 \times 10^{-32}$ | Yes |
| 14 - Topical circulatory preparations | $1.99 \times 10^{-49}$ | Yes |
| 15 - Miscellaneous topical preparations | 0.243496 | No |

The overall variations seen at BNF chapter level, chapter 13 sections show consistently higher prescribing levels in three of the fifteen sections for practices in Metropolitan areas, similar prescribing levels in nine of the fifteen sections and higher prescribing levels in the remaining three sections for practices in Non-Metropolitan areas. In all sections where higher prescribing levels were observed for either archetype, these differences were shown to be statistically significant (Table 6.5).

The main contributor to the variation seen in this chapter was sections 2 (Emollient & barrier preparations) (Figure 6.12).

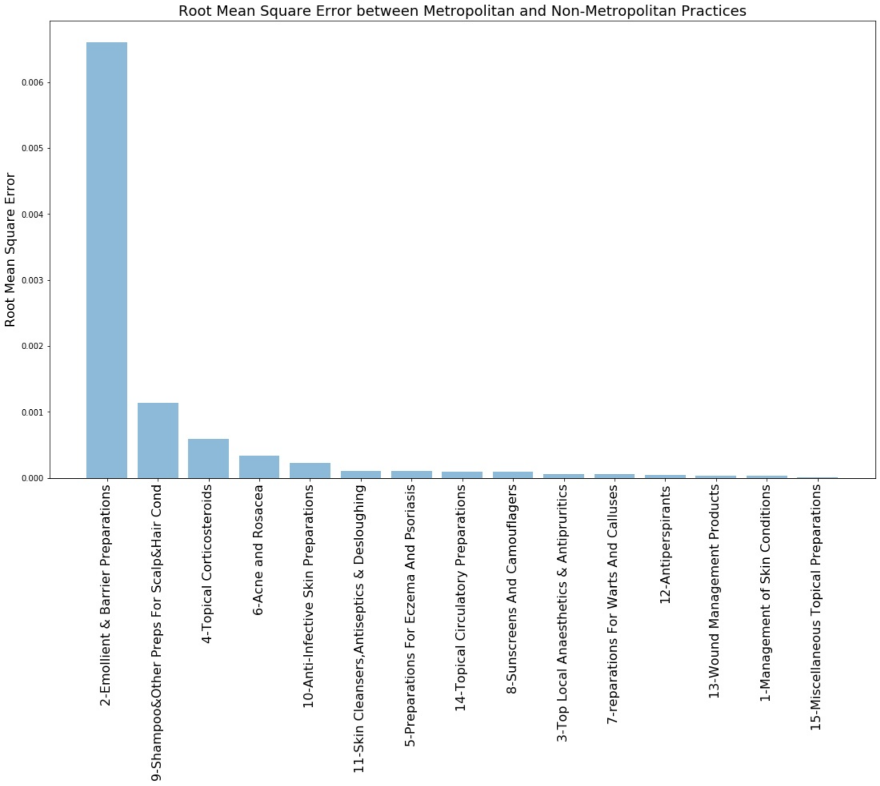


FIGURE 6.12: Comparison of RMSE by cluster at BNF section level (Chapter 13)

## 6.3 Discussion

Breaking down prescribing to chapter and section level shows that the main contributor to the differences in prescribing levels at national level stems from higher prescribing levels in chapter 4 (Central Nervous System medications) of both Analgesics and Antidepressants in Metropolitan practices. It is interesting that these higher levels are not seen in section 1 (Hypnotics and Anxiolytics) also as these are often co-prescribed. This corresponds with the comparisons made with other UK nations in chapter 3 which showed higher levels of prescribing in NI than the other UK nations in the prescribing of Analgesics and Antidepressants. Whilst all the other sections within this chapter show statistically significant differences between prescribing by Metropolitan and Non-Metropolitan practices, Section 10 (Drugs used in substance dependence) shows no significant difference in the prescribing of drugs used in substance abuse between the two clusters. One possible explanation for this is the provision of two main centres for the prescribing of drugs used in substance dependence in Northern Ireland, one in each cluster, which may be obscuring any differences. There is no evidence that any of the features used in the profiling of GP surgeries are contributing to the higher levels of Analgesic and Antidepressant prescribing in Metropolitan practices although research had previously linked both the Northern Ireland 'Troubles' and the resultant residential segregation to mental health problems in NI (O'Reilly and Stevenson, 2003; French, 2009).

Prescribing in BNF chapter 2 (Respiratory System) shows prescribing in Metropolitan practices to be higher in four of the nine sections of which two of the sections, 1 (Bronchodilators) and 4 (Antihistamines, hypo sensitisation and allergic emergencies) contribute the most to the differences seen in overall prescribing within the chapter. Again, none of the features used to profile GP practices can explain these differences but Bronchodilators are a type of medication that make breathing easier by relaxing the muscles in the lungs and widening the airways (bronchi) and are often used to treat long-term conditions where the airways may become narrow and inflamed such as asthma (NHS, 2019). Studies have previously shown links between living in cities where there are higher levels of air pollution and the prevalence of asthma cases found there (Sunyer et al., 1997). This is likely to account for the higher levels of prescribing of Bronchodilators in Metropolitan practices. Similarly, studies have shown that factors directly or indirectly related to farming and rural life decreases the risk of developing hay fever (Braun-Fahrländer et al., 1999). For this reason it is likely that the higher levels of prescribing of Antihistamines, hypo sensitisation and allergic emergencies seen in Metropolitan practices is the result of city living where there is a higher risk of developing hay fever.

Whilst prescribing levels are marginally higher on several sections of chapter 6 (Endocrine System) for Non-Metropolitan practices, there is no statistically significant difference in prescribing of drugs used in Diabetes (Section 1), Sex hormones (Section 4), Hypothalamic and pituitary hormones and anti-oestrogens (Section 5) or other endocrine drugs (Section 7). This suggests that location is not a factor in the development of conditions such as diabetes or menopause. The section which accounts for the greatest variation between the two archetypes is section 2 (Thyroid and Antithyroid drugs) where prescribing is significantly higher in Non-Metropolitan practices. Again, none of the features used in the profiling of GP practices can account for this but environmental and socioeconomic factors are likely to be the cause (Hanley et al., 2015).

Prescribing levels for sections within chapter 14 (Skin) are significantly higher for Non-Metropolitan practices in three sections: 1 (Management of skin conditions), 13 (Wound management products) and 14 (Topical circulatory preparations). This is not unsurprising as there is a higher possibility of Non-Metropolitan patients spending more time outdoors than those in Metropolitan areas. The largest variation in prescribing levels was observed in section 2 (Emollient & barrier preparations) with prescribing in Metropolitan practices being significantly higher that of Non-Metropolitan practices. This again is not unexpected as patients from Metropolitan practices are less likely to spend significant time in the outdoors building up a natural resistance to the effects of sunlight. As a result, they are more likely to require medication to protect their skin from the ultraviolet rays of the sun.

## 6.4 Conclusion

In this chapter we have discovered that prescribing patterns were largely similar for each archetype with levels of prescribing higher in approximately half of the BNF chapters for practices in Metropolitan areas. Whilst variations were observed in other BNF chapters these were not considered to be unusual. The major difference observed between Metropolitan and Non-Metropolitan prescribing was observed in BNF chapter 4 (Central Nervous System) with the largest proportion of variation between the identified clusters with sections 7 (Analgesics) and 3 (Antidepressant drugs). This finding corresponds with the results of the overall prescribing trends as compared with the other UK nations and, whilst Northern Ireland's unique history of civil unrest can be used to explain this, other factors must be considered as there has been relative peace in the province in the last 24 years since the Good Friday Agreement. In Chapter 7, the level of deprivation experienced in the ares in which practices are located and the size of the practice in terms of the number of registered doctors will be investigated to determine what effect the have on prescribing and their contribution to the differences observed in prescribing levels.

# Chapter 7

# Analysis of factors contributing to differences observed in prescribing behaviours of GP practices

"Chaos isn't a pit. Chaos is a ladder."

Lord Baelish
Game of Thrones

In the previous chapter, BNF chapter 4 (Central Nervous System) was identified as being the main contributor to the variations seen between Metropolitan and Non-Metropolitan practices in Northern Ireland. Specifically 4.7 (Analgesics) and 4.3 (Antidepressants) were singled out as the two highest contributors within chapter 4. Many factors such as patient demographics (age structure of the populations, ethnic and cultural differences in population composition etc), in practitioner demographics (including age, gender, part-time/full-time status etc), and in patient-full-time equivalent GP ratios (and consultation times) may contribute to the differences observed (Senior et al., 2003; Carter et al., 2021). No open source data are available covering practitioner demographics, patient-full-time equivalent GP rations or consultation times specific to individual GP surgeries, therefore it is not possible, given the current data set, to explore these factors. Northern Ireland's unique history of civil unrest can be used to explain heightened levels of prescribing of Antidepressants and Analgesics historically but the province has been relatively peaceful since the Good Friday agreement twenty four years ago. This means that other factors must be influencing prescribing levels. In this chapter, deprivation and practice size (defined by the number of GPs working in the practice) and their effects on prescribing will be examined.

The objectives of this chapter are to:

- Understand what effect the level of deprivation associated with the areas in which GP surgeries are located has on prescribing levels.
- Investigate prescribing levels associated with deprivation levels by archetype.

- Investigate prescribing levels by archetype where deprivation is low.
- Understand what effect GP practice size has on prescribing levels.
- Investigate prescribing levels associated with GP practice size by archetype.
- Investigate the structure each of the GP practice sizebands using clustering.

## 7.1 Deprivation

### 7.1.1 Background

Several studies have shown that levels of prescribing are higher in areas with high deprivation than those in areas with low deprivation. Public Health England has linked these higher levels to five specific addictive medicines (Public Health England, 2019) whilst similar correlations between deprivation and certain drugs have been discovered within Northern Ireland (Frazer and Frazer, 2020).

Chapter 4 detailed the work examining the link between geolocation and types of General Practitioner practice where it was found that GP practices in Northern Ireland could be classified into two distinct groups (Metropolitan and Non-Metropolitan) with the former group being in and around the city of Belfast and the latter covering the rest of Northern Ireland. This section explores the effects of deprivation on prescribing behaviours within the two archetypes.

The work in this section was presented at, and published in the proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) as a paper entitled "Examining the Effect of Deprivation on Prescribing Behaviours in Northern Ireland" (Booth et al., 2020a) and modified here to fit within the framework and context of this thesis.

### 7.1.2 Methods

Using the novel data store developed for this study (described in Chapter 3), prescription data for the period July 2015 to December 2019 was extracted. In addition to the number of items prescribed each month and the number of patients registered with each practice, the deprivation quartile metric developed from the Northern Ireland Multiple Deprivation Measure 2017, published by the Northern Ireland Statistics Research Agency was also extracted.

Python, in the form of Jupyter notebooks, was used to analyse the data utilising the scipy library to perform independent t-tests to study significant differences between means of groups and the matplotib library for data visualisation.

These practices were then analysed by which quartile they belonged to with quartile 1 being the areas with lowest deprivation and quartile 4 being those with the highest. Prescription trends for each quartile were compared for each GP type with the average number of items prescribed per registered patient being used to estimate the effect deprivation had on prescribing. Finally, prescribing for each GP type for quartile 1, the least deprived areas were compared to establish what the trend would be without the effects of deprivation.

Where multiple hypothesis tests have been employed there is an increased possibility of false discoveries. In order to account for this a corrected Bonferroni alpha value has been used. This has been calculated using the formula: corrected alpha equals standard alpha divided by the number of hypothesis tests. The value of standard alpha being 0.05.

### 7.1.3 Results

In Chapter 6 it was established that overall prescribing levels were higher for Metropolitan practices than those in Non-Metropolitan practices. By matching each practice to the corresponding deprivation level assigned to the area in which it was located, it was found that 52.2% of practices in the Metropolitan area were in the highest quartile for deprivation compared to only 13.9% of Non-Metropolitan practices (Figure 7.1).

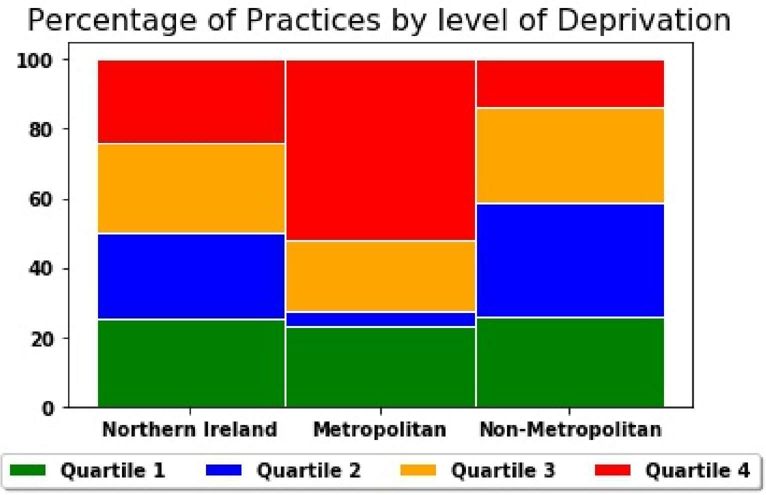


FIGURE 7.1: Percentage of GP practices by level of deprivation in Northern Ireland

Examining prescription levels for Metropolitan practices for each deprivation quartile (Figure 7.2) clearly showed that higher deprivation resulted in higher prescribing levels with an average of 1.44 items prescribed per registered patient in quartile 1, 1.62 in quartile 2, 1.70 in quartile 3 and 2.02 in quartile 4 making prescribing over 40% higher in the most deprived Metropolitan areas than those in the least deprived areas. Performing independent t-tests showed that the differences in all deprivation levels were statistically significant (Table 7.1).

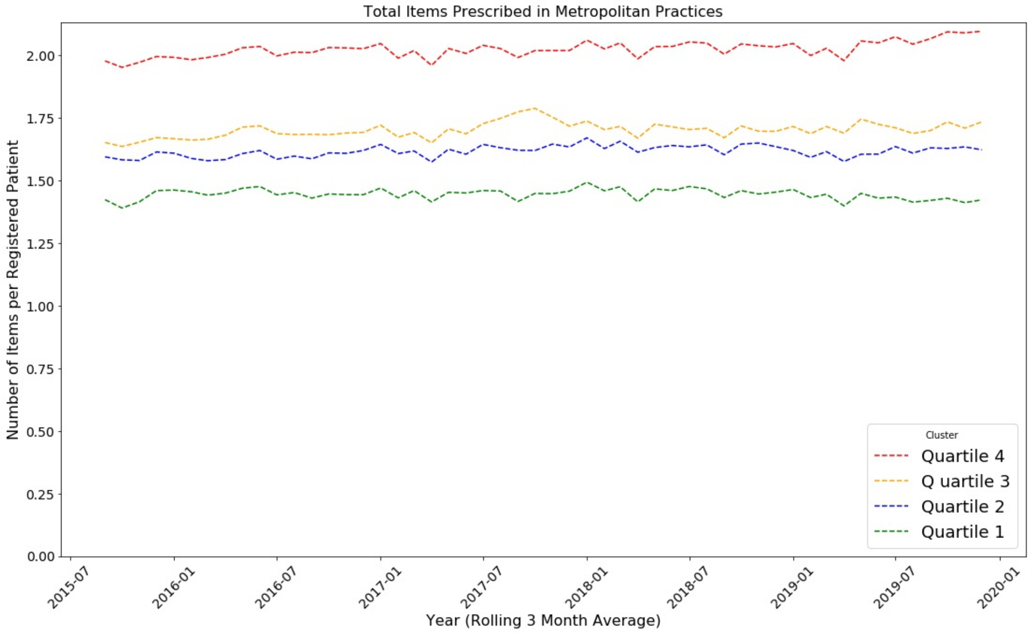


FIGURE 7.2: Prescribing levels for Metropolitan practices by deprivation quartile

TABLE 7.1: Summary of independent t-tests indicating statistical significance of differences observed deprivation levels for Metropolitan practices.

| **Deprivation levels** | **p-value** | **Statistically significant (p<0.02)** |
|---|---|---|
| Quartile 3 and Quartile 4 | 2.50x10$^{-43}$ | Yes |
| Quartile 2 and Quartile 3 | 7.08x10$^{-11}$ | Yes |
| Quartile 1 and Quartile 2 | 8.90x10$^{-30}$ | Yes |

Similarly, examining prescription levels for Non-Metropolitan practices for each deprivation quartile (Figure 7.3) also showed that higher deprivation levels resulted in higher prescribing levels with an average of 1.60 items prescribed per registered patient in quartile 1, 1.71 in quartile 2, 1.74 in quartile 3 and 1.78 in quartile 4 making prescribing more than 11% higher in the most deprived Non-Metropolitan areas than those in the least deprived areas. Performing independent t-tests showed that the differences in prescribing between practices in high deprivation areas (quartile 4) and practices in low areas of deprivation (quartile 1) were statistically significant from those in low/medium (quartile 2) and medium/high (quartile3) areas of deprivation. The difference in prescribing by practices in low/medium and medium/high areas of deprivation were not significantly different statistically (Table 7.2).

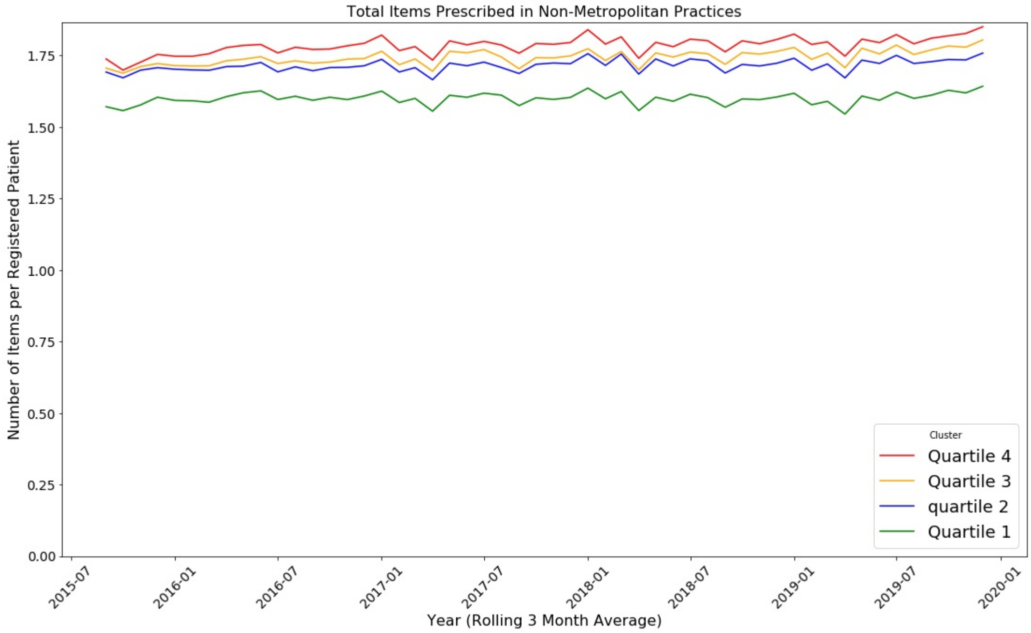


FIGURE 7.3: Prescribing levels for Non-Metropolitan practices by deprivation quartile

TABLE 7.2: Summary of independent t-tests indicating statistical significance of differences observed deprivation levels for Non-Metropolitan practices.

| **Deprivation levels** | **p-value** | **Statistically significant (p<0.01)** |
|---|---|---|
| Quartile 3 and Quartile 4 | 0.00231 | Yes |
| Quartile 2 and Quartile 3 | 0.02655 | No |
| Quartile 1 and Quartile 2 | $1.52\text{x}10^{-16}$ | Yes |
| Quartile 2 and Quartile 4 | $5.48\text{x}10^{-07}$ | Yes |

Comparing the practices in the least deprived areas (quartile 1) of both GP practice types showed that without the effects of deprivation, prescribing levels in Non-Metropolitan practices are actually 11% higher than those in Metropolitan practices (Figure 7.4). Independent t-tests show that the differences observed in prescribing levels are statistically significant (p<0.05).

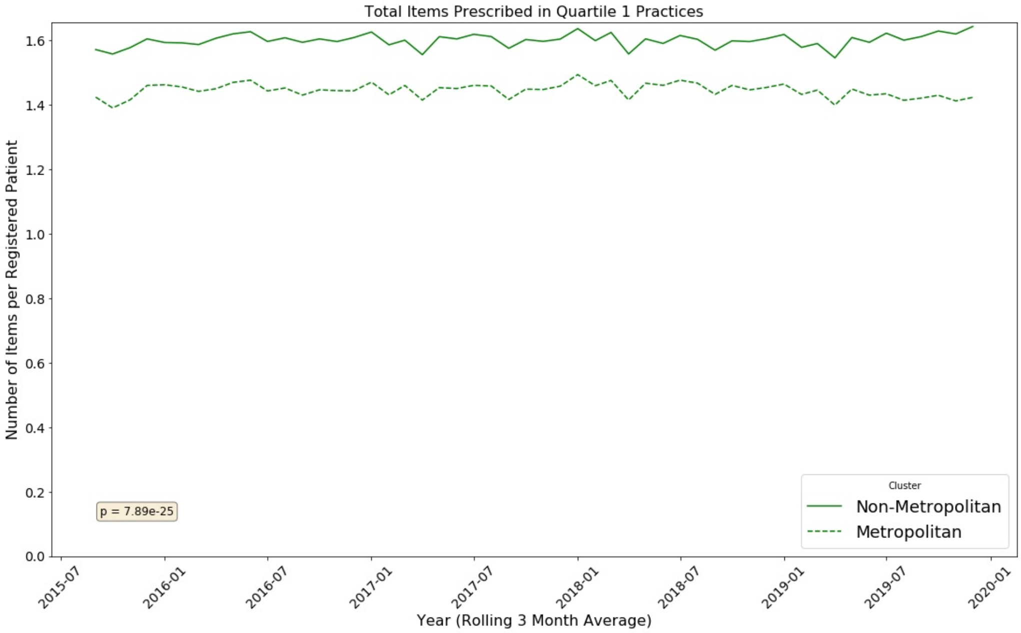


FIGURE 7.4: Prescribing levels for GP practice types in quartile 1 where the effects of deprivation are minimal.

## 7.2 GP practice size

### 7.2.1 Background

Whilst the Institute for Fiscal Studies conducted research on the trends in General Practitioner practice size and the relationship between practice size and the quality of care experienced by patients (Kelly and Stoye, 2014), no studies have been identified examining the relationship between General Practitioner practice size and their associated prescribing levels. This section will examine the relationship between different practice sizes and their prescribing levels and compare these for Metropolitan and Non-Metropolitan practices.

The work in this section was presented at, and published in the proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) as a paper entitled "Examining the Effect of General Practitioner Practice Size on Prescribing Behaviours in Northern Ireland" (Booth et al., 2020b) and modified here to fit within the framework and context of this thesis.

### 7.2.2 Methods

Using the novel data store developed for this study (described in Chapter 3), prescription data for the period July 2015 to December 2019 was extracted. In addition to the number of items prescribed each month and the number of patients registered with each practice, each practice was classified into one of four cohorts: Single-Handed practices (1 doctor), Small practices (2 doctors), Medium sized practices (3-4 doctors) and Large practices (5+ doctors).

Python, in the form of Jupyter notebooks, was used to analyse the data utilising the scipy library to perform independent t-tests to study significant differences between means of groups and the matplotib library for data visualisation.

The number of items prescribed per registered patient was calculated for each practice over the period July 2015 to December 2019 and used to compare the different sizes of practice at both Northern Ireland level and within the two identified archetypes (Metropolitan and Non-Metropolitan).

Each GP practice size was then examined using k-means clustering to discover what archetypes comprised therein using the same six features as previously used (Chapter 4):

- Number of pharmacies dispensing prescriptions issued by a practice.
- Average number of items prescribed per registered patient.
- Average distance traveled from a practice to dispense a prescription.
- Standard deviation of distance traveled from a practice to dispense a prescription.
- Population density of the super output area in which the practice is located.
- Number of registered patients in the practice.

Finally, the average number of Registered patients was calculated for each practice size in Northern Ireland and within each archetype in order to establish whether any differences existed.

### 7.2.3 Results

Comparing prescribing levels for the four sizes of GP practice at Northern Ireland level, it was found that there were no discernible differences in prescribing levels for Single-Handed, Medium and Large practices. Small practices however showed higher prescribing levels than the other three categories (Figure 7.5). Independent t-tests performed showed that there was a statistically significant difference in prescribing levels for single-handed (1 doctor) and small practices (2 doctors) compared to all other categories (Table 7.3).

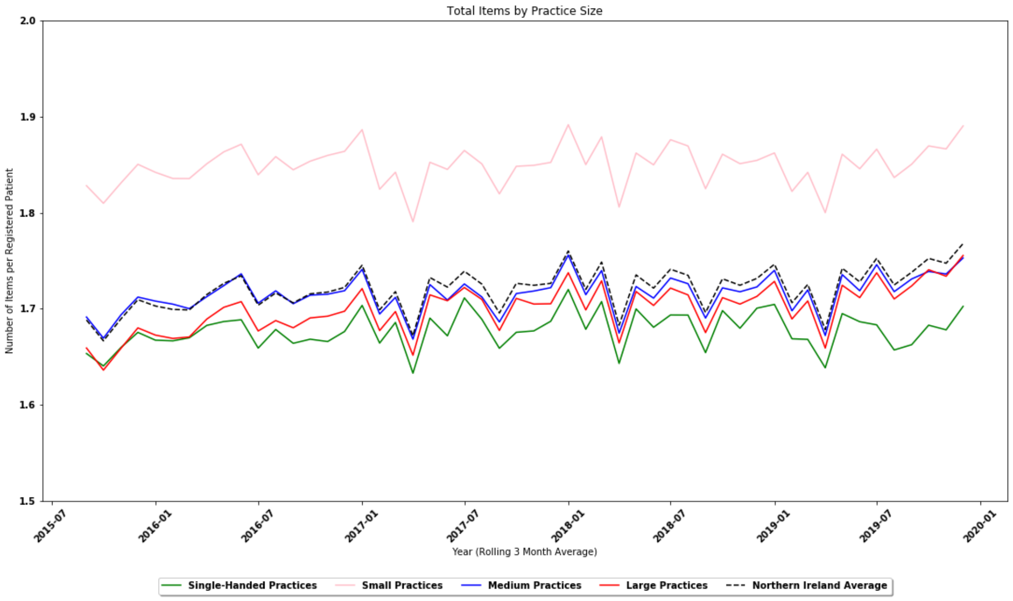


FIGURE 7.5: Prescribing levels in Northern Ireland by GP practice size

TABLE 7.3: Summary of independent t-tests performed on GP practice size categories for Northern Ireland. Statistically significant differences (p<0.005) are in bold.

| | NI Average | Single-Handed | Small | Medium | Large |
|---|---|---|---|---|---|
| NI Average | | **0.00046** | **1.14x10$^{-17}$** | 0.64936 | 0.09352 |
| Single-Handed | **0.00046** | | **9.09x10$^{-26}$** | **0.00186** | 0.06537 |
| Small | **1.14x10$^{-17}$** | **9.09x10$^{-26}$** | | **6.49x10$^{-19}$** | **3.74x10$^{-21}$** |
| Medium | 0.64936 | **0.00186** | **6.49x10$^{-19}$** | | 0.21076 |
| Large | 0.09352 | 0.06537 | **3.74x10$^{-21}$** | 0.21076 | |

Taking the average for each category it was found that Small practices had the highest prescribing levels (1.85 items per registered patient), 10% higher than the lowest being Single-Handed practices (1.68 items per registered patient). Medium and Large practices prescribed 1.72 and 1.70 items respectively.

Comparing the number of practices in each category, it was found that of the 333 practices operating at March 2018 in Northern Ireland, 29 (8.7%) were Single-Handed practices, 61 (18.3%) Small practices, 122 (36.6%) Medium practices and 121 (36.3%) Large practices. Splitting these practices into their behavioural archetypes, there were 90 Metropolitan and 243 Non-Metropolitan practices. Both archetypes were found to be of similar proportions to that seen for Northern Ireland (Figure 7.6).

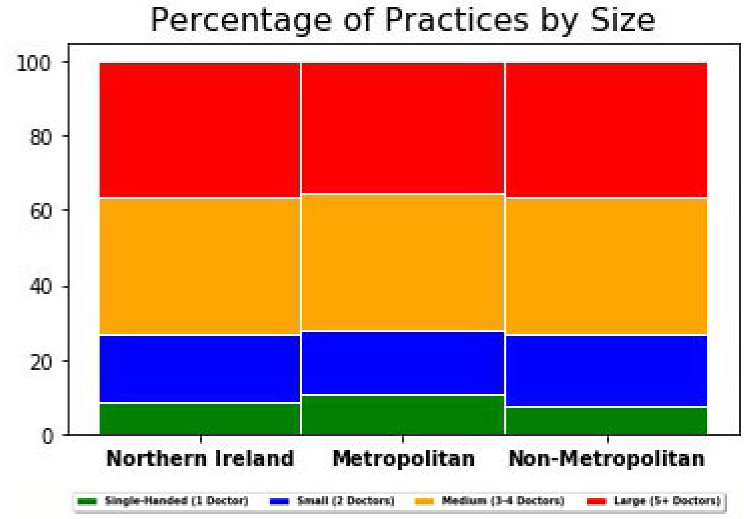


FIGURE 7.6: Percentage of practices by size

Examining each archetype in turn, it was found that within the Metropolitan area prescribing levels were highest in Small practices (2.21 items per registered patient), 32% higher than in Large practices (1.67 items per registered patient). Single-Handed and Medium practices prescribed on average 1.76 and 1.81 items respectively (Figure 7.7). Performing independent t-tests on these categories shows that prescribing levels for all practice types were significantly different statistically than the other categories (Table 7.4).

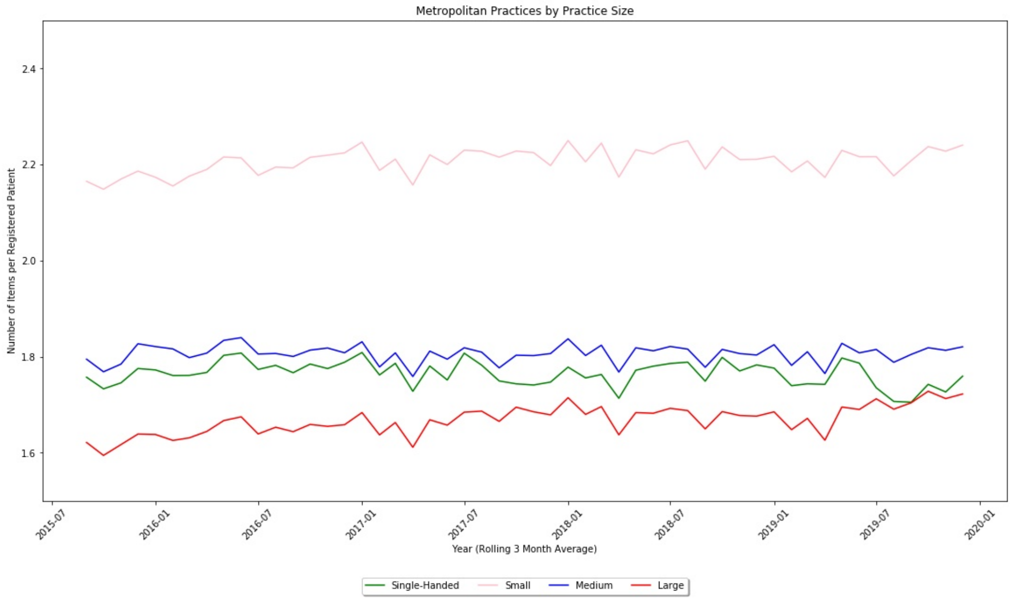


FIGURE 7.7: Prescribing levels for Metropolitan practices by size

TABLE 7.4: Summary of independent t-tests performed on GP practice size categories for Metropolitan practices. Statistically significant differences (p<0.008) are in bold.

| | Single-Handed | Small | Medium | Large |
|---|---|---|---|---|
| Single-Handed | | **5.61x10$^{-56}$** | **0.00128** | **1.92x10$^{-11}$** |
| Small | **5.61x10$^{-56}$** | | **1.88x10$^{-52}$** | **1.65x10$^{-64}$** |
| Medium | **0.00128** | **1.88x10$^{-52}$** | | **3.69x10$^{-19}$** |
| Large | **1.92x10$^{-11}$** | **1.65x10$^{-64}$** | **3.69x10$^{-19}$** | |

Within the Non-Metropolitan area prescribing levels again were highest in Small practices (1.74 items per registered patient), 3.8% higher than in Single-Handed practices (1.64 items per registered patient). Medium and Large practices prescribed on average 1.68 and 1.71 items respectively (Figure 7.8). Independent t-tests showed that prescribing levels were not significantly different statistically between Medium and Large practices (Table 7.8).

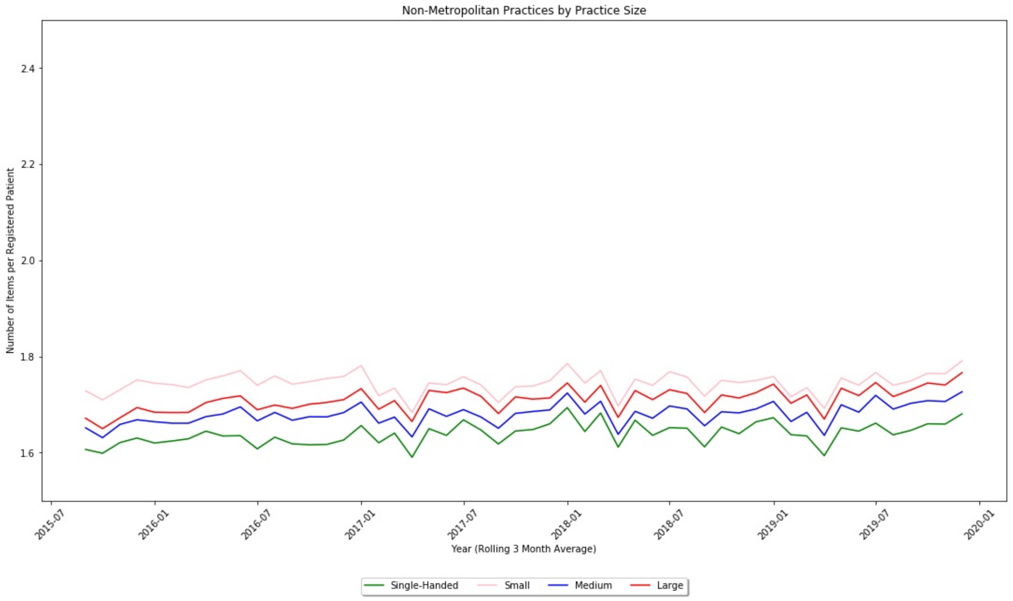


FIGURE 7.8: Prescribing levels for Non-Metropolitan practices by size

TABLE 7.5: Summary of independent t-tests performed on GP practice size categories for Non-Metropolitan practices. Statistically significant differences (p<0.008) are in bold.

| | Single-Handed | Small | Medium | Large |
|---|---|---|---|---|
| Single-Handed | | **1.05x10$^{-13}$** | **0.0011** | **8.60x10$^{-08}$** |
| Small | **1.05x10$^{-13}$** | | **5.69x10$^{-07}$** | **0.00692** |
| Medium | **0.0011** | **5.69x10$^{-07}$** | | 0.01383 |
| Large | **8.60x10$^{-08}$** | **0.00692** | 0.01383 | |

### 7.2.4 Registered Patients per practice type

Examining the number of registered patients for each practice type, the Northern Ireland average is 2,983 for Single-Handed practices, 2,677 for Small practices, 5,098 for Medium practices and 8,600 for Large practices. With the exception of Medium sized practices, practices in Non-Metropolitan areas have more registered patients than their respective counterparts in Metropolitan areas. (Figure 7.9)

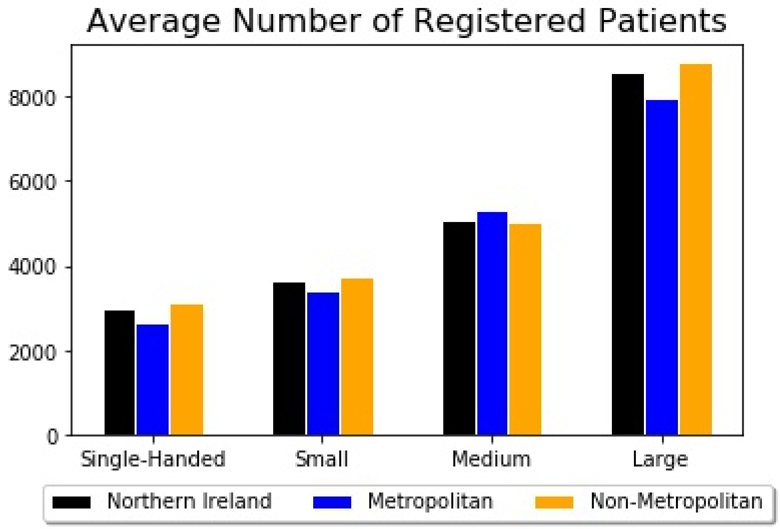


FIGURE 7.9: Average number of registered patients by GP practice size

## 7.3 Clustering of GP practices by sizeband

In Chapter 5, a new methodology was proposed for the classification of GP practices based on prescribing behaviours and geographical attributes. The results of this analysis indicated that there were two main types of GP practice in Northern Ireland - Metropolitan and Non-Metropolitan. Given that GP practices are not all the same size, do these classifications hod true for the different sizes of GP practice?

### 7.3.1 Methods

Splitting the data set into the four practice size bands, Single-Handed (1 registered doctor), Small (2 registered doctors), Medium (3-4 registered doctors) and Large (5 or over registered doctors), the same methodology used in Chapter 5 was applied (Figure 7.10) with each data set being analysed using the k-means algorithm.

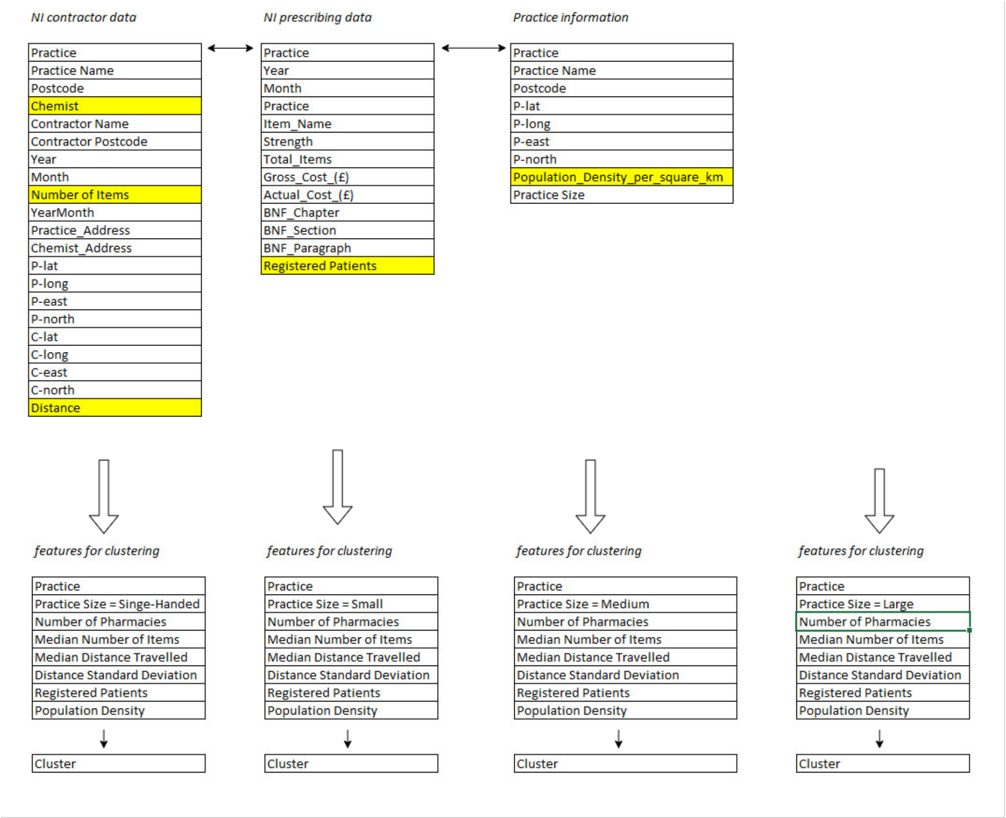


FIGURE 7.10: Workflow for identification of practice types using clustering for Single-Handed, Small, Medium and Large practices

### 7.3.2 Results

In order to ascertain the optimum number of clusters to be used in the k-means algorithm, both the Within Cluster Sum of Squares (Elbow plot) and Silhouette methods were used with the resultant cluster configurations displayed using a Principal Component Analysis plot (Appendix G). In all cases k=2 proved to be optimum.

Of the 29 Single-Handed GP practices, it was fount that 18 were classified as cluster A (Non-Metropolitan) and 11 as cluster B (Metropolitan). Archetipical characteristics were calculated for for each type (Table 7.6).

TABLE 7.6: Archetypical characteristics for each cluster of Single-Handed practices (i.e. Centroid Feature Values) for the period April 2018 - March 2019

| **Feature** | **Cluster A** | **Cluster B** |
|---|---|---|
| Number of Pharmacies | 68.3 (+- 29.3) | 168.1 (+- 46.0) |
| Number of Items per Registered Patient | 240.0 (+- 3.1) | 267.6 (+- 5.9) |
| Distance to Pharmacy (km) | 12.4 (+- 5.0) | 3.8 (+- 1.2) |
| Distance Standard Deviation (km) | 21.4 (+- 6.2) | 11.6 (+- 3.5) |
| Population Density per square km | 989 (+- 1093) | 5587 (+- 2808) |
| Registered Patients | 3142 (+- 1291) | 2680 (+- 1148) |

Of the 61 Small GP practices, it was fount that 46 were classified as cluster A (Non-Metropolitan) and 15 as cluster B (Metropolitan). Archetipical characteristics were calculated for for each type (Table 7.7).

TABLE 7.7: Archetypical characteristics for each cluster of Small practices (i.e. Centroid Feature Values) for the period April 2018 - March 2019

| Feature | Cluster A | Cluster B |
|---|---|---|
| Number of Pharmacies | 75.2 (+- 23.4) | 199.3 (+- 37.7) |
| Number of Items per Registered Patient | 253.2 (+- 4.3) | 315.6 (+- 7.2) |
| Distance to Pharmacy (km) | 14.7 (+- 6.4) | 4.0 (+- 0.9) |
| Distance Standard Deviation (km) | 19.2 (+- 6.1) | 11.8 (+- 2.8) |
| Population Density per square km | 1049 (+- 1184) | 6074 (+- 2153) |
| Registered Patients | 3764 (+- 1285) | 3408 (+- 1453) |

Of the 122 Medium GP practices, it was fount that 89 were classified as cluster A (Non-Metropolitan) and 33 as cluster B (Metropolitan). Archetipical characteristics were calculated for for each type (Table 7.8).

TABLE 7.8: Archetypical characteristics for each cluster of Medium practices (i.e. Centroid Feature Values) for the period April 2018 - March 2019

| Feature | Cluster A | Cluster B |
|---|---|---|
| Number of Pharmacies | 93.5 (+- 30.4) | 211.4 (+- 38.4) |
| Number of Items per Registered Patient | 241.2 (+- 3.6) | 271.2 (+- 4.7) |
| Distance to Pharmacy (km) | 15.2 (+- 5.3) | 4.5 (+- 1.3) |
| Distance Standard Deviation (km) | 18.9 (+- 5.3) | 12.8 (+- 3.2) |
| Population Density per square km | 1276 (+- 1219) | 4988 (+- 2625) |
| Registered Patients | 5026 (+- 1474) | 5291 (+- 1920) |

Of the 121 Large GP practices, it was fount that 86 were classified as cluster A (Non-Metropolitan) and 35 as cluster B (Metropolitan). Archetipical characteristics were calculated for for each type (Table 7.9).

TABLE 7.9: Archetypical characteristics for each cluster of Large practices (i.e. Centroid Feature Values) for the period April 2018 - March 2019

| Feature | Cluster A | Cluster B |
|---|---|---|
| Number of Pharmacies | 120.0 (+- 41.3) | 231.9 (+- 48.9) |
| Number of Items per Registered Patient | 247.2 (+- 2.6) | 246.0 (+- 5.3) |
| Distance to Pharmacy (km) | 15.0 (+- 6.3) | 5.3 (+- 1.7) |
| Distance Standard Deviation (km) | 20.8 (+- 6.2) | 14.5 (+- 3.5) |
| Population Density per square km | 1444 (+- 1279) | 4831 (+- 2641) |
| Registered Patients | 8821 (+- 2397) | 7984 (+- 2262) |

### 7.3.3 Principal Component Explained Variance Ratios

Principal Component Explained Variance Ratios were calculated on each feature for each GP practice size band in order to gain insight into which features contributed most to the variances seen between archetypes. The results of this analysis can be found in Table 7.10.

TABLE 7.10: Principal Component Explained Variance Ratios by GP practice size band

| Feature | Single-Handed | Small | Medium | Large |
|---|---|---|---|---|
| Number of Pharmacies | 0.493 | 0.474 | 0.421 | 0.372 |
| Number of Items per Registered Patient | 0.218 | 0.196 | 0.210 | 0.212 |
| Distance to Pharmacy (km) | 0.128 | 0.163 | 0.151 | 0.172 |
| Distance Standard Deviation (km) | 0.076 | 0.078 | 0.115 | 0.132 |
| Population Density per square km | 0.058 | 0.058 | 0.059 | 0.058 |
| Registered Patients | 0.026 | 0.030 | 0.044 | 0.054 |

### 7.3.4 Discussion

**Deprivation** - Comparing prescription levels of the two archetypes it can be seen that the higher the deprivation of the area in which the practice is located, the higher the prescribing levels are. Prescribing levels for Non-Metropolitan practices are 11.2% higher for practices in high deprivation areas than that of those in low deprivation areas. This contrasts with Metropolitan GP practices where prescribing levels are 40.3% higher in high deprivation areas than that of those in low deprivation areas. Whilst the Northern Ireland 'Troubles' officially ended with the Good Friday Agreement, communities in the Metropolitan area are still segregated having 'Peace Walls' separating rival communities (Cunningham and Gregory, 2014) fuelling high deprivation levels. This is likely to account for the higher prescription rates for Metropolitan GP practices in high deprivation areas. It would be untrue to claim that segregation does not exist in Non-Metropolitan communities but these communities are not as close physically and the 'Peace Walls' which exist in Metropolitan areas are not evident.

Comparing prescribing levels for both archetypes where deprivation is not a major factor (i.e. Quartile 1 - Low deprivation levels) shows that prescribing levels in Metropolitan practices are lower than those in Non-Metropolitan practices. This means that patients in low deprivation areas within the Metropolitan area experience lower levels of sickness that their counterparts in Non-Metropolitan areas.

**Practice Size** - Comparing prescription levels of different sized General Practitioner practices at Northern Ireland level shows that the highest prescribing cohort are Small practices with two registered doctors (1.85 items per registered patient). This cohort prescribes on average 10% more than Single-Handed practices with only one registered doctor being the lowest prescribing cohort (1.68 items per registered patient). Medium and Large practices (3-4 and 5+ doctors respectively) had similar prescribing levels of 1.72 and 1.70 items per registered patient respectively. These results disprove the general theory that Single-Handed practices are more likely to have higher prescription rates due to the pressures on an individual doctor running a practice. These low prescription rates could also reflect the number of patients the individual GP is able to deal with.

Comparing the two archetypes (Metropolitan and Non-Metropolitan) it is evident that Small practices are the highest prescribing cohorts in both archetypes with Large practices having the lowest prescribing levels in Metropolitan areas and Single-Handed practices in Non-Metropolitan areas. It is interesting to note that the difference in prescribing levels is not as pronounced in practices in Non-Metropolitan areas with highest prescribing being 3.8% higher than the lowest. In contrast, the highest prescribing levels for practices in Metropolitan areas is 32% higher than the lowest. Analysis of each of the four sizes of practice showed that these conformed to the two archetypes previously established reinforcing those findings. Further research is needed to explain why a larger difference is seen in practices in Metropolitan areas although deprivation is a possible factor. The lowest prescribing levels in practices in Non-Metropolitan areas are seen in Single-Handed practices. This may be due to the relatively larger number of registered patients in these practices along with a greater knowledge of their patients that will influence prescribing. Prescribing in Small and Medium practices allows the possibility of less tight control on prescribing and they also lack the advantages of scale that Large practices have. The lowest prescribing levels in Metropolitan areas are seen in Large practices which have the advantage of scale and availability of extra services e.g. Cognitive Behaviour Therapy.

**Registered Patients per practice type** - It is not surprising that larger practices have more registered patients than their smaller counterparts given their ability to see more patients each week. It is generally accepted that for Safe working conditions, GPs should be offering 72 appointments per 1,000 patients each week, having an average list size of 1,600 patients (per GP). This means that GPs should be offering 115 appointments per week, an average of 23 per day over five days (PracticeIndex, 2017). Taking this into account it is clear that Single-Handed practices in Northern Ireland are over subscribed placing more pressure on these GPs. Whether this added pressure is detrimental to the overall running of the practice is unclear although anecdotal evidence suggests that single-handed practices are declining with GPs instead opting to join larger partnerships. All other sizes of GP practice conform to the safe working guidelines. The higher ratio of registered patients to GP in singe-handed practices would also result in the lower number of items per registered patient being recorded skewing the results. This would suggest that Large practices in both archetypes are the most efficient in terms of prescribing.

**Clustering of GP practices by practice size** - Having previously discovered two types of GP practice (Metropolitan and Non-Metropolitan) based on prescribing behaviours along with location attributes, it was of interest to discover whether these archetypes held true for differing sizes of GP practice. In all four practice size categories, it was found that the optimal number of clusters remained the same (k=2) reinforcing the previous findings. It was also not surprising to find that within each archetype, the influence of a GP practice grew with its size. For example, the typical number of pharmacies servicing a Non-Metropolitan practice (Cluster A) increased with the size of the GP practice - Singe-Handed (68.3), Small (75.2), Medium (93.5), Large (120.0). This pattern holds true for all features. Calculating Principal Component Explained Variance ratios for each GP practice size, it was found that in all cases, the number of pharmacies servicing a GP practice contributed most to the variances observed with the number of items per registered patient contributing the second highest variation. In all cases, the population density attributed to the location in which the GP practice was located contributed less than 6% to the variations observed. This indicates that whilst the influence of a GP practice increases with its size in terms of the number of doctors within the practice, the size of the local population has little effect on the overall profile of the practice.

### 7.3.5 Limitations

Practices have been categorised based on the number of registered doctors working in the practice. The assumption has been made that all of these doctors work full-time which in reality is probably not the case. For example, a practice with 3 registered doctors would be categorised as a medium sized practice (3-4 registered doctors) but if two of those doctors were part time this would chance the calculation to 2 full time equivalent doctors (1 full time, 2 part time) categorising the practice to that of being a Small practice (2 registered doctors). Also, as no data is available on the number of locums working in any practice, these have been ignored but could potentially affect the categorisation of each practice.

## 7.4 Conclusion

This chapter has investigated two of the possible factors affecting the prescribing behaviours of Northern Ireland GP practices and examined their effect on both Metropolitan and Non-Metropolitan practices. It was established that higher prescribing levels could be associated with high deprivation and that as there were a higher proportion of GP surgeries in high deprivation areas in Metropolitan areas, this had the effect of increasing prescribing levels in these areas. Examining surgeries in low deprivation areas for both archetypes showed that without the effects of deprivation, Metropolitan practices had lower prescribing rates than Non-Metropolitan practices. Therefore whilst deprivation is a factor affecting prescribing levels, it is socio-economic in nature and should not be considered as a feature when categorising GP practices. Practice size was also a contributing factor to the differences seen in prescribing levels with Large practices having lower prescribing rates in Metropolitan areas and Single-Handed practices in Non-Metropolitan areas. Whilst both these factors have been shown to influence prescribing levels, there is no indication that either should be considered as a feature when profiling practices. Investigating whether the size of a GP practice affected the classification of the practice showed that the original categories of Metropolitan and Non-Metropolitan held true no matter what size band a practice was in further reinforcing the conclusions in Chapter 4. In early 2020, a new unexpected factor occurred in the form of a global pandemic. The COVID-19 pandemic forced nations to lockdown all businesses and impose a 'stay-at-home' policy in order to combat the spread of the virus. Chapter 8 examines the effects the COVID-19 pandemic and the first UK national lockdown on prescribing behaviours in Northern Ireland and compares these to those seen in England at the same time.

# Chapter 8

# Analysis of prescribing behaviours during the COVID-19 pandemic

> "There are no happy endings in history, only crisis points that pass."
>
> Isaac Asimov,
> The Gods Themselves

The work presented in this chapter was presented at, and published in the proceedings of the 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI) as a paper entitled "COVID-19 and lockdown: The highs and lows of general practitioner prescribing" (Booth et al., 2021a) and modified here to fit within the framework and context of this thesis.

The objectives of this chapter are to:

- Investigate the effects, if any, that the national lockdown enforced during the COVID-19 pandemic had on prescribing behaviours.
- Compare the trends observed in Northern Ireland with those in England.
- Examine the prescribing behaviours at BNF chapter level to gain insights into any effects on specific types of medication.
- Compare the prescribing behaviours of the two identified archetypes (Metropolitan and Non-Metropolitan).

## 8.1 Background

The COVID-19 virus reached the United Kingdom in early 2020. In an attempt to slow its spread and avoid overwhelming the National Health Service, a national “lockdown” was imposed by the UK government in March 2020. Only essential services were allowed to operate with the majority of the population working from home or on furlough. Older people and those with chronic medical conditions were advised to self-isolate. Those considered to be extremely vulnerable were urged to shield, depending on friends and family to shop for food and other essential supplies.

GPs had to take precautions to limit the spread of the virus among staff as well as patients and, to this end, strategies to limit unnecessary footfall in their surgeries were implemented. Due to the limiting of footfall in surgeries, it was expected that this would change prescribing behaviour.

In the absence of previous studies comparing the overall prescribing trends in England and Northern Ireland, this chapter details the work to investigate the effect that the COVID-19 pandemic and the first UK national lockdown had on GP prescribing behaviour in Northern Ireland and England.

## 8.2 Methods

Using the novel data store developed for this study (described in Chapter 3), prescription data for the period January 2019 to December 2020 was extracted for practices in Northern Ireland. Comparable prescription data for the same period was extracted from the English Prescribing dataset provided by the NHS and population estimates for 2019 and 202 from the Mid Year Population Estimates published by the Office for National Statistics.

Python, in the form of Jupyter notebooks, was used to analyse the data utilising the scipy library to perform independent t-tests to study significant differences between means of groups and the matplotib library for data visualisation.

The number of items prescribed per head of population each month for both regions was then aggregated at regional, BNF Chapter and BNF Section levels and presented graphically. The trends for 2020 were compared to both the previous year and between regions and T-Tests applied to the data to gauge the statistical significance of the year-on-year change. As the UK lockdown started in March 2020 and lasted for approximately two months, the percentage rise in the number of items prescribed from February to March 2020 was calculated for each BNF chapter to gauge the effect lockdown had on the prescribing of each type of medication and provide comparisons. These were compared by BNF Chapter with the results resented graphically and ordered from highest to lowest. The resulting p value produced by T-Tests being interpreted as being statistically significant if it was less than 0.05.

## 8.3 Results

A pattern of 'peak, trough and recovery' of prescribing levels can be seen over the period of the UK wide lockdown for England and Northern Ireland (Figure 8.1). In England, the number of items prescribed between February and March 2020 rose from 87,203,155 to 99,876,063 nearly twice as great an increase as in the same period the previous year (14.5% vs 7.6%). Northern Ireland saw a much greater relative increase of 20.7% in prescribed items in the period February to March 2020 (from 3,252,942 items to 3,924,921 items). This represented a four-fold increase in the rise seen during the same period in 2019 (5.1%). Both regions saw prescribing subsequently fall during April and May 2020 to levels below that seen in 2019 before starting to recover. The percentage change in prescribing between February and March 2020, showed an increase across most BNF categories for both regions (Figure 8.2). Respiratory medications (Chapter 3) had the greatest rise (60.4% in Northern Ireland and 42.1% in England), but prescribing fell dramatically in Immunological Products & Vaccines (Chapter 14). T-Tests on the whole years prescribing compared to the previous year, showed that there was no statistically significant change in prescribing in either England (p=.819) or Northern Ireland (p=.920).

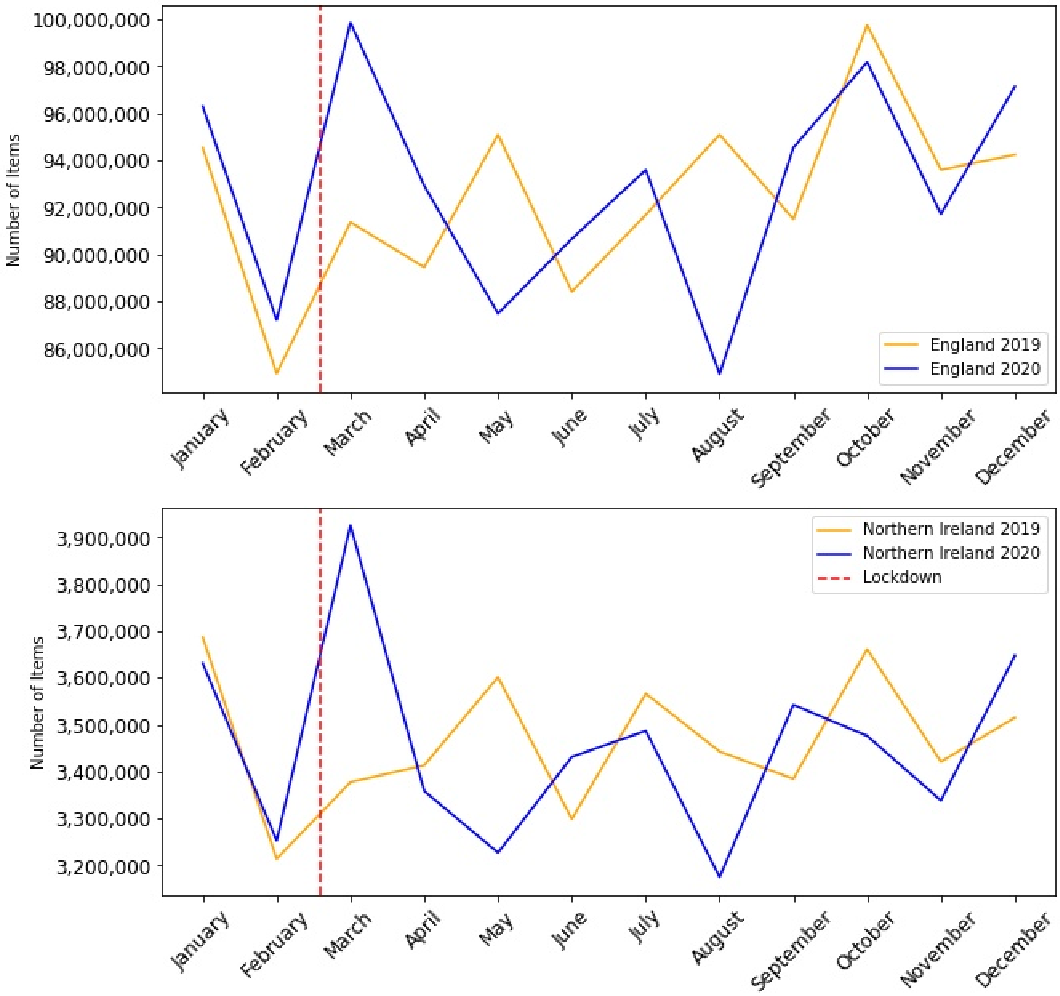


FIGURE 8.1: Prescription trends in England and Northern Ireland during the COVID-19 pandemic including the UK national lockdown (March - May 2020). Approximate date of the start of the lockdown is indicated as a red dotted line. Number of items prescribed for each month are reported on the last day of the month.

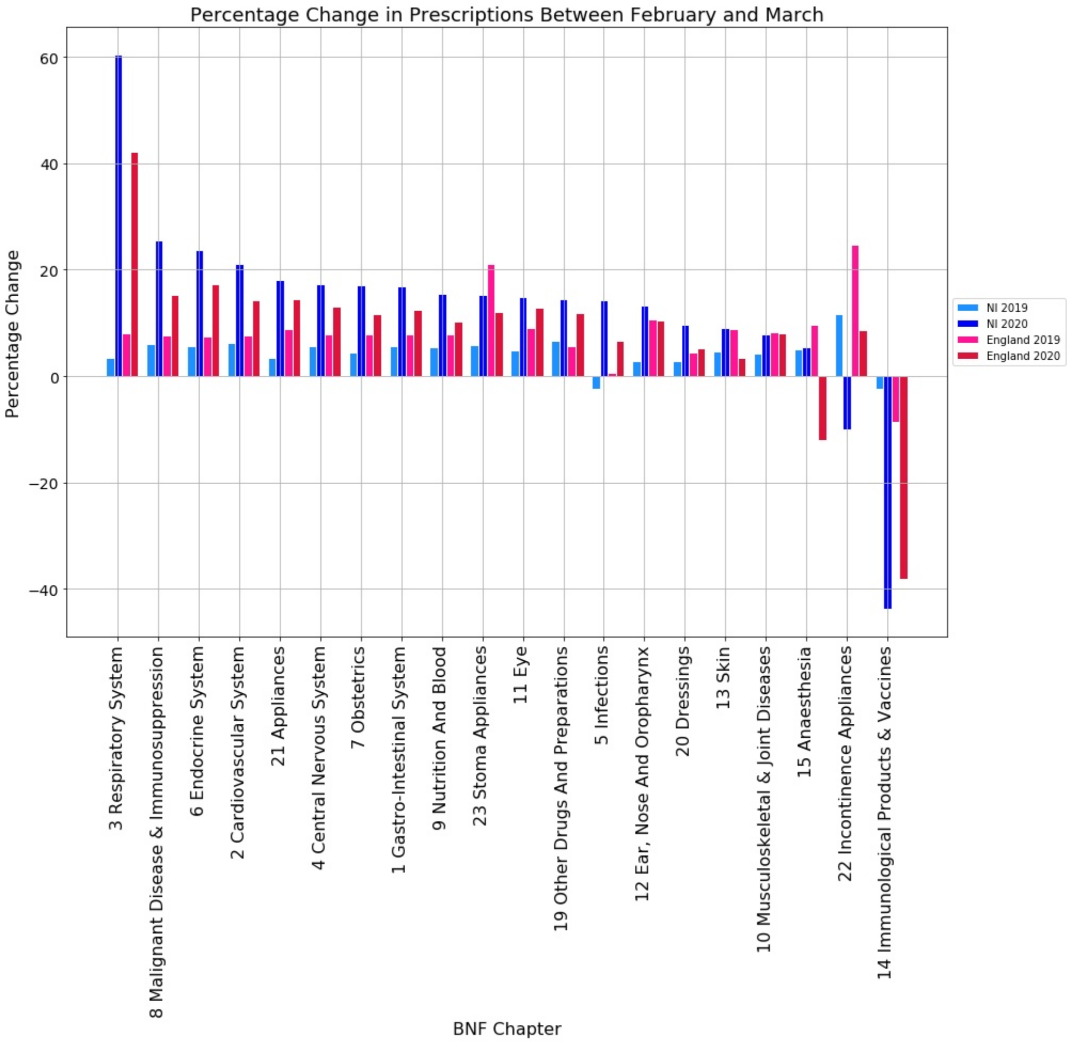


FIGURE 8.2: Percentage change in items prescribed between February and March of 2019 and 2020 by BNF Chapter, ordered by highest to lowest.

BNF Chapter 4 Section 3 (Antidepressant drugs) followed the same 'Peak, Trough, Recovery' pattern observed in most other BNF chapters (Figure 8.3). T-Tests show that the rise in prescribing rate from 2019 to 2020 is not statistically significant in Northern Ireland (p=.0571) whilst the rise in England is (p=0.00568). In order to gauge whether the rise in Antidepressant prescribing was a result direct result of the COVID-19 pandemic and subsequent lockdown, data for Chapter 4 Section 3 was compared for both regions over a five year period (Figure 8.4). This showed that antidepressant prescribing had risen consistently over this period with no unusual increase observed in 2020.

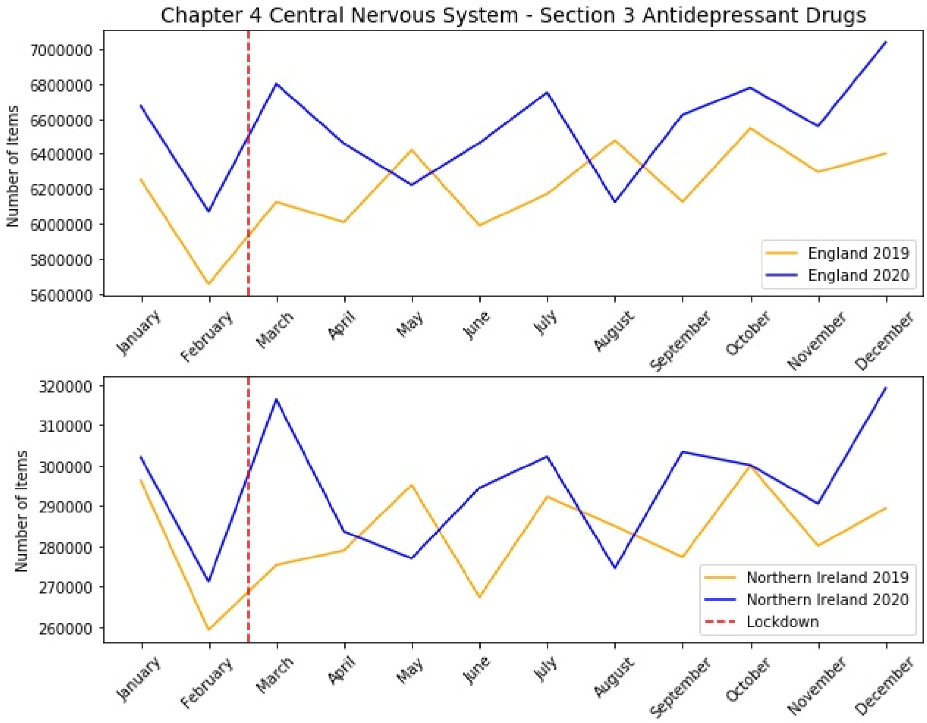


FIGURE 8.3: Prescription trends for BNF Chapter 4 Section 3 (Antidepressants) in England and Northern Ireland. Approximate date of the start of the lockdown is indicated as a red dotted line. The Number of items prescribed for each month are reported on the last day of the month.

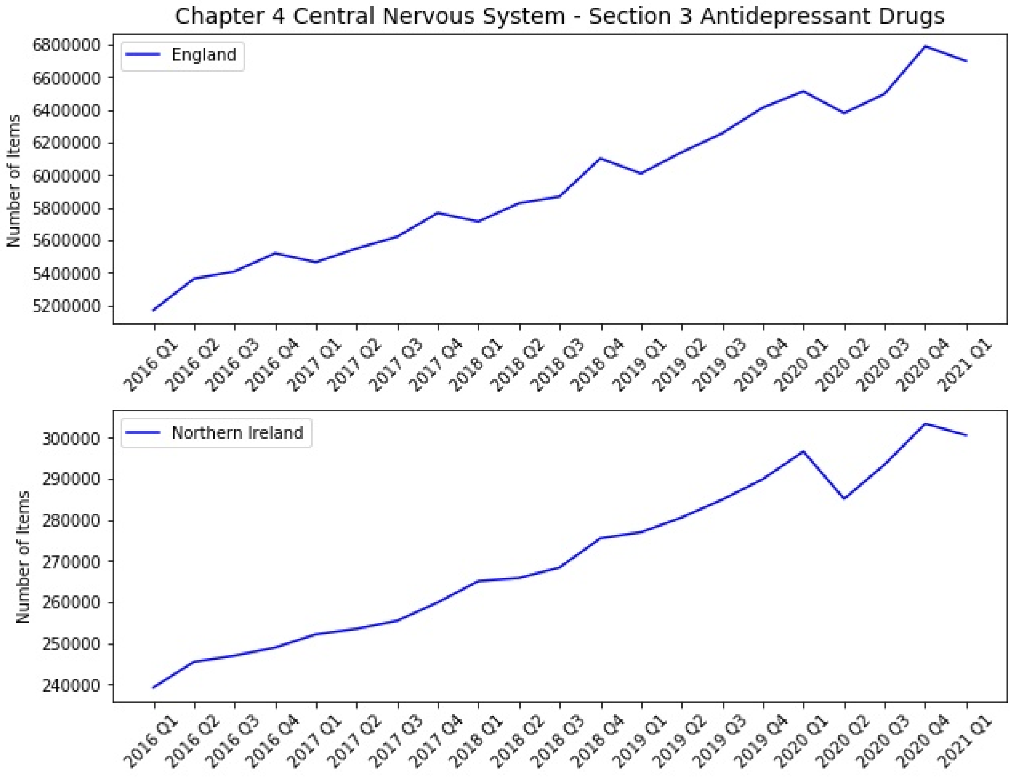


FIGURE 8.4: Prescription trends for BNF Chapter 4 Section 3 (Antidepressants) in England and Northern Ireland over 5 year period (2016-2020)

BNF Chapter 5 (Infections) did not follow the same 'Peak, Trough, Recovery' pattern seen in most other BNF chapters (Figure 8.5). Whilst having the initial peak followed by a trough in both regions, prescribing in this chapter did not recover to previous levels instead plateauing at a lower level than the previous year - Antibacterial drugs were the main contributors to this trend. Chi Squared tests on the rise in prescribing from February to March each year showed that these rises were also statistically significant for both regions. T-Tests on the whole years prescribing compared to the previous year, showed that there was no statistically significant change in prescribing in England (p=.476) whilst prescribing in Northern Ireland was observed as significantly lower statistically than that of the previous year (p=.00835).

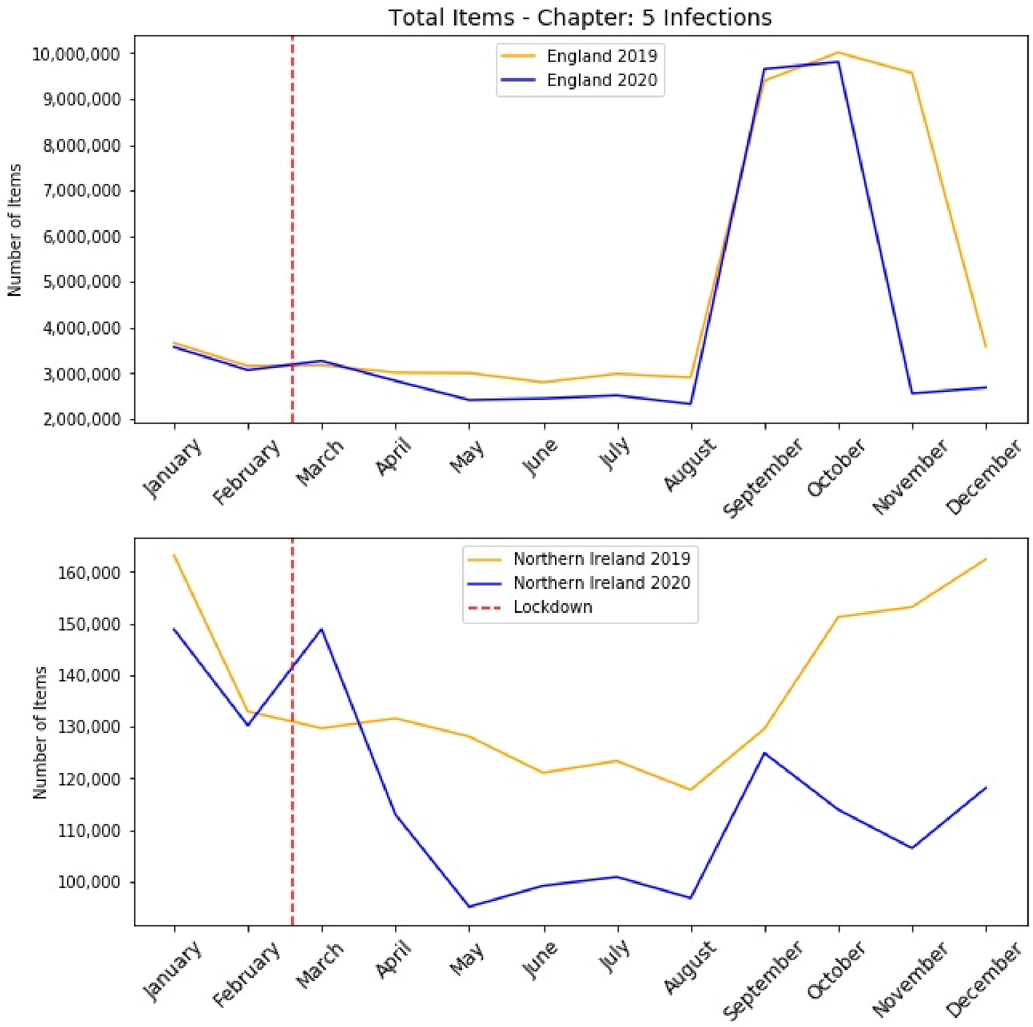


FIGURE 8.5: Prescription trends for BNF Chapter 5 (Infections) in England and Northern Ireland. Approximate date of the start of the lockdown is indicated as a red dotted line. The Number of items prescribed for each month are reported on the last day of the month.

## 8.4 Metropolitan and Non-Metropolitan practices

Examining the effect of the lockdown on Metropolitan and Non-Metropolitan practices (Figure 8.6) it is evident that both archetypes experienced the same peak, trough, recovery pattern seen at national level. In comparing the peaks, the number of items prescribed in Metropolitan practices rose from 852,957 in February 2020 to 1,028,973 in March, a rise of 176,016 (20.64%). The number of items prescribed in Non-Metropolitan practices rose from 2,399,985 in February to 2,895,948, a rise of 495,963 (20.67%). These figures show that the lockdown had the same effect on both archetypes at national level.

Breaking the figures down further, similar trends can be observed for both archetypes at BNF chapter level (Appendix H). The peaks in prescribing in March 2020 are also similar in magnitude for both archetypes (Table 8.1) for most chapters. The exceptions to this are chapters 3 (Respiratory System), 14 (Immunological Products & Vaccines) and 20 (Dressings) which had higher prescribing peaks in Non-Metropolitan practices than in Metropolitan practices. Metropolitan practices showed a higher peak in prescribing in chapter 19 (Other Drugs and Preparations).

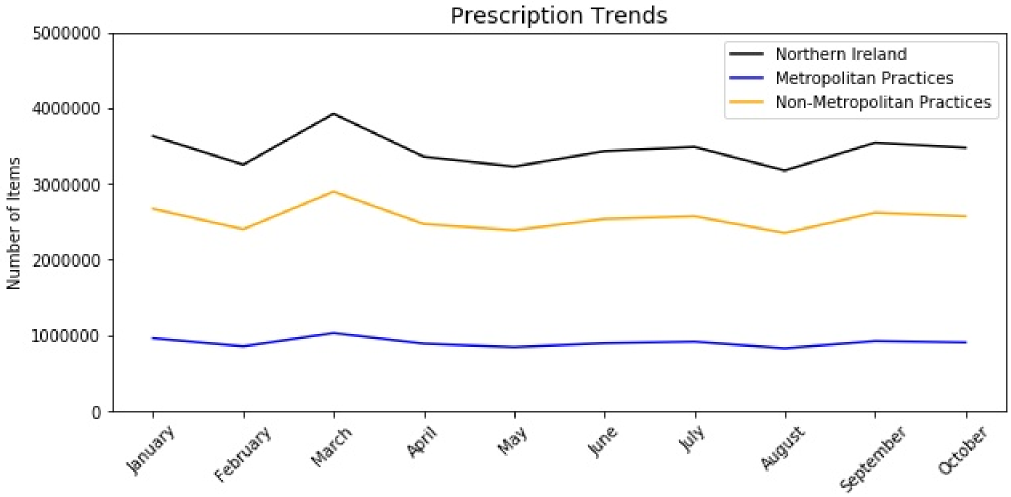


FIGURE 8.6: Prescription trends for Metropolitan and Non-Metropolitan practices as a result of the COVID-19 pandemic and first national lockdown.

TABLE 8.1: Percentage change in prescribing between February and March 2020 for Metropolitan and Non-Metropolitan practices by BNF chapter

| BNF Chapter | Metropolitan | Non-Metropolitan |
|---|---|---|
| 3 Respiratory System | +55.06% | +62.42% |
| 8 Malignant Disease & Immunosuppression | +26.90% | +25.07% |
| 6 Endocrine System | +24.99% | +23.24% |
| 2 Cardiovascular System | +21.53% | +20.96% |
| 19 Other Drugs and Preparations | +20.33% | +12.09% |
| 21 Appliances | +18.09% | +18.08% |
| 7 Obstetrics | +17.87% | +16.73% |
| 23 Stoma Appliances | +17.71% | +14.62% |
| 4 Central Nervous System | +17.64% | +17.13% |
| 1 Gastro-Intestinal System | +17.33% | +16.61% |
| 5 Infections | +15.91% | +13.82% |
| 9 Nutrition and Blood | +14.96% | +15.60% |
| 11 Eye | +14.66% | +15.05% |
| 12 Ear, Nose and Oropharynx | +10.48% | +14.13% |
| 10 Musculoskeletal & Joint Diseases | +9.10% | +7.43% |
| 13 Skin | +7.93% | +9.43% |
| 15 Anaesthesia | +3.45% | +6.06% |
| 20 Dressings | +3.14% | +11.73% |
| 22 Incontinence Appliances | -7.27% | -10.83% |
| 14 Immunological Products & Vaccines | -49.66% | -40.84% |

## 8.5 Discussion

Although it was tacitly accepted at the start of the COVID-19 pandemic in the UK that prescribing patterns and in particular antibiotic prescribing was likely to be altered, there has been little published work on how this has panned out. This is the first study to look at changes in general practice prescribing across all BNF categories associated with the COVID-19 pandemic in the UK focusing on the period February – June 2020 when both Northern Ireland and England were following the same lockdown measures.

We found that the 'Peak, Trough, Recovery' trend did not hold for antibiotic prescribing. A recent study in Scotland (Malcolm et al., 2020) looked solely at antibiotic prescribing and identified the same pattern we did of a spike in antibiotic prescribing during March 2020 followed by sustained decline to below 2019 levels. They found that the number of prescriptions, primarily for respiratory infections, fell by 34% compared with the corresponding week in 2019. The authors suggested that the peak was due to 'just in case' prescriptions and proposed a series of explanations for the subsequent fall to below 2019 levels. They suggested, as we do that factors in this sustained reduction in antibiotic prescribing include improved hygiene, reduced transmission of infection due to lockdown and a reluctance to attend GP surgeries.

Similarly, another article suggested that the rise seen in the prescribing of Antidepressants in England during 2020 and since lockdown supported the belief that predictions of the effect of the lockdown on mental health were correct (Armitage, 2020) although the methodology used has been challenged (Goldacre et al., 2021) suggesting that the growth seen is a normal progression following previous years trends. We found that the overall rise in the prescribing of antidepressants, year on year, was not statistically significant in either region, and although the rise in prescribing in Northern Ireland was greater than that seen in England, it was not evident that overall prescribing could be attributed to the COVID pandemic.

A consistent 'Peak, Trough, Recovery' pattern of prescribing has been identified, following the first Covid-19 lockdown, across the BNF categories and suggest this is not altogether unexpected. GPs, in anticipation of the lockdown, planned to prescribe enough medication to keep their patients supplied for the following two months or more. This not only gave patients peace of mind, but also reduced the number of contacts and the footfall in surgeries during the lockdown. This preplanning contributed to the peak observed in March 2020 and to the trough seen in the following two months when patients presumably had sufficient supplies of medication.

The March peak in prescribing was twice as high in Northern Ireland than in England when compared with the same period the previous year. It is possible that the widespread use of e-Prescribing in England (Adeley, 2006), whereby prescriptions issued at GP surgeries are printed directly at nominated pharmacies, played a significant part in this. GP surgeries in England would not have the added pressure of reducing footfall for this reason as experienced by their Northern Ireland counterparts.

The largest percentage rise in prescribing between February and March 2020 in both England and Northern Ireland was seen in BNF Chapter 3 (Respiratory System) with Sections 1 (Bronchodilators), 2 (Corticosteroids (Respiratory)) and 3 (Cromoglycate and related therapy and leukotriene receptor antagonists) being the main contributors. It is possible that this was due to patients who had at one time been on inhalers, but were no longer taking them, requesting them again "just in case". There was also an issue with a much-circulated myth that patients who had previously suffered from respiratory problems should ask for "rescue packs" containing a 5-day supply of a corticosteroid and an antibiotic which could be started if the patient developed breathing issues (Reuters, 2020). This may also have contributed to the rise seen in March 2020.

An initial increase in the number of antimicrobial drugs prescribed in March (more pronounced in Northern Ireland) was followed by a large and sustained fall in antimicrobial prescribing over the following 5 months compared to the 2019 levels. The initial peak in antimicrobial prescribing may have been due to patients requesting "rescue packs" and patients with chronic respiratory conditions such as COPD obtaining a stock of antibiotics in reserve. The subsequent sustained reduction in antibiotic prescribing is likely to be multifactorial; first because of patients now having a reserve supply and second as a result of the lockdown itself with fewer infections through reduced socialisation and greater emphasis on personal hygiene from campaigns such as the 'Wash Your Hands' advice issued to fight COVID-19 (Maillard et al., 2020). Patients may also have avoided going to their GP during this period as they were afraid of contracting COVID-19 whilst attending their local surgery.

Contrary to the belief that "antibiotic stewardship" (a co-ordinated approach to promote the appropriate use of antibiotics given the loss of effectiveness caused by overuse) would be an early casualty of new working practices in health care during the lockdown, with doctors having to "play safe" during telephone consultations, antimicrobial prescribing decreased after an initial spike in March and prescribing remained consistently lower than 2019 levels until at least August 2020. Malcolm et al. (2020) found a similar pattern in Scotland. Studies into the effect of remote consultations on antibiotic prescribing have so far been inconclusive (Han et al., 2020). In addition to COVID-19 hygiene measures and the move toward teleconsultations (Li et al., 2020), patients deliberately staying away from their GP surgery for fear of catching COVID-19 may have also played a part in the reduction of antimicrobial prescribing. The marked reduction in the prescribing of items listed in Chapter 14 of the BNF (largely vaccines) may reflect this reduced attendance in person at GP surgeries as well as a reduction in travel.

Electronic prescribing may have reduced the need for large quantities of drugs to be prescribed in England at the outset of the lockdown as happened in Northern Ireland. The larger quantities of drugs prescribed in Northern Ireland had the potential to put strain on stocks held in pharmacies and could constitute a waste of resources with patients receiving large supplies of medication they no longer require. This should provide further impetus for Northern Ireland to introduce an e-Prescribing system. Respiratory System medications had the highest peak in March 2020 reflecting public anxiety with the respiratory nature of the COVID-19 symptoms. Misinformation about "rescue packs" may also have stimulated demand. The reduction in antimicrobial prescribing despite fears the opposite would happen is interesting in terms of patient need and will inform the debate about our over-reliance on these agents in the past.

## 8.6 Limitations

Although all records in the English data sets have BNF Codes, the NI data sets have approximately 0.1% of the BNF Chapter data missing. Another limitation is that the explanations and interpretations of the observed patterns and trends require further research to validate them.

T-Tests were performed on two years monthly data for each region which limits the statistical power of the tests however they are useful to look at the relative differences between the regions.

## 8.7 Conclusion

Examining prescription trends during the start of the COVID-19 pandemic, and in particular the months during the first national lockdown (March - June 2020), revealed a pattern of peak, trough and recovery in almost all BNF chapters in both Northern Ireland and England. One exception to this trend was chapter 5 (Infections) which covers most antibiotics. Contrary to the belief that antibiotic stewardship would suffer as a result of the pandemic, prescribing in this chapter did not return to the same levels seen previously. Prescribing of antidepressants (Chapter 4, Section 3) did follow the peak, trough, recovery model and levels were seen to rise over the year. Contrary to speculation that this rise in prescribing was the result of the pandemic and subsequent lockdown, examination of prescribing trends over the previous 5 years showed the rise in prescribing to be part of an ongoing trend. Prescribing trends of the two identified archetypes in Northern Ireland (Metropolitan and Non-Metropolitan) generally followed the same overall trends observed at national level.

The overall aim of this study was to assess the extent to which open data could provide insight and value. Using open prescription data, this study has shown that GP practices can be categorised not only by their location but by their prescribing trends and sphere of influence shown by the distance traveled by patients to dispense prescriptions. Using these new categorisations it has been possible to compare the different prescribing trends and investigate factors contributing to these differences and ultimately investigate the effect of a national lockdown on prescribing. These insights have been possible using data analytics and programming skills learnt over time and reflect the inquisitive nature of academic researchers in collaboration with medical professionals. In Chapter 9 the possibility of opening these data to a wider audience by providing a data science tool to facilitate the analysis of the data by citizens will be explored.

# Chapter 9

# Development and evaluation of a 'Citizen Science' dashboard

> "A common mistake that people make when trying to design something completely foolproof is to underestimate the ingenuity of complete fools ."
>
> Douglas Adams,
> Mostly Harmless

The previous chapters of this thesis have dealt with the creation of a novel data set in order to examine the prescribing trends of GP practices in Northern Ireland and to develop a method of categorising these practices based on not only their geographical location but by their influence as described by their relationship with pharmacies dispensing prescriptions. This analysis has been performed using open prescription data which is available to anyone who wants to access it. Whilst this is a first step in the opening up of health data, the next logical step is the provision of a software solution to facilitate the analysis of these data. To this end, this chapter details the development of a prototype dashboard facilitate this and to analyse the data gathered during a 1 month survey of volunteers using the dashboard to analyse GP prescription data.

The objectives of this chapter are to:

- Develop a prototype data science tool to enable citizens to analyse GP prescription data.
- Recruit citizens to contribute to a study using the developed data science tool.
- Analyse the user data captured during the study on the searched performed and feedback left by participants to assess the extent to which citizens contribute to citizen science.
- Analyse the responses captured from an exit survey to assess participants attitudes to citizen science and the data analysis tool.

## 9.1 Background

Citizen science is defined as a form of open collaboration where members of the public participate in the scientific process to address real-world problems in ways that include identifying research questions, collecting, and analysing data, interpreting results, making new discoveries, developing technologies and applications, and solving complex problems (Haklay et al., 2021).

Citizen science has become more prevalent in all areas of research ranging from health and biomedical research (Wiggins and Wilbanks, 2019) to the assessment of sub-tropical reefs in Australia (Roelfsema et al., 2016) and astronomy with Zooniverse (Simpson et al., 2014) gathering information from citizens all over the globe. More recently citizen science has been used to gather information on the emotional responses to the Coronavirus pandemic (Branova, 2020).

Citizen science projects fall into one of three categories - contributory (i.e., led by experts), community-led or co-created all of which benefit scientists and citizens equally. Scientists benefit from increased research capacity with both parties benefiting from better knowledge (Den Broeder et al., 2018). In fact, evidence suggests that citizens participating in citizen science not only contribute to scientific knowledge but through the process learn about the content and processes of science itself. Participants in the Zooniverse project were asked to participate in a science quiz to test this theory. Results showed that more actively engaged participants performed better (Masters et al., 2016). There is also some evidence that citizen science can contribute positively to social well-being by influencing the questions that are being asked thus leading to new directions for research (Bonney et al., 2016).

Whilst Personal and Public Involvement (PPI) is becoming standard across Digital Health and Wellbeing research in the UK, in general Citizen Science is rare in public health with the largest part of Citizen Science work being carried out in the fields of biology, conservation, and ecology. One area of research affecting public health is research into antimicrobial resistance. With the overuse of antibiotics there is a need for new antimicrobial compounds to be identified. Millions of microbes exist but due to the low probability of any of these antimicrobial compounds being useful in medicine, there is a relatively low amount of funding available for the initial discovery of products compared with later stage antimicrobial development projects. In his thesis, Ethan Dury employed citizen science to isolate antibiotic producing bacteria from soil (Drury, 2020). To expand initial antimicrobial discovery activity beyond soil microbes, with limited funding being available, researchers created a dynamic, long-term, public-engagement activity in the form of the citizen science project 'Swab and Send'. This was launched in early 2015 and was designed to enable individuals to decide where and what to sample to try and find bacteria or fungi, with the potential to produce antimicrobial products against a range of bacterial and fungal indicator strains (Roberts, 2020).

## 9.2 Methods

No previous studies were identified using citizen science as a means of analysing open prescription data however there has been previous work that has used prescription data to corroborate citizen science (Vigo et al., 2018). Permission to undertake this survey was sought from the Ulster University Research Ethics Filter Committee and approved on 2 May 2021.

### 9.2.1 Development of SQL database

Ideally the entire Local Data Store would have been made available for interrogation using the developed dashboard interface. Unfortunately, budget constraints made this impossible, therefore a snapshot of one year's data was chosen for the study.

To develop a prototype data science tool, it was important to consider the data available and the possible ways in which it could be interrogated. The local data store developed for this project was examined to identify entity types suitable for the development of a relational database to be used as the backbone of the tool. The LDS data was then transformed into suitable tables for use in a relational database. Figure 9.1 illustrates the transformation.

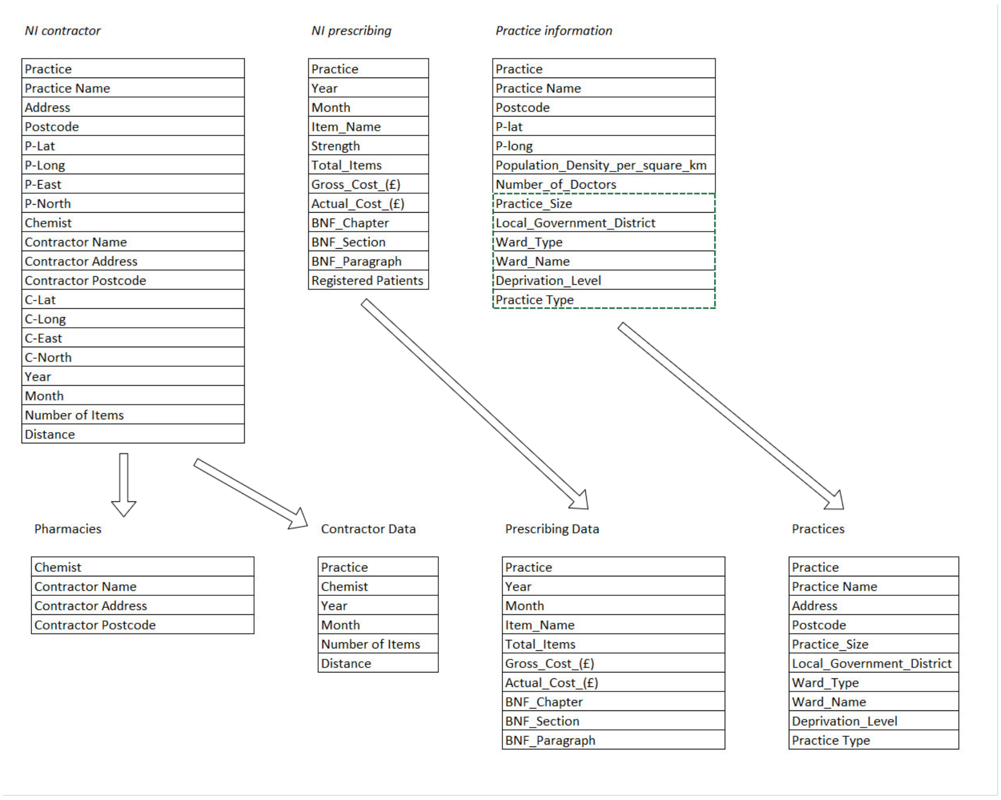


FIGURE 9.1: Workflow for extraction of Local Data Store data into tables for upload to SQL database

Five entities were identified: practices, pharmacies, contractor data, prescription data and British National Formulary (BNF) Categories. BNF categories were added to provide descriptions of the various chapters and sections for users not familiar with BNF categorisations. Whilst the number of items prescribed could be obtained from either the contractor or prescription data sets, each provided different data options for the user to explore. In addition to these five entities, unrelated tables were used to capture any inputs from participants' interaction with the tool and responses to the exit survey. The resulting Entity-Relationship diagram (Figure 9.2) formed the basis of the final MySQL database on which the dashboard interface was built and which was hosted on commercial web space. The full data dictionary is available at Appendix I.

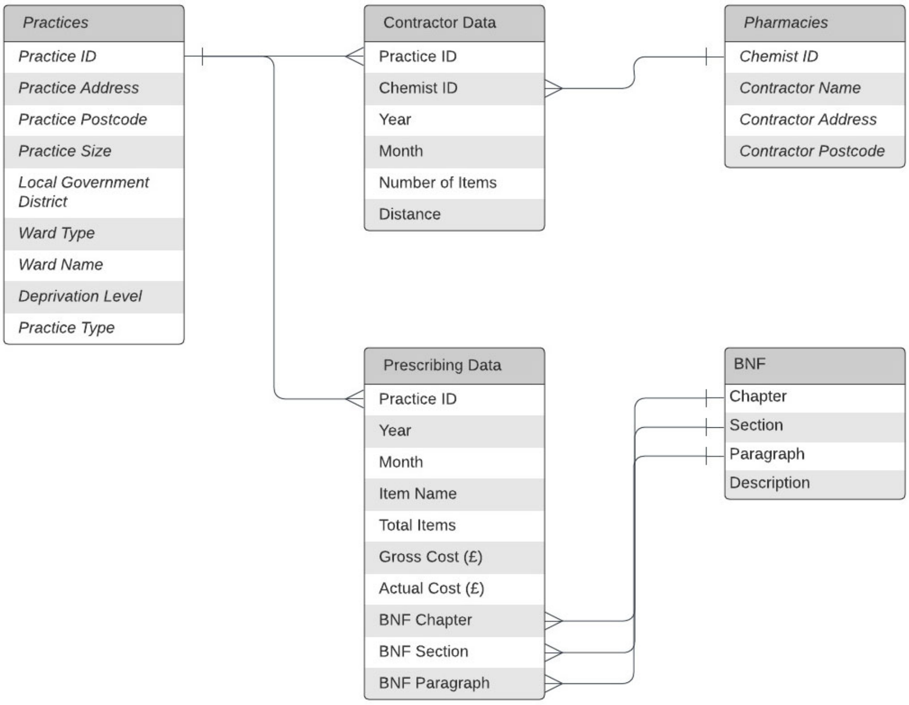


FIGURE 9.2: Entity Relationship diagram of MySQL database table structure.
Graphic created using the Free version of LucidChart [1]

### 9.2.2 Development of dashboard interface

The user interface / dashboard was developed using PHP and was designed to be as user friendly as possible. Initial wire-frame mock-ups for the dashboard and results page formed the basis of the final solution. In addition to these pages, an index page, demographics page, exit survey and thank you page were created. In order to ensure anonymity, a unique session number was assigned to each session and this was used to record the participants' searches using the dashboard and their responses to the exit survey. Figure 9.3 shows the web pages along with the flow of

[1] LucidChart Available at: https://www.lucidchart.com

data with examples of the pages available at Appendix J. Due to time constraints, not all of the available variables were used for the resulting dashboard interface.

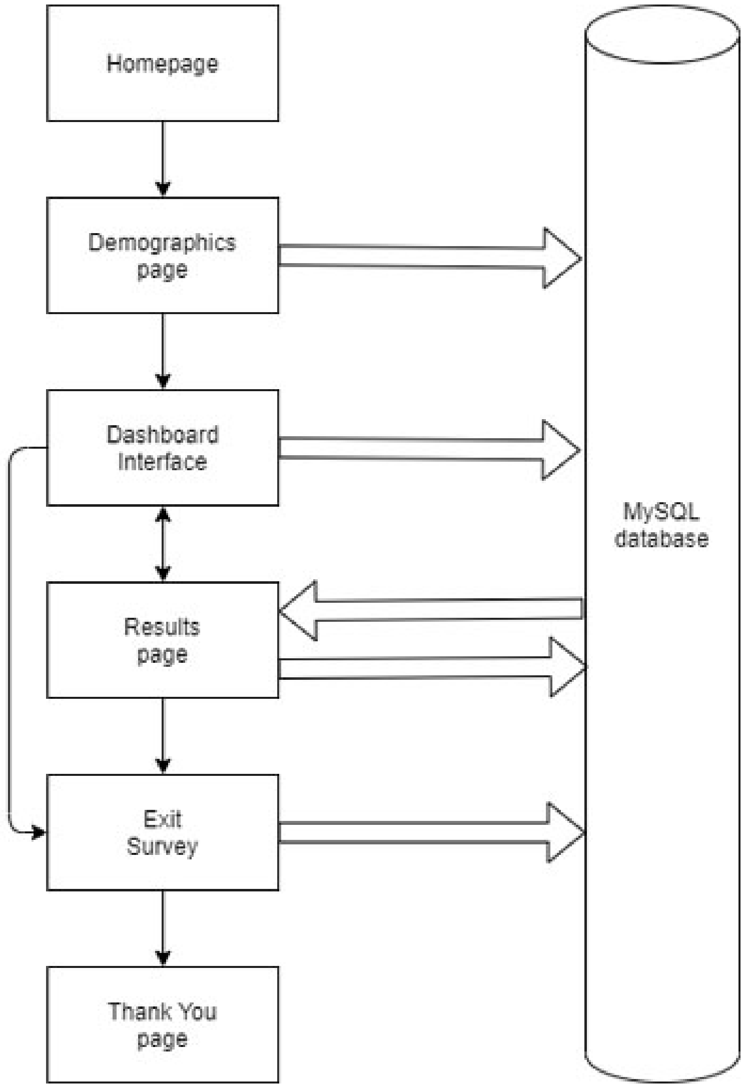
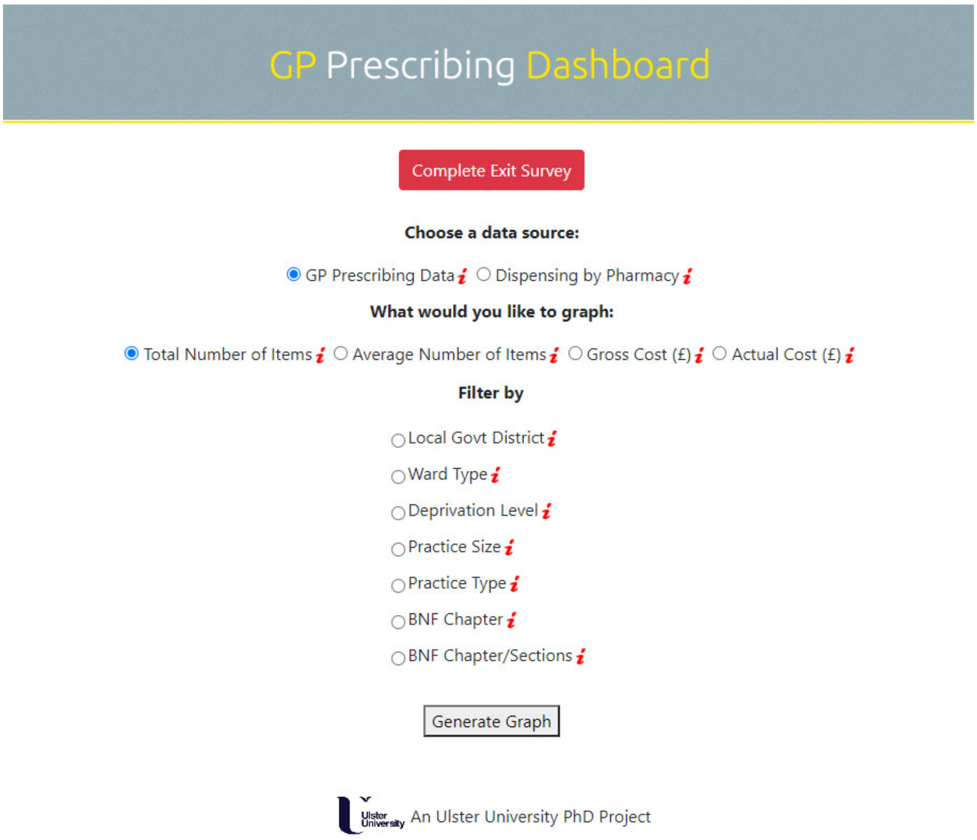


FIGURE 9.3: Structure and data flow of GP prescribing dashboard application.

**Homepage** - This page introduced a potential user to the survey giving them some information on what the study was about and who they could contact for more information. A link to an external Youtube video was included on this page with the video demonstrating the use of the dashboard and some of the features. At the bottom of this page a consent button allowed participants to proceed. Once the consent button had been pressed, a unique SessionID was allocated and the participant was directed to the demographics page.

**Demographics page** - This page asked participants to enter their age and sex. Age was defaulted to blank and sex to 'Prefer not to say'. Completion of these metrics was completely optional although only ages 18 to 65 were allowed as valid entries to the age question. A 'Proceed to dashboard' button sent the captured metrics to the MySQL database before displaying the dashboard.

**Dashboard** - The dashboard (Figure 9.4) consisted of three main sections: the data source, the metric to be graphed and the filters available to be used on that metric. An information icon was available beside every option providing a pop up window explaining the background to the variable.



GP Prescribing Dashboard

**Choose a data source:**

● GP Prescribing Data ℹ ○ Dispensing by Pharmacy ℹ

**What would you like to graph:**

● Total Number of Items ℹ ○ Average Number of Items ℹ ○ Gross Cost (£) ℹ ○ Actual Cost (£) ℹ

**Filter by**

○ Local Govt District ℹ
○ Ward Type ℹ
○ Deprivation Level ℹ
○ Practice Size ℹ
○ Practice Type ℹ
○ BNF Chapter ℹ
○ BNF Chapter/Sections ℹ

An Ulster University PhD Project

FIGURE 9.4: Screen capture of GP prescribing dashboard interface.

Participants were asked to 'Choose data source' with GP prescribing data and Dispensing by pharmacy being the two options. The default choice was GP prescribing data. Each source provided different options for what variables were available to graph; GP prescribing data allowed Total number of items, Average number of items, Gross costs and Actual costs to be graphed whilst Dispensing by pharmacy allowed Total number of items, Average number of items, Number of pharmacies, Number of practices and Distance traveled to be graphed. The Total number of items and Average number of items figures from the different data sources are almost identical and produce similar graphs. For both data sources, Total number of items is the default option to be graphed. On choosing what the user wants to be graphed a number of filters can, if required, be chosen. An option to proceed to the Exit survey is also available.

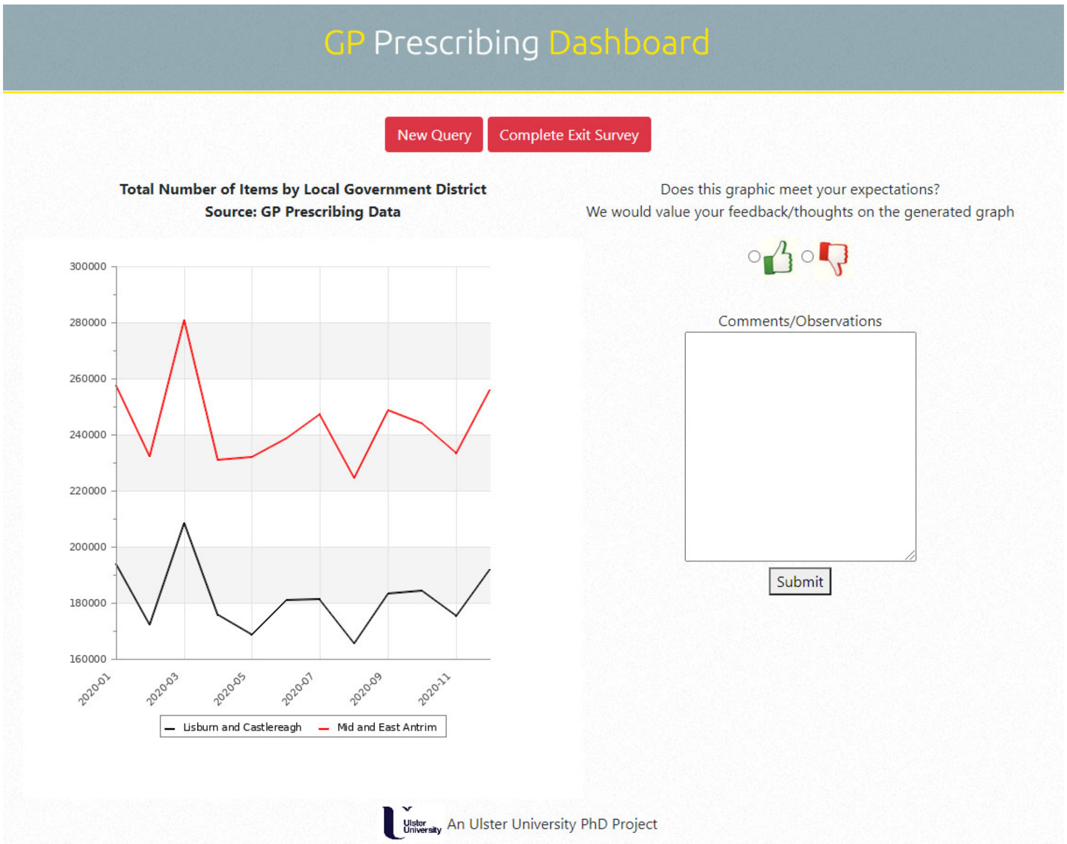
On choosing a filter, the user is presented with the available options associated with it. Table 9.1 lists the filters and options available. Each of the filter options chosen would be represented as a line on the resulting graph. If no filters were used a singe line representing the graph option (e.g. Total number of items) would be displayed. On pressing the ’Generate Graph’ button the chosen options were translated into a JavaScript Object Notation (JSON) object and along with the SessionID and current date and time uploaded to the feedback table in the database before being passed to the results page.

TABLE 9.1: Filters available on dashboard with associated options.

| Filter | Options |
|---|---|
| Local Government District | Belfast, Lisburn & Castlereagh, Newry Mourne & Down, Ards & N. Down, Antrim & Newtownabbey, Mid & East Antrim, Causeway Coast & Glens, Mid Ulster, Armagh Banbridge & Craigavon, Fermanagh & Omagh, Derry & Strabane |
| Ward Type | Rural, Urban, Mixed - rural/urban |
| Deprivation Level | Low (Q1), Low/Medium (Q2), Medium/High (Q3), High (Q4) |
| Practice Size | Single-Handed (1 Doctor), Small (2 Doctors), Medium (3-4 Doctors), Large (5+ Doctors) |
| Practice Type | Metropolitan, Non-Metropolitan |
| BNF Chapter | Choice of multiple BNF chapters (See Appendix D) |
| BNF Chapter and Sections | Choice of single BNF chapter with multiple sections associated with that chapter (See Appendix D) |

**Results page** - The graph displayed on the results page was created using JPgraph Version 4.3.4[2], free software for creating PHP driven charts. The JSON object passed from the dashboard was decoded and a SQL query was generated for each filter option chosen. Where a filter was not selected, a single SQL query was constructed relating to the graph option chosen. Each SQL query was sent to the database with the reply forming the data for a line on the graph. This process is described using pseudo-code in Appendix K. Along with the resulting graph, the user was presented with options to comment on the graph (Figure 9.5), generate a new query or complete the exit survey. No further ’drill down’ of the data was possible at this point although choosing the new query option allowed the user to adjust the query options in order to analyse the data from a different view point. No limit was set on the number of queries generated during each session.

[2] JPGraph, Available at: https://jpgraph.net/



GP Prescribing Dashboard

Total Number of Items by Local Government District
Source: GP Prescribing Data

Does this graphic meet your expectations?
We would value your feedback/thoughts on the generated graph

An Ulster University PhD Project

FIGURE 9.5: Screen capture of GP prescribing dashboard results page.

**Exit Survey** - The exit survey used the Likert scale to gauge participants' opinions on the concept of citizen science, the dashboard design, the results generated, categories available, how often they would use the filters available and how often they would potentially use additional filters. The Likert scale is often referred to as a satisfaction scale measuring the degree to which the participant agrees with a given statement (McLeod, 2019). In this case a five point scale was used with an additional 'Not Applicable' option. Appendix J, Figure J.6 provides a full list of questions on the exit survey. On completion of the exit survey, the responses were uploaded to the survey table in the database and the user was directed to the Thank you page ending the session.

## 9.3 Results

A total of 152 user sessions were recorded during the month of September 2021. Of these, 106 users completed the exit survey. Where any question was not completed by a user, this was treated as a return of 'Not Applicable'. Figure 9.6 shows the breakdown of participants by sex. Almost half 52 (49.1%) of the user sessions were completed by female participants with 43 (40.6%) by male participants and 11 (10.4%) by participants who preferred not to disclose their sex.

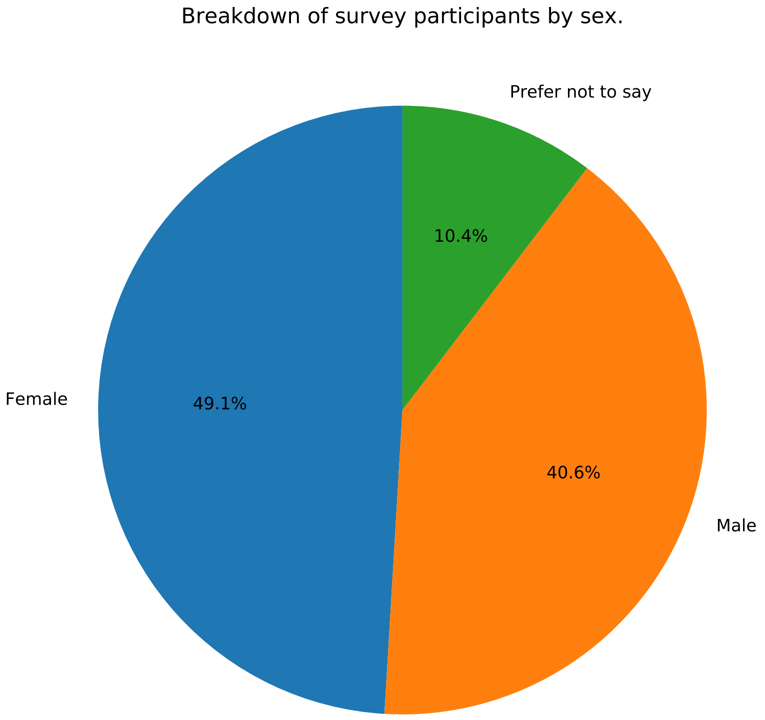


FIGURE 9.6: Breakdown of survey participants by sex.

The age of participants ranged between the lower limit of 18 and the upper limit of 65 with a concentration of users in the 18-30 year range. As the main advertising of the study consisted of an email to staff and students of the university, it is not surprising that this concentration in the lower age range exists. The mean age of users was calculated as approximately 36 years old.

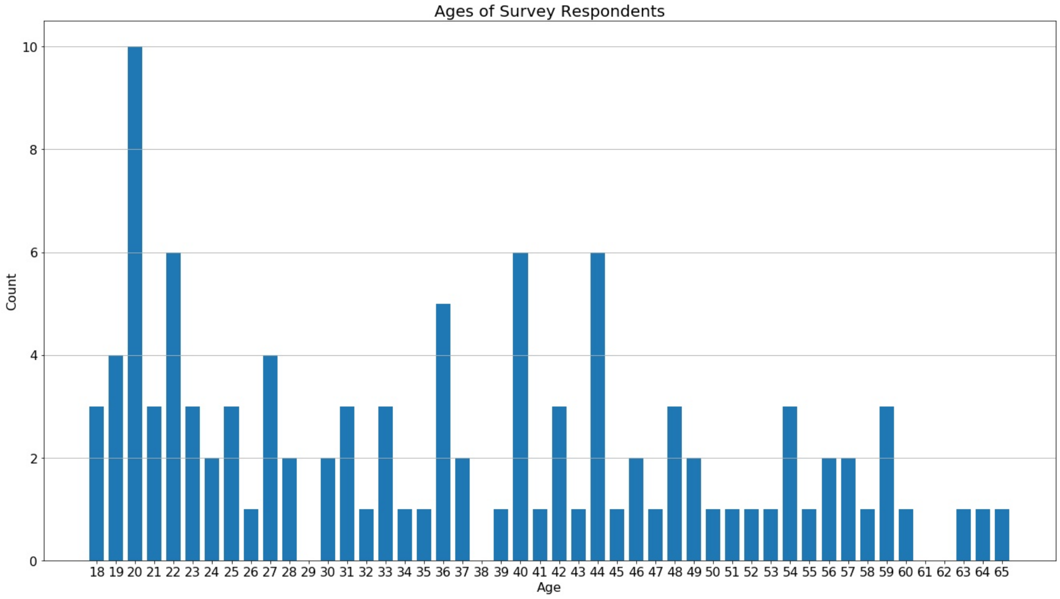


FIGURE 9.7: Breakdown of survey participants by age.

Analysis of participation of the survey showed that the majority of participants took part on a Wednesday (Figure 9.8) which coincided with the launch of the survey. Participation was almost entirely during normal working hours and early evening (Figure 9.9).

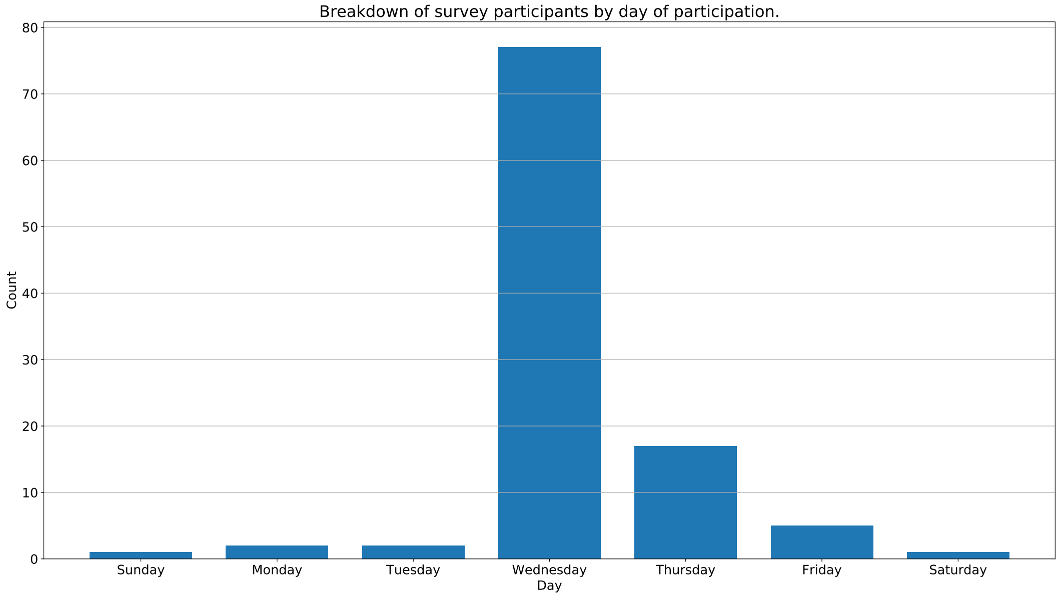


FIGURE 9.8: Breakdown of survey participants by day of participation.

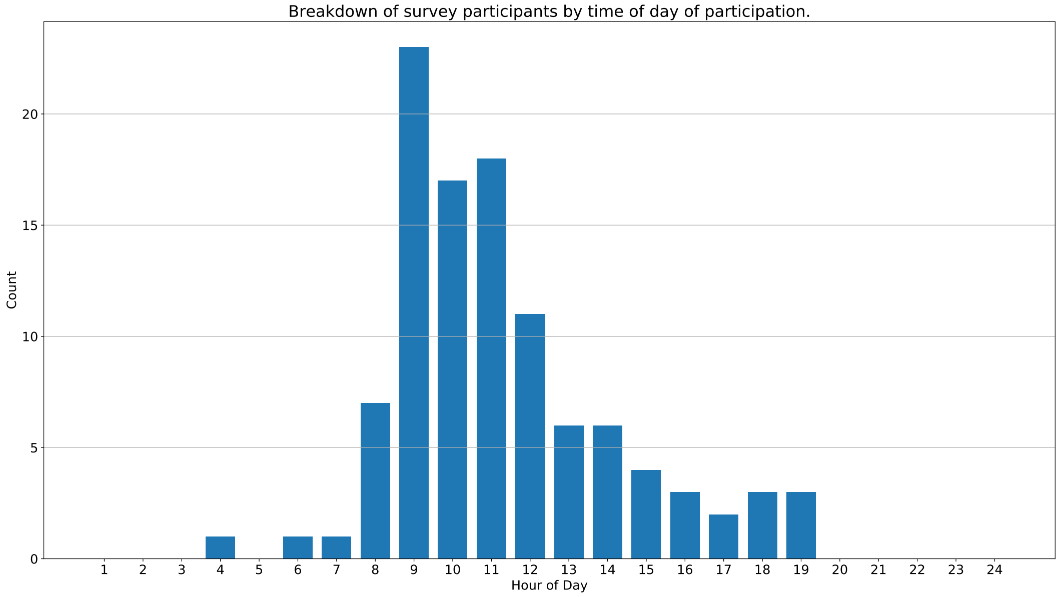


FIGURE 9.9: Breakdown of survey participants by hour of day.

The date and time was recorded for each session when the user pressed the consent button on the dashboard. Similarly, the date and time was recorded when the user accessed the exit survey. From these two date stamps, it was possible to calculate the time each user spent interrogating the dashboard. Figure 9.10 shows the distribution of time spent on the dashboard. This ranged from just over 1 minute to almost 25 minutes with the majority of users staying between one and two minutes.

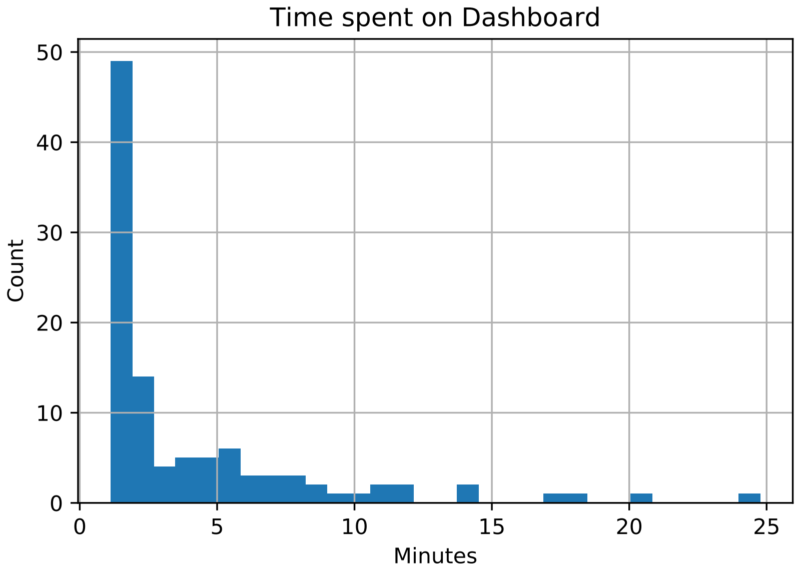


FIGURE 9.10: Distribution of time spent by users on the dashboard.

### 9.3.1 Analysis of query log

A total of 238 searches were made using the dashboard interface during the study with an average of 2.3 searches per user session and a maximum of 22 searches during a single session. Of these, 193 (81.1%) were performed on the prescribing data set and 45 (18.9%) on the dispensing by contractor data set (Figure 9.11). A bias may have been introduced into these figures by the fact that GP prescribing data was the default option on the dashboard interface.

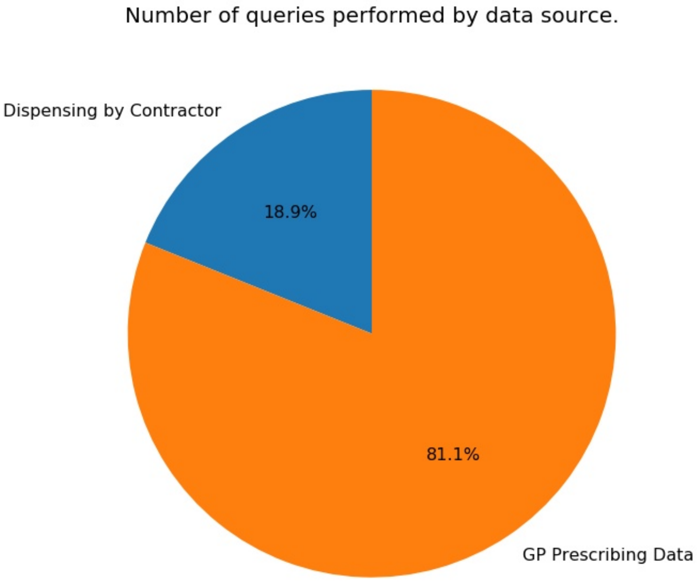


FIGURE 9.11: Number of queries performed by data source.

Analysis of the categories used to produce graphs (Figure 9.12) showed that participants showed most interest in the count of items prescribed with total number of items being graphed 118 times (49.6%) and average number of items 47 times (19.7%). There was also reasonable interest in how much medications were costing with actual costs being graphed on 34 occasions (14.3%) and gross cost on 22 occasions (9.2%). Again, a bias may have been introduced into these results as Number of items was the default option for choice of metric to be graphed.

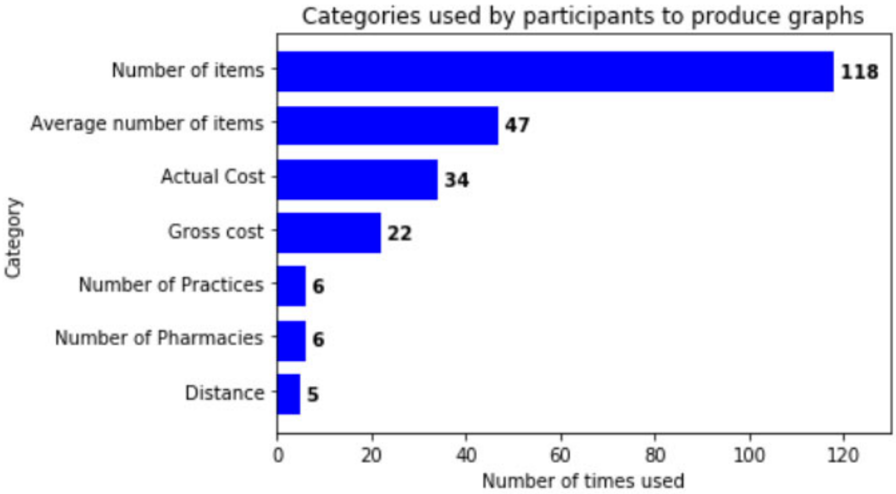


FIGURE 9.12: Categories used by participants to produce graphs.

Of the 238 searches performed during the study, 201 of these used an additional filter to examine trends. Figure 9.13 shows the number of times each filter was used. The filter that was used the most was Local Government District (43.5%) followed by deprivation level (19.5%). No default was set for filters on the dashboard interface so no bias can be attributed to these results from that perspective. Local Government District was the first filter available to choose and that, along with the fact that this was probably the most familiar division to users, may have contributed to its popularity. It was of note that BNF Chapter was only used for 21 (10.5%) searches and BNF section for 6 (3.0%). This may reflect the fact that citizens are not familiar with this system of categorising medications.

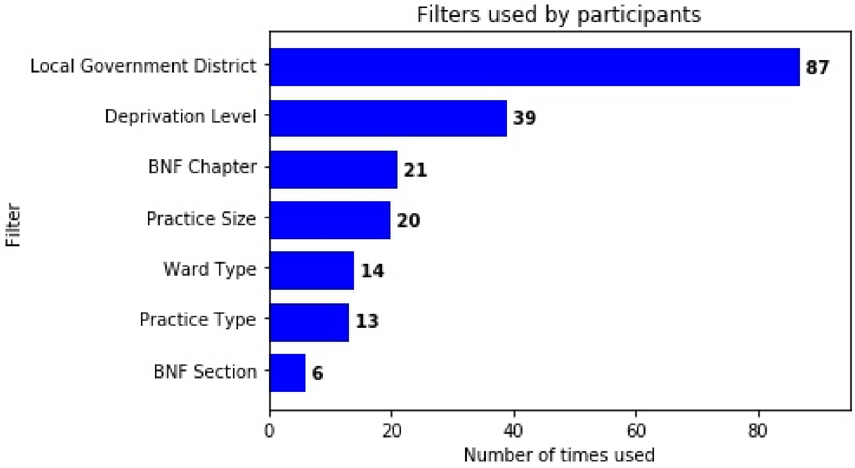


FIGURE 9.13: Filters used by participants.

### 9.3.2 Active / Non-Active participants

Two types of participants were identified in this study, defined by whether or not they left comments/observations on the resultant graphical output from the dashboard. The first are active participants, being those users who contributed comments/observations, the second being non-active participants, being those who did not leave comments. Of the 106 participants in the study, there were 16 (15.1%) active participants and 90 (84.9%) non-active participants. Comparing the time spent on the dashboard and the number of searches performed by the two categories (Figure 9.14) showed that active participants spent longer and performed more searches than non-active participants. On average, active participants spent 10.9 minutes (10 minutes 54 seconds) on the dashboard performing 5.4 searches whilst non-active participants spent an average of 3 minutes performing 1.7 searches. As multiple hypothesis tests have were employed a corrected Bonferroni alpha value was calculated as 0.05/2 = 0.025. The subsequent independent t-tests indicated that the two categories (active and inactive participants) were significantly different statistically (p<0.0025).

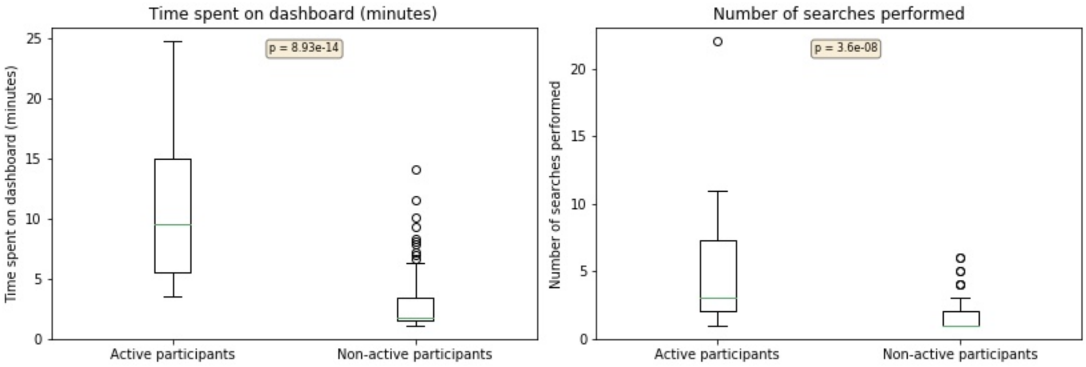


FIGURE 9.14: Boxplots comparing time on dashboard and number of searched performed by active and non-active participants

Analysis of both types of participant by sex and age (Figure 9.15), where younger participants were classified as being 18-25 years old and older being over 25 years old, showed that in the younger category, there was little difference in the active participation between male (4.7%) and female (3.8%) whilst there was a higher proportion of older male active participants (14.0%) than female (9.6%). Active female participants did however spend more time on the dashboard than their male counterparts with the average active female spending 13 minutes compared with almost 10 minutes for the active male participant (Figure 9.16). Active older participants spent slightly longer on the dashboard (11 minutes) than their younger counterparts (9 minutes) (Figure 9.17).

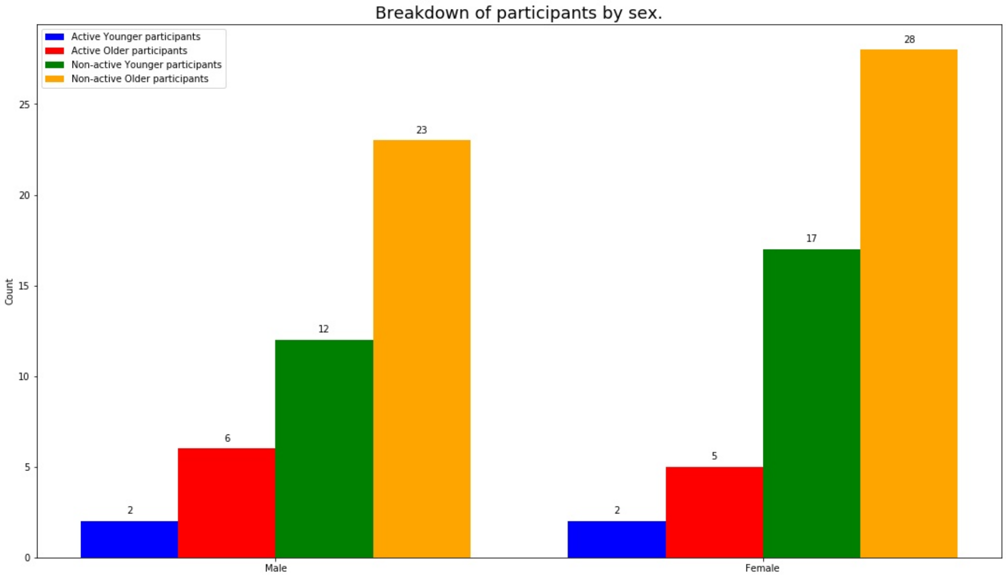


FIGURE 9.15: Breakdown of active and non-active participants by age and sex.

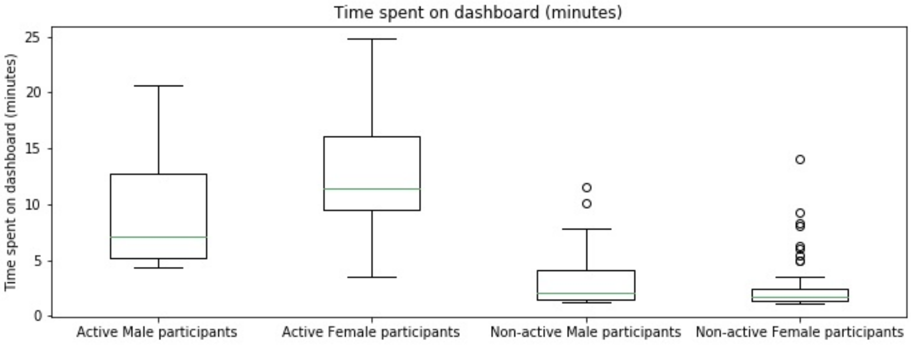


FIGURE 9.16: Boxplots of time spent on the dashboard of active and non-active participants by sex.

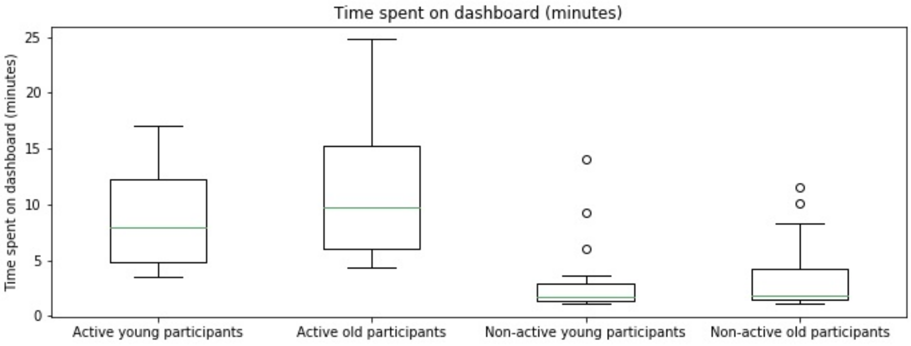


FIGURE 9.17: Boxplots of time spent on the dashboard of active and non-active participants by age.

Comments left by active participants related to 32 for the searches performed during the study (Appendix L). The majority of these comments related to design enhancements being requested for the results page. Of the comments left, 43.8% (14) were left by male participants, 46.9% (15) by female participants and 9.4% (3) by participants who preferred not to reveal their sex. The majority of comments left in each group were left by older participants (92.6% (13) Male over 25, 73.3% (11) female over 25 and 100% (3) of 'prefer not to say' over 25).

> "This looks great. I will suggest that both horizontal and vertical axis are label to make the graph more descriptive." (Male, 40)
>
> "The Y axis should be formatted 9,999,999 would be helpful" (Male, 49)
>
> "Could it be possible to make interactive graphs, where you can zoom in to reduce the axis range? I glanced at this graph quickly and assumed the decline in item number in 2020-08 was at zero." (Female, 22)
>
> "A combined graph with deprivation levels and BNF sub-types would be useful so that population trends can be seen. This could give an early indication of evolving trends e.g. in drug misuse." (Female, 44)

Others commented on the trends observed on resulting graphs.

> "can see the lockdown trends, panic in march 2020" (Male, 50)
>
> "I thought it would be fairly consistent from month to month." (Female, 22)

Whilst there were some comments requesting that additional clarification be added to explain the trends observed.

> "Would maybe be good to see potential explanations for peaks and troughs in data e.g. 2020-03 spike due to COVID-19 cases rising etc." (Female, 22)
>
> "Not clear what the average refers to - average number per GP? Per person?" (Female, 45)

### 9.3.3 Analysis of survey responses

Analysis of the responses recorded from the exit survey were very encouraging with an overall positive reaction being conveyed regarding citizen science and the GP prescribing dashboard. Participants were categorised as younger (18-25) or older (25+) to observe differing perspectives of the two age groups. Participants' attitudes to the concept of citizen science (Figure 17) were quite positive with 5 of the six questions receiving most ratings as 5 (Very Much) agreeing with the given statements. The final question received an overall majority rating of 4. Analysis of the breakdown by age group showed that the positive attitude to citizen science did not depend on age with the older age group (over 25) only slightly higher proportionally than the younger age group (18-25).

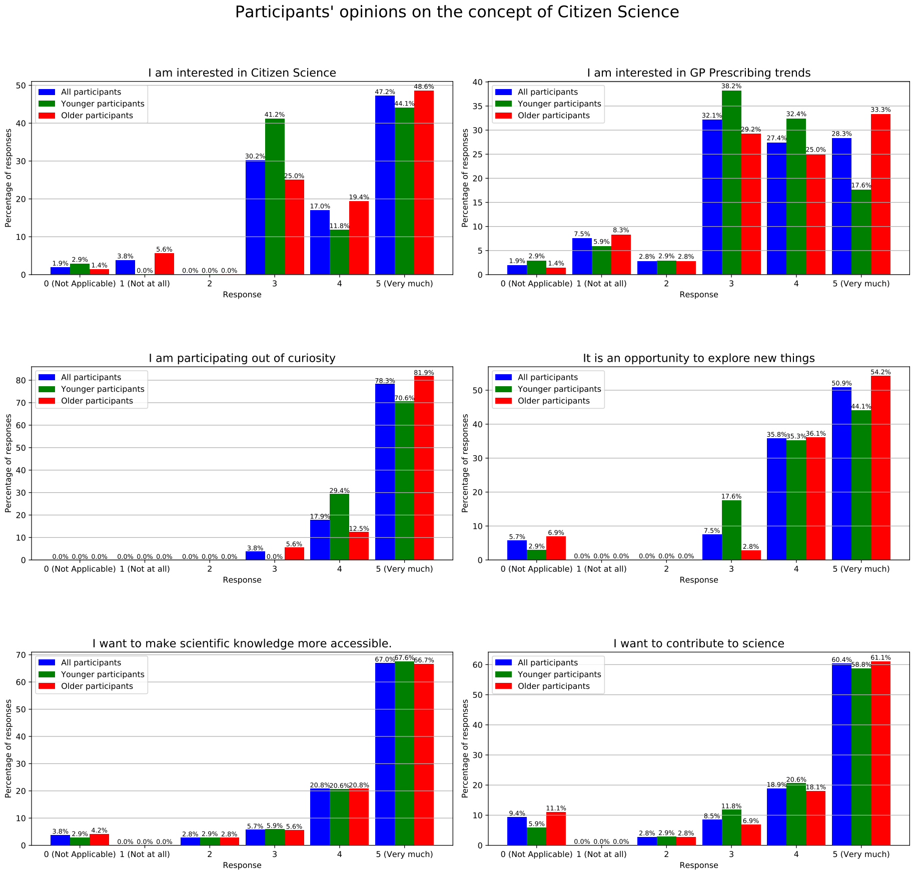


FIGURE 9.18: Participant's opinions on the concept of Citizen Science.

The general reaction to the dashboard interface (Figure 18) was also positive with most users agreeing that it was easy to use, would be useful for the analysis of data and could provide valuable insights into prescribing trends. Users were split on the question of the level of explanations provided regarding the variables. More description and less technical descriptions were requested in response to the "What features could be added" question. Analysis of the responses showed that younger users generally found the dashboard easier to use than older users.

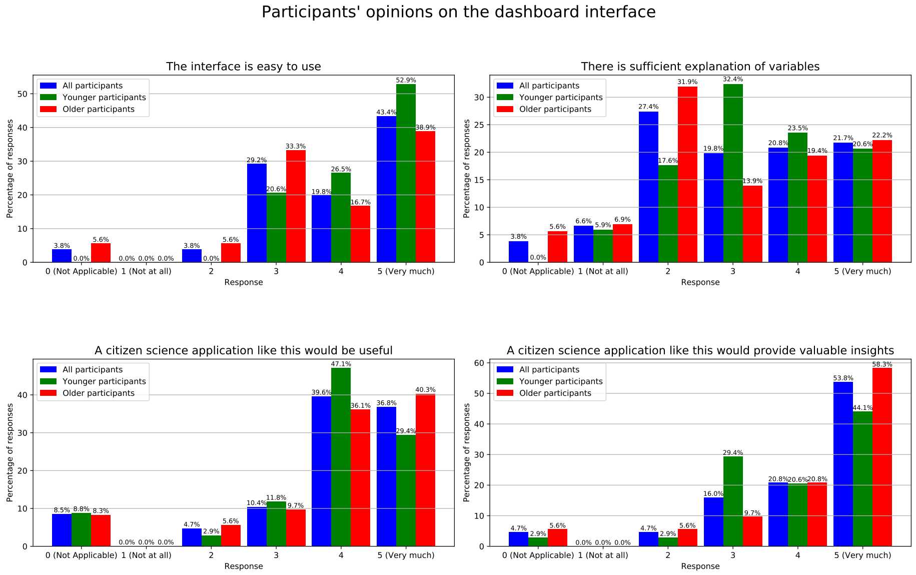


FIGURE 9.19: Participant's opinions on the dashboard interface.

In considering the results produced by the dashboard (Figure 19), users generally found that the resulting graphs were self-explanatory, and the majority did not feel that they would need assistance with interpreting the results although over 27% of participants felt that there was not enough explanation of the variables used. Most users felt that the results were trustworthy but also felt that they would like access to the raw data. Notably, despite the interest shown in the provision of raw data, most users did not want to perform the calculations themselves. As with the other categories, younger users were slightly more represented in the majority responses than the older participants.

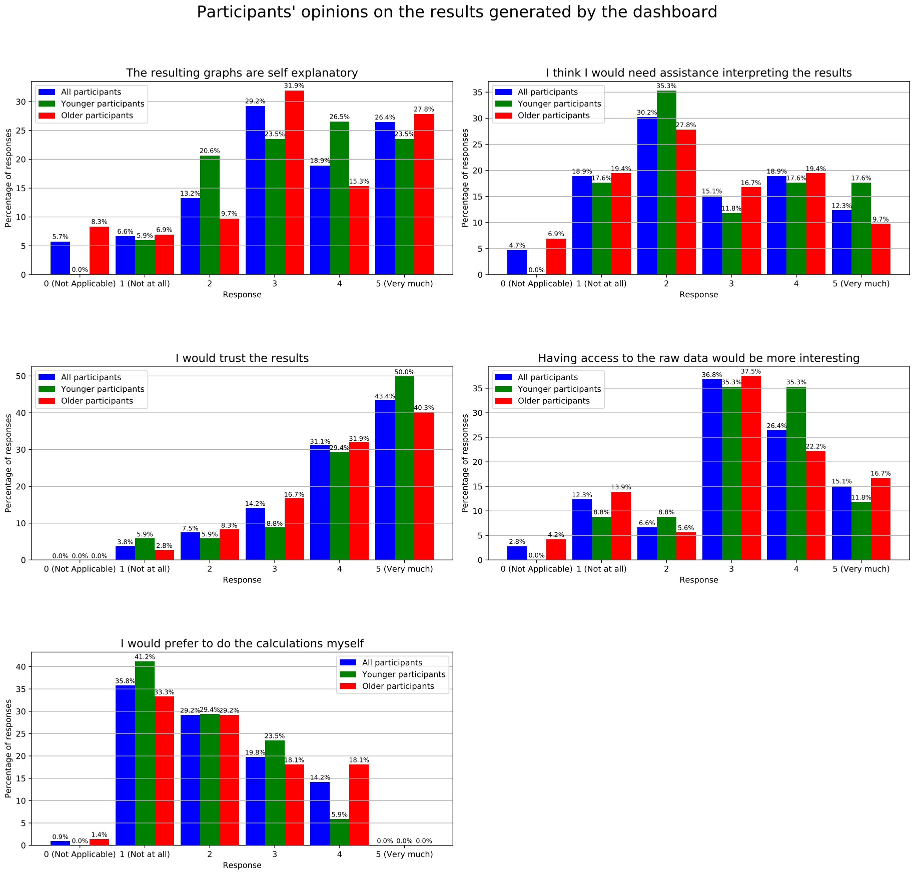


FIGURE 9.20: Participant's opinions on the results generated by the dashboard.

In considering the usefulness of the categories available to interrogate (Figure 20) most users rated all the categories as 5 (very) useful. This shows that despite the low counts for searches on the "Number of Pharmacies", "Number of practices" and "Average distance travelled" seen in the analysis of actual searches performed during the study that users felt they were still useful. Younger users were slightly more represented in the majority responses than the older participants.

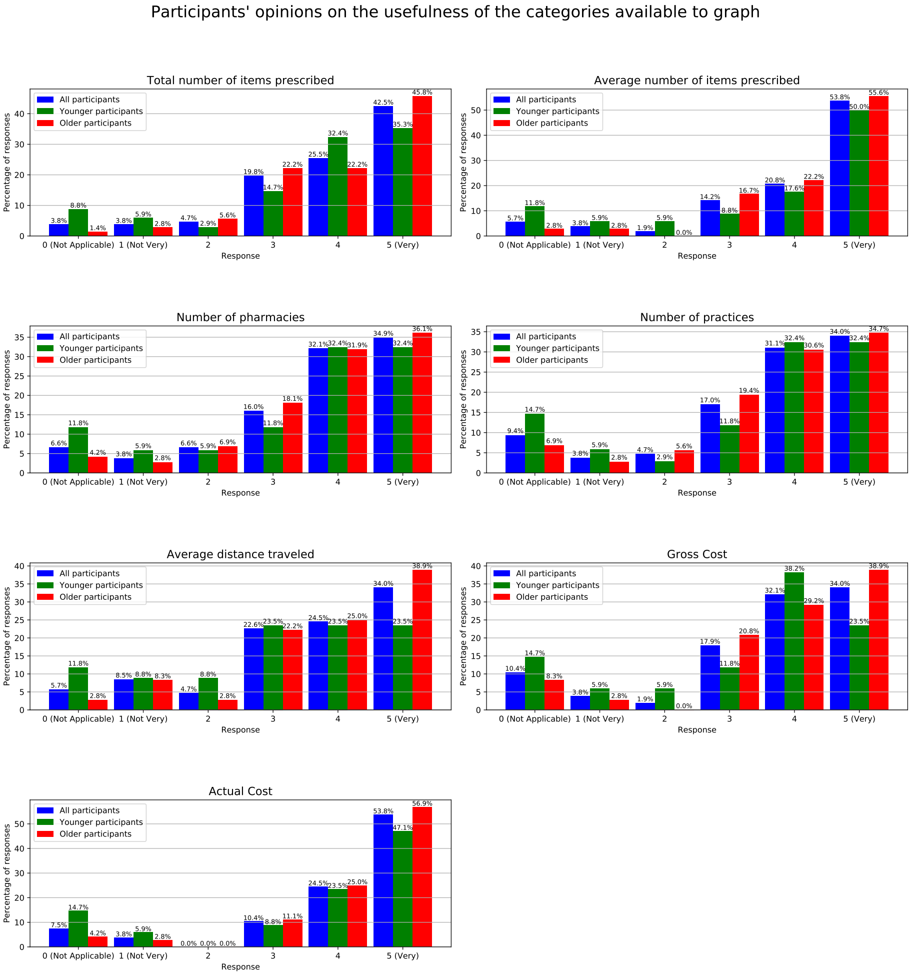


FIGURE 9.21: Participant's opinions on the usefulness of the categories available to graph.

The available filters (Figure 21) generally received a positive reaction to how often users would use them. Local Government District was seen to be the filter that users would use the most and probably reflects the familiarity of the category. Lower ratings, whilst still positive, for the other filter categories most likely reflect the fact that these categories are not as familiar with users. Younger users were more represented in the majority responses than the older participants.

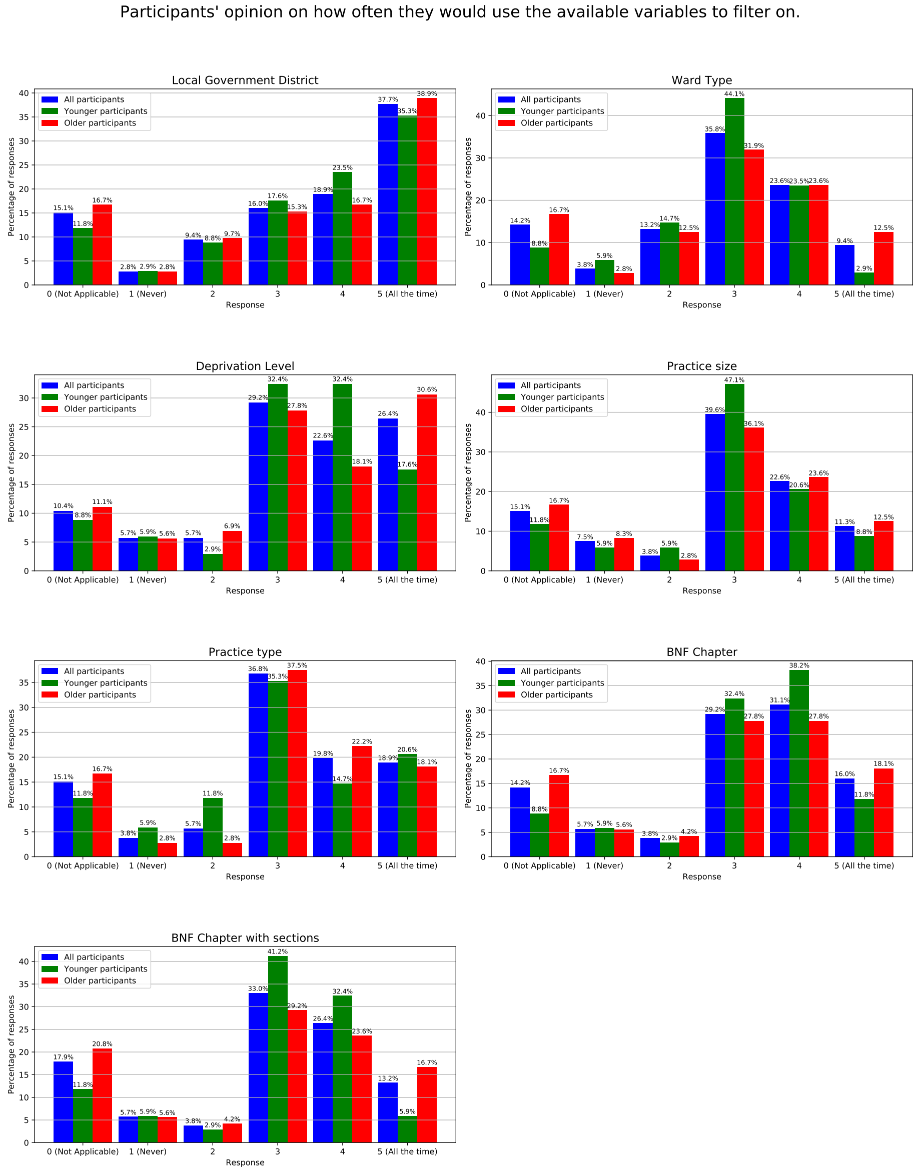


FIGURE 9.22: Participant's opinion on how often they would use the available variables to filter on.

Regarding the provision of additional filters (Figure 22), opinions were mixed. Whilst the general opinion was that the offered categories; Ward (by name), Practice (by name), Pharmacy (by name) and Postcode would be useful there was a high proportion of responses regarding these categories being "not applicable".

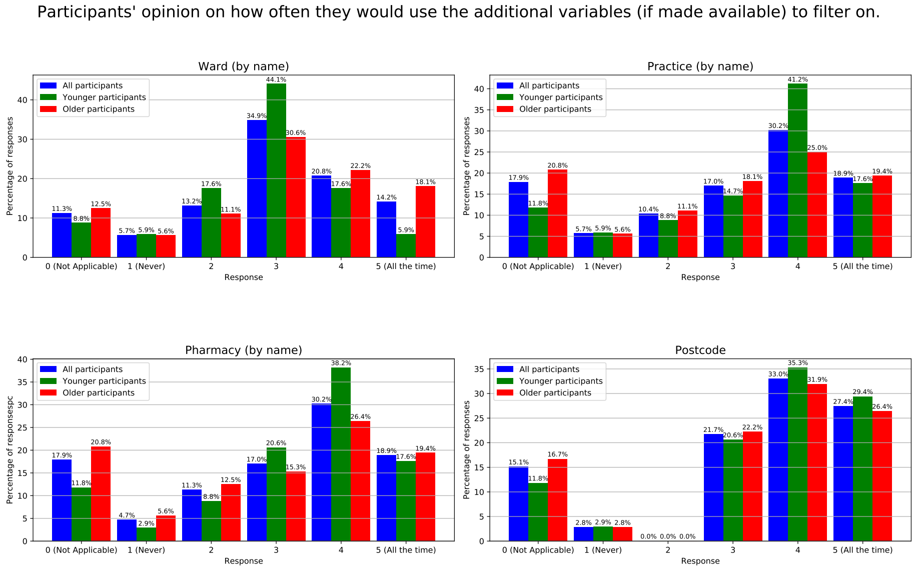


FIGURE 9.23: Participant's opinion on how often they would use the additional variables (if made available) to filter on.

**Do you feel there are any risks associated with citizen science?** - There were 24 responses to this question, the full text of which can be found at Appendix M. Almost half (45.8%) of the participants responded that they felt there were no risks. Of the remaining responses, risks were identified regarding the misinterpretation of the data,

> "Generalised findings can be used to misrepresent service level issues. For example, prescribing trends are higher in Belfast because it has a bigger population, if this is not explicitly stated, citizens may perceive this as Belfast is more reactive in prescribing medications compared to other areas for example. Likewise, general prescription trends do not indicate differences in levels of prescriptions for pain medication, treatment of psychopathology, chronic or acute illness etc. That information would let citizens see if there is a difference in prescription trends for issues relating to their personal health concerns increasing their investment in the process. " (Female, 31)

> "Jumping to conclusions, the general public may not know how to interpret data in a systematic, unbiased way and take from it what they want to see." (Female, 22)

> "Misunderstanding/uninformed interpretation and spread of misinformation/disinformation as a result" (Female, 65)

and the possibility of the data changing the behaviours of the public.

> "Can change behaviours, e.g. patients demanding something from their GP when they have seen the trends elsewhere." (Male, 56)

**What features could be added?** - There were 20 responses to this question with a variety of requests (Appendix M). These ranged from simple design issues;

> "Comma separator for amounts on the y-axis greater than £999 i.e. '3,750,000'." (Male, 39)
>
> "More description of the results graph. Needs axis labels and graph title needs to explain the results better" (Female, 23)
>
> "Manual re-scaling of graphs might be helpful. Having the variables named directly on the plots would make them easier to use at a glance" (Male, 27)

requests for advanced features,

> "Interactive graphs i.e. ability to adjust axis ranges or compare over a longer period of time, annotations of potential reasons for changes in a graph." (Female, 22)
>
> "Alerts to prescription requests among the local population, this may provide and early warning to prescribers as to trends in misused drugs eg cyclizine." (Female, 44)
>
> "It would be nice not to have to look up each graph separately, and have it more along the lines of a Power BI dashboard." (Female, 45)

assistance with the interpretation of the data,

> "Explanations of the data sets and trends" (Male, 50)
>
> "A guide to interpret data" (Male, 18)

and more in depth analysis options.

> "Classifications by drug" (Male, 56)
>
> "More friendly UI. Better descriptions of variables. Variables that give more useful information such as a breakdown of what is being prescribed by category for example. Costs per category etc. Let people know where the money is being spent and possibly give insight into where the major problems are in healthcare." (Male, 44)
>
> "Breakdown by age group" (Male, 59)

## 9.4 Discussion

Analysis of the searches performed using the GP prescribing dashboard show that there was considerably more interest in the GP prescribing data set allowing users to examine prescribing trends for numbers of items prescribed and associated costs. Whilst there was little interest in the "Dispensing by contractor" data set, these data could be leveraged to facilitate requests for more granular analysis at GP surgery level. Whilst no default was set on the filter options on the dashboard, the first option (Local Government District) was the filter option most used. This may reflect its position at the top of the filter list or because it is a category that users are familiar with. The second most used filter (Deprivation Level) most likely reflects the current interest in the effects of deprivation on health in general. Whilst the "BNF Chapter" and "BNF Chapter with sections" filters were not used to a large degree, this may reflect the lack of knowledge regarding the classifications of medicine in the UK. As only 15.1% of participants recorded comments relating to the graphical output from their searches, large numbers of users would be required to gain useful feedback. In retrospect, the 'Thumbs Up, Thumbs Down' element of the results page did not provide any useful data and would be removed from any future version of the dashboard. Almost half of the participants felt that there was no risk in opening this type of data to the public in the form of a citizen science application. Of the other participants, the general feeling was that the risks were that the data could be misinterpreted due to lack of domain knowledge or that patient behaviour could possibly change with demands being made for medications based on trends in other locations. The dashboard was designed as a 'proof of concept' and users reacted positively to using it with several requests for additional features. Design 'fixes' such as better labelling of axes and the use of commas in the presentation of numeric labels over 999 could be implemented easily. Requests for more advanced features such as interactive graphs, multiple graphs, and manual re-scaling of graphs, whilst possible, would require a complete redesign of the dashboard. Whilst a guide to interpreting the data and explanation of the data sets is possible, the provision of an explanation of the trends would require prior analysis and possibly introduce a bias to any analysis performed thereafter. Requests for a deeper level of analysis such as "Classifications by Drug" and "Cost per category" are possible although the former would require additional information to be added to the database. Requests for metrics such as "Breakdown by age group" are not possible with the current data sources. With 15.1% of the participants of this study being classified as active (i.e. left comments / observations relating to the searches they performed), only 6 of the 32 (18.75%) could be considered as scientific observations on the results with the rest being observations on the format of the results page itself and requests for clarification. Analysis of the feedback indicates the need for more data literacy and scientific literacy in the public domain to maximise the chances of citizens providing useful contributions when engaging in citizen science. Active participants spent over three times longer engaging with the dashboard, performing over three times more searches. Analysis of participants by age and sex showed that these were not major factors contributing to whether participants were classified as active or non-active. In order to understand the reasons why people participate in web surveys, Florian Keusch (Keusch, 2015) explores various theoretical frameworks in order to understand participants' motivations. Self-Perception Theory proposes that participants will engage with a topic if they feel that it is consistent with their view of themselves. Participants who consider themselves as having scientific minds are therefore more likely to engage with the dashboard and contribute observations on the resultant graphical output. Social Exchange theory proposed that every interaction can be traced to the perception of rewards or cost to the individual. Rewards are measured in the pleasure or satisfaction the participant feels in contributing to a survey whilst the costs are generally measured in the time spent completing a survey and the amount of 'brain power' involved. Active participants are therefore more likely to be those with the time to interact fully with the dashboard deriving satisfaction from contributing observations.

## 9.5 Limitations

It is important to recognise that any survey relying on voluntary participation will naturally be biased towards those who were interested enough to participate. As this survey was promoted within the University and on social media, it is likely that this also contributes to any bias in the results. The design of the dashboard interface may also have contributed to any bias as some of the dashboard options such as GP prescribing data as the source of data and Total number of items as the metric to be graphed were set as defaults.

## 9.6 Conclusion

The results of this study show that there is an interest in the concept of Citizen Science and relating to GP prescribing data. The proof-of-concept dashboard which was developed for this study received very positive feedback confirming that a data science tool such as this would be useful in the advancement of knowledge. Whilst, in general, users indicated their willingness to contribute to citizen science and felt that they understood the resulting graphs with no help needed to interpret them, only 15.1% left observations on the resulting graphs. This indicates that, if the purpose of the dashboard is to gather observations from citizens, it needs to attract a large number of users. Alternatively, if the dashboard is but a tool to be used by citizens, then the number of observations recorded is not as important as the access the dashboard provides for each citizen to GP prescription data. Similarly, most users spent between one and two minutes on the dashboard which, although indicates an interest in the subject, is unlikely to result in major insights. Users who spent 10 minutes and over interrogating the dashboard are more likely to provide observations on the resulting graphs.

Future work has already been indicated by the requests received by users during the study. The next version of the tool should provide more interactivity with the possibility of displaying more than one graph as output allowing direct comparisons to be made. The ability to adjust the scale has been requested along with more guidance on how to interpret the data. The prototype allowed access to one year's data (2020) and users were keen that this be expanded to multiple years. With this expansion of data, users also requested the ability to select specific time periods to be analysed. Some users felt that the dashboard interface could be more user friendly and further research into what this means in real terms should be carried out.

The next, and last, chapter will bring together the knowledge gained throughout this process, summarising the conclusions, suggesting further areas of investigation and presenting recommendations.

# Chapter 10

# Conclusions, future work and recommendations

"It doesn't matter what we want. Once we get it, then we want something else."

Lord Baelish,
Game of Thrones

The broad aim of this PhD was to provide new analysis on Open Health data combining Open Data sources with geographical data to identify patterns or trends which could be used by healthcare professionals to inform policy and clinical decisions. The specific objectives and associated research questions were:

**Objectives**

- OBJ1 - to identify Open Source health Data containing geographical references.
- OBJ2 - to identify and link relevant supplementary published data to create a novel data set.
- OBJ3 - to identify the types of GP practice using geographical location and relationship with dispensing pharmacies.
- OBJ4 - to analyse any differences in prescribing behaviours between identified types of GP practice.
- OBJ5 - to explore possible factors contributing to differences in prescribing behaviours of identified types of GP practice.
- OBJ6 - to explore the effect the COVID-19 pandemic, and in particular, the first national lockdown had on prescribing behaviours.
- OBJ7 - to assess public opinion on the usefulness of a GP prescribing dashboard in relation to being used as a citizen science tool.

**Research questions**

- RQ1 - Can GP practices be classified in terms of their location with respect to the location of the pharmacies dispensing their prescriptions?

  It was found that using features relating to both a GP practice's geographical location and their influence in the surrounding area in terms of their prescribing trends that GP practices in Northern Ireland could be classified as Metropolitan or Non-Metropolitan.

- RQ2 - If GP practices can be classified in terms of their location, are there differences in prescription behaviour between classes?

  Prescribing patterns were found to be similar for both Metropolitan and Non-Metropolitan practices although higher levels of prescribing was found in half of the BNF chapters for Metropolitan practices with the largest variation between archetypes resulting from the prescribing of Analgesics and Antidepressants.

- RQ3 - Is it possible to identify contributing factors to differences in classes?

  Deprivation and GP practice size were examined as possible factors contributing to the differences in prescribing levels observed. It was found that in general, higher prescribing levels were associated with higher levels of deprivation and, as there were more practices located in high deprivation areas within the Metropolitan area, this contributed to the overall higher prescribing seen in this archetype. Comparing practices in low deprivation areas for both archetypes showed that Non-Metropolitan practices had higher prescribing levels than their Metropolitan counterparts. Higher prescribing levels were also associated with small practices (2 doctors) in both Metropolitan and Non-Metropolitan practices. Large practices (5+ doctors) had the lowest prescribing levels within the Metropolitan area with Single-Handed practices having the lowest prescribing levels in Non-Metropolitan areas.

- RQ4 - What effect has COVID-19 and the National Lockdown had on prescribing behaviours?

  A pattern of 'peak, trough, recovery' was observed in the prescribing trends for almost all BNF chapters in both England and Northern Ireland in the months following the National lockdown imposed in March 2020. The peak in prescribing observed in March 2020 was attributed to the forward planning of GPs in issuing prescriptions of sufficient quantity to cover patients over the lockdown thus reducing footfall in their surgeries. Lower peaks observed in England were possibly due to the e-prescribing system available to English practices. The exception to the 'peak, trough, recovery' pattern was the prescribing of Antibiotics which failed to recover to their pre-lockdown levels.

- RQ5 - What interest do citizens have in prescribing behaviours?

  In order to study the general interest in both the concept of citizen science and the specific interest in GP prescribing data, a bespoke data analysis tool was developed in the form of an interactive dashboard to allow users to interrogate GP prescribing data for 2020. Results showed that there was a general interest in both the concept and the actual dashboard with users requesting that the underlying scope of data be extended and that additional features would improve their overall experience.

## 10.1 Limitations

The limitations identified during this study are:

- The data set on which categorisation was performed only became available from April 2018 limiting the available data to a 1-year period (April 2018 - March 2019) with only GP practices that operated during the whole period being included. No provision made for practices which closed or those opening during the period.
- This study uses number of items per registered patient as a proxy for the levels of sickness experienced. This may not be accurate as some GPs may be over prescribing or prescribing where anther GP would ask the patient to buy over the counter (e.g. paracetamol).
- It is likely that the majority of registered patients do not reside in the same super output area as the practice they attend. As this does not necessarily reflect the actual residential location of the patient receiving the prescription, it is assumed that patients will dispense their prescriptions at their local pharmacy meaning that distance traveled can be used as a proxy.
- The population density used as a feature in clustering practices is the population density of the super output area in which the Practice is located.
- In comparing prescribing levels at BNF chapter level, although all records in the English data sets have BNF Codes, the NI data sets have approximately 0.1% of the BNF Chapter data missing.

- The number of patients registered with a practice is reported each quarter. For the purpose of this analysis it was assumed that this figure would remain the same for the following two months which in reality is unlikely.
- An accurate count of the population is only taken every 10 years during the census. The figures used for comparison in this study are mid year estimates produced by the Office for National Statistics taking into account births, deaths and net migration since the previous year.
- In comparing prescribing levels by practice size, practices were categorised based on the number of registered doctors working in the practice. The assumption was been made that all of these doctors worked full-time which in reality is probably not the case. Also, as no data was available on the number of locums working in any practice, these were been ignored.
- Multiple hypothesis testing employed in this study increased the possibility of false discoveries. Whilst a corrected bonferonni alpha value would address this issue and may have reduced the number of statistically significant results, a reduced alpha value of 0.01 was used in stead of the generally accepted value of 0.05 instead.
- T-Tests performed on the two years monthly data for each region when comparing prescribing during the COVID-19 pandemic limited the statistical power of the tests however they were useful to look at the relative differences between the regions.

## 10.2 Strengths

An important strength of this study was that it employed national and regional data from robust electronic databases which collected all GP prescribing data and that it covered all of Northern Ireland, England, Wales and Scotland.

In considering the effects of the COVID-19 pandemic, comparisons were made for the only comparable period during the pandemic, i.e. around the first national lockdown (February – June 2020) when both countries were operating under the same lockdown rules. Following this period, England and Northern Ireland set their own timetables and rules regarding COVID-19 restrictions with England employing a Tier system and ultimately starting their second lockdown three weeks earlier than Northern Ireland.

## 10.3 Policy and practice implications

A summary of the main findings of this study along with the policy / practice implications are listed in Table 10.1.

TABLE 10.1: Summary of main study findings with implications for policy and practice.

| **Study finding** | **Policy / practice implication** |
|---|---|
| Comparing Northern Ireland with the other UK nations it was observed that overall NI had the second highest prescribing rates per head of population. | Further research is needed to understand why this is the case. Is there more sickness in NI, and if so, what are the contributing factors? |
| GP practices can be categorised using both their location and their prescribing profiles. | The ability to categorise GP practices will allow individual practices to be benchmarked against others in the same category. This will allow researchers to identify anomalous prescribing patterns and help to standardise prescribing. |
| Prescribing of Antidepressants and Analgesics is considerably higher in Metropolitan areas than those in Non-Metropolitan areas | Further research is needed to understand why this is the case. High deprivation may be a contributing factor. |
| There are a larger proportion of GP practices in areas of high deprivation in Metropolitan areas than in Non-Metropolitan areas. | As deprivation is a likely factor contributing to higher levels of prescribing, Government must make the issue of high deprivation a priority. |
| Small practices (2 doctors) have the highest prescribing rates for practices in both Metropolitan and Non-Metropolitan areas. | Small GP practices should be encouraged to grow in order to benefit from economies of scale. |
| The national lockdown imposed at the start of the COVID-19 pandemic resulted in a peak in prescribing. This peak was greater in NI than in England. | E-prescribing, available in England but not in NI, was likely responsible for the lower peak in prescribing seen in England. The development of a similar system in NI would reduce the footfall in NI GP practices and reduce the strain on resources. |
| Although a pattern of 'peak, trough, recovery' was observed during the lockdown, the prescribing of Antibiotics did not recover to pre lockdown levels. | Renewed efforts should be made to educate the public that antibiotics are not always needed with GPs continuing to monitor the use of antibiotics. |

Continued on next page

**Table 10.1 – continued from previous page**

| Study finding | Policy / practice implication |
|---|---|
| The concept of citizen science, and specifically, the GP prescribing dashboard received an overall positive response from most participants. | The development of data science tools to aid citizens in the analysis of data sets should be the next step in the open data movement providing citizen the opportunity to contribute to scientific study. |
| Of the participants, only 15% could be considered as active (i.e. leaving comments relating to the graphs produced). | Citizen science projects will need to attract large numbers of participants in order to gain sufficient feedback. |
| Of the comments left relating to graphs produced, only 18% could be considered to be scientific observations. | Renewed effort to increase the level of data literacy and scientific literacy in the public domain will increase the possibility of citizens providing useful contributions to citizen science. |

## 10.4 Conclusions and future work

### 10.4.1 Exploration of Open Prescription Data

In order to meet the objectives OBJ1 (to identify Open Source health Data containing geographical references) and OBJ2 (to identify and link relevant supplementary published data to create a novel data set), several open data sources were identified and linked to create a novel data set. The details of this process are in Chapter 3.

Initial analysis showed that the number of GP practices had declined between April 2018 and June 2021 with no new practices being established. The number of patients registered with practices had grown in proportion with the population growth. Figures showed that 50% of patients traveled over 10 kilometres to dispense their prescriptions with only 42.88% dispensing their prescriptions within 5 kilometres of the issuing practice. The average distance traveled to dispense prescriptions had seen a steady decrease since April 2018 with a marked fall in travelling distance seen as a result of the COVID-19 pandemic and subsequent lockdown although there were indications that travelling distances had started rising again. Comparing Northern Ireland with the other UK nations it was observed that overall NI had the second highest prescribing rates per head of population. Examining prescribing levels at BNF chapter level, NI had the highest prescribing levels per head of population in six of the twenty BNF chapters, namely chapter 4 (Central Nervous System), Chapter 5 (Infections), Chapter 10 (Musculoskeletal & Joint Diseases), Chapter 13 (Skin), Chapter 15 (Anatesia) and Chapter 20 (Dressings).

As the number of GP practices have declined, further research is needed to establish if this indicates that the number of GPs has also declined. Linking this research to the increase in registered patients further research into how the number of patients per doctor has risen or fallen has affected prescribing trends.

Northern Ireland had the highest prescribing levels per head of population in 6 BNF chapters over the observed period. Further research at section or paragraph level to compare the nations may shed light on where, if any, problems lie.

### 10.4.2 Analysis of General Practice Archetypes

In order to meet objective OBJ3 (to identify the types of GP practice using geographical location and relationship with dispensing pharmacies) and answer research question RQ1 (Can GP practices be classified in terms of their location with respect to the location of the pharmacies dispensing their prescriptions?), six key features relating to GP practices and their relationship with pharmacies dispensing prescriptions were chosen from the local data store. One year's data (April 2018 - March 2019) was subjected to k-means clustering in order to discover types of GP practice. This work is detailed in chapter 5.

This work provided a taxonomy for discovering GP practice types that could be used on dashboards for comparing or bench-marking different practices allowing the possibility of applying standardisation to prescribing practices. These dashboards could be used by government or health authorities. In addition, the provision of a general archetype would allow for the identification of anomalous behaviours indicating the possibility of lack of services, overloading of services or fraudulent behaviour within specific geographical areas or practices. Analysis of specific medications would provide a clearer picture of the local population's health highlighting the geographic location of high-risk areas for specific illnesses and providing a steppingstone for further research into the reasons for higher levels of illness observed in these locations.

Further to the work already done, study of Northern Ireland GP practices using different years of data may show changes to the structure seen with practices moving from one archetype to another. Further research on the method itself using different clustering algorithms could provide new perspective on the method with additional features as deprivation, practice size or even age of prescribing doctor helping to refine the process.

### 10.4.3 Analysis of the prescriptive behaviours of GP Practices

In order to meet objective OBJ4 (analyse any differences in prescribing behaviours between identified types of GP practice) and answer research question RQ2 (If GP practices can be classified in terms of their location, are there differences in prescription behaviour between classes?), time series analysis was performed on the discovered archetypes (Metropolitan and Non-Metropolitan) to compare them. A detailed account of this research is available in chapter 6.

It was discovered that prescribing patterns were largely similar for each archetype with levels of prescribing higher in approximately half of the BNF chapters for practices in Metropolitan areas. It was found that BNF chapter 4 (Central Nervous System) accounted for the largest proportion of variation between the identified clusters with sections 7 (Analgesics) and 3 (Antidepressant Drugs) being the main contributors.

### 10.4.4 Analysis of factors contributing to differences observed in prescribing behaviours of GP Practices

In order to meet objective OBJ5 (to explore possible factors contributing to differences in prescribing behaviours of identified types of GP practice) and answer research question RQ3 (Is it possible to identify contributing factors to differences in classes?), relevant literature on the subject was reviewed with the following conclusion.

Many factors such as patient demographics (age structure of the populations, ethnic and cultural differences in population composition etc), in practitioner demographics (including age, gender, part-time/full-time status etc), and in patient-full-time equivalent GP ratios (and consultation times) may contribute to the differences observed (Senior et al., 2003; Carter et al., 2021).

Given the data set available, it was not possible, to explore these factors but we could examine the effects deprivation and practice size (defined by the number of GPs working in the practice) had on prescribing levels. A detailed account of this research is available in chapter 7.

**Deprivation**

Comparing the prescribing levels of the two previously identified archetypes of GP practice in Northern Ireland (Metropolitan and Non-Metropolitan) showed that higher prescribing levels could be associated with practices located in areas with higher deprivation levels. It was also found that the increase in prescribing was greater for practices in the Metropolitan areas than in the Non-Metropolitan areas and whilst there was a larger proportion of Metropolitan practices in high deprivation areas it was unclear whether this could account for the larger increase observed in prescribing levels. Analysis of practices within the two archetypes in areas with low deprivation levels showed a completely different picture to those with all practices across all levels of deprivation. Prescribing levels for Non-Metropolitan practices were higher than Metropolitan practices where deprivation was not a factor while the opposite was true when examining all practices for both archetypes. This lead to the conclusion that deprivation was and is a major factor affecting prescribing levels and that it has a greater effect on Metropolitan practices than on Non-Metropolitan ones.

The findings of this study were based on the deprivation level of the area in which the GP practice was located. Based on the assumption that patients dispense their prescriptions close to where they live, future analysis of the effects deprivation on prescribing levels should focus more on the deprivation experienced by the patient than that of the area in which the GP practice is located.

Further investigation at British National Formulary (BNF) chapter level would provide insight into the types of medications affected by the level of deprivation.

**Practice size**

Comparing prescribing levels of different sizes of GP practice at Northern Ireland level and within the Metropolitan and Non-Metropolitan archetypes showed that higher prescribing levels were consistently associated with Small practices (two registered doctors). Whilst these practices' prescribing levels were, on average, 10% higher than the lowest prescribing practices in Northern Ireland, levels in Metropolitan areas were 32% higher with Non-Metropolitan areas being 3.8%. Overall the lowest prescribing levels in Northern Ireland were seen in Single-Handed practices with only one registered doctor and may be attributed to a better knowledge of patients influencing prescribing. This was also seen in practices in Non-Metropolitan areas but interestingly, the lowest prescribing practices in Metropolitan areas were Large practices with five or more registered doctors which benefited from the economies of scale and the possible provision of extra services being available.

Further research into the higher prescribing levels seen in Metropolitan areas is needed to shed light on this phenomenon. Studying these high deprivation areas in combination with the sizes of practice operating within them may provide insight.

### 10.4.5 Analysis of prescribing behaviours during the COVID-19 pandemic

In order to meet objective OBJ6 (to explore the effect the COVID-19 pandemic, and in particular, the first national lockdown had on prescribing behaviours) and answer research question RQ4 (What effect has COVID-19 and the National Lockdown had on prescribing behaviours?), data for England and Northern Ireland were compared for the years 2019 (pre-pandemic) and 2020. As both countries entered the first national lockdown on the same date with the same levels of restrictions, it was possible to compare prescribing for both nations during this lockdown. A detailed account of this research can be found in chapter 8.

It was found that prescribing in both nations followed a 'peak, trough, recovery' model overall and for most BNF chapters. It was found that the pattern of antibiotic prescribing was very different with levels not recovering to their previous levels. The peak observed in March 2020 at the start of lockdown showed more pronounced spikes in prescribing in Northern Ireland compared to England. A possible reason for this may be due to the lack of a system of electronic transfer of prescriptions to pharmacies in Northern Ireland. This would be an important implication for practice in Northern Ireland and a potential driver to introduce electronic prescribing in that region.

Further research into whether this pattern of reduced antibiotic prescribing will be sustained in the future is necessary as this may have implications affecting policies on better antibiotic stewardship in the future.

### 10.4.6 Development and Evaluation of a 'Citizen Science' Dashboard

In order to meet objective OBJ7 (to assess public opinion on the usefulness of a GP prescribing dashboard in relation to being used as a citizen science tool) and answer research question RQ5 (What interest does the ordinary citizen have in prescribing behaviours?), a bespoke data science tool in the form of a dashboard was developed. Volunteers used the dashboard to interrogate GP prescribing data, view the results and finally participate in a survey to capture their views on citizen science and the dashboard itself. A detailed account of this research is available in chapter 9.

The results of this study show that there is an interest in the concept of Citizen Science and in particular relating to GP prescribing data. The 'proof of concept' dashboard which was developed for this study received very positive feedback confirming that a data science tool such as this would be useful in the advancement of knowledge. Whilst, in general, users indicated their willingness to contribute to citizen science and felt that they understood the resulting graphs with on help needed to interpret them, less than 10% left observations on the resulting graphs. This indicates that, if the purpose of the dashboard is to gather observations from citizens, it needs to attract a large number of users. Alternatively, if the dashboard is but a tool to be used by citizens, then the number of observations recorded is not as important as the access the dashboard provides for each citizen to GP prescription data. Similarly, the majority of users spent between one and two minutes on the dashboard which, although indicates an interest in the subject, is unlikely to result in major insights. Users who spent 10 minutes and over interrogating the dashboard are more likely to provide insight into the resulting graphs.

Future work has already been indicated by the requests received by users during the study. The next version of the tool should provide more interactivity with the possibility of displaying more than one graph as output allowing direct comparisons to be made. The ability to adjust the scale has been requested along with more guidance on how to interpret the data. The prototype allowed access to one years data (2020) and users were keen that this be expanded to multiple years. With this expansion of data, users also requested the ability to select specific time periods to be analysed. Some users felt that the dashboard interface could be more user friendly and further research into what this means in real terms should be carried out.

## 10.5 Recommendations

The following recommendations are the result of the information gained and experiences had during the course of this study,

### 10.5.1 Open Data

**RM1 - Adoption of csv as standard fie format.** Whilst the majority of contributors of data to the OpendataNI portal provide both csv and excel versions of their files, in terms of data analysis and potential linking of files, csv provides the best solution. Excel files often contain multiple worksheets and these need to be decomposed into flat file csv format before being linked to other data sets. It is therefore recommended that the csv file format should be adopted for all open source data.

**RM2 - Standardisation of column headings.** Contributors of open data should ensure that they are consistent in the naming of column headings. For example, the number assigned to identify each practice within one series of open data files was found under 'Practice', 'PracticeNo', 'PRACTICE' and 'Practno'. This meant that these files could not be linked automatically with the data scientist constantly checking the headings of uploaded data sets to ensure uniformity. It is therefore recommended that Column headings be standardised within each series of data and, where possible, across data sets provided by each contributor.

**RM3 - Standardisation of data types.** Contributors of open data should ensure that the types of data provided are consistent. For example, when reporting the number of items prescribed overall, this was always reported as an integer and could be used in calculations without any problems. Within the same file, the breakdown of items at BNF Chapter, BNF Section and BNF Paragraph levels often resulted in reports of 0 items. This was often found as entries of '-' or blank instead of zero meaning that this column of data did not read in as a numeric automatically, needed wrangling to ensure zero entries were recorded correctly and finally converted to the integer datatype. Similarly, the recording of BNF Chapter was recorded as 1 or "01" in many cases. Again this meant that care had to be taken to ensure that all linked data conformed to the same data types before analysis could begin. It is therefore recommended that data types should be standardised within each series of data and, where possible, across data sets provided by each contributor.

**RM4 - Provision of guidance for data files.** The OpendataNI portal currently facilitates the downloading of open data files providing a link to the contributor's website. In order to gain an understanding of the column headings used it was necessary to navigate to the contributors website to find this information. In a several cases it was necessary to contact the contributor directly via email in order to locate this information. It is therefore recommended that a guidance document should be provided on the OpendataNI portal for each series of data detailing the column headings including an explanation of what they represent and the data type associated with each column.

### 10.5.2 Open Data - analysis

**RM5 - Development of data analysis toolkit.** The OpendataNI portal currently provides a repository of data sets made public by different organisations. In its current format, it does nothing to facilitate the analysis of the data. With the standardisation of data sets recommended in RM1 - RM4, the portal would be in the position to take the next step providing in providing open access to the analysis of the data. Research into the development of a data science tool capable of reading multiple data files from the portal along with the linking of variables which share the same data type should be undertaken. Provision of analysis capabilities within the tool allowing the user to interrogate the linked data and produce graphical output could provide valuable insights into the data not originally considered. The tool in itself could provide added benefit to research as data scientists would be able to see what combinations of data other researchers, including the citizen scientist, have been interested in. Provision of a feedback system may also generate research questions not previously considered. It is therefore recommended that, along with the standardisation of data sources, a data science toolkit be developed to enable more people to interrogate and analyse the open data provided.

### 10.5.3 e-prescribing to be introduced to NI

**RM6 - Development of e-prescribing system for use in NI.** It has been seen that due to the COVID-19 pandemic a radical change in behaviours was forced onto ordinary citizens in order to minimise infection, hospitalisations and deaths. GP surgeries were required to minimise the footfall of patients and this was initially achieved by GPs prescribing medications to keep patients stocked over the initial months of lockdown. Two years later and the situation has not changed dramatically with surgeries still limiting access to patients, often requiring their staff to deliver prescriptions to the patient waiting at the surgery entrance. This has inevitable added to the strain on both manpower and resources within GP surgeries. Some mitigation has been put in place with pharmacies collecting prescriptions on a daily basis, dispensing them and informing the patient when they are ready for collection. This process has helped reduce footfall in surgeries but can be quite a slow process with the time from requesting medication to collection often taking almost 1 week. It is recommended that NI should develop an e-prescribing system similar to that used in England whereby GPs can issue electronic prescriptions straight to the relevant pharmacy for dispensing. This would have the effect of reducing both the amount of paper used in printing physical prescriptions and the amount of petrol used by those collecting them and ultimately contribute to lowering of the countries carbon footprint.

## 10.6 Reflections

"And now the end is near, and so I face the final curtain". I would love to report that I did it my way but in truth I have benefited by the guidance and expertise of my supervisors, Dr McGlade and my co-authors and, unlike Mr Sinatra, I have more than a few regrets. I started this journey with no real knowledge of the state of open data and was saddened by the realisation that no standardisation was evident in the contributions being uploaded by different organisations. If anything, the uploaded data feels more like a duty than a contribution. In a previous job, I had been in charge with publishing monthly data files onto the company website and prided myself in the job making sure that each conformed to preset standards. This is not evident in some of the open data that I have observed. I do however feel that open data is a good thing and believe that further research should be undertaken into making data open more detailed. For example, the data used in this study is collected from pharmacist requests for payment for dispensed medications. The addition of one variable indicating which items were prescribed on the same prescription would open the possibility of research into the treatment of comorbidities. The COVID-19 pandemic struck the UK as I started the second year of my thesis. Prior to this my plan had been to develop links with health service professionals in order to work towards developing a software solution to aid them in their day to day activities. This aid could, in theory, use real time data with a predictive machine learning model in order to predict the need for future stocks of medications. Unfortunately, it did not feel right to put additional pressure on our health service colleagues during what was an already trying time for them. My one attempt at predictive modeling using the open data available to me yielded results with a low accuracy and was abandoned. I do however feel that new ground has been broken with regard to the classification of GP practices, leaving behind the old method of classifying them solely based on the surrounding population and basing it instead on the influence and prescribing patterns of each surgery. I feel that this new method of categorisation could be developed further with additional features such as derivation, practice size or age of prescribing GP enabling the current two categories to be broken down into sub-categories. I believe that in this aspect alone, I have contributed my grain of sand to the beach of knowledge, but with all courses of study, I am left with more questions than answers. GP prescribing trends in Northern Ireland has proved to be a fascinating topic for research with the potential, I believe, to provide further insights into both the overall health of the nation, and more specifically the health of patients registered with specific surgeries and whether differing medications are prescribed.

## Appendix A

# GP Prescribing Data Validation Results

TABLE A.1: Validation Results for GP Prescribing Data 2015-07

| **2015-07** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 463569 | 463569 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.44 | 7.44 |
| Total Items - Max | 1140.00 | 1140.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 81.53 | 81.53 |
| Gross Cost (£) - Maximum | 25143.54 | 25143.54 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 75.59 | 75.59 |
| Actual Cost (£) - Maximum | 24555.47 | 24555.47 |
| Rows with Missing BNF Category | 21409 | 21409 |
| Percentage of Items not Categorised | 0.62 | 0.62 |
| File Size (kb) | 64320 | 36110 |

TABLE A.2: Validation Results for GP Prescribing Data 2015-08

| **2015-08** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 449768 | 449768 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 6.95 | 6.95 |
| Total Items - Max | 983.00 | 983.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 77.06 | 77.06 |
| Gross Cost (£) - Maximum | 20371.56 | 20371.56 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 71.52 | 71.52 |
| Actual Cost (£) - Maximum | 20044.38 | 20044.38 |
| Rows with Missing BNF Category | 18619.00 | 18619.00 |
| Percentage of Items not Categorised | 0.60 | 0.60 |
| File Size (kb) | 61197 | 33140 |

TABLE A.3: Validation Results for GP Prescribing Data 2015-09

| **2015-09** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 455729 | 455729 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.20 | 7.20 |
| Total Items - Max | 1017.00 | 1017.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 79.14 | 79.14 |
| Gross Cost (£) - Maximum | 14953.36 | 14953.36 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 73.39 | 73.39 |
| Actual Cost (£) - Maximum | 14163.87 | 14163.87 |
| Rows with Missing BNF Category | 19854.00 | 19854.00 |
| Percentage of Items not Categorised | 0.61 | 0.61 |
| File Size (kb) | 63254 | 33582 |

TABLE A.4: Validation Results for GP Prescribing Data 2015-10

| **2015-10** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 461063 | 461063 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.22 | 7.22 |
| Total Items - Max | 1037.00 | 1037.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 79.62 | 79.62 |
| Gross Cost (£) - Maximum | 18570.73 | 18570.73 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 73.82 | 73.82 |
| Actual Cost (£) - Maximum | 18190.96 | 18190.96 |
| Rows with Missing BNF Category | 20941.00 | 20941.00 |
| Percentage of Items not Categorised | 0.63 | 0.63 |
| File Size (kb) | 63155 | 36224 |

TABLE A.5: Validation Results for GP Prescribing Data 2015-11

| **2015-11** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 457661 | 457661 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.15 | 7.15 |
| Total Items - Max | 1061.00 | 1061.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 78.73 | 78.73 |
| Gross Cost (£) - Maximum | 15588.80 | 15588.80 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 72.98 | 72.98 |
| Actual Cost (£) - Maximum | 14533.03 | 14533.03 |
| Rows with Missing BNF Category | 20100.00 | 20100.00 |
| Percentage of Items not Categorised | 0.61 | 0.61 |
| File Size (kb) | 62672 | 35936 |

TABLE A.6: Validation Results for GP Prescribing Data 2015-12

| **2015-12** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 456818 | 456818 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.44 | 7.44 |
| Total Items - Max | 1062.00 | 1062.00 |
| Gross Cost (£) - Minimum | 0.02 | 0.02 |
| Gross Cost (£) - Mean | 81.80 | 81.80 |
| Gross Cost (£) - Maximum | 12578.62 | 12578.62 |
| Actual Cost (£) - Minimum | 0.02 | 0.02 |
| Actual Cost (£) - Mean | 75.79 | 75.79 |
| Actual Cost (£) - Maximum | 12458.85 | 12458.85 |
| Rows with Missing BNF Category | 19552.00 | 19552.00 |
| Percentage of Items not Categorised | 0.58 | 0.58 |
| File Size (kb) | 62555 | 35858 |

TABLE A.7: Validation Results for GP Prescribing Data 2016-01

| **2016-01** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 456739 | 456739 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.22 | 7.22 |
| Total Items - Max | 1054.00 | 1054.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 77.66 | 77.66 |
| Gross Cost (£) - Maximum | 16712.96 | 16712.96 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 72.08 | 72.08 |
| Actual Cost (£) - Maximum | 16259.05 | 16259.05 |
| Rows with Missing BNF Category | 20688.00 | 20688.00 |
| Percentage of Items not Categorised | 0.63 | 0.63 |
| File Size (kb) | 62114 | 35431 |

TABLE A.8: Validation Results for GP Prescribing Data 2016-02

| **2016-02** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 455235 | 455235 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.16 | 7.16 |
| Total Items - Max | 1045.00 | 1045.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 76.69 | 76.69 |
| Gross Cost (£) - Maximum | 14745.91 | 14745.91 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 71.18 | 71.18 |
| Actual Cost (£) - Maximum | 14498.59 | 14498.59 |
| Rows with Missing BNF Category | 20225.00 | 20225.00 |
| Percentage of Items not Categorised | 0.62 | 0.62 |
| File Size (kb) | 61888 | 35309 |

TABLE A.9: Validation Results for GP Prescribing Data 2016-03

| **2016-03** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 458523 | 458523 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.43 | 7.43 |
| Total Items - Max | 1096.00 | 1096.00 |
| Gross Cost (£) - Minimum | 0.02 | 0.02 |
| Gross Cost (£) - Mean | 79.79 | 79.79 |
| Gross Cost (£) - Maximum | 12836.97 | 12836.97 |
| Actual Cost (£) - Minimum | 0.02 | 0.02 |
| Actual Cost (£) - Mean | 74.03 | 74.03 |
| Actual Cost (£) - Maximum | 12659.07 | 12659.07 |
| Rows with Missing BNF Category | 20729.00 | 20729.00 |
| Percentage of Items not Categorised | 0.61 | 0.61 |
| File Size (kb) | 62386 | 35559 |

TABLE A.10: Validation Results for GP Prescribing Data 2016-04

| **2016-04** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 459668 | 459668 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.38 | 7.38 |
| Total Items - Max | 1076.00 | 1076.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 79.08 | 79.08 |
| Gross Cost (£) - Maximum | 13764.88 | 13764.88 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 73.37 | 73.37 |
| Actual Cost (£) - Maximum | 13409.97 | 13409.97 |
| Rows with Missing BNF Category | 22332.00 | 22332.00 |
| Percentage of Items not Categorised | 0.66 | 0.66 |
| File Size (kb) | 63743 | 36584 |

TABLE A.11: Validation Results for GP Prescribing Data 2016-05

| **2016-05** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 457559 | 457559 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.27 | 7.27 |
| Total Items - Max | 1002.00 | 1002.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 78.12 | 78.12 |
| Gross Cost (£) - Maximum | 16226.57 | 16226.57 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 72.53 | 72.53 |
| Actual Cost (£) - Maximum | 15531.08 | 15531.08 |
| Rows with Missing BNF Category | 22192.00 | 22192.00 |
| Percentage of Items not Categorised | 0.67 | 0.67 |
| File Size (kb) | 62331 | 36449 |

TABLE A.12: Validation Results for GP Prescribing Data 2016-06

| **2016-06** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 463944 | 463944 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.45 | 7.45 |
| Total Items - Max | 1124.00 | 1124.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 78.83 | 78.83 |
| Gross Cost (£) - Maximum | 11568.20 | 11568.20 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 73.16 | 73.16 |
| Actual Cost (£) - Maximum | 11250.99 | 11250.99 |
| Rows with Missing BNF Category | 23784.00 | 23784.00 |
| Percentage of Items not Categorised | 0.69 | 0.69 |
| File Size (kb) | 64410 | 36989 |

TABLE A.13: Validation Results for GP Prescribing Data 2016-07

| **2016-07** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 450297 | 450297 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.15 | 7.15 |
| Total Items - Max | 1031.00 | 1031.00 |
| Gross Cost (£) - Minimum | 0.03 | 0.03 |
| Gross Cost (£) - Mean | 75.90 | 75.90 |
| Gross Cost (£) - Maximum | 14340.77 | 14340.77 |
| Actual Cost (£) - Minimum | 0.03 | 0.03 |
| Actual Cost (£) - Mean | 70.53 | 70.53 |
| Actual Cost (£) - Maximum | 13682.80 | 13682.80 |
| Rows with Missing BNF Category | 20775.00 | 20775.00 |
| Percentage of Items not Categorised | 0.64 | 0.64 |
| File Size (kb) | 61373 | 33212 |

TABLE A.14: Validation Results for GP Prescribing Data 2016-08

| **2016-08** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 459372 | 459372 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.41 | 7.41 |
| Total Items - Max | 1110.00 | 1110.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 79.12 | 79.12 |
| Gross Cost (£) - Maximum | 19954.36 | 19954.36 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 73.47 | 73.47 |
| Actual Cost (£) - Maximum | 19167.13 | 19167.13 |
| Rows with Missing BNF Category | 22607.00 | 22607.00 |
| Percentage of Items not Categorised | 0.66 | 0.66 |
| File Size (kb) | 62674 | 36665 |

TABLE A.15: Validation Results for GP Prescribing Data 2016-09

| **2016-09** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 460150 | 460150 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.40 | 7.40 |
| Total Items - Max | 1032.00 | 1032.00 |
| Gross Cost (£) - Minimum | 0.02 | 0.02 |
| Gross Cost (£) - Mean | 78.96 | 78.96 |
| Gross Cost (£) - Maximum | 16016.19 | 16016.19 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 73.31 | 73.31 |
| Actual Cost (£) - Maximum | 15193.89 | 15193.89 |
| Rows with Missing BNF Category | 24069.00 | 24069.00 |
| Percentage of Items not Categorised | 0.71 | 0.71 |
| File Size (kb) | 63967 | 36456 |

TABLE A.16: Validation Results for GP Prescribing Data 2016-10

| **2016-10** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 455029 | 455029 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.21 | 7.21 |
| Total Items - Max | 1026.00 | 1026.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 76.87 | 76.87 |
| Gross Cost (£) - Maximum | 16611.08 | 16611.08 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 71.41 | 71.41 |
| Actual Cost (£) - Maximum | 15918.95 | 15918.95 |
| Rows with Missing BNF Category | 22525.00 | 22525.00 |
| Percentage of Items not Categorised | 0.69 | 0.69 |
| File Size (kb) | 62565 | 36751 |

TABLE A.17: Validation Results for GP Prescribing Data 2016-11

| **2016-11** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 459652 | 459652 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.45 | 7.45 |
| Total Items - Max | 1075.00 | 1075.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 78.51 | 78.51 |
| Gross Cost (£) - Maximum | 14745.15 | 14745.15 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 72.87 | 72.87 |
| Actual Cost (£) - Maximum | 13603.98 | 13603.98 |
| Rows with Missing BNF Category | 22731.00 | 22731.00 |
| Percentage of Items not Categorised | 0.66 | 0.66 |
| File Size (kb) | 63195 | 37086 |

TABLE A.18: Validation Results for GP Prescribing Data 2016-12

| **2016-12** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 461722 | 461722 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.44 | 7.44 |
| Total Items - Max | 1091.00 | 1091.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 79.04 | 79.04 |
| Gross Cost (£) - Maximum | 15874.55 | 15874.55 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 73.40 | 73.40 |
| Actual Cost (£) - Maximum | 15276.00 | 15276.00 |
| Rows with Missing BNF Category | 15178.00 | 15178.00 |
| Percentage of Items not Categorised | 0.44 | 0.44 |
| File Size (kb) | 63598 | 37333 |

TABLE A.19: Validation Results for GP Prescribing Data 2017-01

| **2017-01** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 466981 | 466981 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.32 | 7.32 |
| Total Items - Max | 1072.00 | 1072.00 |
| Gross Cost (£) - Minimum | 0.02 | 0.02 |
| Gross Cost (£) - Mean | 78.05 | 78.05 |
| Gross Cost (£) - Maximum | 15295.21 | 15295.21 |
| Actual Cost (£) - Minimum | 0.02 | 0.02 |
| Actual Cost (£) - Mean | 72.48 | 72.48 |
| Actual Cost (£) - Maximum | 14962.62 | 14962.62 |
| Rows with Missing BNF Category | 15544.00 | 15544.00 |
| Percentage of Items not Categorised | 0.45 | 0.45 |
| File Size (kb) | 65094 | 37409 |

TABLE A.20: Validation Results for GP Prescribing Data 2017-02

| **2017-02** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 456440 | 456440 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 6.91 | 6.91 |
| Total Items - Max | 964.00 | 964.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 73.64 | 73.64 |
| Gross Cost (£) - Maximum | 10007.87 | 10007.87 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 68.40 | 68.40 |
| Actual Cost (£) - Maximum | 9785.67 | 9785.67 |
| Rows with Missing BNF Category | 11993.00 | 11993.00 |
| Percentage of Items not Categorised | 0.38 | 0.38 |
| File Size (kb) | 62425 | 36524 |

TABLE A.21: Validation Results for GP Prescribing Data 2017-03

| **2017-03** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 475230 | 475230 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.48 | 7.48 |
| Total Items - Max | 1081.00 | 1081.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 79.57 | 79.57 |
| Gross Cost (£) - Maximum | 20303.33 | 20303.33 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 73.83 | 73.83 |
| Actual Cost (£) - Maximum | 19637.85 | 19637.85 |
| Rows with Missing BNF Category | 14953.00 | 14953.00 |
| Percentage of Items not Categorised | 0.42 | 0.42 |
| File Size (kb) | 66323 | 38120 |

TABLE A.22: Validation Results for GP Prescribing Data 2017-04

| **2017-04** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 452296 | 452296 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 6.95 | 6.95 |
| Total Items - Max | 973.00 | 973.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 72.99 | 72.99 |
| Gross Cost (£) - Maximum | 16932.14 | 16932.14 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 67.79 | 67.79 |
| Actual Cost (£) - Maximum | 16416.22 | 16416.22 |
| Rows with Missing BNF Category | 8531.00 | 8531.00 |
| Percentage of Items not Categorised | 0.27 | 0.27 |
| File Size (kb) | 61942 | 36177 |

TABLE A.23: Validation Results for GP Prescribing Data 2017-05

| **2017-05** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 474301 | 474301 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.39 | 7.39 |
| Total Items - Max | 1094.00 | 1094.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 78.93 | 78.93 |
| Gross Cost (£) - Maximum | 14440.86 | 14440.86 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 73.25 | 73.25 |
| Actual Cost (£) - Maximum | 14337.07 | 14337.07 |
| Rows with Missing BNF Category | 9675.00 | 9675.00 |
| Percentage of Items not Categorised | 0.28 | 0.28 |
| File Size (kb) | 65167 | 38157 |

TABLE A.24: Validation Results for GP Prescribing Data 2017-06

| **2017-06** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 474851 | 474851 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.35 | 7.35 |
| Total Items - Max | 1070.00 | 1070.00 |
| Gross Cost (£) - Minimum | 0.02 | 0.02 |
| Gross Cost (£) - Mean | 79.60 | 79.60 |
| Gross Cost (£) - Maximum | 11466.18 | 11466.18 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 73.82 | 73.82 |
| Actual Cost (£) - Maximum | 10872.39 | 10872.39 |
| Rows with Missing BNF Category | 9301.00 | 9301.00 |
| Percentage of Items not Categorised | 0.27 | 0.27 |
| File Size (kb) | 66480 | 38248 |

TABLE A.25: Validation Results for GP Prescribing Data 2017-07

| **2017-07** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 460970 | 460970 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.05 | 7.05 |
| Total Items - Max | 966.00 | 966.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 77.78 | 77.78 |
| Gross Cost (£) - Maximum | 10585.25 | 10585.25 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 72.12 | 72.12 |
| Actual Cost (£) - Maximum | 9478.08 | 9478.08 |
| Rows with Missing BNF Category | 8292.00 | 8292.00 |
| Percentage of Items not Categorised | 0.25 | 0.25 |
| File Size (kb) | 63242 | 36995 |

TABLE A.26: Validation Results for GP Prescribing Data 2017-08

| **2017-08** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 470328 | 470328 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.29 | 7.29 |
| Total Items - Max | 1105.00 | 1105.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 76.73 | 76.73 |
| Gross Cost (£) - Maximum | 12679.18 | 12679.18 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 71.25 | 71.25 |
| Actual Cost (£) - Maximum | 12213.63 | 12213.63 |
| Rows with Missing BNF Category | 8130.00 | 8130.00 |
| Percentage of Items not Categorised | 0.24 | 0.24 |
| File Size (kb) | 64598 | 37831 |

TABLE A.27: Validation Results for GP Prescribing Data 2017-09

| **2017-09** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 466180 | 466180 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.10 | 7.10 |
| Total Items - Max | 1071.00 | 1071.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 74.42 | 74.42 |
| Gross Cost (£) - Maximum | 10047.21 | 10047.21 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 69.15 | 69.15 |
| Actual Cost (£) - Maximum | 9872.23 | 9872.23 |
| Rows with Missing BNF Category | 8052.00 | 8052.00 |
| Percentage of Items not Categorised | 0.24 | 0.24 |
| File Size (kb) | 64026 | 37478 |

TABLE A.28: Validation Results for GP Prescribing Data 2017-10

| **2017-10** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 473115.00 | 473115.00 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.26 | 7.26 |
| Total Items - Max | 1053.00 | 1053.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 77.81 | 77.81 |
| Gross Cost (£) - Maximum | 12655.84 | 12655.84 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 72.22 | 72.22 |
| Actual Cost (£) - Maximum | 11757.53 | 11757.53 |
| Rows with Missing BNF Category | 9081.00 | 9081.00 |
| Percentage of Items not Categorised | 0.26 | 0.26 |
| File Size (kb) | 65483 | 38576 |

TABLE A.29: Validation Results for GP Prescribing Data 2017-11

| **2017-11** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 473571 | 473571 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.24 | 7.24 |
| Total Items - Max | 1091.00 | 1091.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 74.04 | 74.04 |
| Gross Cost (£) - Maximum | 9985.36 | 9985.36 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 68.88 | 68.88 |
| Actual Cost (£) - Maximum | 9667.60 | 9667.60 |
| Rows with Missing BNF Category | 7725.00 | 7725.00 |
| Percentage of Items not Categorised | 0.23 | 0.23 |
| File Size (kb) | 67103 | 41351 |

TABLE A.30: Validation Results for GP Prescribing Data 2017-12

| **2017-12** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 459732 | 459732 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.26 | 7.26 |
| Total Items - Max | 1024.00 | 1024.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 70.98 | 70.98 |
| Gross Cost (£) - Maximum | 11317.84 | 11317.84 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 66.09 | 66.09 |
| Actual Cost (£) - Maximum | 10891.13 | 10891.13 |
| Rows with Missing BNF Category | 7246.00 | 7246.00 |
| Percentage of Items not Categorised | 0.22 | 0.22 |
| File Size (kb) | 65080 | 40000 |

TABLE A.31: Validation Results for GP Prescribing Data 2018-01

| **2018-01** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 479890 | 479890 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.58 | 7.58 |
| Total Items - Max | 1137.00 | 1137.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 75.06 | 75.06 |
| Gross Cost (£) - Maximum | 11841.84 | 11841.84 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 69.83 | 69.83 |
| Actual Cost (£) - Maximum | 11584.65 | 11584.65 |
| Rows with Missing BNF Category | 8726.00 | 8726.00 |
| Percentage of Items not Categorised | 0.24 | 0.24 |
| File Size (kb) | 67674 | 40545 |

TABLE A.32: Validation Results for GP Prescribing Data 2018-02

| **2018-02** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 460275 | 460275 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 6.94 | 6.94 |
| Total Items - Max | 972.00 | 972.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 69.10 | 69.10 |
| Gross Cost (£) - Maximum | 13860.25 | 13860.25 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 64.33 | 64.33 |
| Actual Cost (£) - Maximum | 13602.90 | 13602.90 |
| Rows with Missing BNF Category | 7382.00 | 7382.00 |
| Percentage of Items not Categorised | 0.23 | 0.23 |
| File Size (kb) | 64775 | 38763 |

TABLE A.33: Validation Results for GP Prescribing Data 2018-03

| **2018-03** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 473219 | 473219 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.39 | 7.39 |
| Total Items - Max | 1089.00 | 1089.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 74.79 | 74.79 |
| Gross Cost (£) - Maximum | 11660.87 | 11660.87 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 69.55 | 69.55 |
| Actual Cost (£) - Maximum | 11531.76 | 11531.76 |
| Rows with Missing BNF Category | 7713.00 | 7713.00 |
| Percentage of Items not Categorised | 0.22 | 0.22 |
| File Size (kb) | 66708 | 39967 |

TABLE A.34: Validation Results for GP Prescribing Data 2018-04

| **2018-04** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 465119 | 465119 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 6.99 | 6.99 |
| Total Items - Max | 1011.00 | 1011.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 70.19 | 70.19 |
| Gross Cost (£) - Maximum | 18323.43 | 18323.43 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 65.39 | 65.39 |
| Actual Cost (£) - Maximum | 18207.86 | 18207.86 |
| Rows with Missing BNF Category | 8008.00 | 8008.00 |
| Percentage of Items not Categorised | 0.25 | 0.25 |
| File Size (kb) | 65559 | 39273 |

TABLE A.35: Validation Results for GP Prescribing Data 2018-05

| **2018-05** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 476872 | 476872 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.36 | 7.36 |
| Total Items - Max | 1044.00 | 1044.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 74.07 | 74.07 |
| Gross Cost (£) - Maximum | 11000.64 | 11000.64 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 68.91 | 68.91 |
| Actual Cost (£) - Maximum | 10902.66 | 10902.66 |
| Rows with Missing BNF Category | 7854.00 | 7854.00 |
| Percentage of Items not Categorised | 0.22 | 0.22 |
| File Size (kb) | 67283 | 40329 |

TABLE A.36: Validation Results for GP Prescribing Data 2018-06

| **2018-06** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 473861 | 473861 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.24 | 7.24 |
| Total Items - Max | 1015.00 | 1015.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 72.98 | 72.98 |
| Gross Cost (£) - Maximum | 8855.86 | 8855.86 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 67.93 | 67.93 |
| Actual Cost (£) - Maximum | 8785.77 | 8785.77 |
| Rows with Missing BNF Category | 7482.00 | 7482.00 |
| Percentage of Items not Categorised | 0.22 | 0.22 |
| File Size (kb) | 66837 | 40059 |

TABLE A.37: Validation Results for GP Prescribing Data 2018-07

| **2018-07** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 465221 | 465221 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.26 | 7.26 |
| Total Items - Max | 1014.00 | 1014.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 71.35 | 71.35 |
| Gross Cost (£) - Maximum | 10380.00 | 10380.00 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 66.48 | 66.48 |
| Actual Cost (£) - Maximum | 10380.00 | 10380.00 |
| Rows with Missing BNF Category | 7518.00 | 7518.00 |
| Percentage of Items not Categorised | 0.22 | 0.22 |
| File Size (kb) | 65527 | 40148 |

TABLE A.38: Validation Results for GP Prescribing Data 2018-08

| **2018-08** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 471790 | 471790 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.37 | 7.37 |
| Total Items - Max | 1104.00 | 1104.00 |
| Gross Cost (£) - Minimum | 0.03 | 0.03 |
| Gross Cost (£) - Mean | 74.81 | 74.81 |
| Gross Cost (£) - Maximum | 11150.94 | 11150.94 |
| Actual Cost (£) - Minimum | 0.03 | 0.03 |
| Actual Cost (£) - Mean | 69.64 | 69.64 |
| Actual Cost (£) - Maximum | 11140.63 | 11140.63 |
| Rows with Missing BNF Category | 8167.00 | 8167.00 |
| Percentage of Items not Categorised | 0.23 | 0.23 |
| File Size (kb) | 66932 | 40807 |

TABLE A.39: Validation Results for GP Prescribing Data 2018-09

| **2018-09** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 458180 | 458180 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.00 | 7.00 |
| Total Items - Max | 959.00 | 959.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 70.67 | 70.67 |
| Gross Cost (£) - Maximum | 9928.69 | 9928.69 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 65.82 | 65.82 |
| Actual Cost (£) - Maximum | 9918.96 | 9918.96 |
| Rows with Missing BNF Category | 7109.00 | 7109.00 |
| Percentage of Items not Categorised | 0.22 | 0.22 |
| File Size (kb) | 65332 | 39605 |

TABLE A.40: Validation Results for GP Prescribing Data 2018-10

| **2018-10** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 476794 | 476794 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.54 | 7.54 |
| Total Items - Max | 1088.00 | 1088.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 74.53 | 74.53 |
| Gross Cost (£) - Maximum | 17868.18 | 17868.18 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 69.42 | 69.42 |
| Actual Cost (£) - Maximum | 16698.63 | 16698.63 |
| Rows with Missing BNF Category | 8281.00 | 8281.00 |
| Percentage of Items not Categorised | 0.23 | 0.23 |
| File Size (kb) | 68630 | 41796 |

TABLE A.41: Validation Results for GP Prescribing Data 2018-11

| **2018-11** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 468699 | 468699 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.36 | 7.36 |
| Total Items - Max | 1033.00 | 1033.00 |
| Gross Cost (£) - Minimum | 0.02 | 0.02 |
| Gross Cost (£) - Mean | 71.70 | 71.70 |
| Gross Cost (£) - Maximum | 12587.58 | 12587.58 |
| Actual Cost (£) - Minimum | 0.02 | 0.02 |
| Actual Cost (£) - Mean | 66.91 | 66.91 |
| Actual Cost (£) - Maximum | 12393.33 | 12393.33 |
| Rows with Missing BNF Category | 7606.00 | 7606.00 |
| Percentage of Items not Categorised | 0.22 | 0.22 |
| File Size (kb) | 67336 | 40102 |

TABLE A.42: Validation Results for GP Prescribing Data 2018-12

| **2018-12** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 452898 | 452898 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.20 | 7.20 |
| Total Items - Max | 1012.00 | 1012.00 |
| Gross Cost (£) - Minimum | 0.02 | 0.02 |
| Gross Cost (£) - Mean | 70.17 | 70.17 |
| Gross Cost (£) - Maximum | 10449.10 | 10449.10 |
| Actual Cost (£) - Minimum | 0.02 | 0.02 |
| Actual Cost (£) - Mean | 65.48 | 65.48 |
| Actual Cost (£) - Maximum | 10432.88 | 10432.88 |
| Rows with Missing BNF Category | 6897.00 | 6897.00 |
| Percentage of Items not Categorised | 0.21 | 0.21 |
| File Size (kb) | 64922 | 38550 |

TABLE A.43: Validation Results for GP Prescribing Data 2019-01

| **2019-01** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 476265 | 476265 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.74 | 7.74 |
| Total Items - Max | 1140.00 | 1140.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 75.46 | 75.46 |
| Gross Cost (£) - Maximum | 12073.51 | 12073.51 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 70.42 | 70.42 |
| Actual Cost (£) - Maximum | 12027.52 | 12027.52 |
| Rows with Missing BNF Category | 8782.00 | 8782.00 |
| Percentage of Items not Categorised | 0.24 | 0.24 |
| File Size (kb) | 68001 | 40287 |

TABLE A.44: Validation Results for GP Prescribing Data 2019-02

| **2019-02** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 455864 | 455864 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.05 | 7.05 |
| Total Items - Max | 982.00 | 982.00 |
| Gross Cost (£) - Minimum | 0.02 | 0.02 |
| Gross Cost (£) - Mean | 68.65 | 68.65 |
| Gross Cost (£) - Maximum | 9382.20 | 9382.20 |
| Actual Cost (£) - Minimum | 0.02 | 0.02 |
| Actual Cost (£) - Mean | 64.12 | 64.12 |
| Actual Cost (£) - Maximum | 8465.31 | 8465.31 |
| Rows with Missing BNF Category | 7806.00 | 7806.00 |
| Percentage of Items not Categorised | 0.24 | 0.24 |
| File Size (kb) | 64922 | 38444 |

TABLE A.45: Validation Results for GP Prescribing Data 2019-03

| **2019-03** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 463260 | 463260 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.29 | 7.29 |
| Total Items - Max | 1032.00 | 1032.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 71.59 | 71.59 |
| Gross Cost (£) - Maximum | 10380.00 | 10380.00 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 66.87 | 66.87 |
| Actual Cost (£) - Maximum | 10380.00 | 10380.00 |
| Rows with Missing BNF Category | 8881.00 | 8881.00 |
| Percentage of Items not Categorised | 0.26 | 0.26 |
| File Size (kb) | 66029 | 39102 |

TABLE A.46: Validation Results for GP Prescribing Data 2019-04

| **2019-04** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 459373 | 459373 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.43 | 7.43 |
| Total Items - Max | 1052.00 | 1052.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 73.64 | 73.64 |
| Gross Cost (£) - Maximum | 9997.80 | 9997.80 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 68.64 | 68.64 |
| Actual Cost (£) - Maximum | 9377.52 | 9377.52 |
| Rows with Missing BNF Category | 6946.00 | 6946.00 |
| Percentage of Items not Categorised | 0.20 | 0.20 |
| File Size (kb) | 65596 | 38750 |

TABLE A.47: Validation Results for GP Prescribing Data 2019-05

| **2019-05** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 468269 | 468269 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.69 | 7.69 |
| Total Items - Max | 1090.00 | 1090.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 77.40 | 77.40 |
| Gross Cost (£) - Maximum | 14474.90 | 14474.90 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 72.10 | 72.10 |
| Actual Cost (£) - Maximum | 14421.78 | 14421.78 |
| Rows with Missing BNF Category | 8060.00 | 8060.00 |
| Percentage of Items not Categorised | 0.22 | 0.22 |
| File Size (kb) | 67016 | 39623 |

TABLE A.48: Validation Results for GP Prescribing Data 2019-06

| **2019-06** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 455879 | 455879 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.24 | 7.24 |
| Total Items - Max | 1016.00 | 1016.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 73.30 | 73.30 |
| Gross Cost (£) - Maximum | 11410.04 | 11410.04 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 68.33 | 68.33 |
| Actual Cost (£) - Maximum | 11403.55 | 11403.55 |
| Rows with Missing BNF Category | 6959.00 | 6959.00 |
| Percentage of Items not Categorised | 0.21 | 0.21 |
| File Size (kb) | 65265 | 38593 |

TABLE A.49: Validation Results for GP Prescribing Data 2019-07

| **2019-07** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 463742 | 463742 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.69 | 7.69 |
| Total Items - Max | 1122.00 | 1122.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 77.52 | 77.52 |
| Gross Cost (£) - Maximum | 21072.59 | 21072.59 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 72.23 | 72.23 |
| Actual Cost (£) - Maximum | 20883.44 | 20883.44 |
| Rows with Missing BNF Category | 7147.00 | 7147.00 |
| Percentage of Items not Categorised | 0.20 | 0.20 |
| File Size (kb) | 66437 | 39308 |

TABLE A.50: Validation Results for GP Prescribing Data 2019-08

| **2019-08** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 458781 | 458781 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.50 | 7.50 |
| Total Items - Max | 1076.00 | 1076.00 |
| Gross Cost (£) - Minimum | 0.02 | 0.02 |
| Gross Cost (£) - Mean | 78.48 | 78.48 |
| Gross Cost (£) - Maximum | 12997.49 | 12997.49 |
| Actual Cost (£) - Minimum | 0.02 | 0.02 |
| Actual Cost (£) - Mean | 73.12 | 73.12 |
| Actual Cost (£) - Maximum | 12989.25 | 12989.25 |
| Rows with Missing BNF Category | 6815.00 | 6815.00 |
| Percentage of Items not Categorised | 0.20 | 0.20 |
| File Size (kb) | 65721 | 38861 |

TABLE A.51: Validation Results for GP Prescribing Data 2019-09

| **2019-09** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 459842 | 459842 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.36 | 7.36 |
| Total Items - Max | 1031.00 | 1031.00 |
| Gross Cost (£) - Minimum | 0.02 | 0.02 |
| Gross Cost (£) - Mean | 76.87 | 76.87 |
| Gross Cost (£) - Maximum | 17660.51 | 17660.51 |
| Actual Cost (£) - Minimum | 0.02 | 0.02 |
| Actual Cost (£) - Mean | 71.67 | 71.67 |
| Actual Cost (£) - Maximum | 17660.51 | 17660.51 |
| Rows with Missing BNF Category | 6340.00 | 6340.00 |
| Percentage of Items not Categorised | 0.19 | 0.19 |
| File Size (kb) | 65857 | 38939 |

TABLE A.52: Validation Results for GP Prescribing Data 2019-10

| **2019-10** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 471258 | 471258 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.77 | 7.77 |
| Total Items - Max | 1138.00 | 1138.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 79.84 | 79.84 |
| Gross Cost (£) - Maximum | 11291.83 | 11291.83 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 74.50 | 74.50 |
| Actual Cost (£) - Maximum | 11159.09 | 11159.09 |
| Rows with Missing BNF Category | 6973.00 | 6973.00 |
| Percentage of Items not Categorised | 0.19 | 0.19 |
| File Size (kb) | 68054 | 40459 |

TABLE A.53: Validation Results for GP Prescribing Data 2019-11

| **2019-11** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 459345 | 459345 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.45 | 7.45 |
| Total Items - Max | 1067.00 | 1067.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 76.00 | 76.00 |
| Gross Cost (£) - Maximum | 11571.00 | 11571.00 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 70.94 | 70.94 |
| Actual Cost (£) - Maximum | 10418.81 | 10418.81 |
| Rows with Missing BNF Category | 6217.00 | 6217.00 |
| Percentage of Items not Categorised | 0.18 | 0.18 |
| File Size (kb) | 66185 | 39325 |

TABLE A.54: Validation Results for GP Prescribing Data 2019-12

| **2019-12** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 456596 | 456596 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.70 | 7.70 |
| Total Items - Max | 1184.00 | 1184.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 79.16 | 79.16 |
| Gross Cost (£) - Maximum | 17900.85 | 17900.85 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 73.88 | 73.88 |
| Actual Cost (£) - Maximum | 17899.39 | 17899.39 |
| Rows with Missing BNF Category | 6689.00 | 6689.00 |
| Percentage of Items not Categorised | 0.19 | 0.19 |
| File Size (kb) | 65820 | 39075 |

TABLE A.55: Validation Results for GP Prescribing Data 2020-01

| **2020-01** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 468562 | 468562 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.75 | 7.75 |
| Total Items - Max | 1131.00 | 1131.00 |
| Gross Cost (£) - Minimum | 0.02 | 0.02 |
| Gross Cost (£) - Mean | 79.70 | 79.70 |
| Gross Cost (£) - Maximum | 12967.50 | 12967.50 |
| Actual Cost (£) - Minimum | 0.02 | 0.02 |
| Actual Cost (£) - Mean | 74.38 | 74.38 |
| Actual Cost (£) - Maximum | 12903.73 | 12903.73 |
| Rows with Missing BNF Category | 7029.00 | 7029.00 |
| Percentage of Items not Categorised | 0.19 | 0.19 |
| File Size (kb) | 67175 | 39720 |

TABLE A.56: Validation Results for GP Prescribing Data 2020-02

| **2020-02** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 450261 | 450261 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.22 | 7.22 |
| Total Items - Max | 1025.00 | 1025.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 74.06 | 74.06 |
| Gross Cost (£) - Maximum | 10541.20 | 10541.20 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 69.17 | 69.17 |
| Actual Cost (£) - Maximum | 10131.49 | 10131.49 |
| Rows with Missing BNF Category | 6014.00 | 6014.00 |
| Percentage of Items not Categorised | 0.18 | 0.18 |
| File Size (kb) | 64441 | 38057 |

TABLE A.57: Validation Results for GP Prescribing Data 2020-03

| **2020-03** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 472325 | 472325 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 8.31 | 8.31 |
| Total Items - Max | 1253.00 | 1253.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 86.97 | 86.97 |
| Gross Cost (£) - Maximum | 13740.80 | 13740.80 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 81.04 | 81.04 |
| Actual Cost (£) - Maximum | 12772.99 | 12772.99 |
| Rows with Missing BNF Category | 6647.00 | 6647.00 |
| Percentage of Items not Categorised | 0.17 | 0.17 |
| File Size (kb) | 67956 | 40156 |

TABLE A.58: Validation Results for GP Prescribing Data 2020-04

| **2020-04** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 448653 | 448653 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.49 | 7.49 |
| Total Items - Max | 1035.00 | 1035.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 79.51 | 79.51 |
| Gross Cost (£) - Maximum | 12824.17 | 12824.17 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 74.32 | 74.32 |
| Actual Cost (£) - Maximum | 12726.33 | 12726.33 |
| Rows with Missing BNF Category | 7155.00 | 7155.00 |
| Percentage of Items not Categorised | 0.21 | 0.21 |
| File Size (kb) | 64082 | 37987 |

TABLE A.59: Validation Results for GP Prescribing Data 2020-05

| **2020-05** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 440780 | 440780 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.32 | 7.32 |
| Total Items - Max | 1065.00 | 1065.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 78.96 | 78.96 |
| Gross Cost (£) - Maximum | 12182.80 | 12182.80 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 73.69 | 73.69 |
| Actual Cost (£) - Maximum | 10983.53 | 10983.53 |
| Rows with Missing BNF Category | 5936.00 | 5936.00 |
| Percentage of Items not Categorised | 0.18 | 0.18 |
| File Size (kb) | 62882 | 37197 |

TABLE A.60: Validation Results for GP Prescribing Data 2020-06

| **2020-06** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 450427 | 450427 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.62 | 7.62 |
| Total Items - Max | 1160.00 | 1160.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 83.59 | 83.59 |
| Gross Cost (£) - Maximum | 11983.30 | 11983.30 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 77.95 | 77.95 |
| Actual Cost (£) - Maximum | 11727.47 | 11727.47 |
| Rows with Missing BNF Category | 6358.00 | 6358.00 |
| Percentage of Items not Categorised | 0.19 | 0.19 |
| File Size (kb) | 64373 | 38120 |

TABLE A.61: Validation Results for GP Prescribing Data 2020-07

| **2020-07** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 449918 | 449918 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.75 | 7.75 |
| Total Items - Max | 1147.00 | 1147.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 84.09 | 84.09 |
| Gross Cost (£) - Maximum | 13073.90 | 13073.90 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 78.37 | 78.37 |
| Actual Cost (£) - Maximum | 11776.85 | 11776.85 |
| Rows with Missing BNF Category | 6380.00 | 6380.00 |
| Percentage of Items not Categorised | 0.18 | 0.18 |
| File Size (kb) | 64301 | 37995 |

TABLE A.62: Validation Results for GP Prescribing Data 2020-08

| **2020-08** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 437307 | 437307 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.26 | 7.26 |
| Total Items - Max | 1114.00 | 1114.00 |
| Gross Cost (£) - Minimum | 0.02 | 0.02 |
| Gross Cost (£) - Mean | 79.42 | 79.42 |
| Gross Cost (£) - Maximum | 18009.13 | 18009.13 |
| Actual Cost (£) - Minimum | 0.02 | 0.02 |
| Actual Cost (£) - Mean | 74.13 | 74.13 |
| Actual Cost (£) - Maximum | 17904.46 | 17904.46 |
| Rows with Missing BNF Category | 5806.00 | 5806.00 |
| Percentage of Items not Categorised | 0.18 | 0.18 |
| File Size (kb) | 62483 | 36891 |

TABLE A.63: Validation Results for GP Prescribing Data 2020-09

| **2020-09** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 456607 | 456607 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.76 | 7.76 |
| Total Items - Max | 1240.00 | 1240.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 84.25 | 84.25 |
| Gross Cost (£) - Maximum | 12878.20 | 12878.20 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 78.55 | 78.55 |
| Actual Cost (£) - Maximum | 11586.99 | 11586.99 |
| Rows with Missing BNF Category | 5538.00 | 5538.00 |
| Percentage of Items not Categorised | 0.16 | 0.16 |
| File Size (kb) | 65445 | 38708 |

TABLE A.64: Validation Results for GP Prescribing Data 2020-10

| **2020-10** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 452756 | 452756 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.68 | 7.68 |
| Total Items - Max | 1124.00 | 1124.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 83.22 | 83.22 |
| Gross Cost (£) - Maximum | 13984.32 | 13984.32 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 77.74 | 77.74 |
| Actual Cost (£) - Maximum | 13870.57 | 13870.57 |
| Rows with Missing BNF Category | 6175.00 | 6175.00 |
| Percentage of Items not Categorised | 0.18 | 0.18 |
| File Size (kb) | 65236 | 38756 |

TABLE A.65: Validation Results for GP Prescribing Data 2020-11

| **2020-11** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 448891 | 448891 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.44 | 7.44 |
| Total Items - Max | 1182.00 | 1182.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 80.48 | 80.48 |
| Gross Cost (£) - Maximum | 12509.12 | 12509.12 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 75.22 | 75.22 |
| Actual Cost (£) - Maximum | 12438.18 | 12438.18 |
| Rows with Missing BNF Category | 5562.00 | 5562.00 |
| Percentage of Items not Categorised | 0.17 | 0.17 |
| File Size (kb) | 64740 | 38464 |

TABLE A.66: Validation Results for GP Prescribing Data 2020-12

| **2020-12** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 457053 | 457053 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.98 | 7.98 |
| Total Items - Max | 1203.00 | 1203.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 86.60 | 86.60 |
| Gross Cost (£) - Maximum | 14644.25 | 14644.25 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 80.88 | 80.88 |
| Actual Cost (£) - Maximum | 13179.65 | 13179.65 |
| Rows with Missing BNF Category | 5734.00 | 5734.00 |
| Percentage of Items not Categorised | 0.16 | 0.16 |
| File Size (kb) | 65918 | 39133 |

TABLE A.67: Validation Results for GP Prescribing Data 2021-01

| **2021-01** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 445848 | 445848 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.46 | 7.46 |
| Total Items - Max | 1181.00 | 1181.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 81.74 | 81.74 |
| Gross Cost (£) - Maximum | 13484.55 | 13484.55 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 76.43 | 76.43 |
| Actual Cost (£) - Maximum | 13111.35 | 13111.35 |
| Rows with Missing BNF Category | 6519.00 | 6519.00 |
| Percentage of Items not Categorised | 0.20 | 0.20 |
| File Size (kb) | 63748 | 37705 |

TABLE A.68: Validation Results for GP Prescribing Data 2021-02

| **2021-02** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 440422 | 440422 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.27 | 7.27 |
| Total Items - Max | 1090.00 | 1090.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 79.75 | 79.75 |
| Gross Cost (£) - Maximum | 11704.00 | 11704.00 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 74.65 | 74.65 |
| Actual Cost (£) - Maximum | 11113.83 | 11113.83 |
| Rows with Missing BNF Category | 5336.00 | 5336.00 |
| Percentage of Items not Categorised | 0.17 | 0.17 |
| File Size (kb) | 62879 | 37200 |

TABLE A.69: Validation Results for GP Prescribing Data 2021-03

| **2021-03** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 458284 | 458284 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 8.03 | 8.03 |
| Total Items - Max | 1215.00 | 1215.00 |
| Gross Cost (£) - Minimum | 0.02 | 0.02 |
| Gross Cost (£) - Mean | 87.08 | 87.08 |
| Gross Cost (£) - Maximum | 16176.77 | 16176.77 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 81.37 | 81.37 |
| Actual Cost (£) - Maximum | 16133.13 | 16133.13 |
| Rows with Missing BNF Category | 5431.00 | 5431.00 |
| Percentage of Items not Categorised | 0.15 | 0.15 |
| File Size (kb) | 65570 | 38811 |

TABLE A.70: Validation Results for GP Prescribing Data 2021-04

| **2021-04** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 450107 | 450107 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.64 | 7.64 |
| Total Items - Max | 1187.00 | 1187.00 |
| Gross Cost (£) - Minimum | 0.02 | 0.02 |
| Gross Cost (£) - Mean | 82.68 | 82.68 |
| Gross Cost (£) - Maximum | 12069.40 | 12069.40 |
| Actual Cost (£) - Minimum | 0.02 | 0.02 |
| Actual Cost (£) - Mean | 77.38 | 77.38 |
| Actual Cost (£) - Maximum | 12049.38 | 12049.38 |
| Rows with Missing BNF Category | 5299.00 | 5299.00 |
| Percentage of Items not Categorised | 0.15 | 0.15 |
| File Size (kb) | 64380 | 38078 |

TABLE A.71: Validation Results for GP Prescribing Data 2021-05

| **2021-05** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 443890 | 443890 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.39 | 7.39 |
| Total Items - Max | 1062.00 | 1062.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 80.76 | 80.76 |
| Gross Cost (£) - Maximum | 13363.30 | 13363.30 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 75.61 | 75.61 |
| Actual Cost (£) - Maximum | 13316.93 | 13316.93 |
| Rows with Missing BNF Category | 6067.00 | 6067.00 |
| Percentage of Items not Categorised | 0.18 | 0.18 |
| File Size (kb) | 63395 | 37451 |

TABLE A.72: Validation Results for GP Prescribing Data 2021-06

| **2021-06** | **Original File** | **Cleaned File** |
|---|---|---|
| Number of Rows | 461846 | 461846 |
| Total Items - Minimum | 1.00 | 1.00 |
| Total Items - Mean | 7.89 | 7.89 |
| Total Items - Max | 1221.00 | 1221.00 |
| Gross Cost (£) - Minimum | 0.01 | 0.01 |
| Gross Cost (£) - Mean | 86.31 | 86.31 |
| Gross Cost (£) - Maximum | 20254.39 | 20254.39 |
| Actual Cost (£) - Minimum | 0.01 | 0.01 |
| Actual Cost (£) - Mean | 80.78 | 80.78 |
| Actual Cost (£) - Maximum | 19040.65 | 19040.65 |
| Rows with Missing BNF Category | 6179.00 | 6179.00 |
| Percentage of Items not Categorised | 0.17 | 0.17 |
| File Size (kb) | 66147 | 39134 |

# Appendix B

# Dispensing by Contractor Validation Results

TABLE B.1: Validation Results for Dispensing by Contractor 2018-04

| **2018-04** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 19433 | 19433 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 166.9 | 166.9 |
| | Maximum | 12675 | 12675 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 14.8 |
| | Maximum | - | 147.2 |
| Size of File (kb) | | 6619 | 7498 |

TABLE B.2: Validation Results for Dispensing by Contractor 2018-05

| **2018-05** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 19551 | 19551 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 178.9 | 178.9 |
| | Maximum | 14587 | 14587 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 14.9 |
| | Maximum | - | 144.9 |
| Size of File (kb) | | 6659 | 7544 |

TABLE B.3: Validation Results for Dispensing by Contractor 2018-06

| **2018-06** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 19379 | 19379 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 175.9 | 175.9 |
| | Maximum | 12841 | 12841 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 14.9 |
| | Maximum | - | 147.2 |
| Size of File (kb) | | 6602 | 7478 |

TABLE B.4: Validation Results for Dispensing by Contractor 2018-07

| **2018-07** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 19433 | 19433 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 173.7 | 173.7 |
| | Maximum | 13704 | 13704 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 15.7 |
| | Maximum | - | 151.2 |
| Size of File (kb) | | 6633 | 7494 |

TABLE B.5: Validation Results for Dispensing by Contractor 2018-08

| **2018-08** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 19708 | 19708 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 176.5 | 176.5 |
| | Maximum | 13246 | 13246 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 15.4 |
| | Maximum | - | 147.2 |
| Size of File (kb) | | 6723 | 7599 |

TABLE B.6: Validation Results for Dispensing by Contractor 2018-09

| **2018-09** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 19122 | 19122 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 167.7 | 167.7 |
| | Maximum | 13325 | 13325 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 14.4 |
| | Maximum | - | 141.7 |
| Size of File (kb) | | 6523 | 7374 |

TABLE B.7: Validation Results for Dispensing by Contractor 2018-10

| **2018-10** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 19691 | 19691 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 182.5 | 182.5 |
| | Maximum | 14468 | 14468 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 14.7 |
| | Maximum | - | 140.6 |
| Size of File (kb) | | 6732 | 7611 |

TABLE B.8: Validation Results for Dispensing by Contractor 2018-11

| **2018-11** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 19406 | 19406 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 177.7 | 177.7 |
| | Maximum | 12122 | 12122 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 14.4 |
| | Maximum | - | 140.3 |
| Size of File (kb) | | 6642 | 7503 |

TABLE B.9: Validation Results for Dispensing by Contractor 2018-12

| **2018-12** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 18588 | 18588 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 175.4 | 175.4 |
| | Maximum | 13379 | 13379 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 13.9 |
| | Maximum | - | 144.9 |
| Size of File (kb) | | 3178 | 5558 |

TABLE B.10: Validation Results for Dispensing by Contractor 2019-01

| **2019-01** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 19443 | 19443 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 189.6 | 189.6 |
| | Maximum | 13921 | 13921 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 14.2 |
| | Maximum | - | 155 |
| Size of File (kb) | | 6635 | 7494 |

TABLE B.11: Validation Results for Dispensing by Contractor 2019-02

| **2019-02** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 18904 | 18904 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 170 | 170 |
| | Maximum | 12390 | 12390 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 14.1 |
| | Maximum | - | 147.2 |
| Size of File (kb) | | 3206 | 5629 |

TABLE B.12: Validation Results for Dispensing by Contractor 2019-03

| **2019-03** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 19119 | 19119 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 176.6 | 176.6 |
| | Maximum | 12865 | 12865 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 14.2 |
| | Maximum | - | 152.3 |
| Size of File (kb) | | 3244 | 5694 |

TABLE B.13: Validation Results for Dispensing by Contractor 2019-04

| **2019-04** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 19031 | 19031 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 179.3 | 179.3 |
| | Maximum | 13040 | 13040 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 14.3 |
| | Maximum | - | 147.2 |
| Size of File (kb) | | 3229 | 5669 |

TABLE B.14: Validation Results for Dispensing by Contractor 2019-05

| **2019-05** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 19175 | 19175 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 187.8 | 187.8 |
| | Maximum | 13313 | 13313 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 14.7 |
| | Maximum | - | 157.5 |
| Size of File (kb) | | 3261 | 5719 |

TABLE B.15: Validation Results for Dispensing by Contractor 2019-06

| **2019-06** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 18706 | 18706 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 176.3 | 176.3 |
| | Maximum | 13410 | 13410 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 14.5 |
| | Maximum | - | 142.6 |
| Size of File (kb) | | 3181 | 5580 |

TABLE B.16: Validation Results for Dispensing by Contractor 2019-07

| **2019-07** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 19140 | 19140 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 186.3 | 186.3 |
| | Maximum | 14395 | 14395 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 15.7 |
| | Maximum | - | 146.4 |
| Size of File (kb) | | 3257 | 5712 |

TABLE B.17: Validation Results for Dispensing by Contractor 2019-08

| **2019-08** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 18878 | 18878 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 182.3 | 182.3 |
| | Maximum | 15419 | 15419 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 15.2 |
| | Maximum | - | 157.7 |
| Size of File (kb) | | 3208 | 5629 |

TABLE B.18: Validation Results for Dispensing by Contractor 2019-09

| **2019-09** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 18641 | 18641 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 181.5 | 181.5 |
| | Maximum | 14114 | 14114 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 14.3 |
| | Maximum | - | 147.2 |
| Size of File (kb) | | 3167 | 5557 |

TABLE B.19: Validation Results for Dispensing by Contractor 2019-10

| **2019-10** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 19050 | 19050 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 192.1 | 192.1 |
| | Maximum | 15067 | 15067 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 14.3 |
| | Maximum | - | 147.3 |
| Size of File (kb) | | 3255 | 5697 |

TABLE B.20: Validation Results for Dispensing by Contractor 2019-11

| **2019-11** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 18580 | 18580 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 183.9 | 183.9 |
| | Maximum | 13407 | 13407 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 14 |
| | Maximum | - | 147.3 |
| Size of File (kb) | | 3178 | 5559 |

TABLE B.21: Validation Results for Dispensing by Contractor 2019-12

| **2019-12** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 18410 | 18410 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 190.9 | 190.9 |
| | Maximum | 14748 | 14748 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 14 |
| | Maximum | - | 147.3 |
| Size of File (kb) | | 3149 | 5509 |

TABLE B.22: Validation Results for Dispensing by Contractor 2020-01

| **2020-01** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 18720 | 18720 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 193.9 | 193.9 |
| | Maximum | 14334 | 14334 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 13.9 |
| | Maximum | - | 155 |
| Size of File (kb) | | 3185 | 5585 |

TABLE B.23: Validation Results for Dispensing by Contractor 2020-02

| **2020-02** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 18306 | 18306 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 177.6 | 177.6 |
| | Maximum | 12125 | 12125 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 13.9 |
| | Maximum | - | 158.7 |
| Size of File (kb) | | 6285 | 7065 |

TABLE B.24: Validation Results for Dispensing by Contractor 2020-03

| **2020-03** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 17803 | 17803 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 220.4 | 220.4 |
| | Maximum | 16727 | 16727 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 14 |
| | Maximum | - | 150.7 |
| Size of File (kb) | | 6116 | 6867 |

TABLE B.25: Validation Results for Dispensing by Contractor 2020-04

| **2020-04** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 14425 | 14425 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 232.8 | 232.8 |
| | Maximum | 11717 | 11717 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 13.7 |
| | Maximum | - | 143.3 |
| Size of File (kb) | | 4952 | 5561 |

TABLE B.26: Validation Results for Dispensing by Contractor 2020-05

| **2020-05** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 13831 | 13831 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 233.3 | 233.3 |
| | Maximum | 11309 | 11309 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 13.4 |
| | Maximum | - | 141.3 |
| Size of File (kb) | | 4744 | 5332 |

TABLE B.27: Validation Results for Dispensing by Contractor 2020-06

| **2020-06** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 14016 | 14016 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 244.7 | 244.7 |
| | Maximum | 14347 | 14347 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 13.1 |
| | Maximum | - | 143.7 |
| Size of File (kb) | | 4812 | 5402 |

TABLE B.28: Validation Results for Dispensing by Contractor 2020-07

| **2020-07** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 13967 | 13967 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 249.6 | 249.6 |
| | Maximum | 13380 | 13380 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 13.4 |
| | Maximum | - | 154.8 |
| Size of File (kb) | | 4795 | 5386 |

TABLE B.29: Validation Results for Dispensing by Contractor 2020-08

| **2020-08** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 13771 | 13771 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 230.5 | 230.5 |
| | Maximum | 12005 | 12005 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 13.2 |
| | Maximum | - | 144.9 |
| Size of File (kb) | | 4730 | 5311 |

TABLE B.30: Validation Results for Dispensing by Contractor 2020-09

| **2020-09** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 13807 | 13807 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 256.5 | 256.5 |
| | Maximum | 14290 | 14290 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 12.8 |
| | Maximum | - | 147.1 |
| Size of File (kb) | | 4739 | 5324 |

TABLE B.31: Validation Results for Dispensing by Contractor 2020-10

| **2020-10** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 13764 | 13764 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 252.5 | 252.5 |
| | Maximum | 13145 | 13145 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 12.9 |
| | Maximum | - | 147.1 |
| Size of File (kb) | | 4742 | 5323 |

TABLE B.32: Validation Results for Dispensing by Contractor 2020-11

| **2020-11** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 13631 | 13631 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 244.9 | 244.9 |
| | Maximum | 13309 | 13309 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 12.6 |
| | Maximum | - | 146.4 |
| Size of File (kb) | | 4696 | 5271 |

TABLE B.33: Validation Results for Dispensing by Contractor 2020-12

| **2020-12** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 13648 | 13648 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 267.2 | 267.2 |
| | Maximum | 13294 | 13294 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 12.5 |
| | Maximum | - | 141.3 |
| Size of File (kb) | | 4705 | 5278 |

TABLE B.34: Validation Results for Dispensing by Contractor 2021-01

| **2021-01** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 13463 | 13463 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 247 | 247 |
| | Maximum | 13637 | 13637 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 12.7 |
| | Maximum | - | 141.5 |
| Size of File (kb) | | 4663 | 5231 |

TABLE B.35: Validation Results for Dispensing by Contractor 2021-02

| **2021-02** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 13345 | 13345 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 240 | 240 |
| | Maximum | 13388 | 13388 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 12.3 |
| | Maximum | - | 142.4 |
| Size of File (kb) | | 4637 | 5205 |

TABLE B.36: Validation Results for Dispensing by Contractor 2021-03

| **2021-03** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 13718 | 13718 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 268.3 | 268.3 |
| | Maximum | 14874 | 14874 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 12.5 |
| | Maximum | - | 143.7 |
| Size of File (kb) | | 4771 | 5354 |

TABLE B.37: Validation Results for Dispensing by Contractor 2021-04

| **2021-04** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 13617 | 13617 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 252.4 | 252.4 |
| | Maximum | 13728 | 13728 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 12.4 |
| | Maximum | - | 146 |
| Size of File (kb) | | 4739 | 5318 |

TABLE B.38: Validation Results for Dispensing by Contractor 2021-05

| **2021-05** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 13675 | 13675 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 239.9 | 239.9 |
| | Maximum | 12626 | 12626 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 12.5 |
| | Maximum | - | 157.6 |
| Size of File (kb) | | 4768 | 5342 |

TABLE B.39: Validation Results for Dispensing by Contractor 2021-06

| **2021-06** | | **Original File** | **Cleaned File** |
|---|---|---|---|
| Number of Rows | | 14030 | 14030 |
| Number of Items | Minimum | 1 | 1 |
| | Mean | 259.8 | 259.8 |
| | Maximum | 15439 | 15439 |
| Distance (km) | Minimum | - | 0 |
| | Mean | - | 13.1 |
| | Maximum | - | 144.3 |
| Size of File (kb) | | 4890 | 5479 |

# Appendix C

# Data Sources

**Dispensing by Contractor**
https://www.opendatani.gov.uk/dataset/dispensing-by-contractor

**GP Prescribing Data**
https://www.opendatani.gov.uk/dataset/gp-prescribing-data

**English Prescribing Dataset (EPD)**
https://digital.nhs.uk/data-and-information/publications/statistical/practice-level-prescribing-data

**NHS Wales General Practice Prescribing Data**
https://nwssp.nhs.wales/ourservices/primary-care-services/general-information/data-and-publications/general-practice-prescribing-data-extract/

**Public Health Scotland Prescriptions in the Community**
https://www.opendata.nhs.scot/dataset/prescriptions-in-the-community

**GP Practice List Sizes**
https://www.opendatani.gov.uk/dataset/gp-practice-list-sizes

**Postcode to Output Area to Lower Layer Super Output Area to Middle Layer Super Output Area to Local Authority District (February 2019)**
http://geoportal.statistics.gov.uk/datasets/c479d770cba14845a0e43db4e3eb5afa

**National Statistics Postcode Lookup (February 2020)**
https://geoportal.statistics.gov.uk/datasets/national-statistics-postcode-lookup-february-2020

**Usual resident population**
https://www.nomisweb.co.uk/census/2011/ks101uk

**Northern Ireland Multiple Deprivation Measure 2017 (NIMDM2017)**

https://www.nisra.gov.uk/statistics/deprivation/northern-ireland-multiple-deprivation-measure-2017-nimdm2017

**Find a GP practice (nidirect website)**

https://www.nidirect.gov.uk/services/gp-practices

**Batch Geocoding Service**

https://www.doogal.co.uk/BatchGeocoding.php

**Mid Year Population Estimates**

https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates

# Appendix D

# British National Formulary (BNF) Structure

All BNF Chapters, Sections and Paragraphs organise in the form Chapter.Section.Paragraph[1]

**1 Gastro-Intestinal System**

- 1.1 Dyspepsia and gastro-oesophageal reflux disease
  - 1.1.1 Antacids and simeticone
  - 1.1.2 Compound alginates and proprietary indigestion preparations
- 1.2 Antispasmodics and other drugs altering gut motility
- 1.3 Antisecretory drugs and mucosal protectants
  - 1.3.1 H2-receptor antagonists
  - 1.3.2 Selective antimuscarinics
  - 1.3.3 Chelates and complexes
  - 1.3.4 Prostaglandin analogues
  - 1.3.5 Proton pump inhibitors
- 1.4 Acute diarrhoea
  - 1.4.1 Adsorbents and bulk-forming drugs
  - 1.4.2 Antimotility drugs
  - 1.4.3 Enkephalinase inhibitors
  - 1.4.4 Tryptophan hydroxylase inhibitors
- 1.5 Chronic bowel disorders
  - 1.5.1 Aminosalicylates
  - 1.5.2 Corticosteroids
  - 1.5.3 Drugs affecting immune response
  - 1.5.4 Food allergy
- 1.6 Laxatives
  - 1.6.1 Bulk-forming laxatives

[1] OenPrescribing All BNF Sections Available at: https://openprescribing.net/bnf/

1.6.2 Stimulant laxatives

1.6.3 Faecal softeners

1.6.4 Osmotic laxatives

1.6.5 Bowel cleansing preparations

1.6.6 Peripheral opioid-receptor antagonists

1.6.7 Other drugs used in constipation

1.7 Local preparations for anal and rectal disorders

1.7.1 Soothing haemorrhoidal preparations

1.7.2 Copound haemorrhoidal preparations with corticosteroid

1.7.3 Rectal sclerosants

1.7.4 Management of anal fissures

1.8 Stoma care

1.8.1 Local care of stoma

1.9 Drugs affecting intestinal secretions

1.9.1 Drugs affecting biliary composition and flow

1.9.4 Pancreatin

**2 Cardiovascular System**

2.1 Positive inotropic drugs

2.1.1 Cardiac glycosides

2.1.2 Phosphodiesterase Type-3 inhibitors

2.2 Diuretics

2.2.1 Thiazides and related diuretics

2.2.2 Loop diuretics

2.2.3 Potassium-sparing diuretics and aldosterone antagonists

2.2.4 Potassium sparing diuretics and compounds

2.2.5 Osmotic diuretics

2.2.8 Diuretics with potassium

2.3 Anti-arrhythmic drugs

2.3.2 Drugs for arrhythmias

2.4 Beta-adrenoceptor blocking drugs

2.5 Hypertension and heart failure

2.5.1 Vasodilator antihypertensive drugs

2.5.2 Centrally-acting antihypertensive drugs

2.5.3 Adrenergic neurone blocking drugs

- 2.5.4 Alpha-adrenoceptor blocking drugs
- 2.5.5 Renin-angiotensin system drugs
- 2.5.8 Other adrenergic neurone blocking drugs

2.6 Nitrates, calcium-channel blockers & other antianginal drugs

- 2.6.1 Nitrates
- 2.6.2 Calcium-channel blockers
- 2.6.3 Other antianginal drugs
- 2.6.4 Peripheral vasodilators and related drugs

2.7 Sympathomimetics

- 2.7.1 Inotropic sympathomimetics
- 2.7.2 Vasoconstrictor sympathomimetics
- 2.7.3 Cardiopulmonary resuscitation

2.8 Anticoagulants and protamine

- 2.8.1 Parenteral anticoagulants
- 2.8.2 Oral anticoagulants
- 2.8.3 Protamine sulfate

2.9 Antiplatelet drugs

2.10 Stable angina, acute coronary syndromes and fibrinolysis

- 2.10.2 Fibrinolytic drugs

2.11 Antifibrinolytic drugs and haemostatics

2.12 Lipid-regulating drugs

2.13 Local sclerosants

**3 Respiratory System**

3.1 Bronchodilators

- 3.1.1 Adrenoceptor agonists
- 3.1.2 Antimuscarinic bronchodilators
- 3.1.3 Theophylline
- 3.1.4 Compound bronchodilator preparations

3.2 Corticosteroids (respiratory)

3.3 Cromoglycate, leukotriene and phosphodesterase type-4 inhib

- 3.3.1 Cromoglycate and related therapy
- 3.3.2 Leukotriene receptor antagonists
- 3.3.3 Phosphodiesterase Type-4 inhibitors

3.4 Antihistamines, hyposensitisation and allergic emergencies

- 3.4.1 Antihistamines
- 3.4.2 Allergen immunotherapy
- 3.4.3 Allergic emergencies

3.5 Respiratory stimulants and pulmonary surfactants

- 3.5.1 Respiratory stimulants

3.6 Oxygen

3.7 Mucolytics

3.8 Aromatic inhalations

3.9 Cough preparations

- 3.9.1 Cough suppressants
- 3.9.2 Expectorant and demulcent cough preparations

3.10 Systemic nasal decongestants

3.11 Antifibrotics

- 3.11.1 Antifibrotics

**4 Central Nervous System**

4.1 Hypnotics and anxiolytics

- 4.1.1 Hypnotics
- 4.1.2 Anxiolytics
- 4.1.3 Barbiturates

4.2 Drugs used in psychoses and related disorders

- 4.2.1 Antipsychotic drugs
- 4.2.2 Antipsychotic depot injections
- 4.2.3 Drugs used for mania and hypomania

4.3 Antidepressant drugs

- 4.3.1 Tricyclic and related antidepressant drugs
- 4.3.2 Monoamine-oxidase inhibitors (maois)
- 4.3.3 Selective serotonin re-uptake inhibitors
- 4.3.4 Other antidepressant drugs

4.4 CNS stimulants and drugs used for ADHD

4.5 Drugs used in the treatment of obesity

- 4.5.1 Gastro-intestinal anti-obesity drugs
- 4.5.2 Centrally-acting appetite suppressants

4.6 Drugs used in nausea and vertigo

4.7 Analgesics

- 4.7.1 Non-opioid analgesics and compound preparations
- 4.7.2 Opioid analgesics
- 4.7.3 Neuropathic pain
- 4.7.4 Antimigraine drugs

4.8 Antiepileptic drugs

- 4.8.1 Control of epilepsy
- 4.8.2 Drugs used in status epilepticus

4.9 Drugs used in parkinsonism and related disorders

- 4.9.1 Dopaminergic drugs used in parkinsonism
- 4.9.2 Antimuscarinic drugs used in parkinsonism
- 4.9.3 Essential tremor,chorea,tics and related disorders

4.10 Drugs used in substance dependence

- 4.10.1 Alcohol dependence
- 4.10.2 Nicotine dependence
- 4.10.3 Opioid dependence

4.11 Drugs for dementia

**5 Infections**

5.1 Antibacterial drugs

- 5.1.1 Penicillins
- 5.1.2 Cephalosporins and other beta-lactams
- 5.1.3 Tetracyclines
- 5.1.4 Aminoglycosides
- 5.1.5 Macrolides
- 5.1.6 Clindamycin and lincomycin
- 5.1.7 Some other antibacterials
- 5.1.8 Sulfonamides and trimethoprim
- 5.1.9 Antituberculosis drugs
- 5.1.10 Antileprotic drugs
- 5.1.11 Metronidazole, tinidazole and ornidazole
- 5.1.12 Quinolones
- 5.1.13 Urinary-tract infections

5.2 Antifungal drugs

- 5.2.1 Triazole antifungals
- 5.2.2 Imidazole antifungals
- 5.2.3 Polyene antifungals

    - 5.2.4 Echinocandin antifungals
    - 5.2.5 Other antifungals
- 5.3 Antiviral drugs
    - 5.3.1 HIV infection
    - 5.3.2 Herpesvirus infections
    - 5.3.3 Viral hepatitis
    - 5.3.4 Influenza
    - 5.3.5 Respiratory syncytial virus
- 5.4 Antiprotozoal drugs
    - 5.4.1 Antimalarials
    - 5.4.2 Amoebicides
    - 5.4.4 Antigiardial drugs
    - 5.4.5 Leishmaniacides
    - 5.4.8 Drugs for pneumocystis pneumonia
- 5.5 Anthelmintics
    - 5.5.1 Drugs for threadworms
    - 5.5.2 Ascaricides
    - 5.5.3 Drugs for tapeworm infections
    - 5.5.5 Schistosomicides
    - 5.5.6 Filaricides

**6 Endocrine System**

- 6.1 Drugs used in diabetes
    - 6.1.1 Insulin
    - 6.1.2 Antidiabetic drugs
    - 6.1.4 Treatment of hypoglycaemia
    - 6.1.6 Diabetic diagnostic and monitoring agents
- 6.2 Thyroid and antithyroid drugs
    - 6.2.1 Thyroid hormones
    - 6.2.2 Antithyroid drugs
- 6.3 Corticosteroids (endocrine)
    - 6.3.1 Replacement therapy
    - 6.3.2 Glucocorticoid therapy
- 6.4 Sex Hormones
    - 6.4.1 Female sex hormones and their modulators

6.4.2 Male sex hormones and antagonists

6.4.3 Anabolic steroids

6.5 Hypothalamic and pituitary hormones and anti, oestrogens

6.5.1 Hypothalamic & anterior pituitary hormone & antioestrogens

6.5.2 Posterior pituitary hormones and antagonists

6.6 Drugs affecting bone metabolism

6.6.1 Calcitonin and parathyroid hormone

6.6.2 Bisphosphonates and other drugs

6.7 Other endocrine drugs

6.7.1 Bromocriptine and other dopaminergic drugs

6.7.2 Drugs affecting gonadotrophins

6.7.3 Metyrapone

6.7.4 Somatomedins

**7 Obstetrics, Gynaecology and Urinary-Tract Disorders**

7.1 Drugs used in obstetrics

7.1.1 Prostaglandins and oxytocics

7.1.2 Mifepristone

7.1.3 Myometrial relaxants

7.2 Treatment of vaginal and vulval conditions

7.2.1 Preparations for vaginal and vulval changes

7.2.2 Vaginal and vulval infections

7.3 Contraceptives

7.3.1 Combined hormonal contraceptives and systems

7.3.2 Progestogen-only contraceptives

7.3.3 Spermicidal contraceptives

7.3.5 Emergency contraception

7.4 Drugs for genito-urinary disorders

7.4.1 Drugs for urinary retention

7.4.2 Drugs for urinary frequency enuresis and incontinence

7.4.3 Drugs used in urological pain

7.4.4 Bladder instillations and urological surgery

7.4.5 Drugs for erectile dysfunction

7.4.6 Drugs for premature ejaculation

**8 Malignant Disease and Immunosuppression**

- 8.1 Cytotoxic drugs
  - 8.1.1 Alkylating drugs
  - 8.1.2 Anthracyclines and cytotoxic antibiotics
  - 8.1.3 Antimetabolites
  - 8.1.4 Vinca alkaloids and etoposide
  - 8.1.5 Other antineoplastic drugs
- 8.2 Drugs affecting the immune response
  - 8.2.1 Antiproliferative immunosuppressants
  - 8.2.2 Corticosteroids and other immunosuppressants
  - 8.2.3 Anti-lymphocyte monoclonal antibodies
  - 8.2.4 Other immunomodulating drugs
- 8.3 Sex hormones and hormone antagonists in malignant disease
  - 8.3.1 Oestrogens
  - 8.3.2 Progestogens
  - 8.3.4 Hormone antagonists

**9 Nutrition and Blood**

- 9.1 Anaemias and some other blood disorders
  - 9.1.1 Iron-deficiency anaemias
  - 9.1.2 Drugs used in megaloblastic anaemias
  - 9.1.3 Hypoplastic,haemolytic and renal anaemias
  - 9.1.4 Drugs used in platelet disorders
  - 9.1.6 Drugs used in neutropenia
- 9.2 Fluids and electrolytes
  - 9.2.1 Oral preparation for fluid and electrolyte imbalance
  - 9.2.2 Parent prepn for fluid and electrolyte imb
- 9.3 Intravenous nutrition
- 9.4 Oral nutrition
  - 9.4.1 Foods for special diets
  - 9.4.2 Enteral nutrition
- 9.5 Minerals
  - 9.5.1 Calcium and magnesium
  - 9.5.2 Phosphorus
  - 9.5.3 Fluoride

- 9.5.4 Zinc and other minerals
- 9.5.5 Selenium

9.6 Vitamins

- 9.6.1 Vitamin A
- 9.6.2 Vitamin B group
- 9.6.3 Vitamin C
- 9.6.4 Vitamin D
- 9.6.5 Vitamin E
- 9.6.6 Vitamin K
- 9.6.7 Multivitamin preparations

9.7 Bitters and tonics

9.8 Metabolic disorders

- 9.8.1 Drugs used in metabolic disorders

9.9 Foods

9.10 Compound vitamin and mineral formulations

9.11 Health supplements

- 9.11.1 Amino acids and nutritional agents
- 9.11.2 Enzymes
- 9.11.3 Glandular
- 9.11.4 Digestive aids

9.12 Other health supplements

**10 Musculoskeletal and Joint Diseases**

10.1 Drugs used in rheumatic diseases and gout

- 10.1.1 Non-steroidal anti-inflammatory drugs
- 10.1.2 Corticosteroids
- 10.1.3 Rheumatic disease suppressant drugs
- 10.1.4 Gout and cytotoxic induced hyperuicaemia
- 10.1.5 Other drugs for rheumatic diseases

10.2 Drugs used in neuromuscular disorders

- 10.2.1 Drugs which enhance neuromuscular transmission
- 10.2.2 Skeletal muscle relaxants

10.3 Drugs for soft-tissue disorders and topical pain relief

- 10.3.1 Enzymes
- 10.3.2 Rubefacients, topical NSAIDs, capsaicin and poultices

**11 Eye**

- 11.3 Anti-infective eye preparations
  - 11.3.1 Antibacterials
  - 11.3.2 Antifungals
  - 11.3.3 Antivirals
- 11.4 Corticosteroids and other anti-inflammatory preparations
  - 11.4.1 Corticosteroids
  - 11.4.2 Other anti-inflammatory preparations
- 11.5 Mydriatics and cycloplegics
- 11.6 Treatment of glaucoma
- 11.7 Local anaesthetics
- 11.8 Miscellaneous ophthalmic preparations
  - 11.8.1 Tear deficiency, eye lubricant and astringent
  - 11.8.2 Ocular diagnostic & peri-operative prepn & photodynamic tt
  - 11.8.3 Other eye preparations

**12 Ear, Nose and Oropharynx**

- 12.1 Drugs acting on the ear
  - 12.1.1 Otitis externa
  - 12.1.3 Removal of ear wax and other substances
- 12.2 Drugs acting on the nose
  - 12.2.1 Drugs used in nasal allergy
  - 12.2.2 Topical nasal decongestants
  - 12.2.3 Nasal preparations for infection
- 12.3 Drugs acting on the oropharynx
  - 12.3.1 Drugs for oral ulceration and inflammation
  - 12.3.2 Oropharyngeal anti-infective drugs
  - 12.3.3 Lozenges and sprays
  - 12.3.4 Mouth-washes, gargles and dentifrices
  - 12.3.5 Treatment of dry mouth

**13 Skin**

- 13.1 Management of skin conditions
  - 13.1.1 Management of skin conditions

- 13.2 Emollient and barrier preparations
  - 13.2.1 Emollients
  - 13.2.2 Barrier preparations
  - 13.2.3 Dusting-powders
- 13.3 Topical local anaesthetics and antipruritics
- 13.4 Topical corticosteroids
- 13.5 Preparations for eczema and psoriasis
  - 13.5.1 Preparations for eczema
  - 13.5.2 Preparations for psoriasis
  - 13.5.3 Drugs affecting the immune response
- 13.6 Acne and rosacea
  - 13.6.1 Topical preparations for acne
  - 13.6.2 Oral preparations for acne
  - 13.6.3 Topical preparation for rosacea
- 13.7 Preparations for warts and calluses
- 13.8 Sunscreens and camouflagers
  - 13.8.1 Sunscreening preparations
  - 13.8.2 Camouflagers
- 13.9 Shampoos and other preparations for scalp and hair conditions
- 13.10 Anti-infective skin preparations
  - 13.10.1 Antibacterial preparations
  - 13.10.2 Antifungal preparations
  - 13.10.3 Antiviral preparations
  - 13.10.4 Parasiticidal preparations
  - 13.10.5 Preparations for minor cuts and abrasions
- 13.11 Skin cleansers,antiseptics and desloughing agents
  - 13.11.1 Alcohols and saline
  - 13.11.2 Chlorhexidine salts
  - 13.11.3 Cationic surfactants and soaps
  - 13.11.4 Chlorine and iodine
  - 13.11.5 Phenolics
  - 13.11.6 Oxidisers and dyes
  - 13.11.7 Desloughing agents
- 13.12 Antiperspirants
- 13.13 Wound management products
  - 13.13.1 Medicated stockings

13.13.8 Gel and colloid dressings

13.14 Topical circulatory preparations

13.15 Miscellaneous topical preparations

**14 Immunological Products and Vaccines**

14.3 Diagnostic vaccines

14.4 Vaccines and antisera

14.5 Immunoglobulins

14.5.1 Normal immunoglobulin

14.5.2 Disease-specific immunoglobulins

14.5.3 Anti-D (Rho) immunoglobulin

**15 Anaesthesia**

15.1 General anaesthesia

15.1.1 Intravenous anaesthetics

15.1.2 Inhalational anaesthetics

15.1.3 Antimuscarinic drugs

15.1.4 Sedative and analgesic peri-operative drgs

15.1.5 Neuromuscular blocking drugs

15.1.6 Anticholinesterases used in anaesthesia

15.1.7 Antagonists for respiratory depression

15.2 Local anaesthesia

15.2.1 Local anaesthetics

**18 Preparations used in Diagnosis**

18.3 X-Ray contrast media

**19 Other Drugs and Preparations**

19.1 Alcohol, wines and spirits

19.2 Selective preparations

19.2.1 Individually formulated preparations bought in

19.2.2 Individually formulated preparations prepared extemp

  - 19.2.3 Homeopathic preparations
  - 19.2.5 Household and other over the counter lines
  - 19.2.7 Poisoning antidotes
- 19.4 Single substances
- 19.5 Other preparations
- 19.6 Acids
  - 19.6.1 Concentrated waters
  - 19.6.3 Extracts
  - 19.6.5 Oils
  - 19.6.6 Tinctures
  - 19.6.7 Syrups
- 19.7 Base,diluent, suspending agents and stabilisers
- 19.8 Colouring,flavouring and sweetening agents
- 19.9 Disinfectants, preservatives and sterilising agents
- 19.13 Cordials and soft drinks
- 19.14 Waters
  - 19.14.1 Sterile water
  - 19.14.2 Purified water
- 19.15 Other gases

**20 Dressings**

- 20.1 Absorbent Cottons
- 20.2 Arm Sling/Bandages
- 20.3 Wound Management & Other Dressings
- 20.4 Gauzes & Gauze Tissue
- 20.5 Tracheostomy & Laryngectomy Appliances
- 20.6 Foam
- 20.7 Lints
- 20.8 Plasters
- 20.9 Stockinette
- 20.10 Surgical Adhesive Tape
- 20.11 Surgical Sutures
- 20.12 Swabs
- 20.13 Unspecified Dressing

20.14 Skin Closure Strips, Sterile

20.15 Skin Adhesive, Sterile

20.16 Tapeless Holders

20.17 Cervical Collar

20.18 Cellulose Wadding BP 1988

20.20 Silk Garments

**21 Appliances**

21.1 Other Appliances

21.2 Catheters

21.3 Chiropody Appliances

21.4 Contraceptive Devices

21.5 Suprapubic Appliances

21.6 Trusses

21.7 Elastic Hosiery

21.8 Oxygen Masks

21.9 Special Sanction Authorisations

21.10 C.A.P.D. Administration Equipment

21.11 Special Authorisation Guernsey

21.12 Peak Flow Meters

21.13 Catheter Maintenance Products

21.14 Lubricant Gels

21.16 Irrigation Solutions

21.17 Nasal Device

21.18 Vacuum Pumps for Erectile Dysfunction

21.19 Oral Film Forming Agents

21.20 Venous Ulcer Compression System

21.21 Dry Mouth Products

21.22 Emollients

21.23 Vaginal Moisturisers

21.24 Nasal Products

21.25 Vaginal Dilators

21.26 Leg Ulcer Wrap

21.27 Lymphoedema Garments

21.28 Anal Irrigation System

21.29 Pressure Offloading Device

21.30 Eye Products

21.31 Cycloidal Vibration Accessories

21.32 Inhalation Solutions

21.33 Indwelling Pleural Cath Drain System

21.34 Vaginal PH Correction Products

21.35 Acne Treatment

21.36 Adhesive Dressing Remover Ster Silicone

21.37 Pelvic Toning Devices

21.38 Low Friction Products

21.39 Prosthetic Adhesives

21.40 Bacterial Decolonisation Products

21.41 Physical Debridement Device

21.42 Jaw Rehabilitation Device

21.43 Micro-Enema - Sodium Citrate

21.44 Debrisoft pad 13cm x 20cm

21.45 Douches

21.46 Hernia Support Garments

21.47 Dev For Fungal Nail Infections

21.48 Detection Sensor Interstitial Fluid/Gluc

21.49 Pulsed Electromagnetic Stimulator

**22 Incontinence Appliances**

22.2 Anal Plugs

22.5 Catheter Valves

22.10 Drainable Dribbling Appliances

22.15 Faecal Collectors

22.20 Incontinence Belts

22.30 Incontinence Sheaths

22.40 Incontinence Sheath Fixing Strips & Adh

22.50 Leg Bags

22.60 Night Drainage Bags

22.70 Suspensory Systems

22.80 Tubing And Accessories

22.85 Insert For Female Stress Incont

22.90 Urinal Systems

**23 Stoma Appliances**

23.5 Adhesive Discs/Rings/Pads/Plasters

23.10 Adhesive (Pastes/Sprays/Solutions)

23.15 Adhesive Removers (Sprays/Liquids/Wipes)

23.20 Bag Closures

23.25 Bag Covers

23.30 Belts

23.35 Colostomy Bags

23.40 Colostomy Sets

23.45 Deodorants

23.46 Discharge Solidifying Agents

23.50 Filters/Bridges

23.55 Flanges

23.60 Ileostomy Bags

23.65 Ileostomy Sets

23.70 Irrigation Washout Appliances

23.75 Pressure Plates/Shields

23.80 Skin Fillers and Protectives

23.85 Skin Protectors

23.90 Stoma Caps/Dressings

23.92 Tubing & Accessories

23.93 Accessories (Guernsey)

23.94 Two Piece Ostomy Systems

23.96 Urostomy Bags

23.98 Urostomy Sets

23.99 Ostomy Appliances R/Sub Allowed Pre 1985

# Appendix E

# Comparison of Northern Ireland Prescribing to that of other UK Nations by BNF Chapter

**Chapter 1 - Gastro-Intestinal System** Analysis of prescribing behaviours of the UK nations for BNF chapter 1 (Gastro-Intestinal System) (Figure E.1) shows Northern Ireland to have the second highest prescribing levels in this category beaten only by Wales. Yearly averages (Table E.1) show that in the period 2015-2020 NI prescribing has risen by 3.47% as did England (3.49%), with prescribing in Scotland falling by 0.96%. There was a strong correlation between NI prescribing and that of the other nations, England (r=.88), Scotland (r=.75), Wales (r=.91).

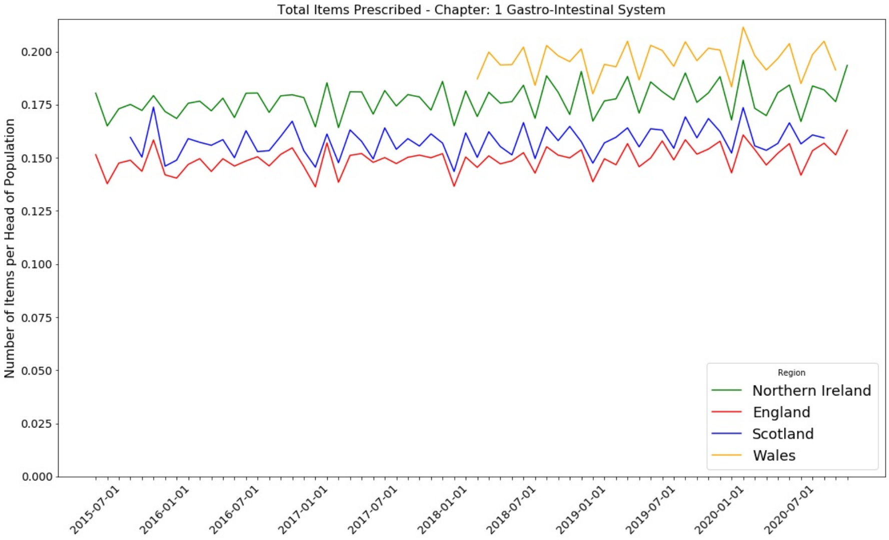


FIGURE E.1: Number of items prescribed per head of population by UK region - BNF Chapter 1 (Gastro-Intestinal System)

TABLE E.1: Yearly average number of items prescribed by head of population - BNF Chapter 1 (Gastro-Intestinal System)

| Year | Northern Ireland | England | Scotland | Wales |
|---|---|---|---|---|
| **2015** | 0.1742 | 0.1480 | 0.1613 | - |
| **2016** | 0.1602 | 0.1475 | 0.1560 | - |
| **2017** | 0.1760 | 0.1481 | 0.1560 | - |
| **2018** | 0.1773 | 0.1486 | 0.1571 | 0.1952 |
| **2019** | 0.1802 | 0.1510 | 0.1600 | 0.1965 |
| **2020** | 0.1802 | 0.1531 | 0.1597 | 0.1968 |

**Chapter 2 - Cardiovascular System** Analysis of prescribing behaviours of the UK nations for BNF chapter 2 (Cardiovascular System) (Figure E.2) shows Northern Ireland having the second lowest prescribing levels in this category with Scotland having the lowest. Wales has the highest prescribing levels with England the second highest. Yearly averages (Table E.2) show that in the period 2015-2020 NI prescribing has risen by 4.78%, whilst prescribing fell in both England and Scotland (-0.04% and -5.84%) respectively. There was a strong correlation between NI prescribing and that of the other nations, England (r=.83), Scotland (r=.63), Wales (r=.90).

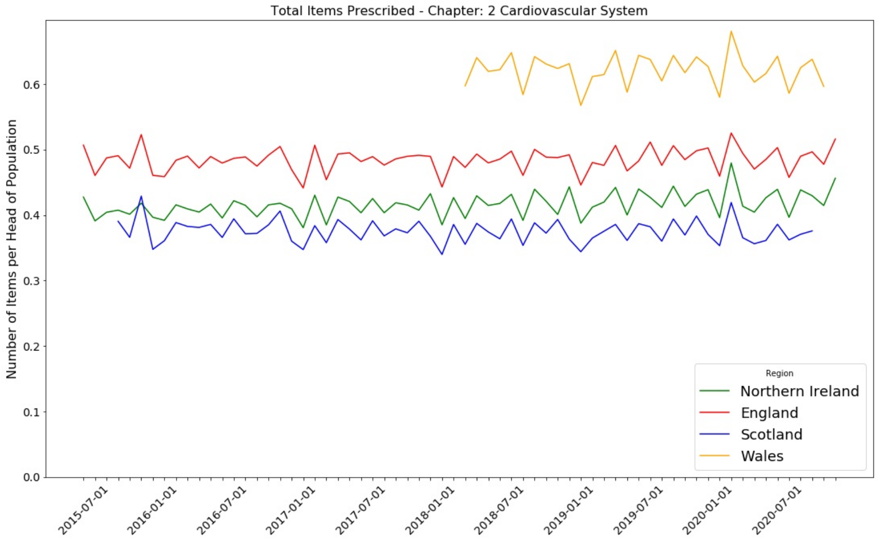


FIGURE E.2: Number of items prescribed per head of population by UK region - BNF Chapter 2 (Cardiovascular System)

TABLE E.2: Yearly average number of items prescribed by head of population - BNF Chapter 2 (Cardiovascular System)

| Year | Northern Ireland | England | Scotland | Wales |
|---|---|---|---|---|
| **2015** | 0.4083 | 0.4899 | 0.3951 | - |
| **2016** | 0.3726 | 0.4816 | 0.3784 | - |
| **2017** | 0.4107 | 0.4811 | 0.3737 | - |
| **2018** | 0.4155 | 0.4823 | 0.3729 | 0.6230 |
| **2019** | 0.4227 | 0.4855 | 0.3738 | 0.6210 |
| **2020** | 0.4278 | 0.4897 | 0.3720 | 0.6202 |

**Chapter 3 - Respiratory System**

Analysis of prescribing behaviours of the UK nations for BNF chapter 3 (Respiratory System) (Figure E.3) shows Northern Ireland having the second highest prescribing levels in this category with Wales having the highest. Yearly averages (Table E.3) show that in the period 2015-2020 NI prescribing has fallen by 2.03%, with prescribing in England and Scotland rising by 1.16% and 1.33% respectively. There was a strong correlation between NI prescribing and that of the other nations, England (r=.87), Scotland (r=.80), Wales (r=.88).

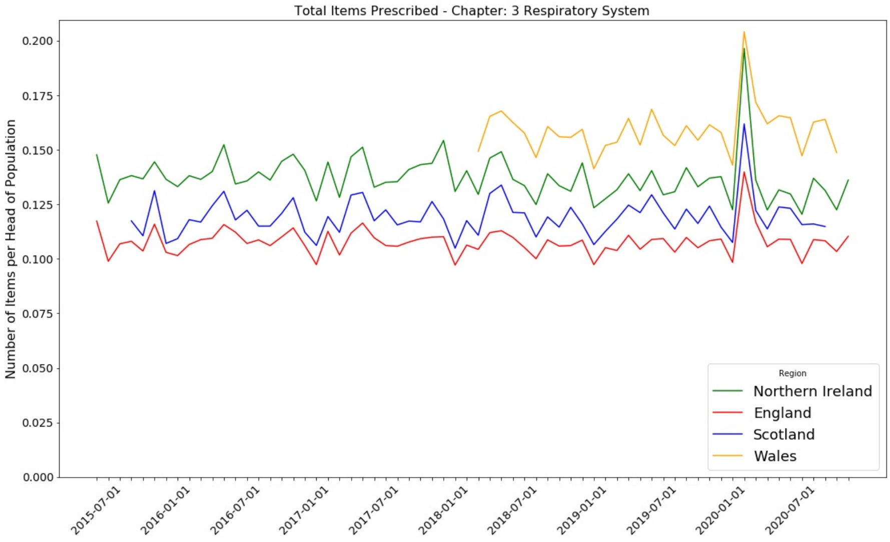


FIGURE E.3: Number of items prescribed per head of population by UK region - BNF Chapter 3 (Respiratory System)

TABLE E.3: Yearly average number of items prescribed by head of population - BNF Chapter 3 (Respiratory System)

| Year | Northern Ireland | England | Scotland | Wales |
|---|---|---|---|---|
| **2015** | 0.1381 | 0.1085 | 0.1198 | - |
| **2016** | 0.1273 | 0.1086 | 0.1188 | - |
| **2017** | 0.1391 | 0.1079 | 0.1188 | - |
| **2018** | 0.1374 | 0.1066 | 0.1188 | 0.1579 |
| **2019** | 0.1341 | 0.1062 | 0.1189 | 0.1564 |
| **2020** | 0.1353 | 0.1097 | 0.1213 | 0.1629 |

**Chapter 4 - Central Nervous System**

Analysis of prescribing behaviours of the UK nations for BNF chapter 4 (Central Nervous System) (Figure E.4) shows Northern Ireland having the highest prescribing levels in this category with Wales coming a close second. Yearly averages (Table E.4) show that in the period 2015-2020 prescribing levels rose in NI (4.33%), England (3.68%) and Scotland (1.09%). There was a strong correlation between NI prescribing and that of the other nations, England (r=.90), Scotland (r=.80), Wales (r=.94).

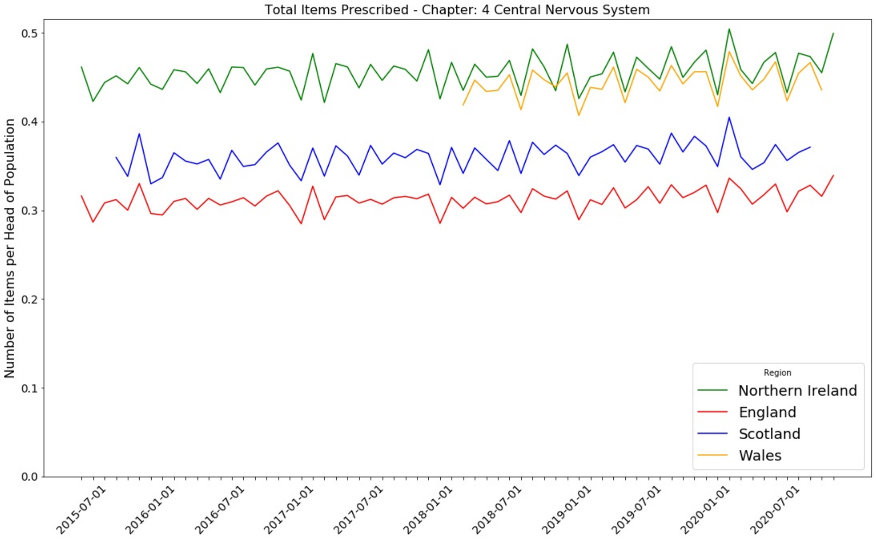


FIGURE E.4: Number of items prescribed per head of population by UK region - BNF Chapter 4 (Central Nervous System)

TABLE E.4: Yearly average number of items prescribed by head of population - BNF Chapter 4 (Central Nervous System)

| Year | Northern Ireland | England | Scotland | Wales |
|---|---|---|---|---|
| **2015** | 0.4471 | 0.3089 | 0.3613 | - |
| **2016** | 0.4125 | 0.3084 | 0.3534 | - |
| **2017** | 0.4518 | 0.3090 | 0.3569 | - |
| **2018** | 0.4542 | 0.3099 | 0.3592 | 0.4383 |
| **2019** | 0.4591 | 0.3139 | 0.3656 | 0.4437 |
| **2020** | 0.4665 | 0.3202 | 0.3653 | 0.4485 |

### Chapter 5 - Infections

Analysis of prescribing behaviours of the UK nations for BNF chapter 5 (Infections) (Figure E.5) shows Northern Ireland having the highest prescribing levels in this category. Yearly averages (Table E.5) show that in the period 2015-2020 prescribing levels fell in NI (24.9%), England (21.65%) and Scotland (20.19%). Whist prescribing levels had already been falling for the three nations between 2015 and 2019, the large fall seen in 2020 is most likely to have been caused by the lack of socialising during the lockdown imposed as a result of the COVID-19 pandemic. There was a strong correlation between NI prescribing and that of the other nations, England (r=.97), Scotland (r=.97), Wales (r=.96).

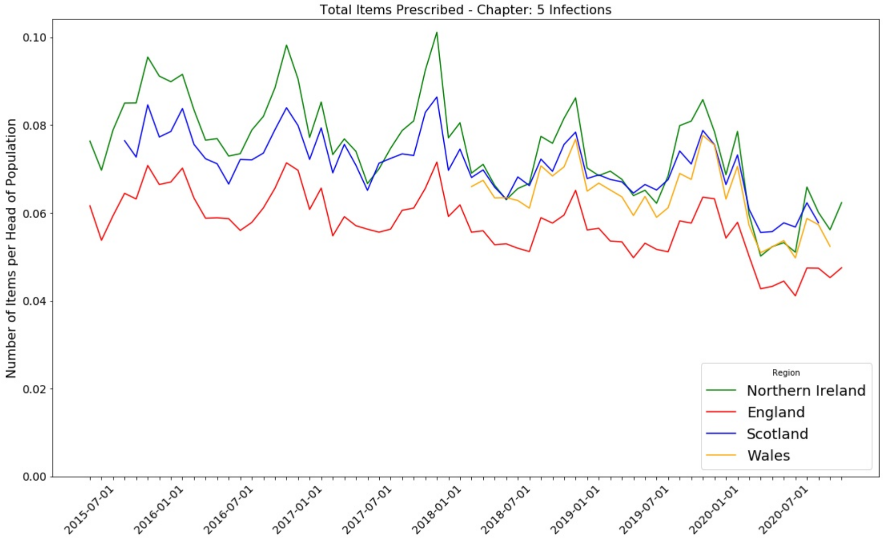


FIGURE E.5: Number of items prescribed per head of population by UK region - BNF Chapter 5 (Infections)

TABLE E.5: Yearly average number of items prescribed by head of population - BNF Chapter 5 (Infections)

| Year | Northern Ireland | England | Scotland | Wales |
|---|---|---|---|---|
| **2015** | 0.0817 | 0.0622 | 0.0779 | - |
| **2016** | 0.0772 | 0.0629 | 0.0755 | - |
| **2017** | 0.0784 | 0.0602 | 0.0738 | - |
| **2018** | 0.0746 | 0.0574 | 0.0708 | 0.0660 |
| **2019** | 0.0724 | 0.0558 | 0.0698 | 0.0662 |
| **2020** | 0.0614 | 0.0487 | 0.0622 | 0.0583 |

### Chapter 6 - Endocrine System

Analysis of prescribing behaviours of the UK nations for BNF chapter 6 (Endocrine System) (Figure E.6) shows Northern Ireland having the second lowest prescribing levels in this category with Scotland having the lowest. Yearly averages (Table E.6) show that in the period 2015-2020 prescribing levels rose in both NI (6.17%) and England (6.54%) with prescribing levels in Scotland falling by 0.08%. There was a strong correlation between NI prescribing and that of the other nations, England (r=.83), Scotland (r=.75), Wales (r=.85).

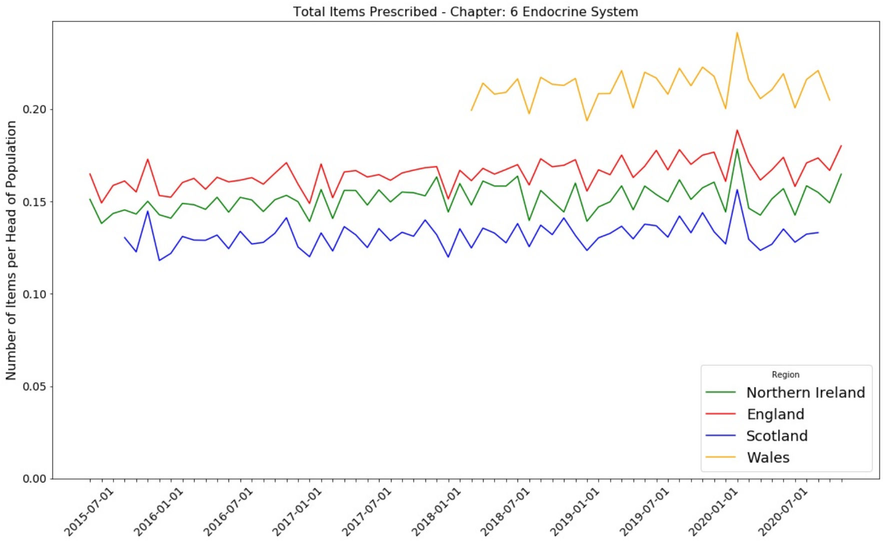


FIGURE E.6: Number of items prescribed per head of population by UK region - BNF Chapter 6 (Endocrine System)

TABLE E.6: Yearly average number of items prescribed by head of population - BNF Chapter 6 (Endocrine System)

| Year | Northern Ireland | England | Scotland | Wales |
|---|---|---|---|---|
| **2015** | 0.1452 | 0.1602 | 0.1326 | - |
| **2016** | 0.1352 | 0.1606 | 0.1289 | - |
| **2017** | 0.1512 | 0.1627 | 0.1302 | - |
| **2018** | 0.1538 | 0.1656 | 0.1317 | 0.2097 |
| **2019** | 0.1526 | 0.1695 | 0.1340 | 0.2125 |
| **2020** | 0.1541 | 0.1707 | 0.1325 | 0.2138 |

### Chapter 7 - Obstetrics

Analysis of prescribing behaviours of the UK nations for BNF chapter 7 (Obstetrics) (Figure E.7) shows Northern Ireland having the second highest prescribing levels in this category although the levels seen are similar to that of England and Scotland. Wales has the highest prescribing levels. Yearly averages (Table E.7) show that in the period 2015-2020 prescribing levels rose in both NI (0.33%) and England (5.03%) with prescribing levels in Scotland falling by 8.73%. There was a strong correlation between NI prescribing and that of the other nations, England (r=.79), Scotland (r=.63), Wales (r=.89).

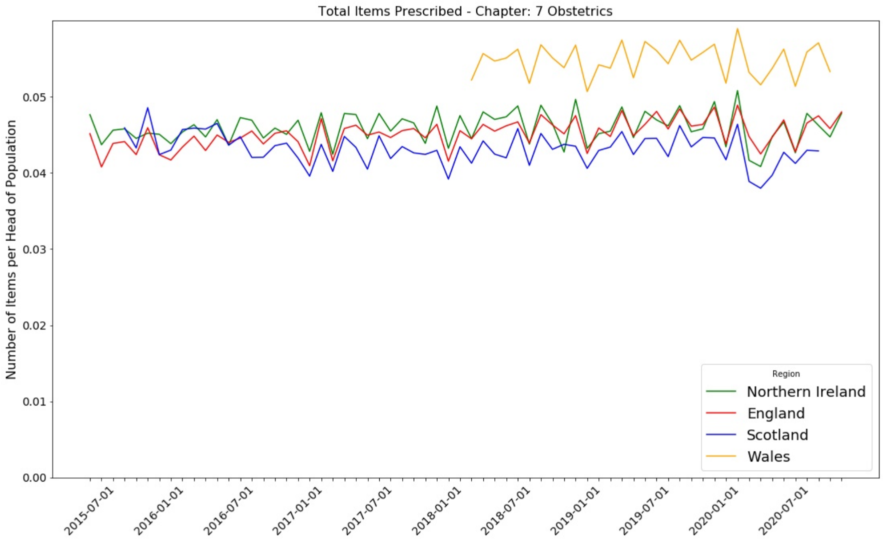


FIGURE E.7: Number of items prescribed per head of population by UK region - BNF Chapter 7 (Obstetrics)

TABLE E.7: Yearly average number of items prescribed by head of population - BNF Chapter 7 (Obstetrics)

| Year | Northern Ireland | England | Scotland | Wales |
|---|---|---|---|---|
| **2015** | 0.0454 | 0.0437 | 0.0459 | - |
| **2016** | 0.0415 | 0.0440 | 0.0441 | - |
| **2017** | 0.0459 | 0.0447 | 0.0424 | - |
| **2018** | 0.0464 | 0.0454 | 0.0428 | 0.0546 |
| **2019** | 0.0465 | 0.0462 | 0.0436 | 0.0551 |
| **2020** | 0.0455 | 0.0459 | 0.0419 | 0.0545 |

**Chapter 8 - Malignant Disease & Immunosuppression**

Analysis of prescribing behaviours of the UK nations for BNF chapter 8 (Malignant Disease & Immunosuppression) (Figure E.8) shows Northern Ireland and Scotland having similar prescribing levels with Wales having higher levels and England having lower levels than both nations. Yearly averages (Table E.8) show that in the period 2015-2020 prescribing levels rose in all nations with NI (14.29%) having the highest rise over the period, Scotland 9.08% and England 2.02%. There was a strong correlation between NI prescribing and that of the other nations, England (r=.75), Scotland (r=.87), Wales (r=.84).

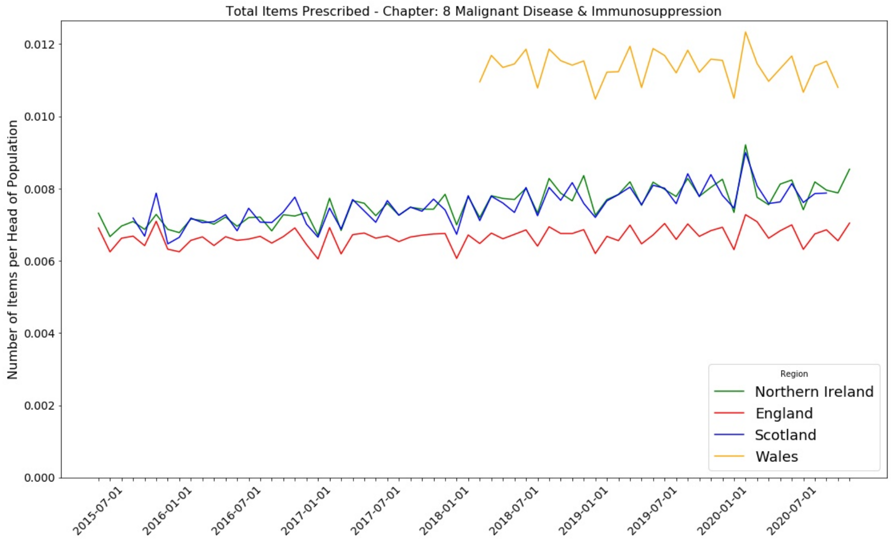


FIGURE E.8: Number of items prescribed per head of population by UK region - BNF Chapter 8 (Malignant Disease & Immunosuppression)

TABLE E.8: Yearly average number of items prescribed by head of population - BNF Chapter 8 (Malignant Disease & Immunosuppression)

| Year | Northern Ireland | England | Scotland | Wales |
|---|---|---|---|---|
| **2015** | 0.0070 | 0.0067 | 0.0072 | - |
| **2016** | 0.0065 | 0.0066 | 0.0071 | - |
| **2017** | 0.0074 | 0.0066 | 0.0073 | - |
| **2018** | 0.0077 | 0.0067 | 0.0076 | 0.0114 |
| **2019** | 0.0079 | 0.0067 | 0.0078 | 0.0114 |
| **2020** | 0.0080 | 0.0068 | 0.0079 | 0.0113 |

**Chapter 9 - Nutrition And Blood**

Analysis of prescribing behaviours of the UK nations for BNF chapter 9 (Nutrition And Blood) (Figure E.9) shows Northern Ireland having lower levels of prescribing than England at the start of the period but reversing this position by 2020. Wales consistently has the highest prescribing levels and Scotland consistently has the lowest prescribing levels. Yearly averages (Table E.9) show that in the period 2015-2020 prescribing levels rose in all nations with NI (18.10%) having the highest rise over the period, Scotland 5.26% and England 0.85%. There was a moderate correlation between NI prescribing and that of England (r=.56), but a high correlation between NI and Scotland (r=.74) and Wales (r=.87).

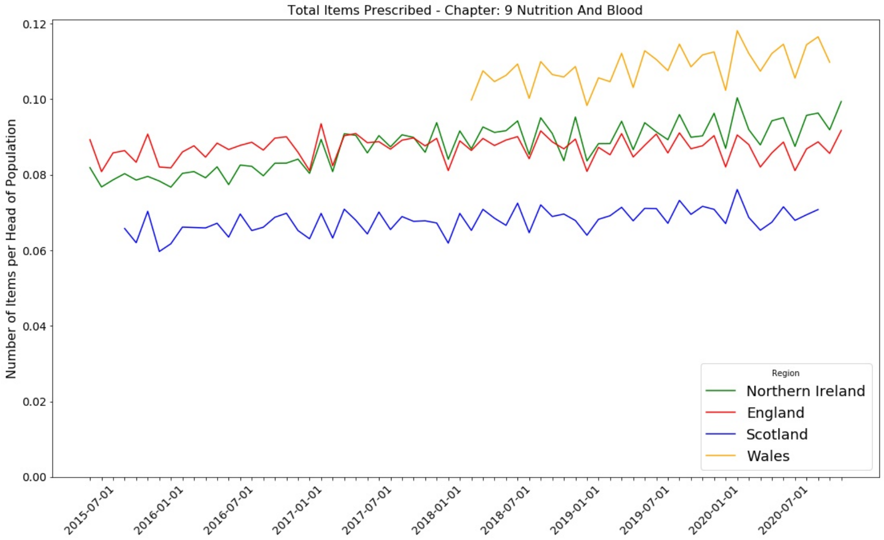


FIGURE E.9: Number of items prescribed per head of population by UK region - BNF Chapter 9 (Nutrition And Blood)

TABLE E.9: Yearly average number of items prescribed by head of population - BNF Chapter 9 (Nutrition And Blood)

| Year | Northern Ireland | England | Scotland | Wales |
|---|---|---|---|---|
| **2015** | 0.0793 | 0.0861 | 0.0661 | - |
| **2016** | 0.0737 | 0.0867 | 0.0658 | - |
| **2017** | 0.0872 | 0.0879 | 0.0671 | - |
| **2018** | 0.0901 | 0.0879 | 0.0682 | 0.1056 |
| **2019** | 0.0906 | 0.0874 | 0.0694 | 0.1082 |
| **2020** | 0.0937 | 0.0868 | 0.0695 | 0.1114 |

**Chapter 10 - Musculoskeletal & Joint Diseases**

Analysis of prescribing behaviours of the UK nations for BNF chapter 10 (Nutrition And Blood) (Figure E.10) shows Northern Ireland having the highest levels of prescribing and England the lowest. Yearly averages (Table E.10) show that in the period 2015-2020 prescribing levels fell steadily in all nations with NI falling by 12.26%, England by 16.34% and Scotland by 12.46%. There was a high correlation between NI prescribing and that of the other nations - England (r=.92), Scotland (r=.83) and Wales (r=.93).

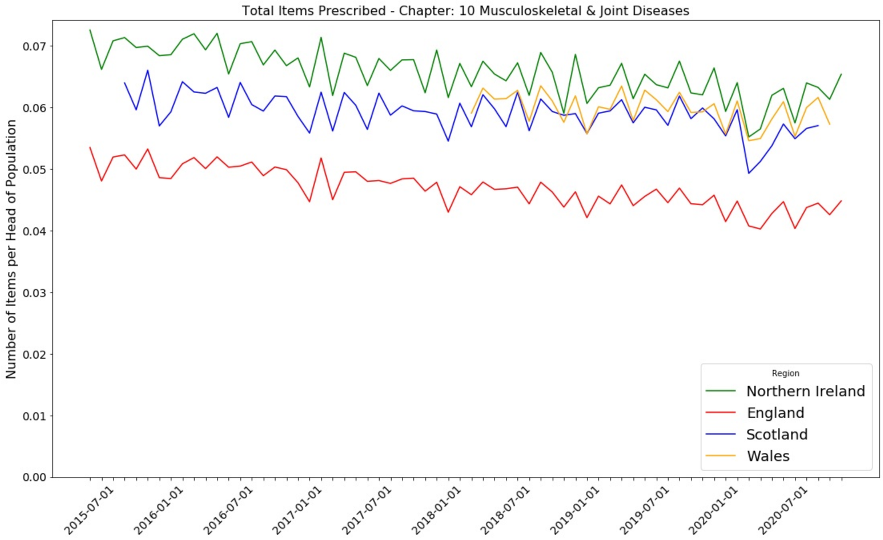


FIGURE E.10: Number of items prescribed per head of population by UK region - BNF Chapter 10 (Musculoskeletal & Joint Diseases)

TABLE E.10: Yearly average number of items prescribed by head of population - BNF Chapter 10 (Musculoskeletal & Joint Diseases)

| Year | Northern Ireland | England | Scotland | Wales |
|---|---|---|---|---|
| **2015** | 0.0701 | 0.0515 | 0.0632 | - |
| **2016** | 0.0632 | 0.0502 | 0.0612 | - |
| **2017** | 0.0664 | 0.0480 | 0.0594 | - |
| **2018** | 0.0651 | 0.0462 | 0.0590 | 0.0609 |
| **2019** | 0.0641 | 0.0452 | 0.0591 | 0.0603 |
| **2020** | 0.0615 | 0.0430 | 0.0553 | 0.0582 |

**Chapter 11 - Eye**

Analysis of prescribing behaviours of the UK nations for BNF chapter 11 (Eye) (Figure E.11) shows Northern Ireland having similar levels of prescribing as England and Scotland until mid 2018 when levels fell dramatically resulting in NI having the lowest levels of prescribing. Wales consistently had the highest prescribing levels. Yearly averages (Table E.11) show that in the period 2015-2020 prescribing levels fell steadily in all nations with NI falling by 30.32%, England by 18.77% and Scotland by 17.01%. There was a high correlation between NI prescribing and that of the other nations - England (r=.88), Scotland (r=.85) and Wales (r=.83).

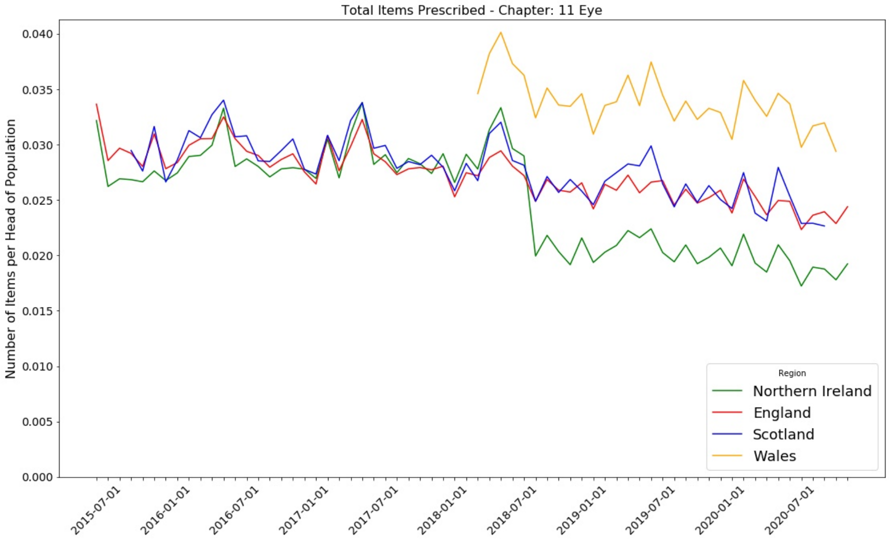


FIGURE E.11: Number of items prescribed per head of population by UK region - BNF Chapter 11 (Eye)

TABLE E.11: Yearly average number of items prescribed by head of population - BNF Chapter 11 (Eye)

| Year | Northern Ireland | England | Scotland | Wales |
|---|---|---|---|---|
| **2015** | 0.0277 | 0.0300 | 0.0296 | - |
| **2016** | 0.0259 | 0.0295 | 0.0302 | - |
| **2017** | 0.0289 | 0.0286 | 0.0295 | - |
| **2018** | 0.0264 | 0.0271 | 0.0278 | 0.0357 |
| **2019** | 0.0207 | 0.0258 | 0.0266 | 0.0339 |
| **2020** | 0.0193 | 0.0244 | 0.0245 | 0.0324 |

**Chapter 12 - Ear, Nose And Oropharynx**

Analysis of prescribing behaviours of the UK nations for BNF chapter 12 (Ear, Nose And Oropharynx) (Figure E.12) shows Northern Ireland having similar levels of prescribing as Scotland and Wales until 2020 when levels fell slightly resulting in NI having the second lowest levels of prescribing. England consistently had the lowest prescribing levels. Yearly averages (Table E.12) show that in the period 2015-2018 prescribing levels were consistent, falling steadily from 2018 onward in all nations. Over the period 2015-2020 NI prescribing levels fell by 17.90%, England by 13.81% and Scotland by 14.07%. There was a high correlation between NI prescribing and that of the other nations - England (r=.91), Scotland (r=.88) and Wales (r=.84).

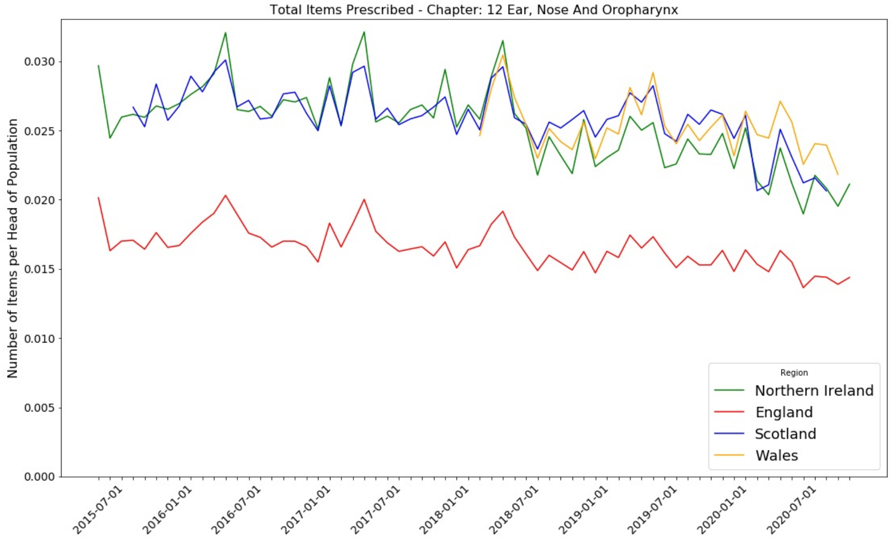


FIGURE E.12: Number of items prescribed per head of population by UK region - BNF Chapter 12 (Ear, Nose And Oropharynx)

TABLE E.12: Yearly average number of items prescribed by head of population - BNF Chapter 12 (Ear, Nose And Oropharynx)

| Year | Northern Ireland | England | Scotland | Wales |
|---|---|---|---|---|
| **2015** | 0.0265 | 0.0174 | 0.0268 | - |
| **2016** | 0.0251 | 0.0177 | 0.0275 | - |
| **2017** | 0.0271 | 0.0171 | 0.0267 | - |
| **2018** | 0.0259 | 0.0164 | 0.0261 | 0.0258 |
| **2019** | 0.0239 | 0.0160 | 0.0261 | 0.0255 |
| **2020** | 0.0218 | 0.0150 | 0.0230 | 0.0245 |

### Chapter 13 - Skin

Analysis of prescribing behaviours of the UK nations for BNF chapter 13 (Skin) (Figure E.13) shows Northern Ireland having the highest of prescribing with Scotland second highest, Wales third highest and England lowest. Yearly averages (Table E.13) show that in the period 2015-2018 prescribing levels fell in all nations with NI prescribing levels falling by 26.54%, England by 30.41% and Scotland by 25.18%. There was a high correlation between NI prescribing and that of the other nations - England (r=.94), Scotland (r=.93) and Wales (r=.93).

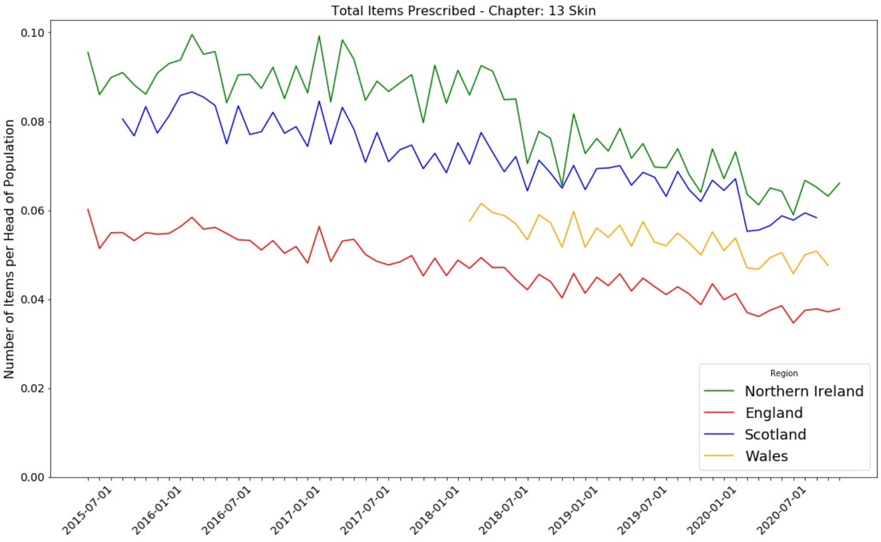


FIGURE E.13: Number of items prescribed per head of population by UK region - BNF Chapter 13 (Skin)

TABLE E.13: Yearly average number of items prescribed by head of population - BNF Chapter 13 (Skin)

| Year | Northern Ireland | England | Scotland | Wales |
|---|---|---|---|---|
| **2015** | 0.0894 | 0.0549 | 0.0802 | - |
| **2016** | 0.0835 | 0.0543 | 0.0810 | - |
| **2017** | 0.0895 | 0.0501 | 0.0759 | - |
| **2018** | 0.0831 | 0.0459 | 0.0706 | 0.0573 |
| **2019** | 0.0728 | 0.0428 | 0.0670 | 0.0541 |
| **2020** | 0.0657 | 0.0382 | 0.0600 | 0.0498 |

**Chapter 14 - Immunological Products & Vaccines**

Analysis of prescribing behaviours of the UK nations for BNF chapter 14 (Immunological Products & Vaccines) (Figure E.14) shows Northern Ireland and Scotland having the low of prescribing with England and Wales showing seasonal spikes in prescribing levels. Yearly averages (Table E.14) show that in the period 2015-2018 prescribing levels fell in all nations with NI prescribing levels falling by 63.06%, England by 43.01% and Scotland by 71.23%. It is likely that the dramatic fall in prescribing levels during 2020 for NI and Scotland were caused by the COVID-19 pandemic and subsequent lockdowns. There was a high correlation between NI prescribing and that of Scotland (r=.91) with a weak negative correlation between NI and England (r=-.21) and Wales (r=-.18).

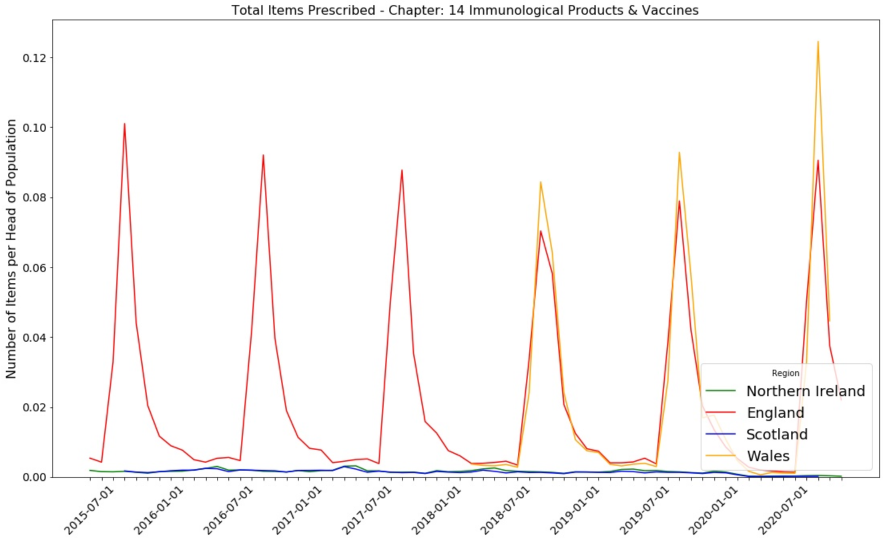


FIGURE E.14: Number of items prescribed per head of population by UK region - BNF Chapter 14 (Immunological Products & Vaccines)

TABLE E.14: Yearly average number of items prescribed by head of population - BNF Chapter 14 (Immunological Products & Vaccines)

| Year | Northern Ireland | England | Scotland | Wales |
|---|---|---|---|---|
| **2015** | 0.0015 | 0.0347 | 0.0015 | - |
| **2016** | 0.0018 | 0.0205 | 0.0019 | - |
| **2017** | 0.0018 | 0.0200 | 0.0017 | - |
| **2018** | 0.0017 | 0.0191 | 0.0014 | 0.0237 |
| **2019** | 0.0016 | 0.0191 | 0.0013 | 0.0198 |
| **2020** | 0.0006 | 0.0198 | 0.0004 | 0.0219 |

**Chapter 15 - Anesthesia**

Analysis of prescribing behaviours of the UK nations for BNF chapter 15 (anesthesia) (Figure E.15) shows Northern Ireland having the highest prescribing rates at the start of the period with Scotland's prescribing rates climbing over the period to match those of NI. Prescribing rates in England and Wales fell dramatically at the end of 2019 before starting to pick up again. Both England and Wales ending the period with much lower rates than those seen in NI and Scotland. Yearly averages (Table E.15) show that in the period 2015-2018 prescribing levels rose in NI by 20.62% and Scotland by 42.41%. Prescribing levels in England were steady until the end of 2019 before plummeting resulting in an overall fall of 8.80%, likely due to the COVID-19 pandemic. There was a high correlation between NI prescribing and that of Scotland (r=.77) with a weak negative correlation between NI and England (r=-.01) and a weak correlation between NI and Wales (r=.29).

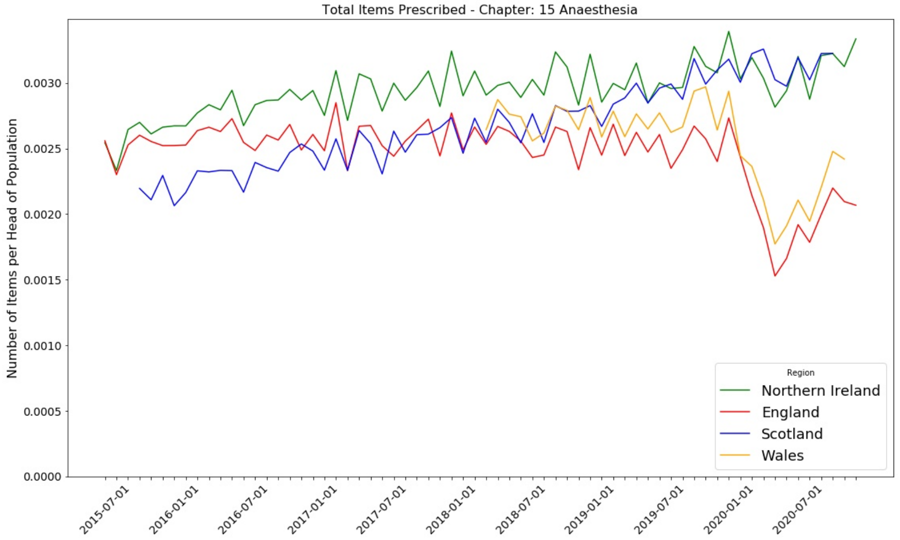


FIGURE E.15: Number of items prescribed per head of population by UK region - BNF Chapter 15 (Anesthesia)

TABLE E.15: Yearly average number of items prescribed by head of population - BNF Chapter 15 (anesthesia)

| Year | Northern Ireland | England | Scotland | Wales |
|---|---|---|---|---|
| **2015** | 0.0026 | 0.0025 | 0.0022 | - |
| **2016** | 0.0026 | 0.0026 | 0.0023 | - |
| **2017** | 0.0029 | 0.0026 | 0.0025 | - |
| **2018** | 0.0030 | 0.0026 | 0.0027 | 0.0027 |
| **2019** | 0.0030 | 0.0025 | 0.0029 | 0.0027 |
| **2020** | 0.0031 | 0.0020 | 0.0031 | 0.0022 |

**Chapter 19 - Other Drugs And Preparations**

Due to a change in the reporting of medications for BNF chapter 19 (Other Drugs And Preparations) (Figure E.16) in Northern Ireland during 2018 it is not clear whether all nations are reporting on the same basis.

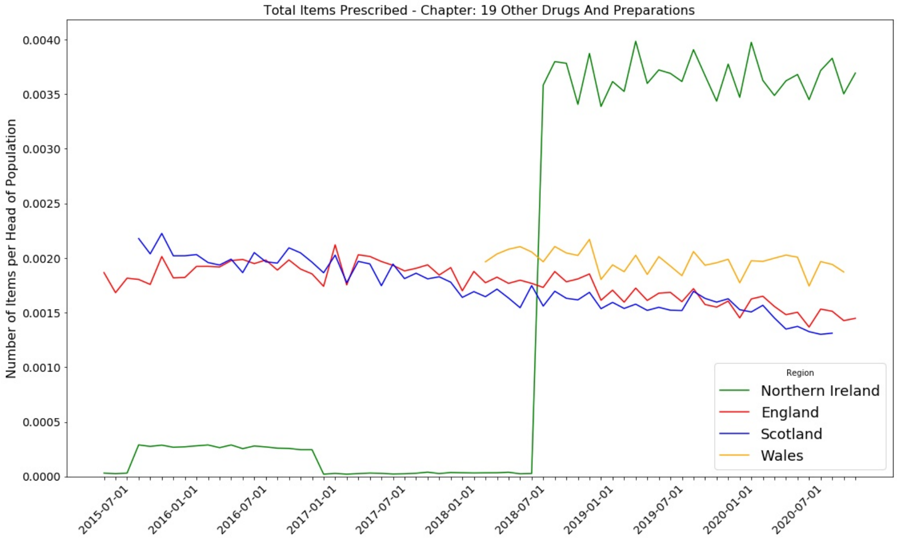


FIGURE E.16: Number of items prescribed per head of population by UK region - BNF Chapter 19 (Other Drugs And Preparations)

**Chapter 20 - Dressings**

Analysis of prescribing behaviours of the UK nations for BNF chapter 20 (Dressings) (Figure E.17) shows Northern Ireland having the highest prescribing rates and England the lowest. Prescribing rates fell in all nations over the period 2015-202 with yearly averages (Table E.16) showing that levels fell in NI by 26.65%, England by 36.18% and in Scotland by 18.03%. There was a high correlation between NI prescribing and that of all nations - England (r=.92), Scotland (r=.82) and Wales (r=.76).

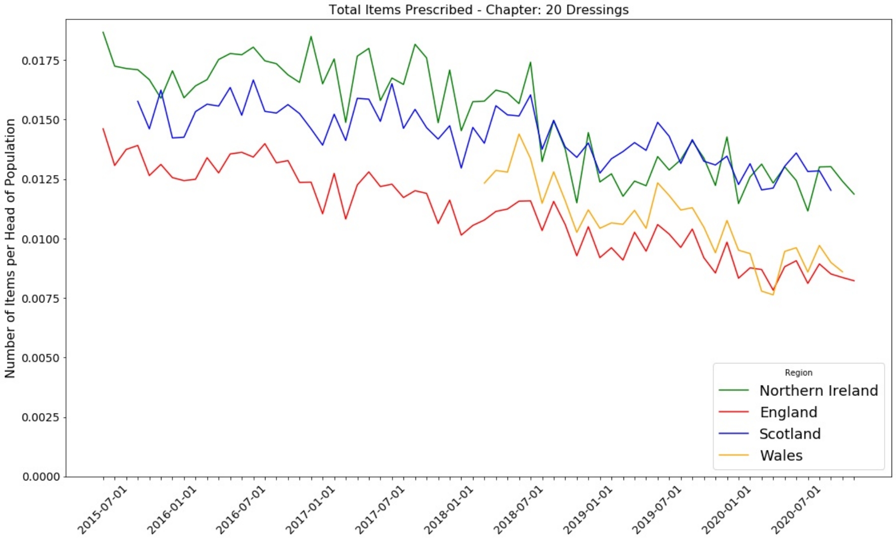


FIGURE E.17: Number of items prescribed per head of population by UK region - BNF Chapter 20 (Dressings)

TABLE E.16: Yearly average number of items prescribed by head of population - BNF Chapter 20 (Dressings)

| Year | Northern Ireland | England | Scotland | Wales |
|---|---|---|---|---|
| **2015** | 0.0171 | 0.0135 | 0.0155 | - |
| **2016** | 0.0156 | 0.0131 | 0.0154 | - |
| **2017** | 0.0169 | 0.0119 | 0.0150 | - |
| **2018** | 0.0152 | 0.0109 | 0.0145 | 0.0124 |
| **2019** | 0.0129 | 0.0097 | 0.0137 | 0.0109 |
| **2020** | 0.0126 | 0.0086 | 0.0127 | 0.0091 |

### Chapter 21 - Appliances

Due to a change in the reporting of medications for BNF chapter 21 (Appliances) (Figure E.18) in Northern Ireland during 2018 it is not clear whether all nations are reporting on the same basis.

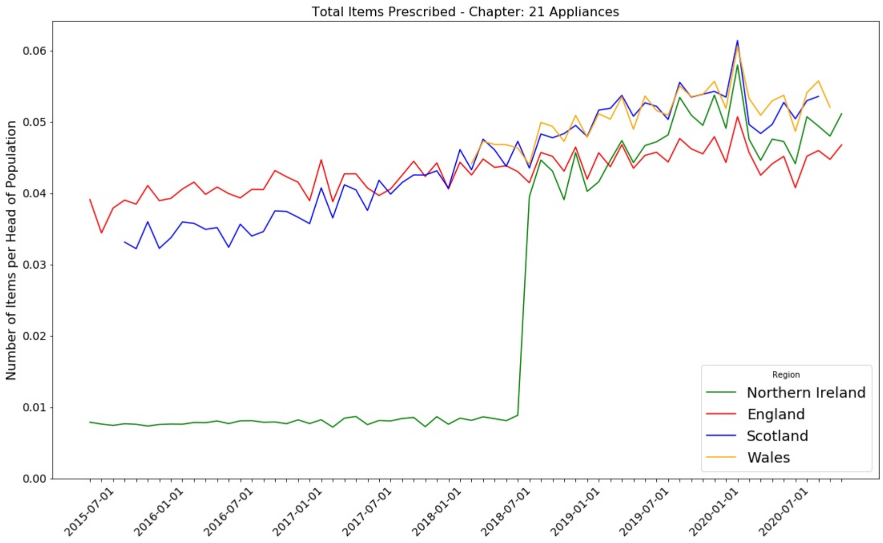


FIGURE E.18: Number of items prescribed per head of population by UK region - BNF Chapter 21 (Appliances)

### Chapter 22 - Incontinence Appliances

Analysis of prescribing behaviours of the UK nations for BNF chapter 22 (Incontinence Appliances) (Figure E.19) shows Northern Ireland having the joint lowest prescribing rates along with Wales. Scotland has the highest with England second highest. Yearly averages (Table E.17) shows that levels rose in NI by 3.25% and in England by 3.65%. Scotland saw a fall in prescribing of 5.42%. There was a moderate correlation between NI prescribing and that of England (r=.50) and weak correlations with Scotland (r=.28) and Wales (r=.37).

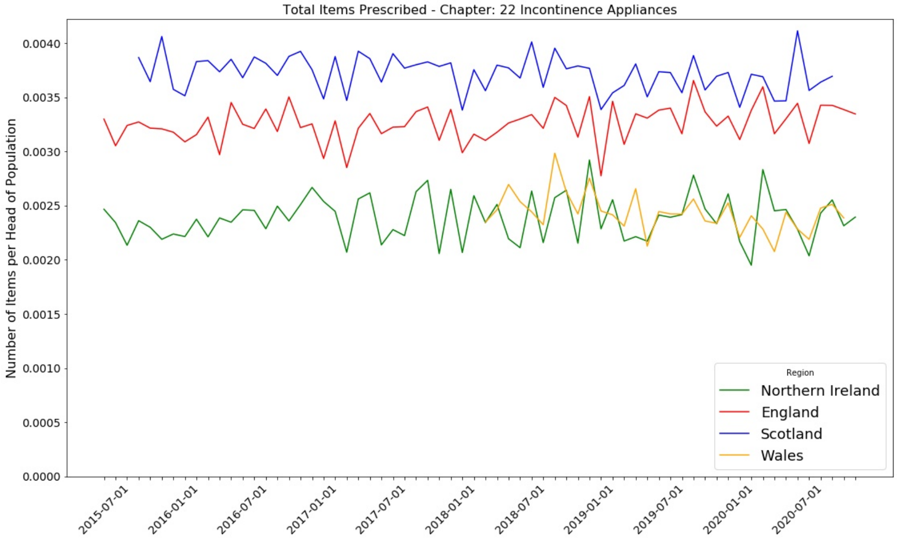


FIGURE E.19: Number of items prescribed per head of population by UK region - BNF Chapter 22 (Incontinence Appliances)

TABLE E.17: Yearly average number of items prescribed by head of population - BNF Chapter 22 (Incontinence Appliances)

| Year | Northern Ireland | England | Scotland | Wales |
|---|---|---|---|---|
| **2015** | 0.0023 | 0.0032 | 0.0039 | - |
| **2016** | 0.0022 | 0.0032 | 0.0038 | - |
| **2017** | 0.0024 | 0.0032 | 0.0038 | - |
| **2018** | 0.0024 | 0.0032 | 0.0037 | 0.0025 |
| **2019** | 0.0024 | 0.0033 | 0.0036 | 0.0024 |
| **2020** | 0.0024 | 0.0033 | 0.0036 | 0.0023 |

### Chapter 23 - Stoma Appliances

Analysis of prescribing behaviours of the UK nations for BNF chapter 23 (Stoma Appliances) (Figure E.20) shows Northern Ireland starting the period with the lowest prescribing rates of all the nations but climbing steadily to replace Scotland as the second highest by the end of the period. England had the highest prescribing levels during the whole period. Yearly averages (Table E.18) shows that prescribing levels rose in NI by 61.12%, in England by 14.43% and Scotland by 7.19%. There was a high correlation between NI prescribing and that of England (r=.72) and Scotland (r=.60) and a weak correlation with prescribing in Wales (r=.38).

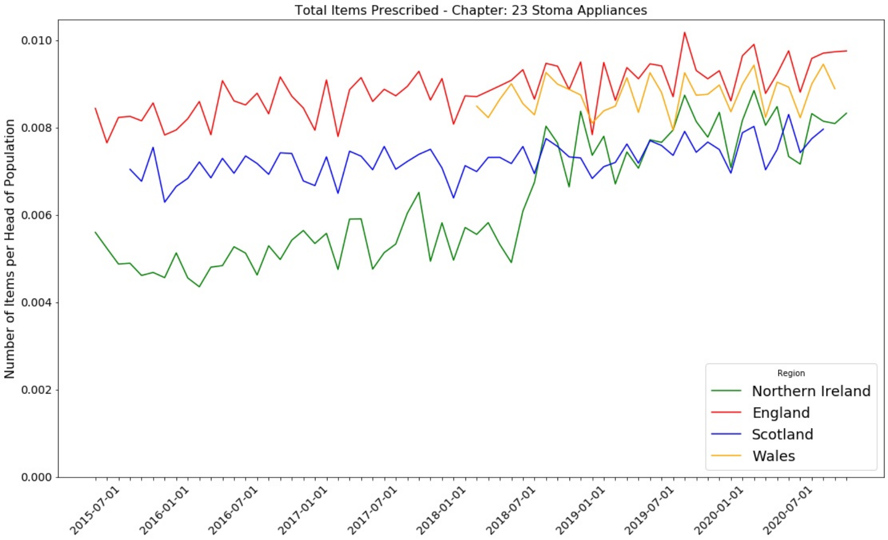


FIGURE E.20: Number of items prescribed per head of population by UK region - BNF Chapter 23 (Stoma Appliances)

TABLE E.18: Yearly average number of items prescribed by head of population - BNF Chapter 23 (Stoma Appliances)

| Year | Northern Ireland | England | Scotland | Wales |
|---|---|---|---|---|
| **2015** | 0.0050 | 0.0082 | 0.0071 | - |
| **2016** | 0.0044 | 0.0085 | 0.0070 | - |
| **2017** | 0.0055 | 0.0087 | 0.0072 | - |
| **2018** | 0.0061 | 0.0089 | 0.0072 | 0.0087 |
| **2019** | 0.0077 | 0.0092 | 0.0074 | 0.0087 |
| **2020** | 0.0080 | 0.0094 | 0.0076 | 0.0089 |

# Appendix F

# Prescribing trends by BNF chapter of Metropolitan and Non-Metropolitan practices

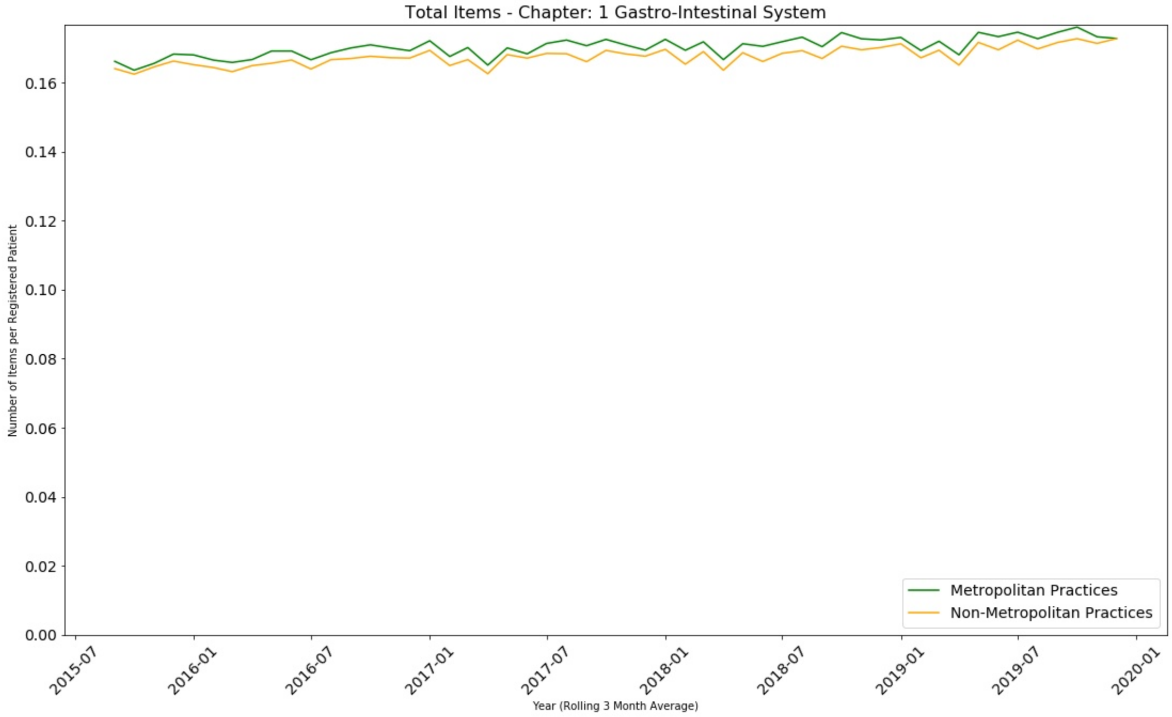


FIGURE F.1: Prescribing by Archetype - BNF chapter 1 (Gastro-Intestinal System)

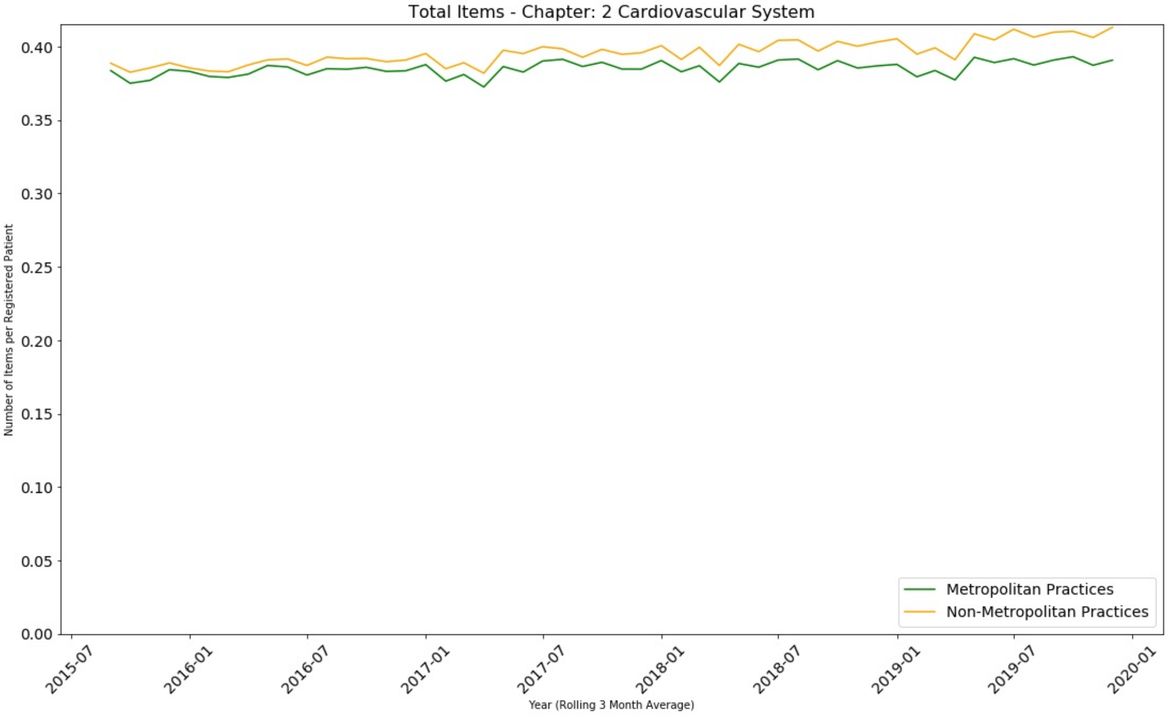


FIGURE F.2: Prescribing by Archetype - BNF chapter 2 (Cardiovascular System)

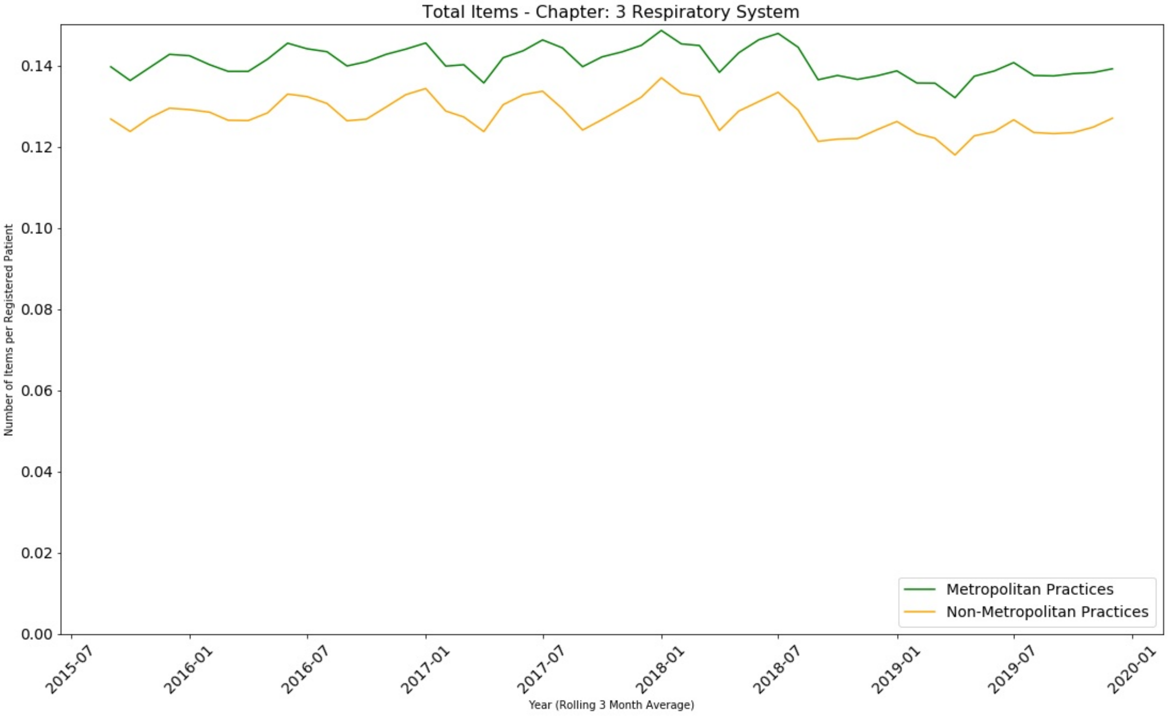


FIGURE F.3: Prescribing by Archetype - BNF chapter 3 (Respiratory System)

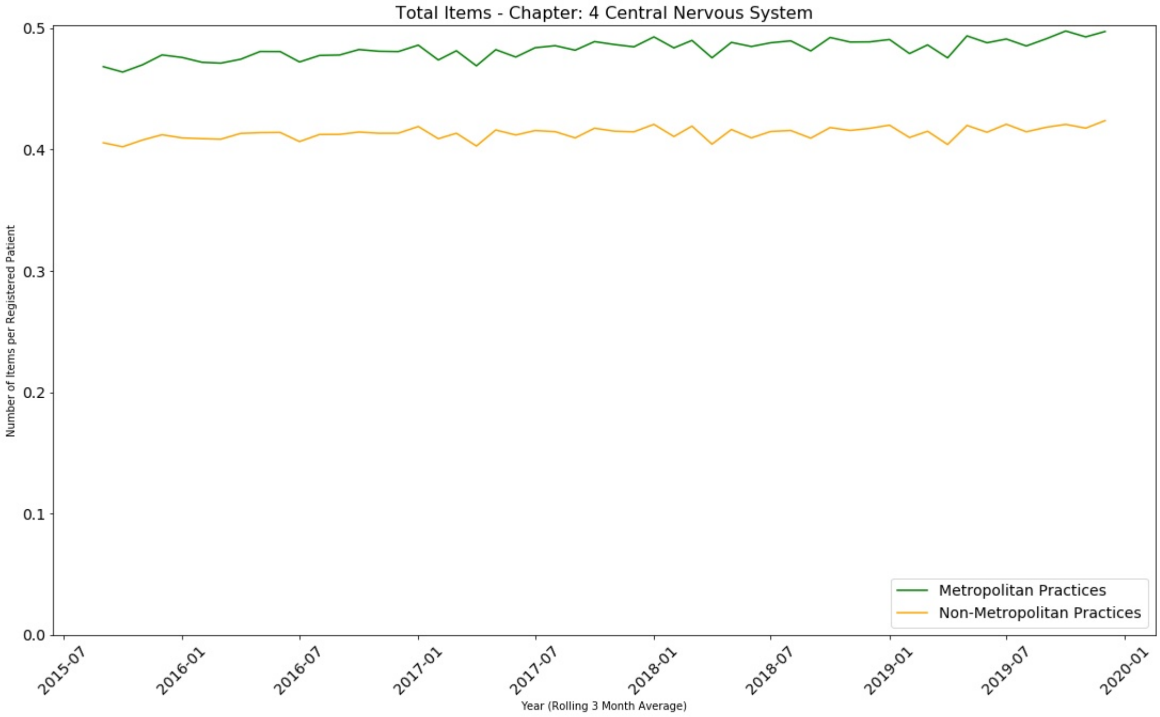


FIGURE F.4: Prescribing by Archetype - BNF chapter 4 (Central Nervous System)

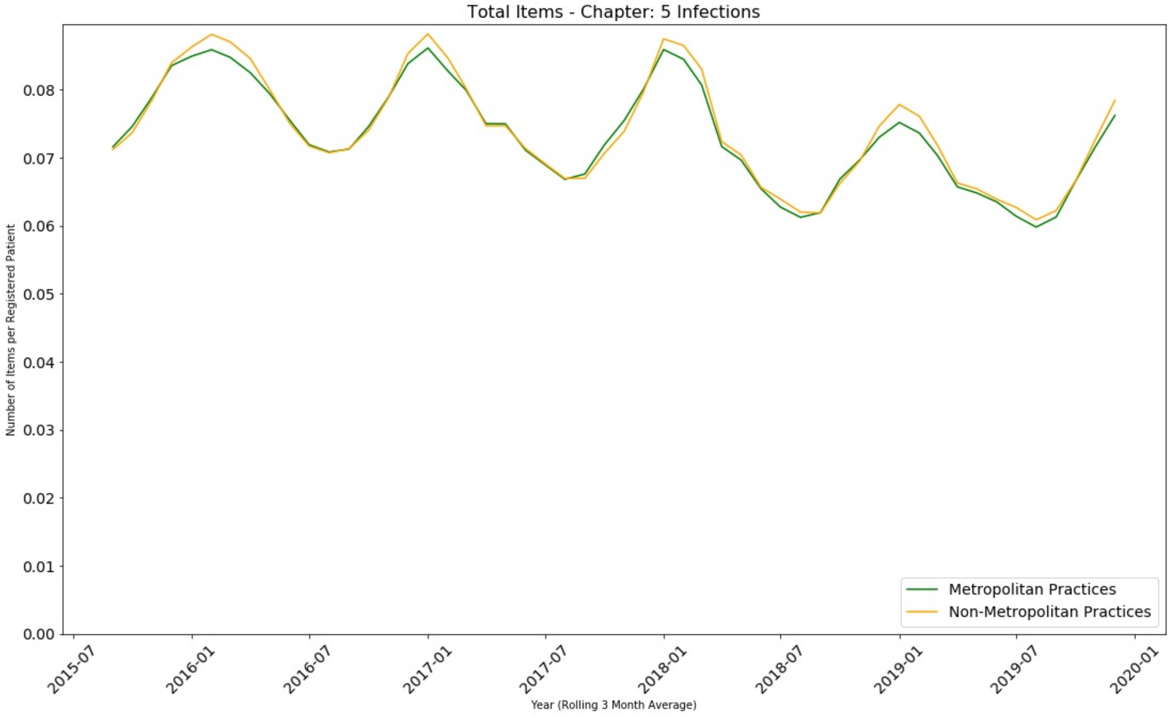


FIGURE F.5: Prescribing by Archetype - BNF chapter 5 (Infections)

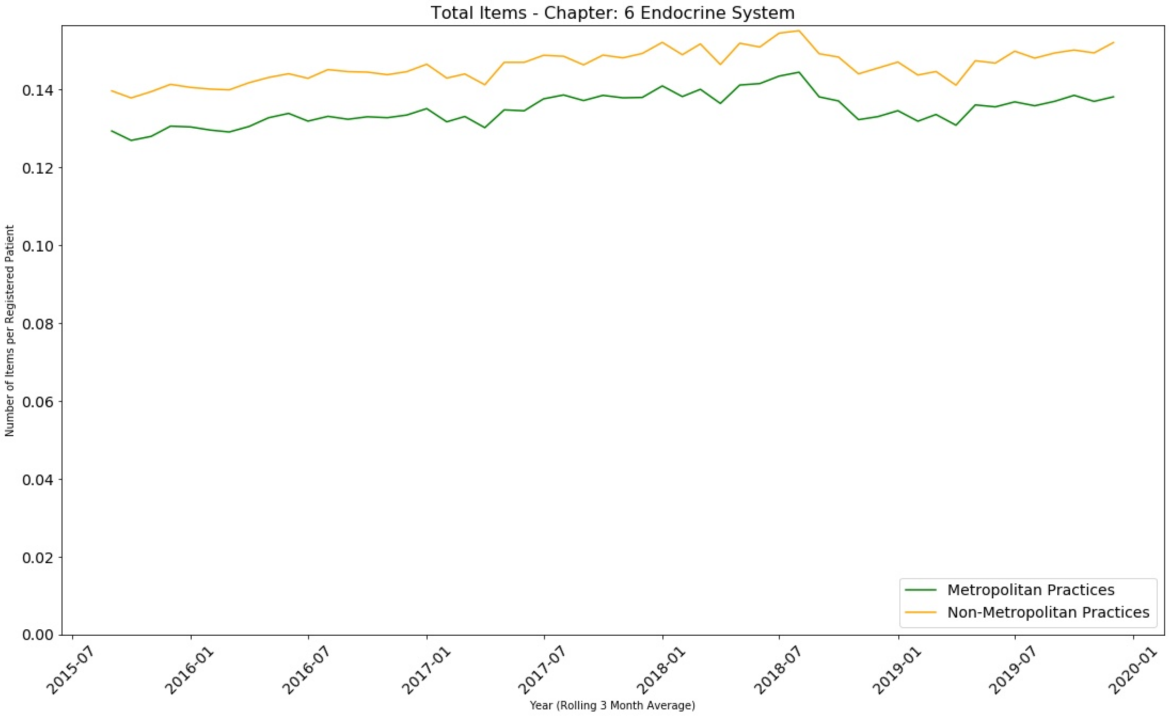


FIGURE F.6: Prescribing by Archetype - BNF chapter 6 (Endocrine System)

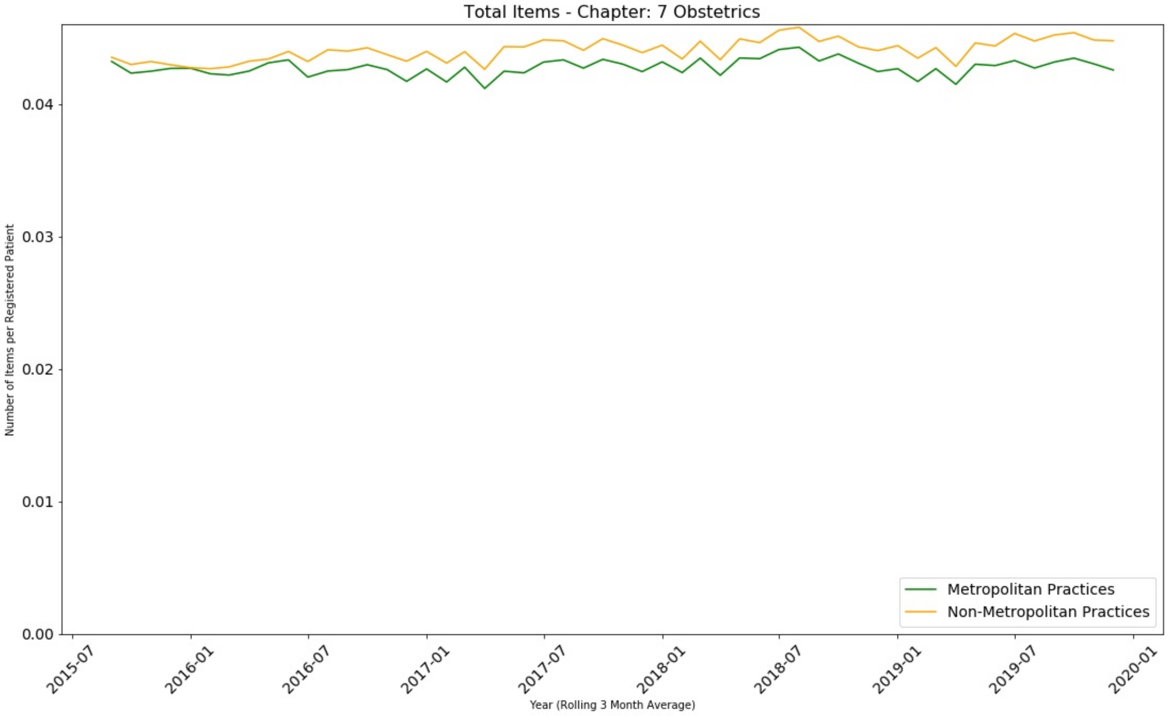


FIGURE F.7: Prescribing by Archetype - BNF chapter 7 (Obstetrics)

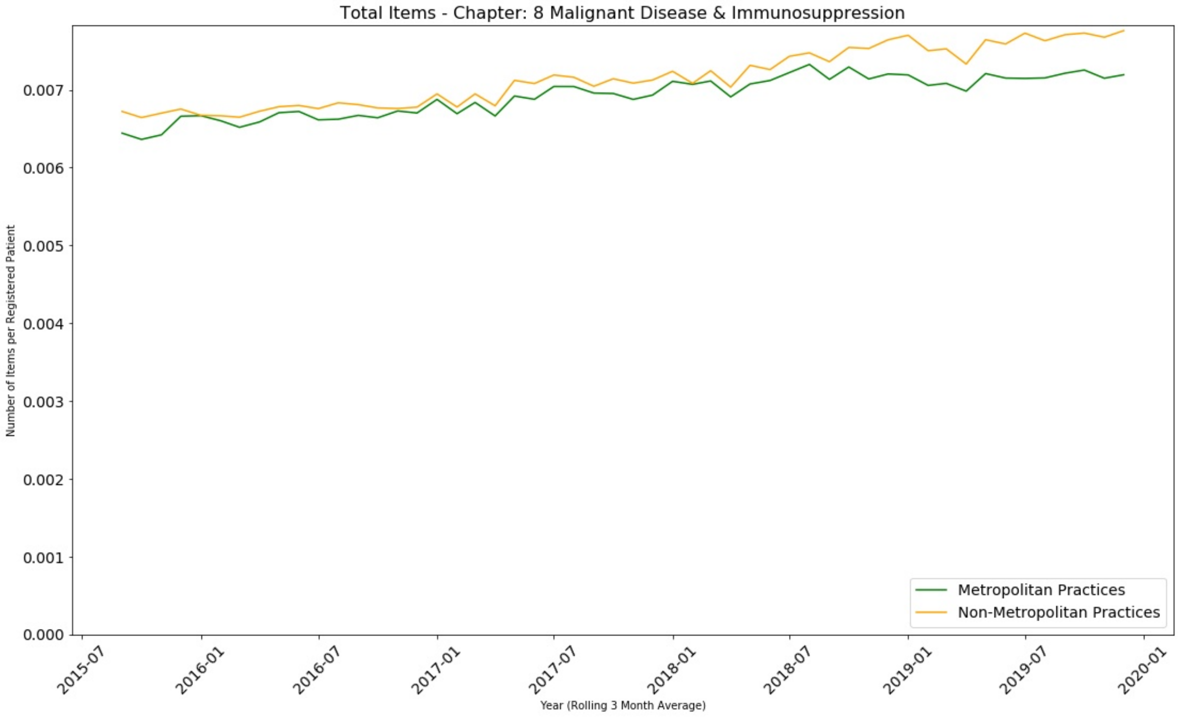


FIGURE F.8: Prescribing by Archetype - BNF chapter 8 (Malignant Disease & Immunosuppression)

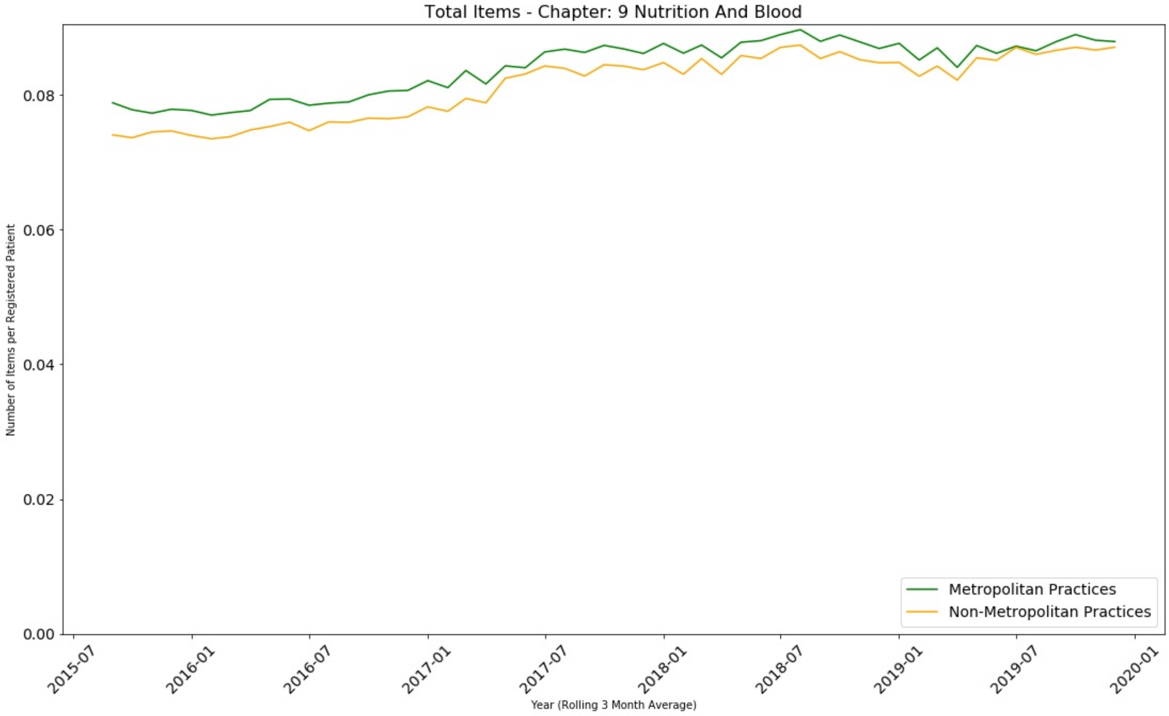


FIGURE F.9: Prescribing by Archetype - BNF chapter 9 (Nutrition And Blood)

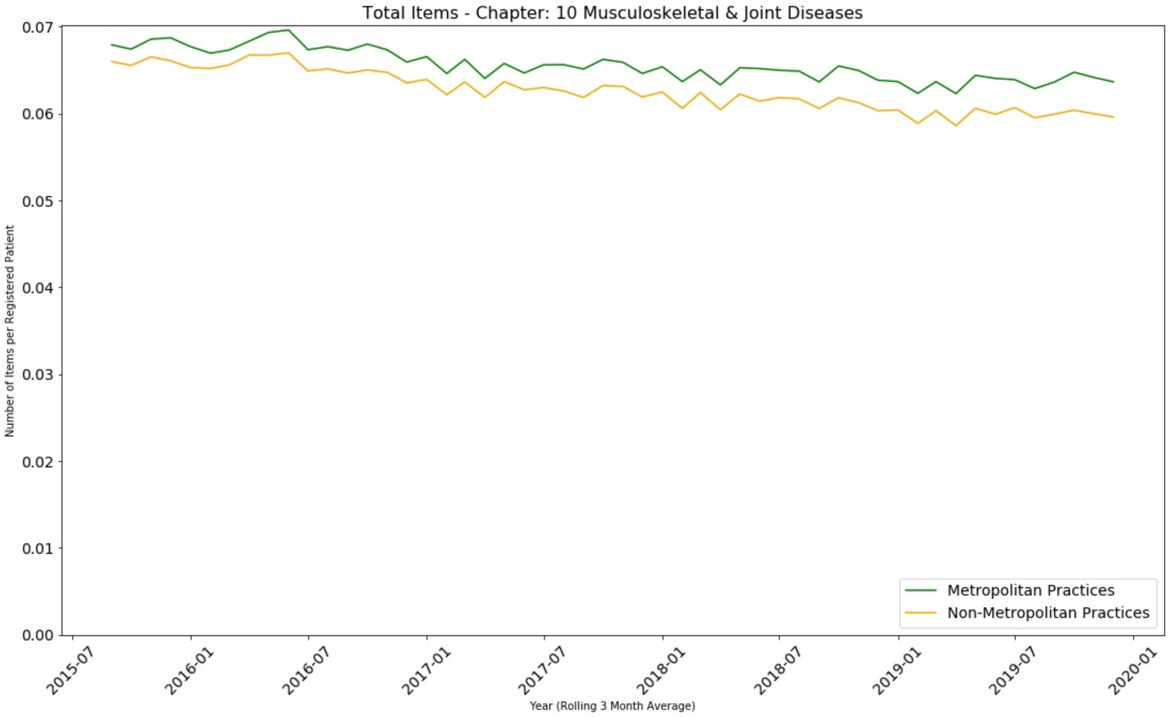


FIGURE F.10: Prescribing by Archetype - BNF chapter 10 (Musculoskeletal & Joint Diseases)

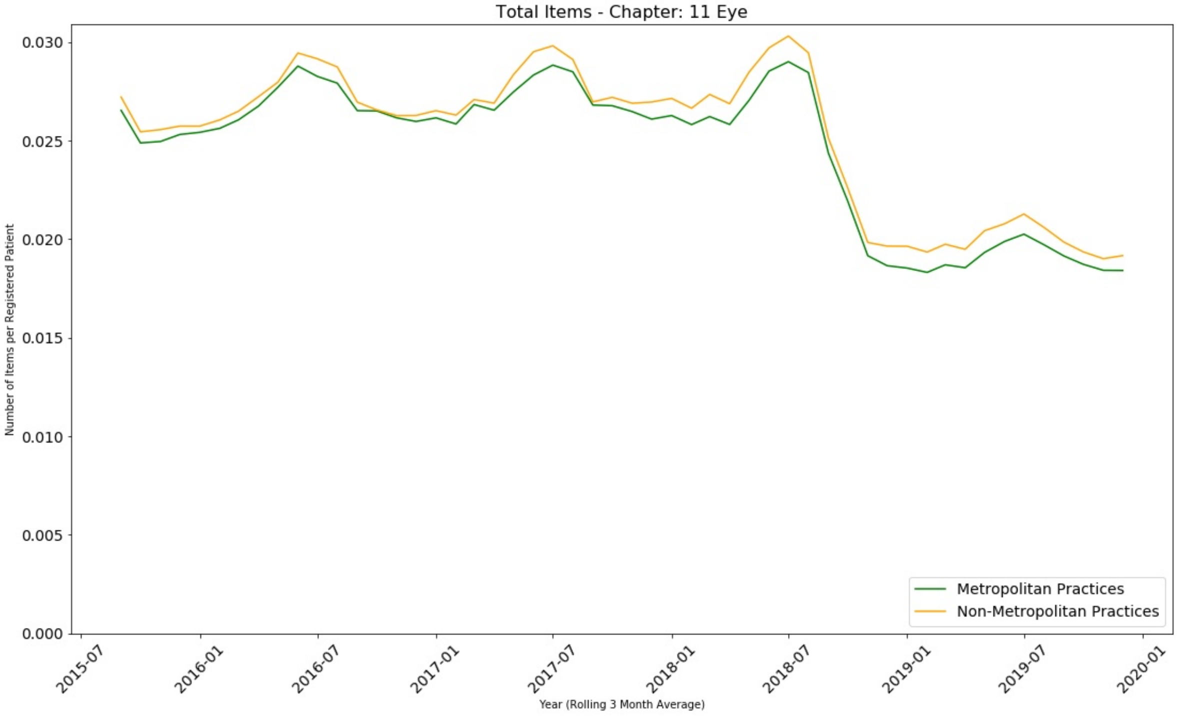


FIGURE F.11: Prescribing by Archetype - BNF chapter 11 (Eye)

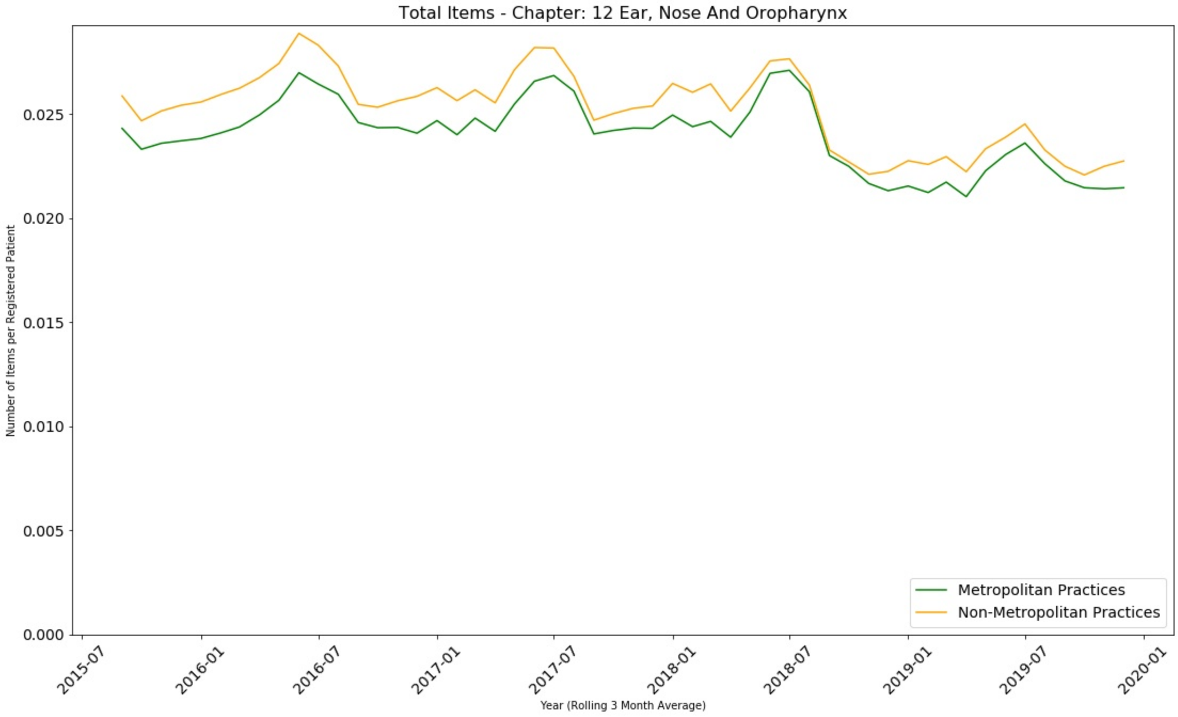


FIGURE F.12: Prescribing by Archetype - BNF chapter 12 (Ear, Nose And Oropharynx)

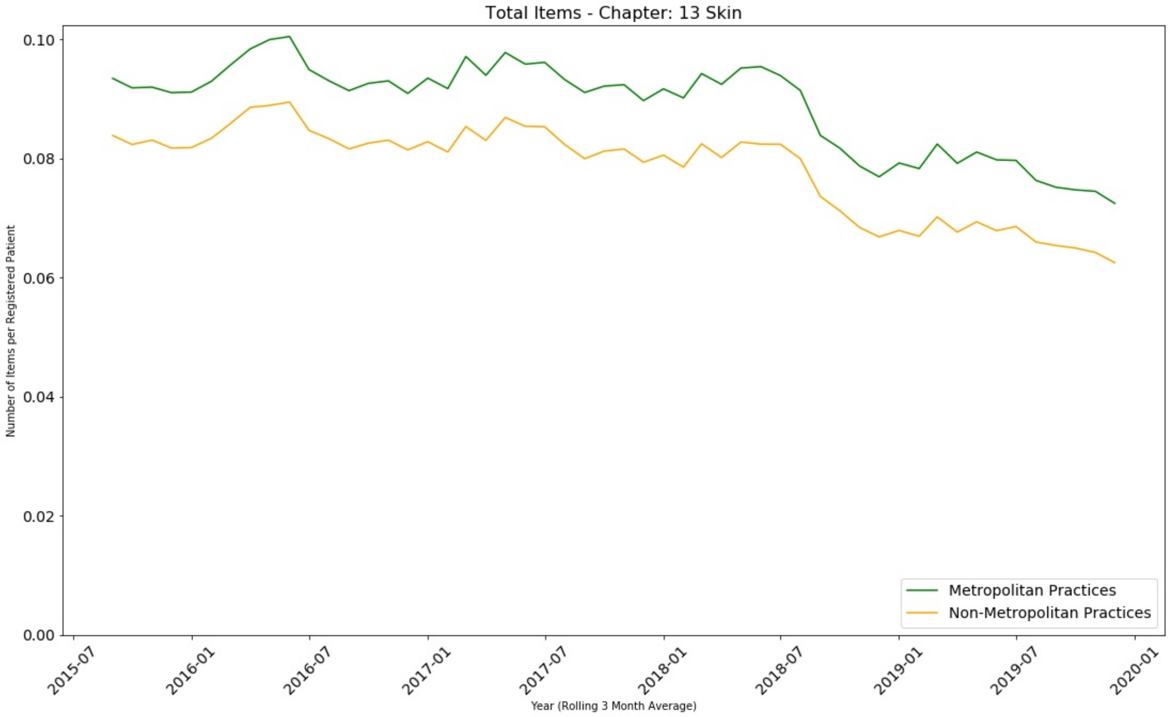


FIGURE F.13: Prescribing by Archetype - BNF chapter 13 (Skin)

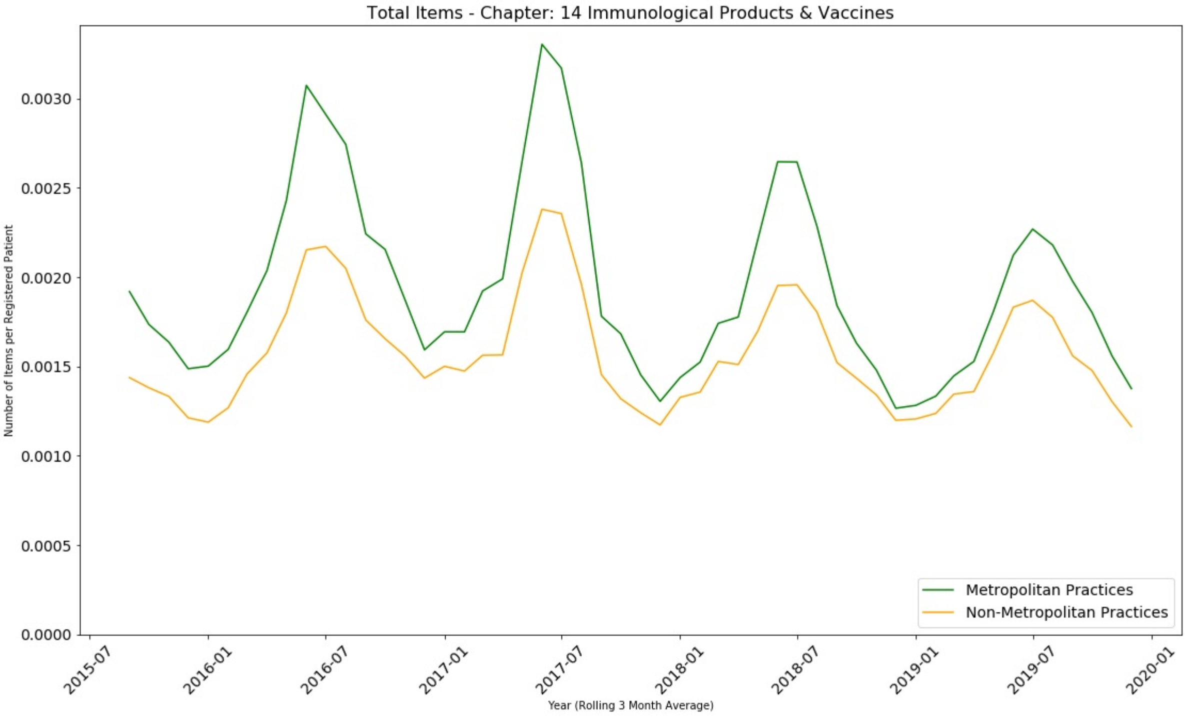


FIGURE F.14: Prescribing by Archetype - BNF chapter 14 (Immunological Products & Vaccines)

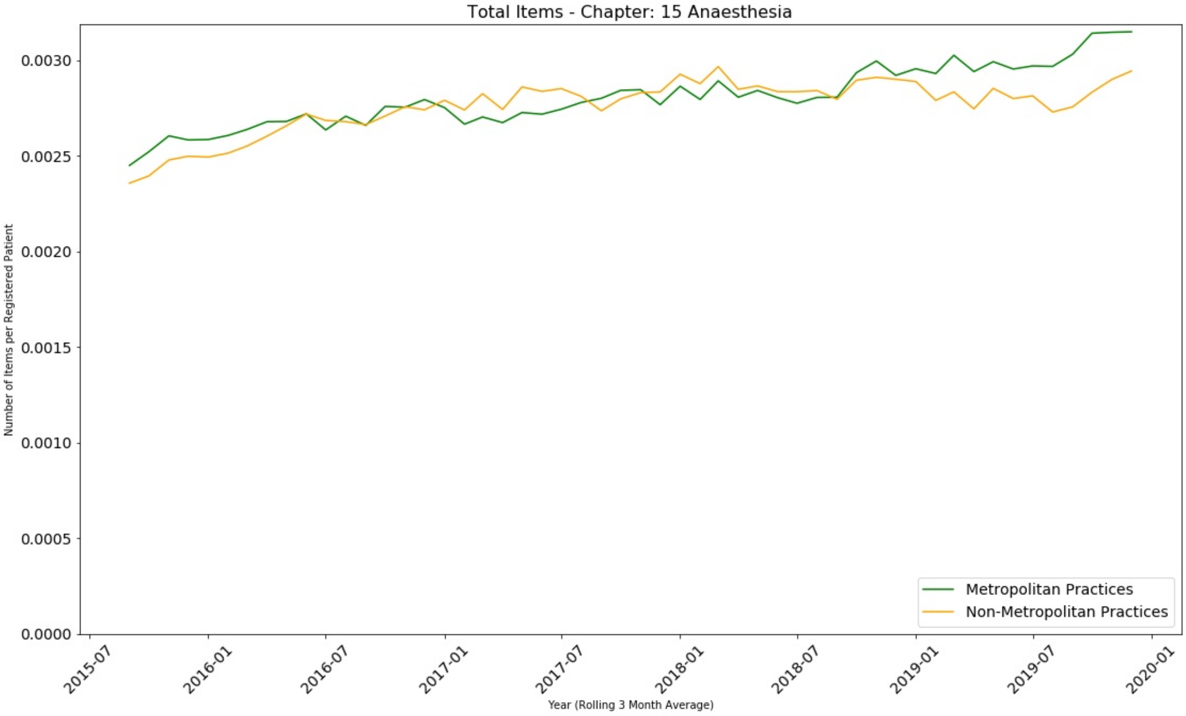


FIGURE F.15: Prescribing by Archetype - BNF chapter 15 (Anaesthesia)

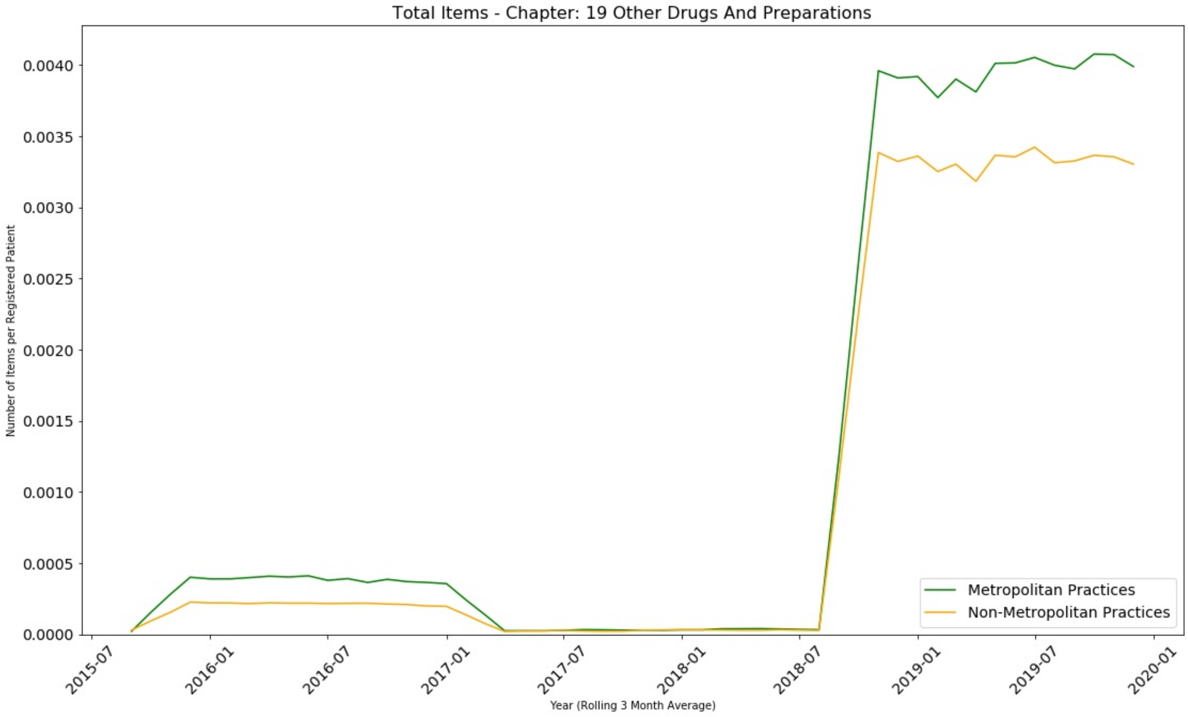


FIGURE F.16: Prescribing by Archetype - BNF chapter 19 (Other Drugs And Preparations)

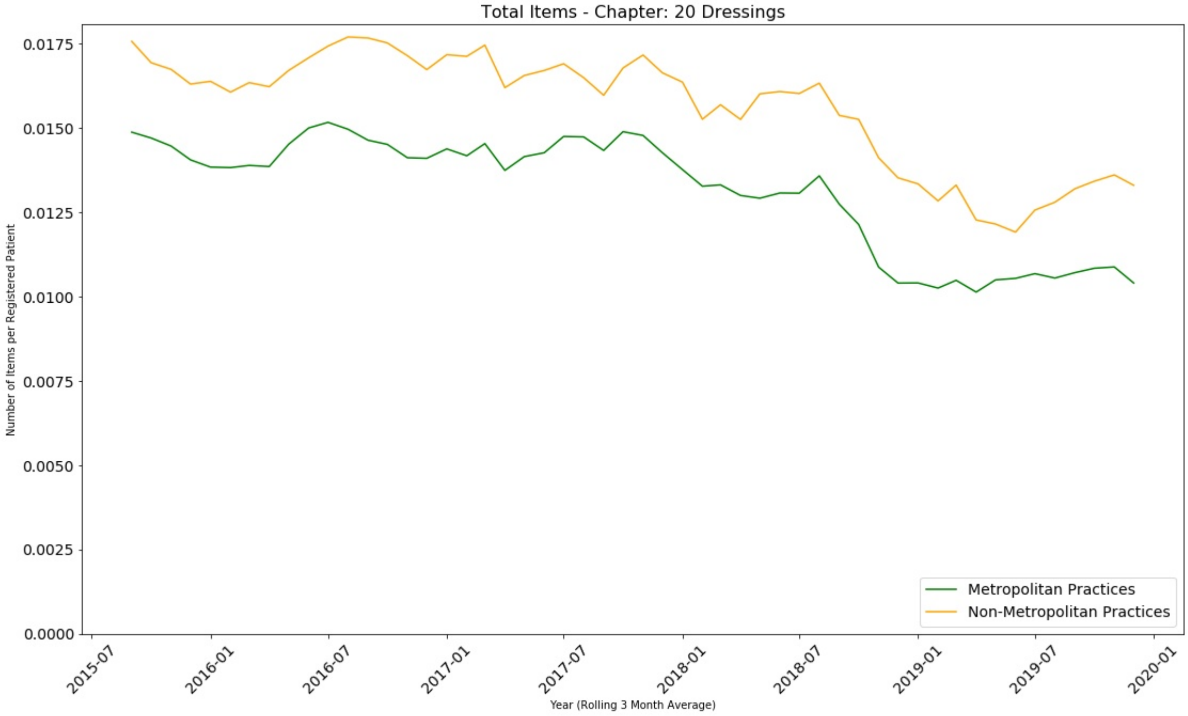


FIGURE F.17: Prescribing by Archetype - BNF chapter 20 (Dressings)

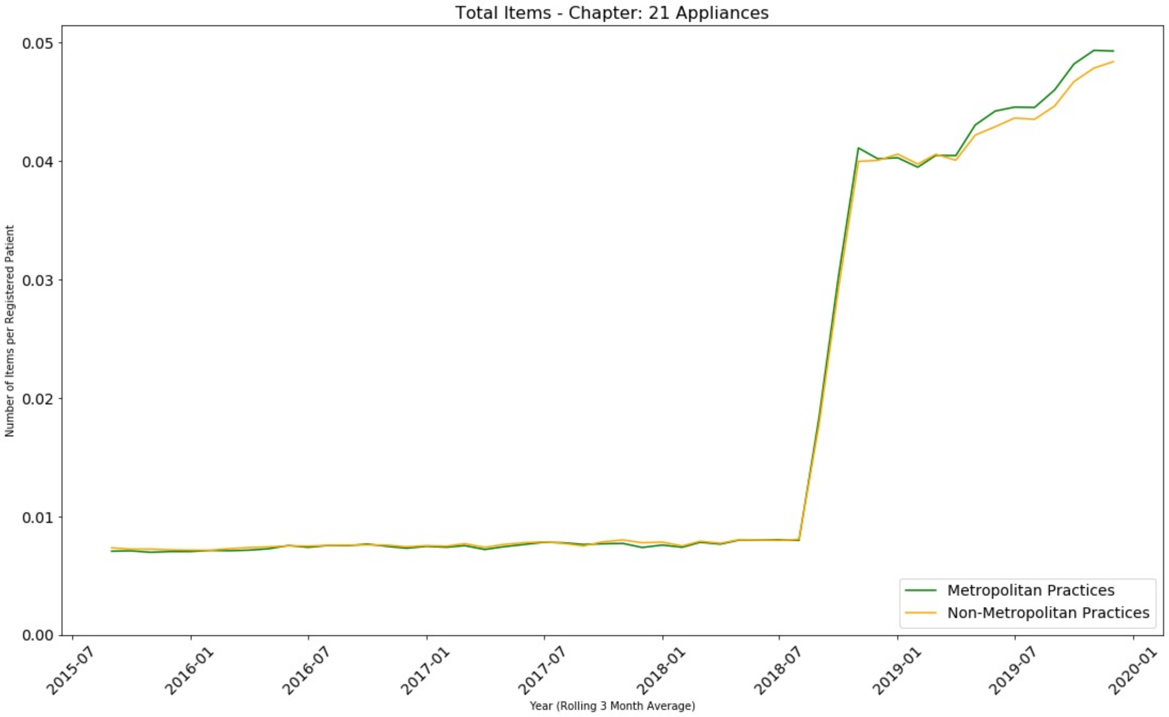


FIGURE F.18: Prescribing by Archetype - BNF chapter 21 (Appliances)

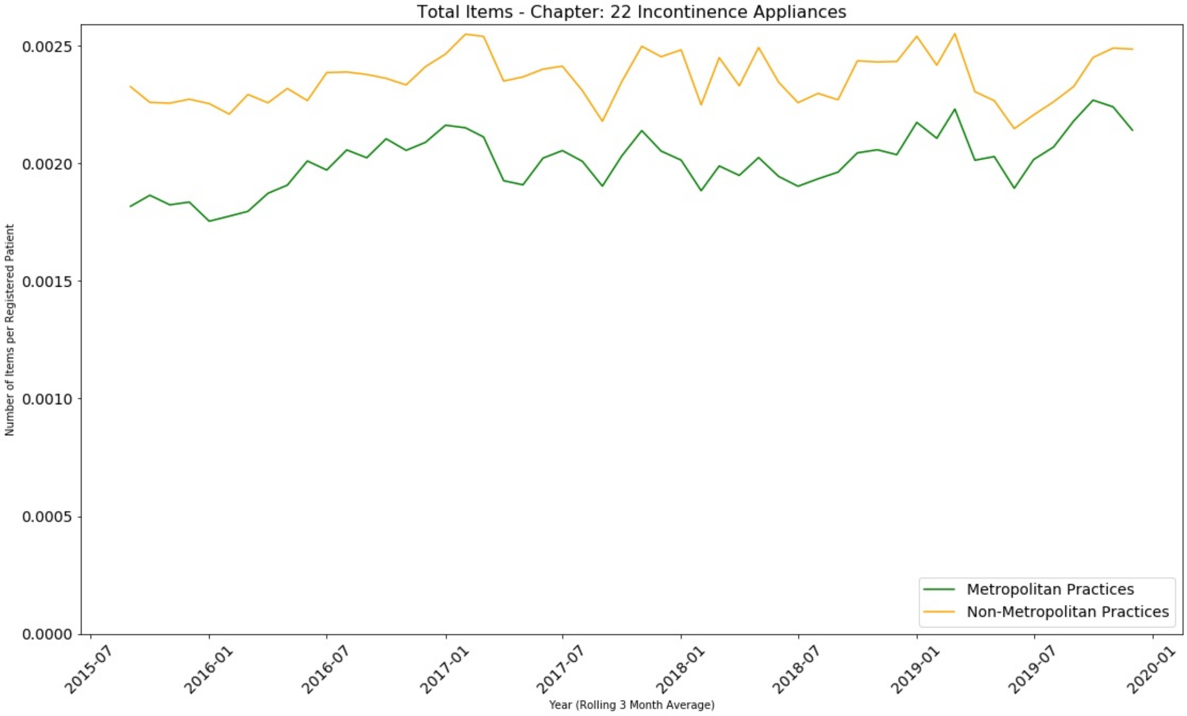


FIGURE F.19: Prescribing by Archetype - BNF chapter 22 (Incontinence Appliances)

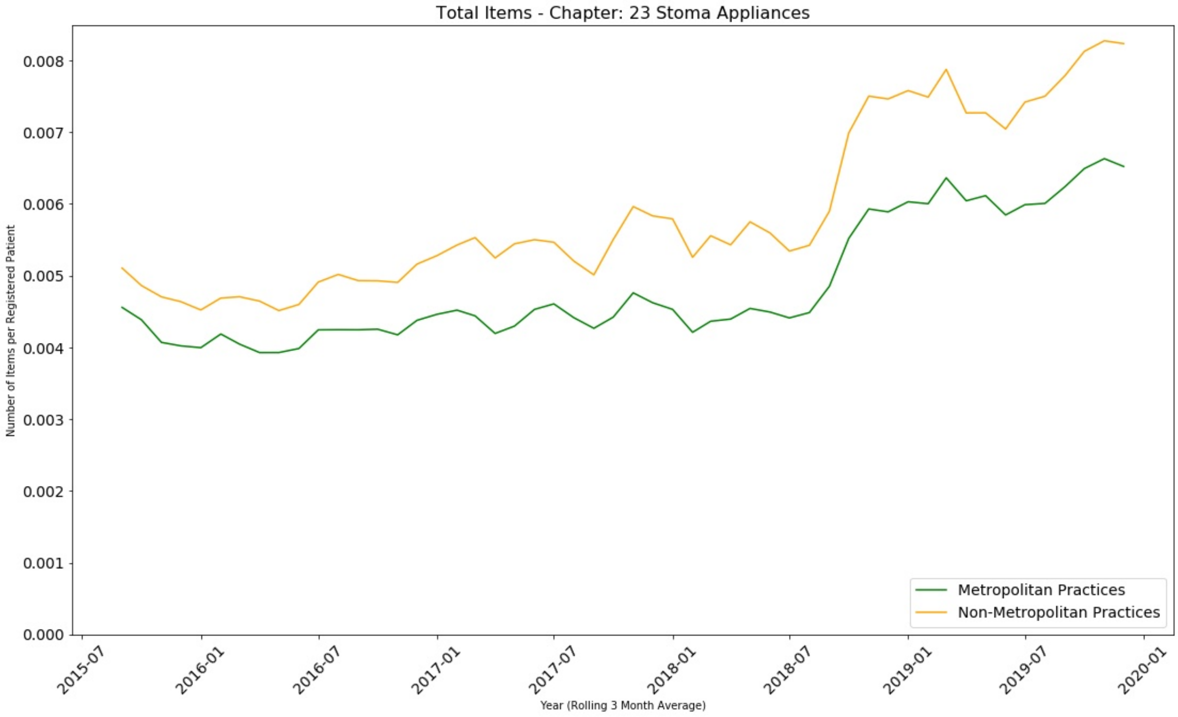


FIGURE F.20: Prescribing by Archetype - BNF chapter 23 (Stoma Appliances)

# Appendix G

# Clustering of GP practices by practice size

## G.1 Single-Handed Practices

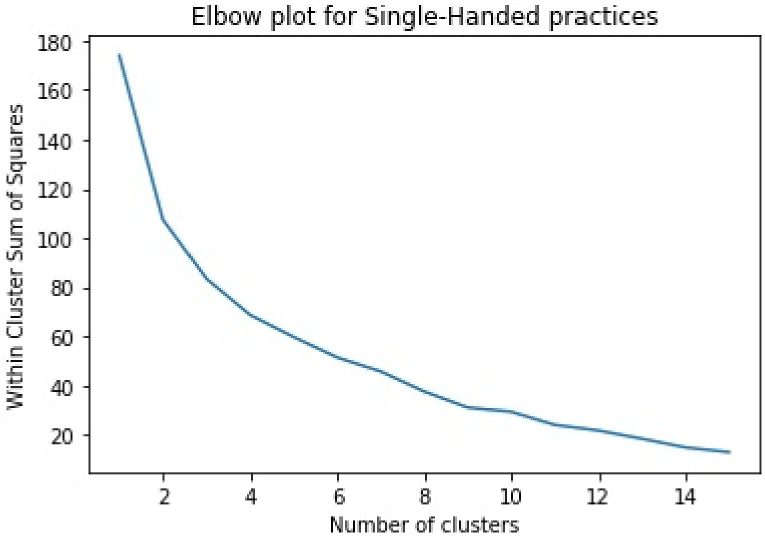


FIGURE G.1: Elbow plot for Single-Handed practices

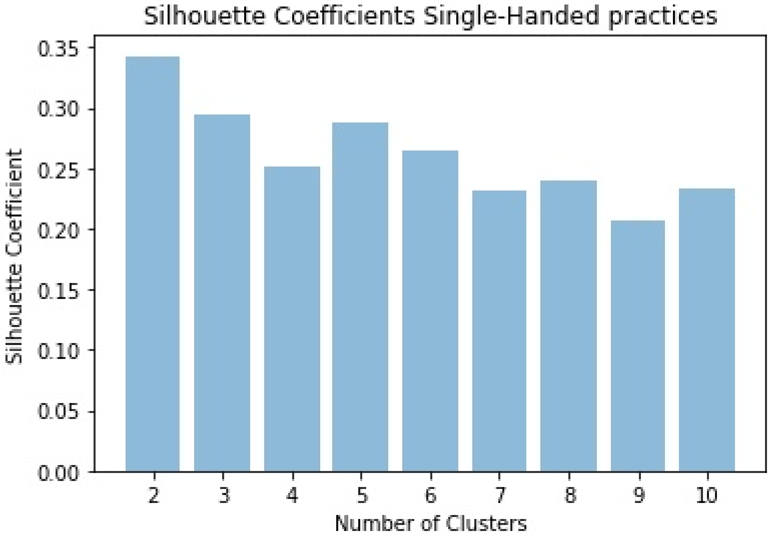


FIGURE G.2: Silhouette Coefficient graph for Single-Handed practices

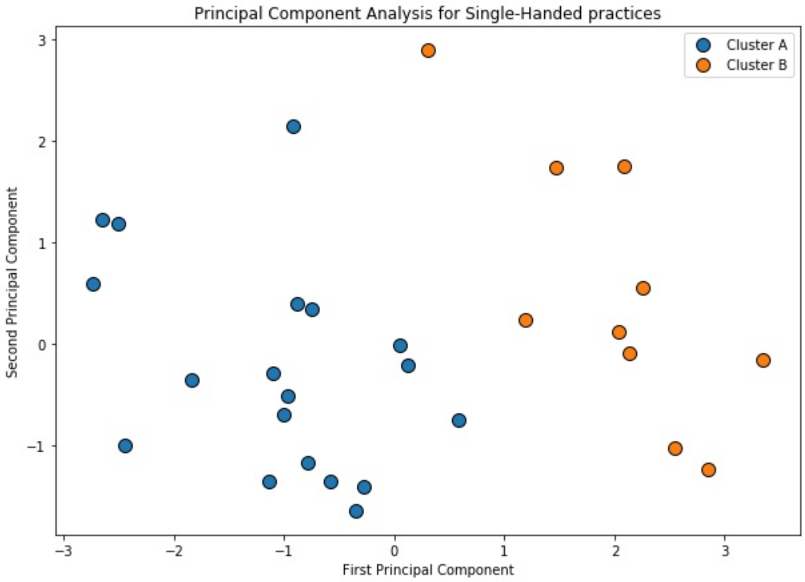


FIGURE G.3: Principal Components plot for Single-Handed practices

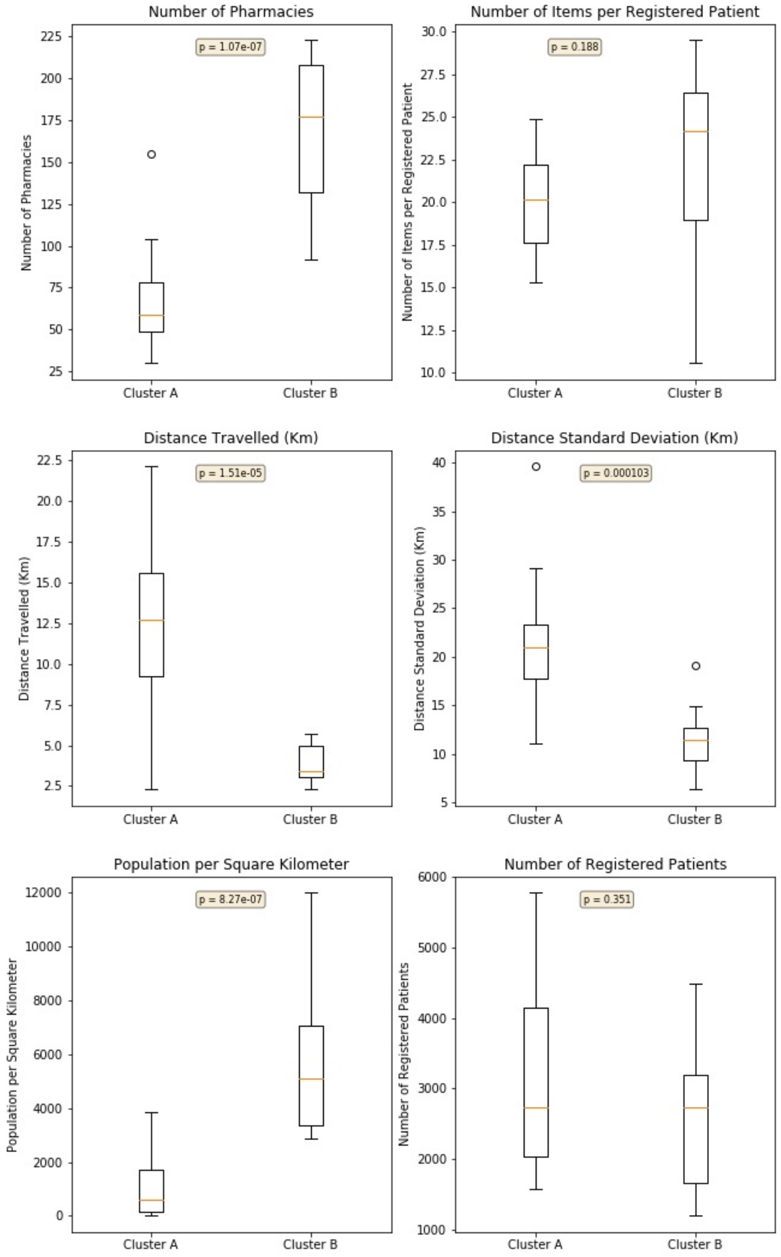


FIGURE G.4: Box plots of feature statistics for Single-Handed practices

## G.2 Small Practices

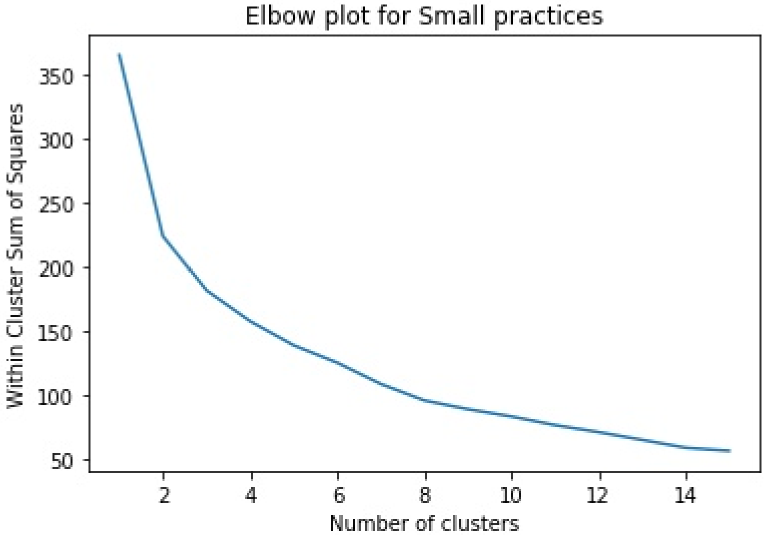


FIGURE G.5: Elbow plot for Small practices

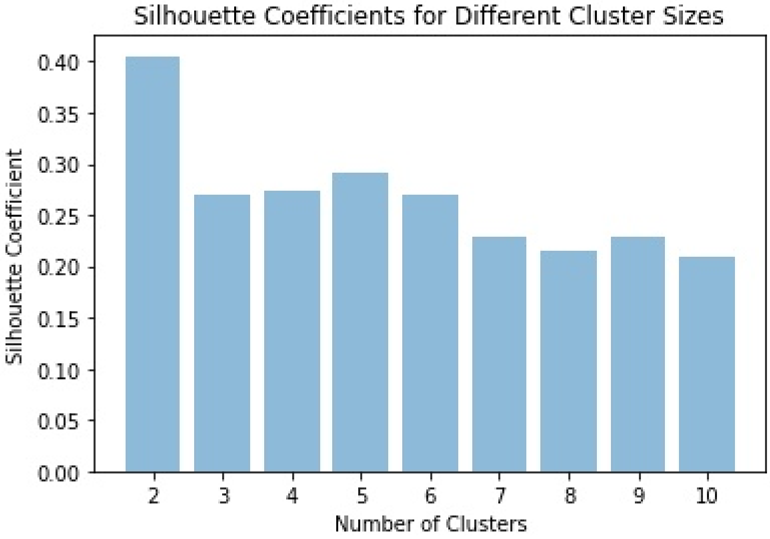


FIGURE G.6: Silhouette Coefficient graph for Small practices

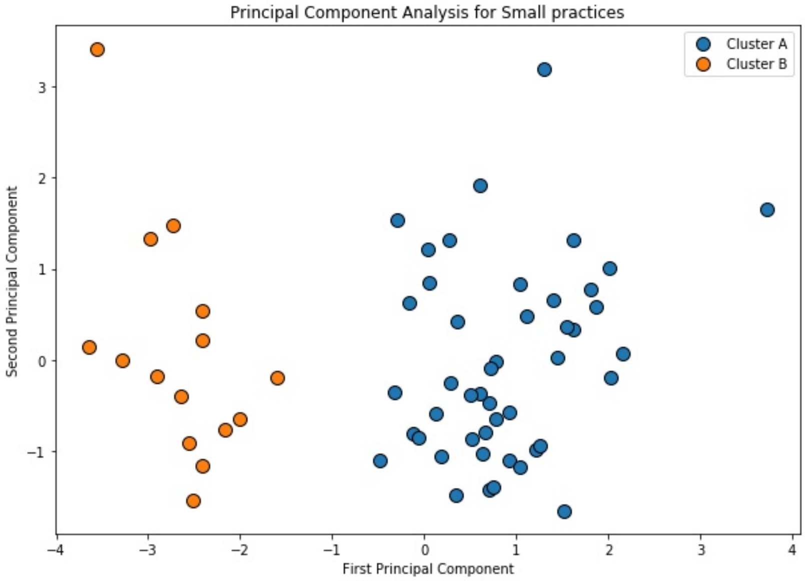


FIGURE G.7: Principal Components plot for Small practices

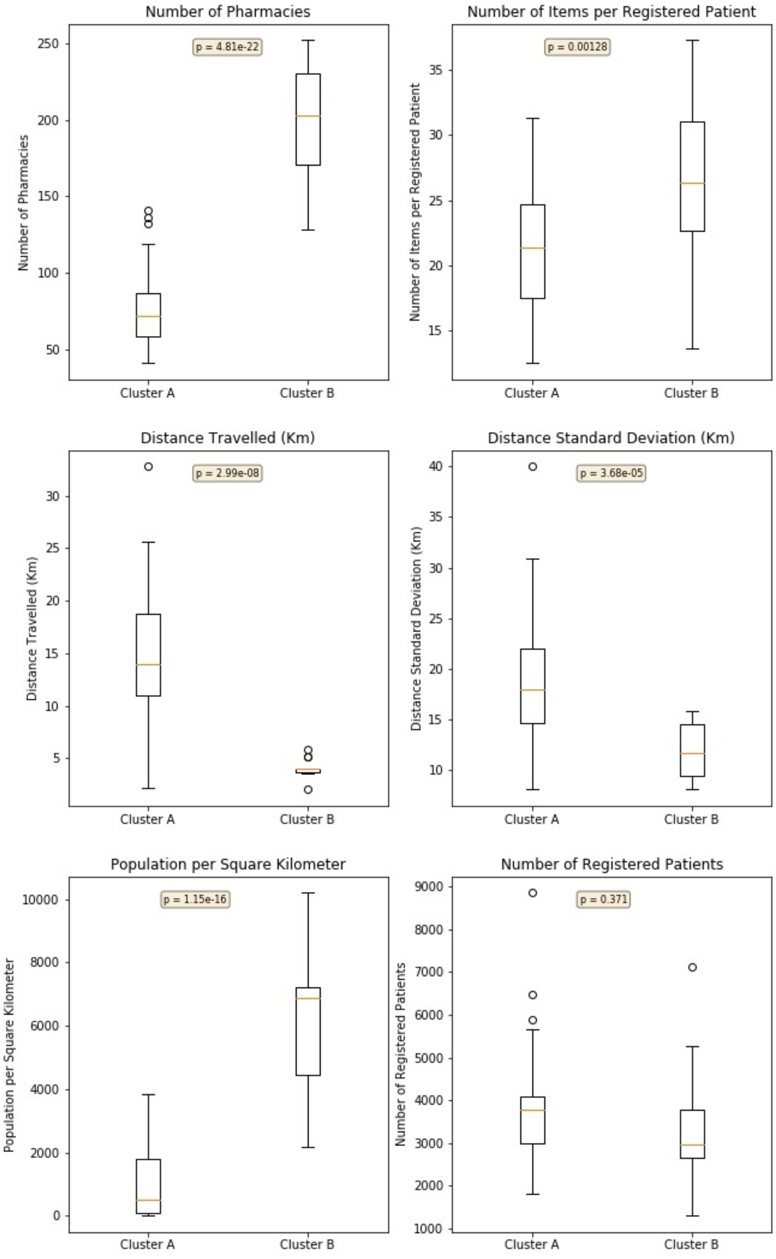


FIGURE G.8: Box plots of feature statistics for Small practices

## G.3 Medium Practices

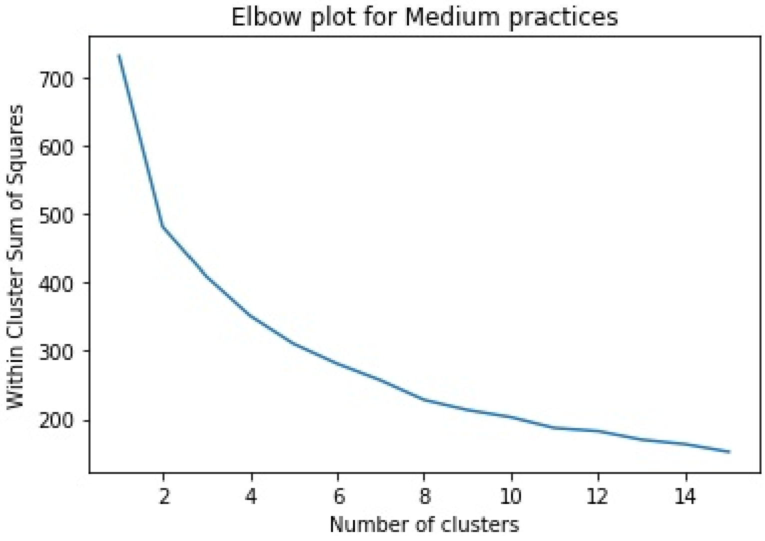


FIGURE G.9: Elbow plot for Medium practices

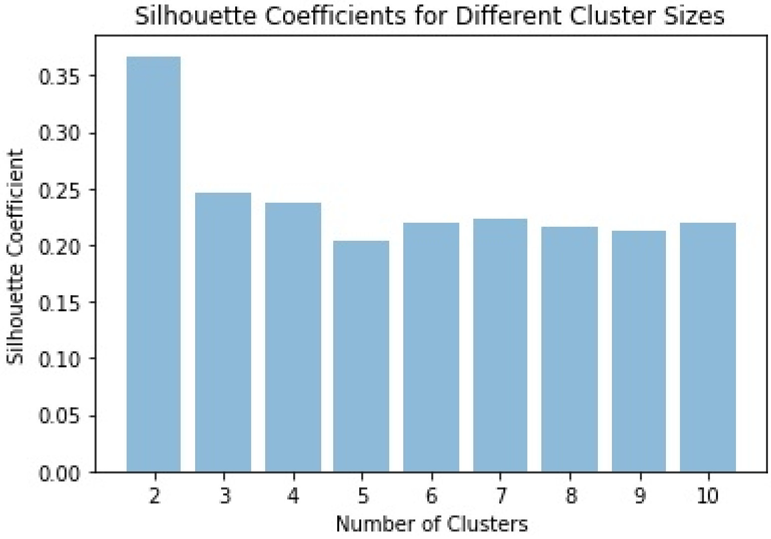


FIGURE G.10: Silhouette Coefficient graph for Medium practices

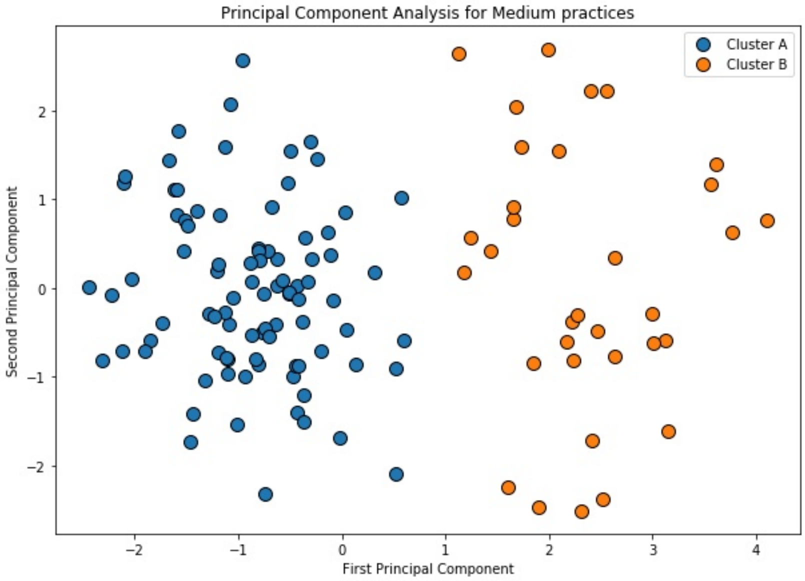


FIGURE G.11: Principal Components plot for Medium practices

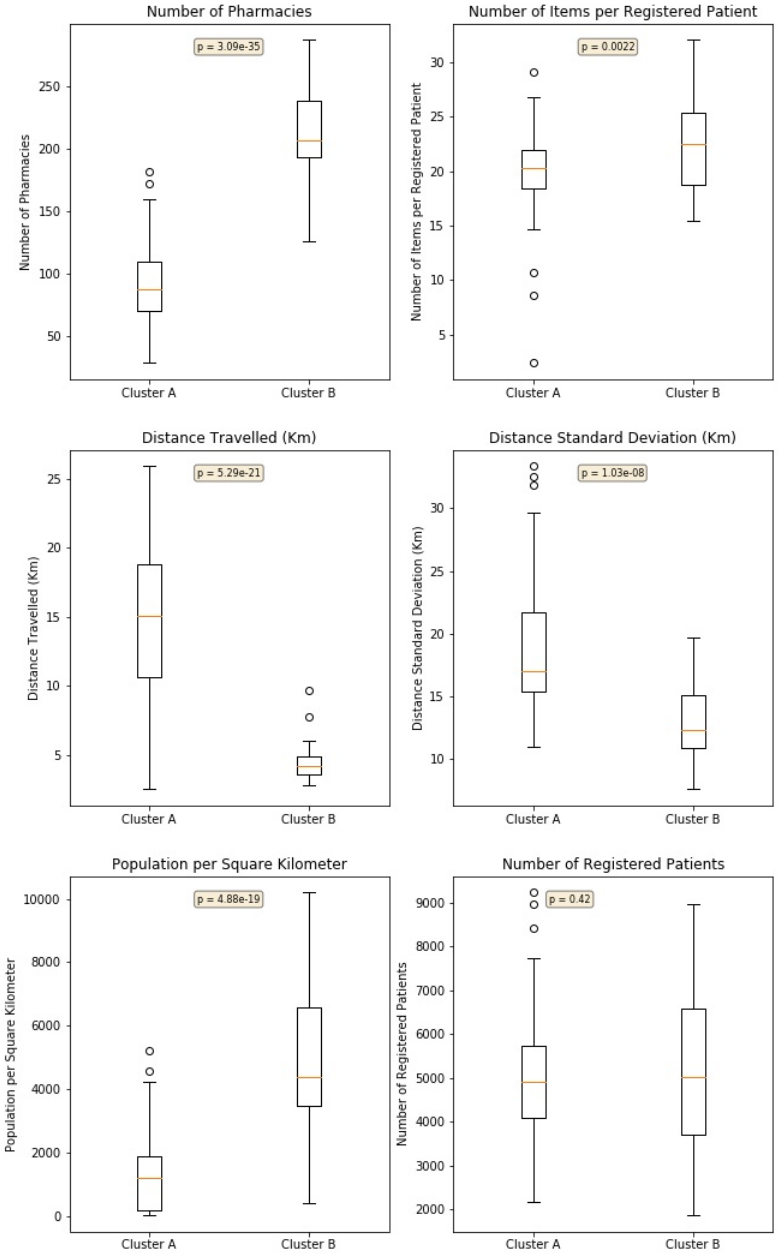


FIGURE G.12: Box plots of feature statistics for Medium practices

## G.4 Large Practices

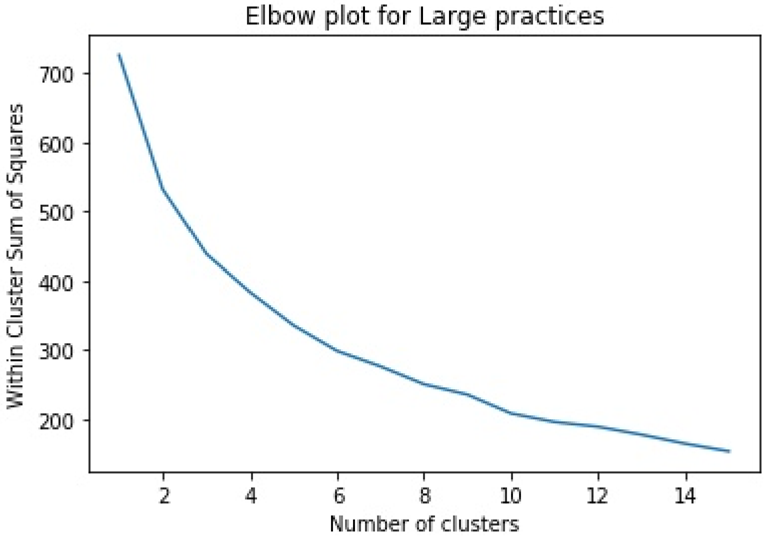


FIGURE G.13: Elbow plot for Large practices

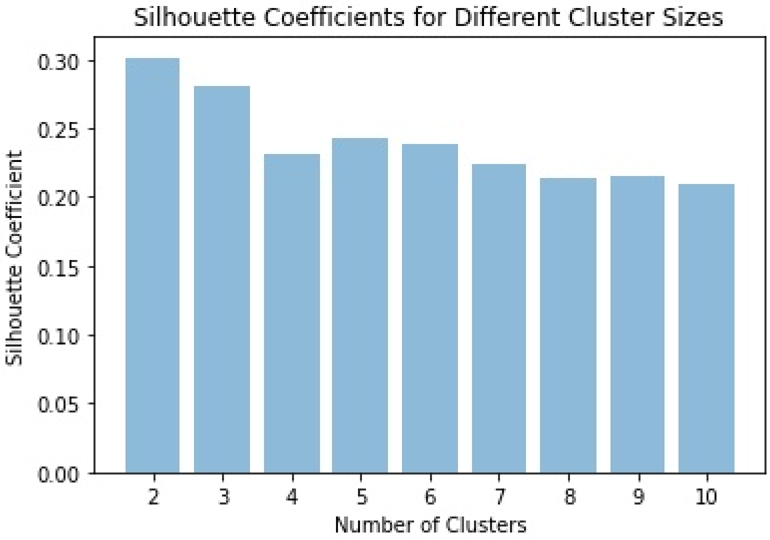


FIGURE G.14: Silhouette Coefficient graph for Large practices

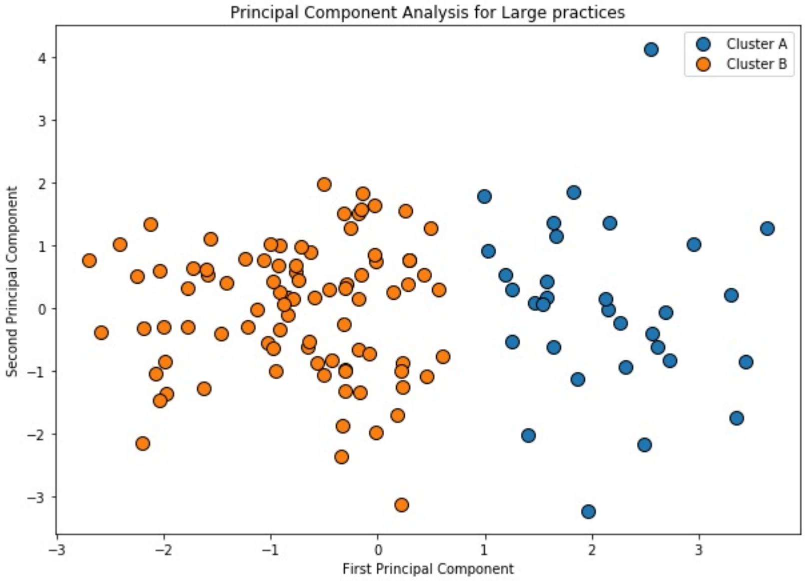


FIGURE G.15: Principal Components plot for Large practices

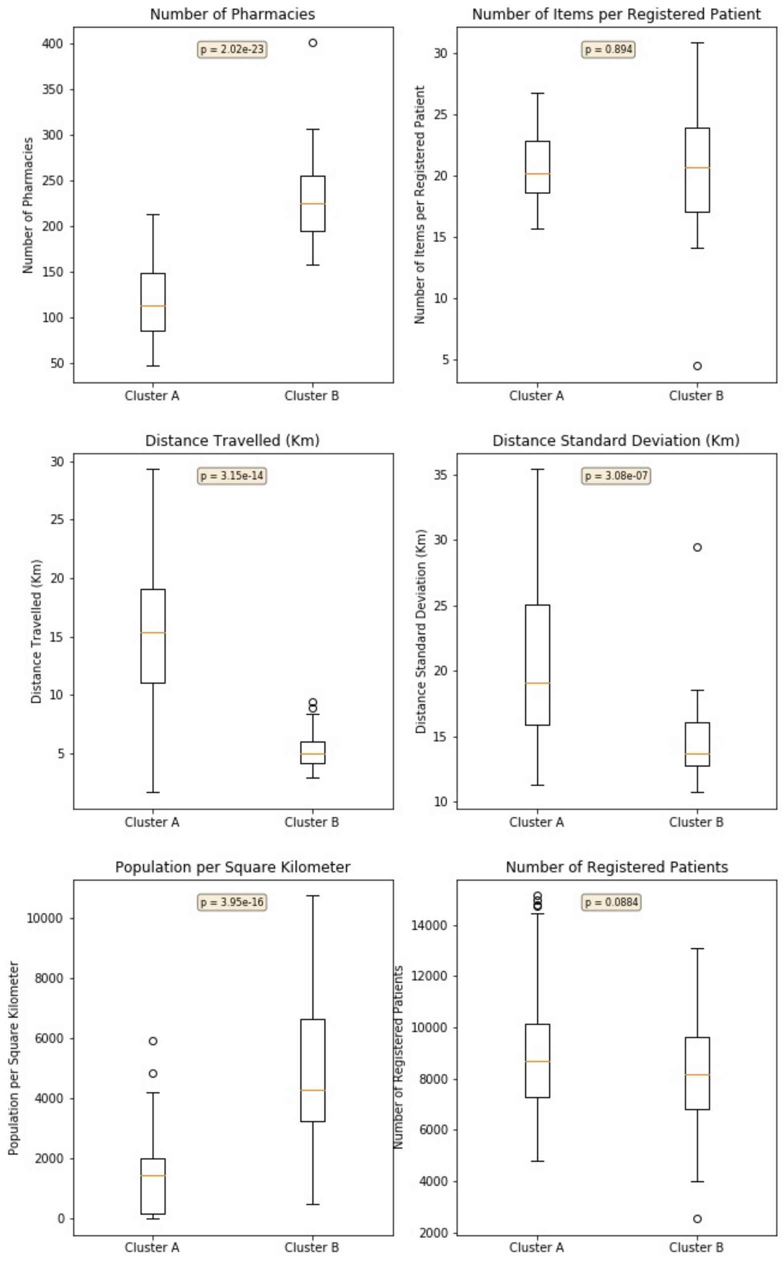


FIGURE G.16: Box plots of feature statistics for Large practices

# Appendix H

# Prescribing trends of NI GP practices by Archetype during COVID-19 and lockdown

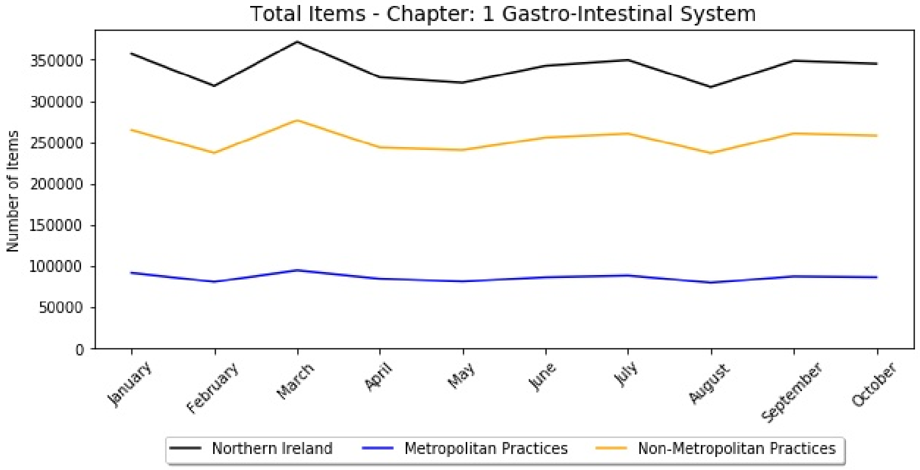


FIGURE H.1: Prescribing by Archetype during the COVID-19 pandemic and first national lockdown - BNF chapter 1 (Gastro-Intestinal System)

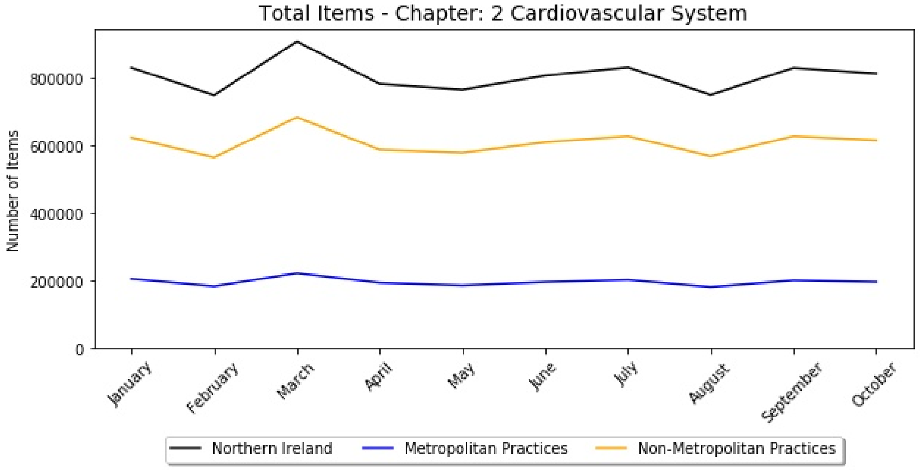


FIGURE H.2: Prescribing by Archetype during the COVID-19 pandemic and first national lockdown - BNF chapter 2 (Cardiovascular System)

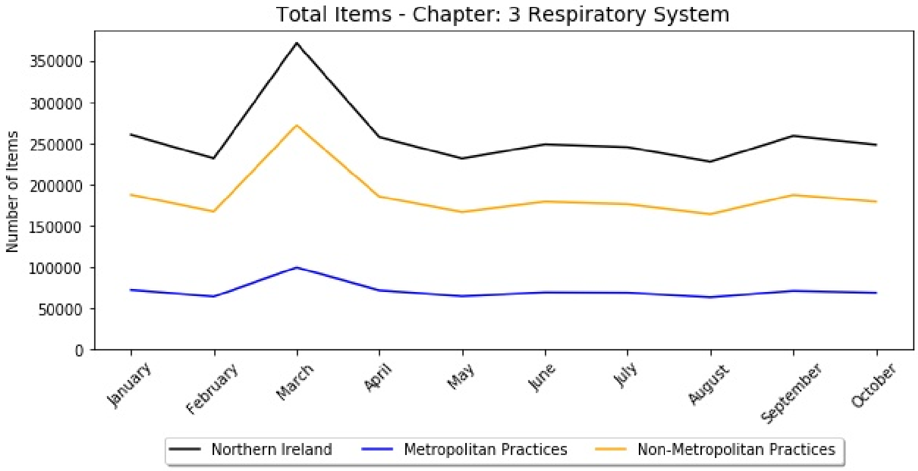


FIGURE H.3: Prescribing by Archetype during the COVID-19 pandemic and first national lockdown - BNF chapter 3 (Respiratory System)

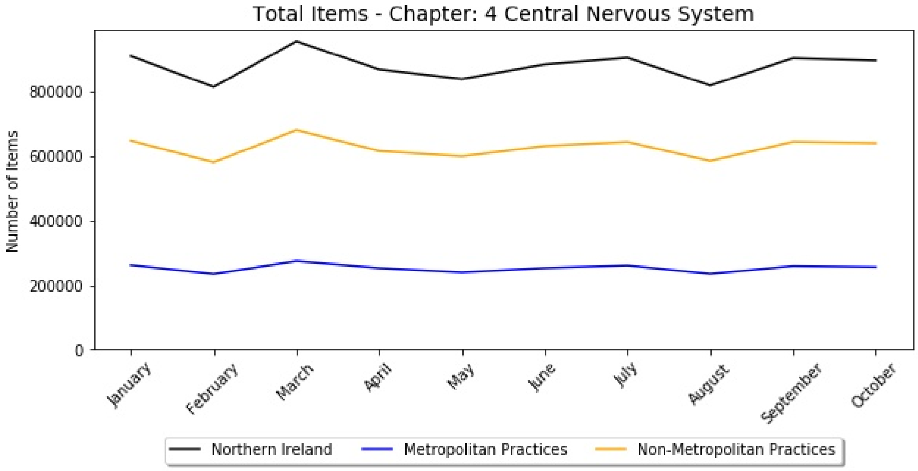


FIGURE H.4: Prescribing by Archetype during the COVID-19 pandemic and first national lockdown - BNF chapter 4 (Central Nervous System)

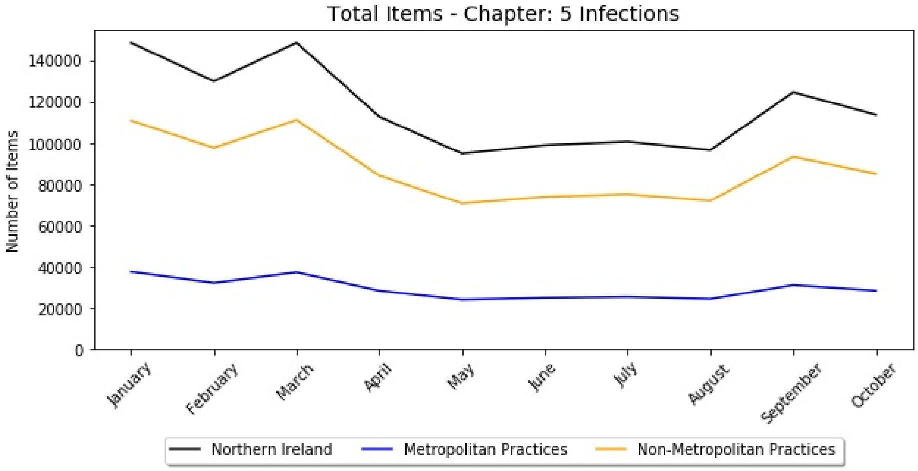


FIGURE H.5: Prescribing by Archetype during the COVID-19 pandemic and first national lockdown - BNF chapter 5 (Infections)

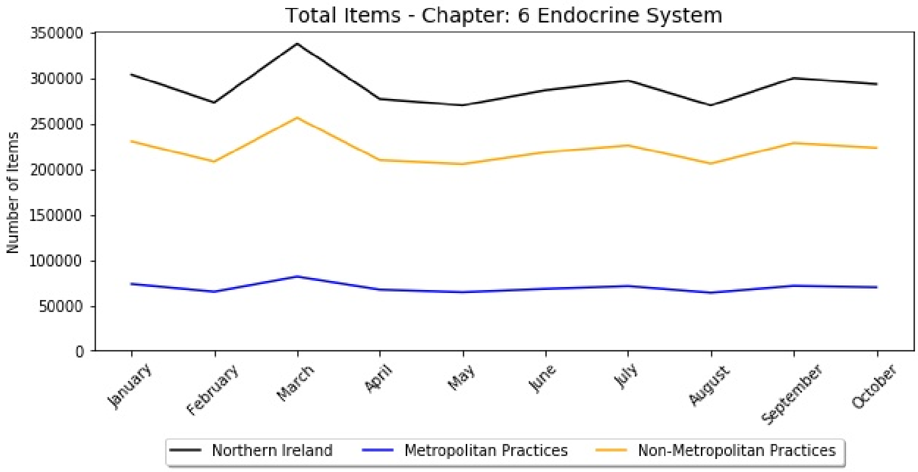


FIGURE H.6: Prescribing by Archetype during the COVID-19 pandemic and first national lockdown - BNF chapter 6 (Endocrine System)

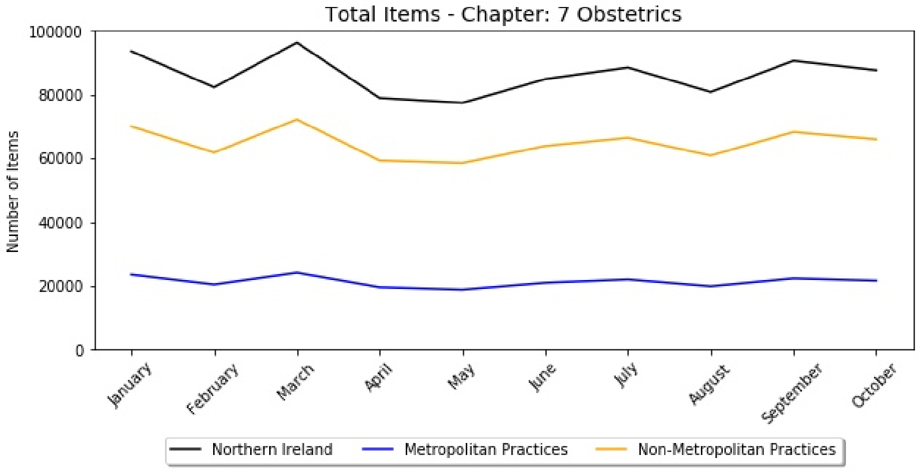


FIGURE H.7: Prescribing by Archetype during the COVID-19 pandemic and first national lockdown - BNF chapter 7 (Obstetrics)

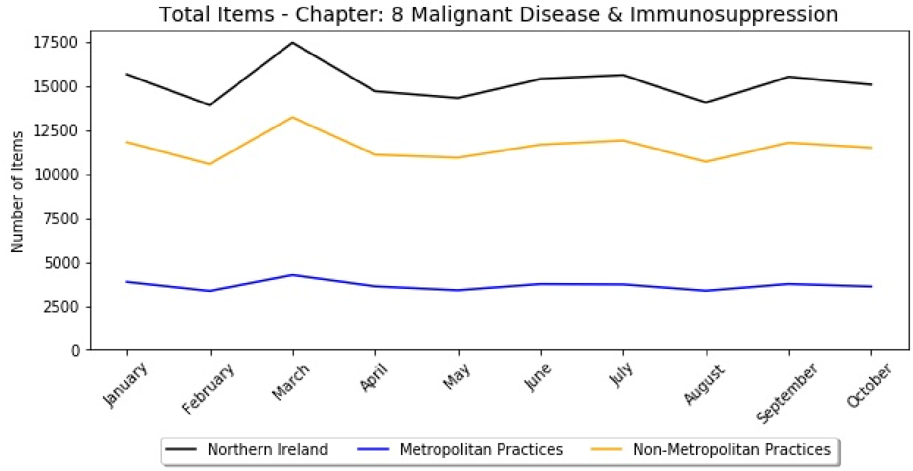


FIGURE H.8: Prescribing by Archetype during the COVID-19 pandemic and first national lockdown - BNF chapter 8 (Malignant Disease & Immunosuppression)

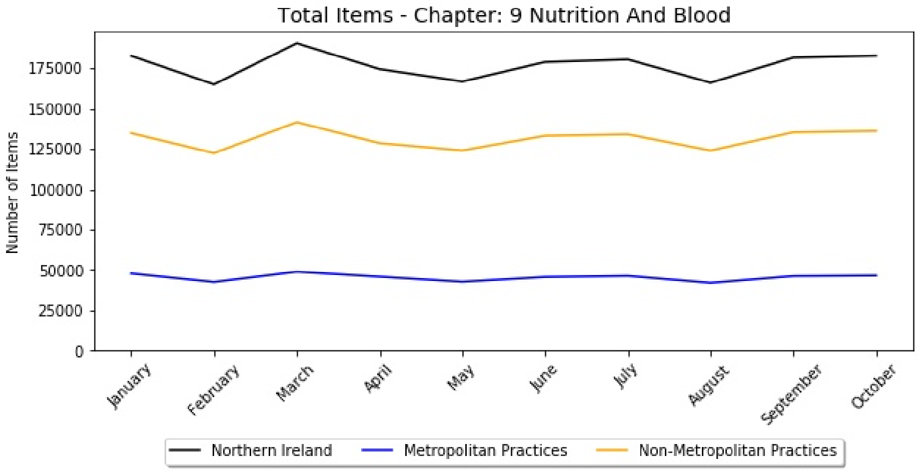


FIGURE H.9: Prescribing by Archetype during the COVID-19 pandemic and first national lockdown - BNF chapter 9 (Nutrition and Blood)

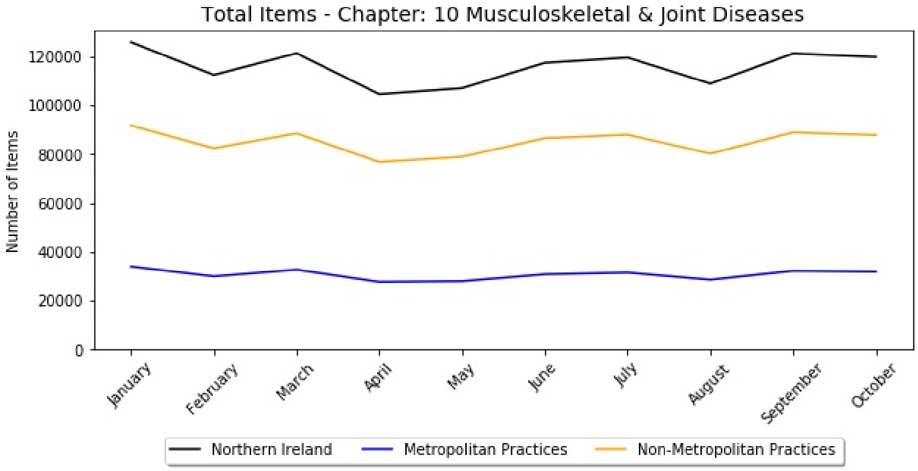


FIGURE H.10: Prescribing by Archetype during the COVID-19 pandemic and first national lockdown - BNF chapter 10 (Musculoskeletal & Joint Diseases)

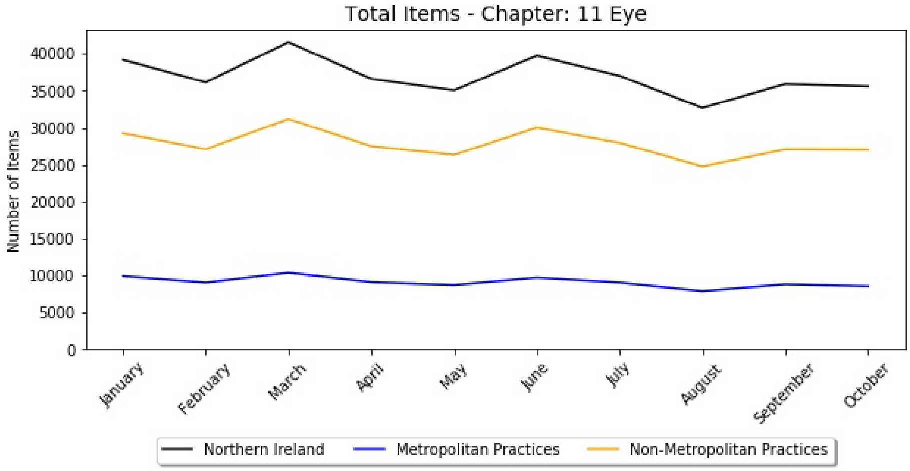


FIGURE H.11: Prescribing by Archetype during the COVID-19 pandemic and first national lockdown - BNF chapter 11 (Eye)

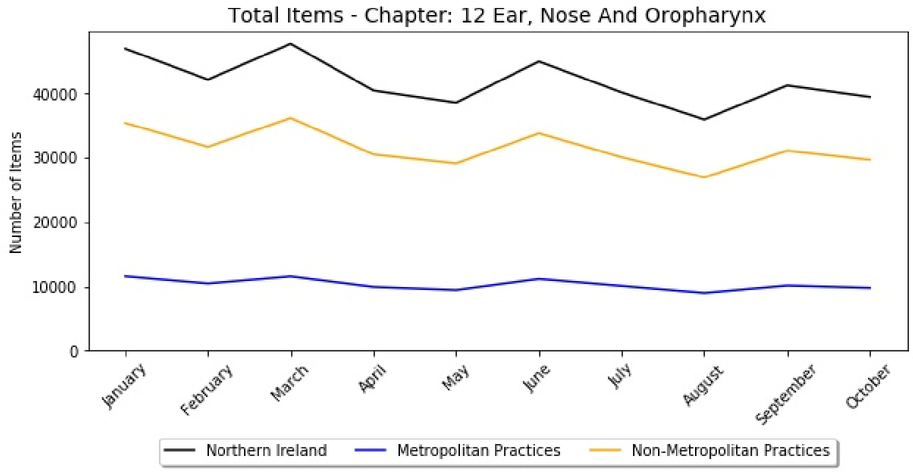


FIGURE H.12: Prescribing by Archetype during the COVID-19 pandemic and first national lockdown - BNF chapter 12 (Ear, Nose and Oropharynx)

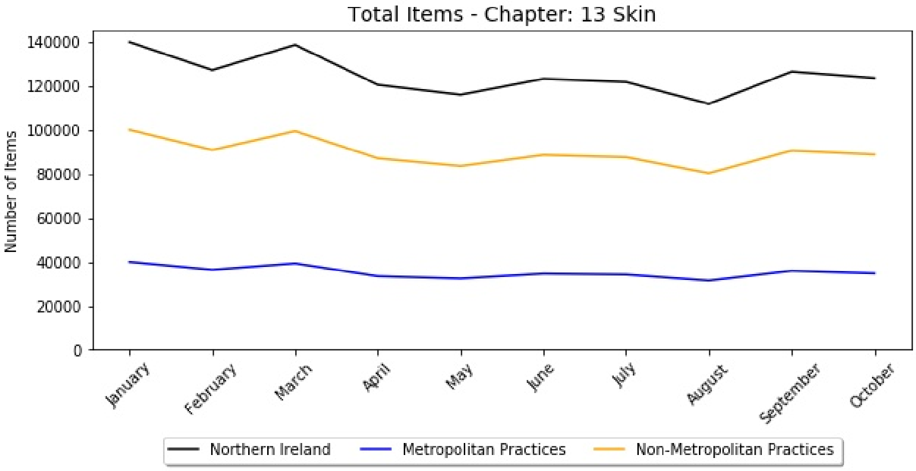


FIGURE H.13: Prescribing by Archetype during the COVID-19 pandemic and first national lockdown - BNF chapter 13 (Skin)

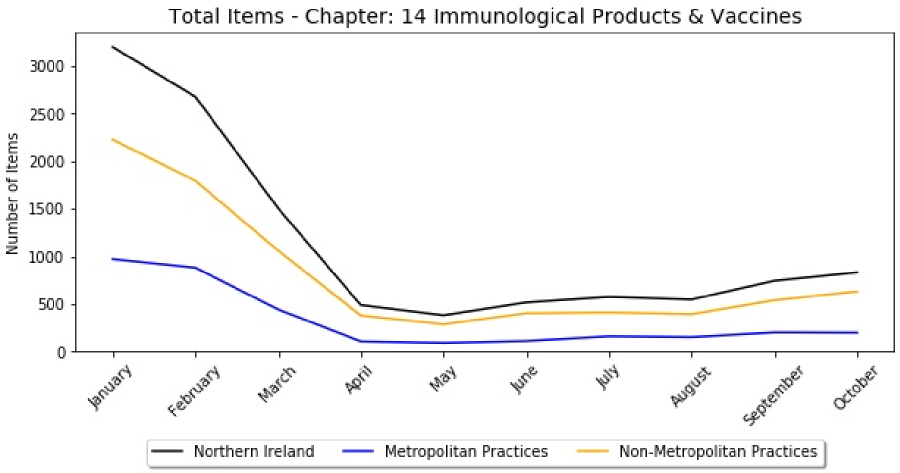


FIGURE H.14: Prescribing by Archetype during the COVID-19 pandemic and first national lockdown - BNF chapter 14 (Immunological Products & Vaccines)

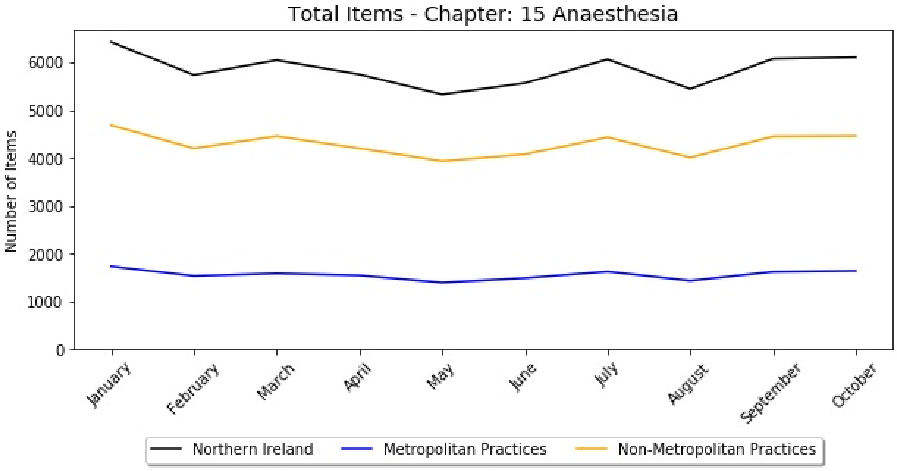


FIGURE H.15: Prescribing by Archetype during the COVID-19 pandemic and first national lockdown - BNF chapter 15 (Anaesthesia)

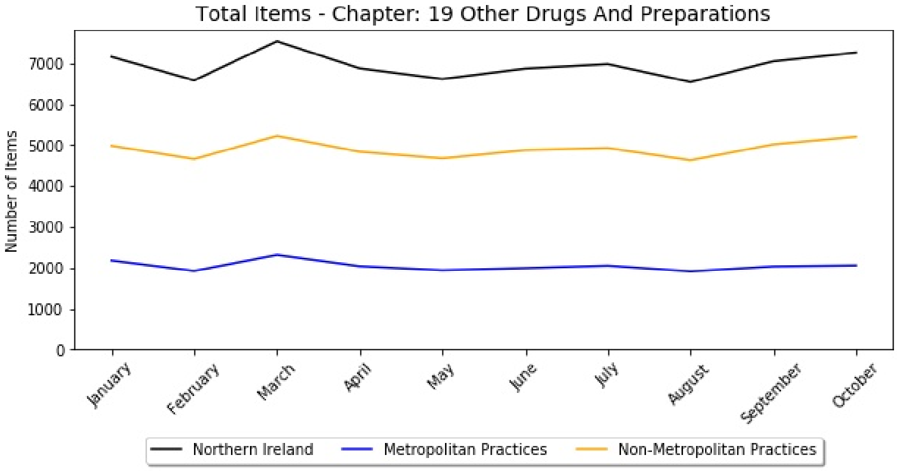


FIGURE H.16: Prescribing by Archetype during the COVID-19 pandemic and first national lockdown - BNF chapter 19 (Other Drugs and Preparations)

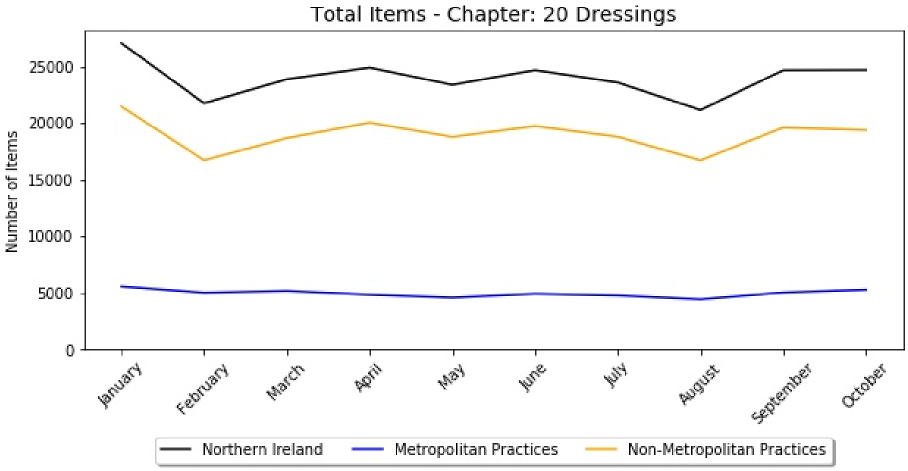


FIGURE H.17: Prescribing by Archetype during the COVID-19 pandemic and first national lockdown - BNF chapter 20 (Dressings)

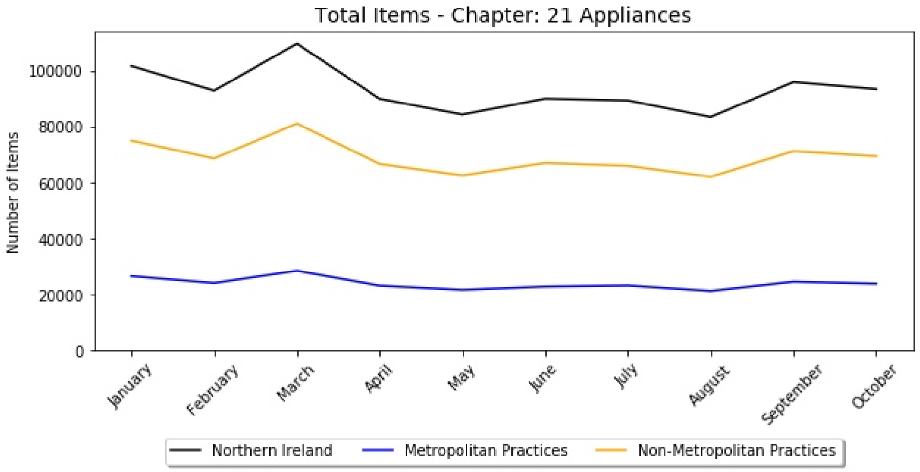


FIGURE H.18: Prescribing by Archetype during the COVID-19 pandemic and first national lockdown - BNF chapter 21 (Appliances)

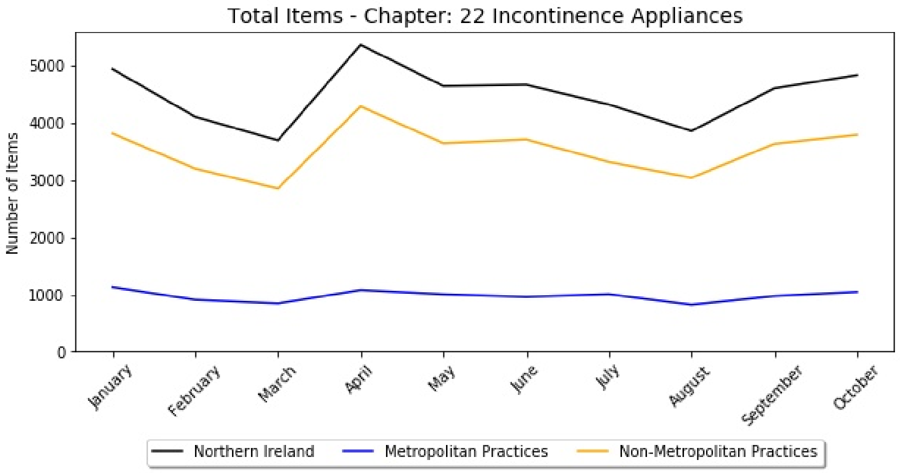


FIGURE H.19: Prescribing by Archetype during the COVID-19 pandemic and first national lockdown - BNF chapter 22 (Incontinence Appliances)

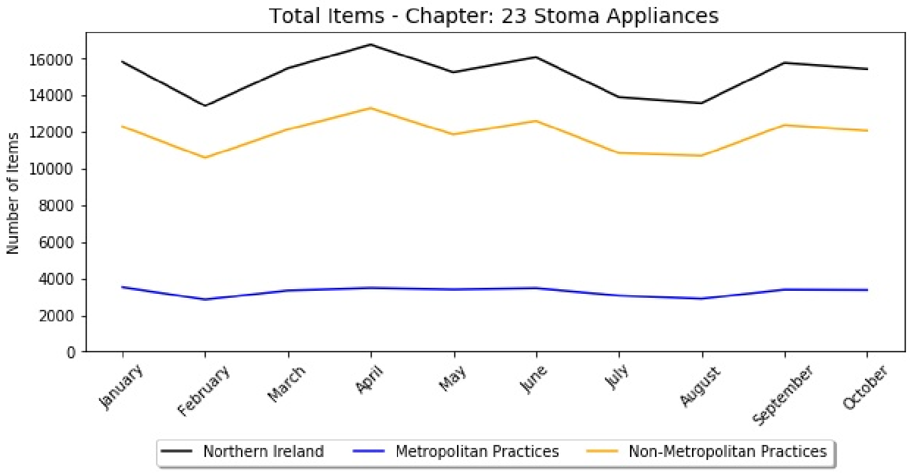


FIGURE H.20: Prescribing by Archetype during the COVID-19 pandemic and first national lockdown - BNF chapter 23 (Stoma Appliances)

# Appendix I

# GP prescribing dashboard MySQL Database - Data Dictionary

TABLE I.1: Data dictionary for SQL Table: bnf

| Column | Type | Null | Comments |
|---|---|---|---|
| chapter | int(2) | No | Chapter number |
| section | int(2) | No | Section number |
| Description | varchar(40) | No | Chapter / Section description |

TABLE I.2: Data dictionary for SQL Table: contractor data

| Column | Type | Null | Comments |
|---|---|---|---|
| Practice | int(3) | No | Practice ID number |
| Pharmacy | int(4) | No | Pharmacy ID number |
| YearMonth | char(7) | No | e.g. 2020-01 |
| Year | int(4) | No | e.g. 2020 |
| Month | int(2) | No | e.g. 1 |
| Number_of_Items | int(8) | No | |
| Distance | float | No | in kilometres |

TABLE I.3: Data dictionary for SQL Table: feedback

| Column | Type | Null | Comments |
|---|---|---|---|
| SessionID | varchar(60) | No | 32 characters generated randomly |
| queryID | char(20) | Yes | generated from date and time of query |
| querystring | json | Yes | JSON object created from dashboard selections |
| liked | char(10) | Yes | |
| comments | text | Yes | |

TABLE I.4: Data dictionary for SQL Table: pharmacies

| Column | Type | Null | Comments |
|---|---|---|---|
| Pharmacy | int(4) | Yes | Pharmacy ID number |
| Pharmacy Name | varchar(100) | Yes | |
| Pharmacy_Address | varchar(83) | Yes | |
| Postcode | varchar(8) | Yes | Full postcode |
| Postcode4 | char(4) | No | 4 digit postcode |

TABLE I.5: Data dictionary for SQL Table: practices

| Column | Type | Null | Comments |
|---|---|---|---|
| Practice | int(3) | Yes | Practice ID number |
| Practice Name | varchar(36) | Yes | |
| Practice_Address | varchar(98) | Yes | |
| Postcode | varchar(8) | Yes | Full postcode |
| Postcode4 | char(4) | No | 4 digit pstcode |
| LGD2014NAME | varchar(31) | Yes | Local Government district |
| Ward_Type | varchar(17) | Yes | Taken from NISRA classification |
| Ward_Name | varchar(27) | Yes | |
| Deprivation_Qartile | varchar(2) | Yes | |
| Practice_Size | varchar(13) | Yes | |
| Archetype | varchar(6) | Yes | Practice type |

TABLE I.6: Data dictionary for SQL Table: prescription data

| Column | Type | Null | Comments |
|---|---|---|---|
| Practice | int(3) | No | Practice ID number |
| YearMonth | char(7) | No | e.g. 2020-01 |
| Year | int(4) | No | e.g. 2020 |
| Month | int(2) | No | e.g. 1 |
| Number_of_Items | int(8) | No | |
| GrossCost | float | No | in £ |
| ActualCost | float | No | in £ |
| BNFchapter | int(2) | No | Chapter number |
| BNFsection | int(2) | No | Section number |

TABLE I.7: Data dictionary for SQL Table: survey

| Column | Type | Null | Comments |
|---|---|---|---|
| SessionID | char(60) | No | 32 characters generated randomly |
| consent | char(20) | No | Date/Time consent button pressed |
| age | int(3) | No | |
| sex | char(1) | No | |
| surveystart | char(20) | No | Date/Time survey page accessed. |
| q1 | int(1) | No | I am interested in Citizen Science |
| q2 | int(1) | No | I am interested in GP Prescribing trends |
| q3 | int(1) | No | I am participating out of curiosity |
| q4 | int(1) | No | It is an opportunity to explore new things |
| q5 | int(1) | No | I want to make scientific knowledge more accessible. |
| q6 | int(1) | No | I want to contribute to science |
| q7 | int(1) | No | The interface is easy to use |
| q8 | int(1) | No | There is sufficient explanation of variables |
| q9 | int(1) | No | The resulting graphs are self explanatory |
| q10 | int(1) | No | I think I would need assistance interpreting the results |
| q11 | int(1) | No | A citizen science application like this would be useful |
| q12 | int(1) | No | A citizen science application like this would provide valuable insights. |
| q13 | int(1) | No | I would trust the results. |
| q14 | int(1) | No | Having access to the raw data would be more interesting. |
| q15 | int(1) | No | I would prefer to do the calculations myself |
| q16 | int(1) | No | Total Number of Items Prescribed |

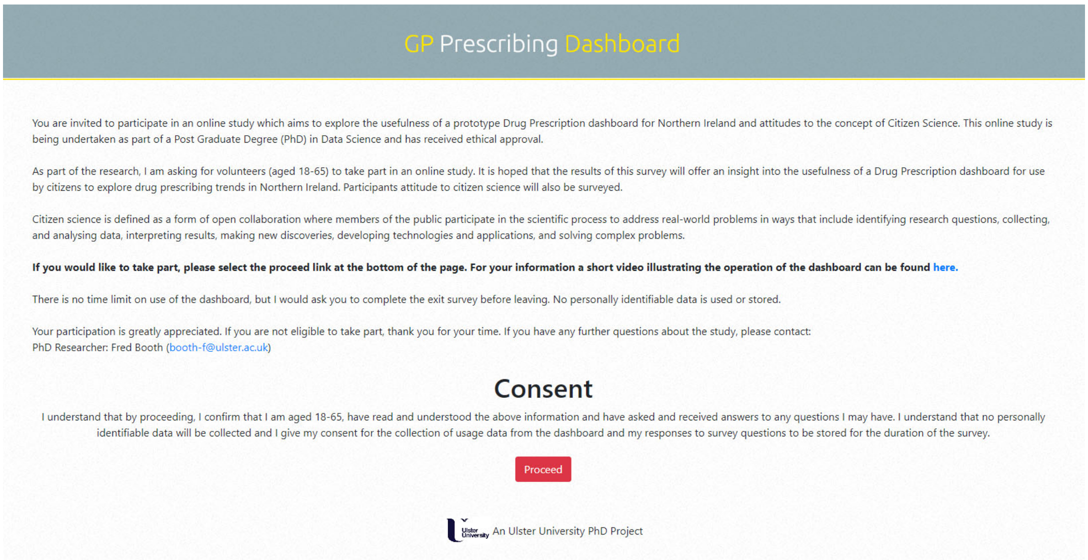
Continued on next page

**Table I.7 – continued from previous page**

| Column | Type | Null | Comments |
|---|---|---|---|
| q17 | int(1) | No | Average Number of Items Prescribed |
| q18 | int(1) | No | NOT USED |
| q19 | int(1) | No | Number of Pharmacies |
| q20 | int(1) | No | Number of Practices |
| q21 | int(1) | No | Average Distance traveled |
| q22 | int(1) | No | Gross Cost |
| q23 | int(1) | No | Actual Cost |
| q24 | int(1) | No | Local Government District |
| q25 | int(1) | No | Ward Type |
| q26 | int(1) | No | Ward (by Name) |
| q27 | int(1) | No | Deprivation Level |
| q28 | int(1) | No | Practice Size |
| q29 | int(1) | No | Practice Type |
| q30 | int(1) | No | Practice (by Name) |
| q31 | int(1) | No | Pharmacy (by Name) |
| q32 | int(1) | No | Postcode |
| q33 | int(1) | No | BNF Chapter |
| q34 | int(1) | No | BNF Chapter with Sections |
| q35 | text | No | Do you feel there are any risks associated with citizen science? |
| q36 | text | No | What features could be added? |

# Appendix J

# GP prescribing dashboard - sample pages



GP Prescribing Dashboard

You are invited to participate in an online study which aims to explore the usefulness of a prototype Drug Prescription dashboard for Northern Ireland and attitudes to the concept of Citizen Science. This online study is being undertaken as part of a Post Graduate Degree (PhD) in Data Science and has received ethical approval.

As part of the research, I am asking for volunteers (aged 18-65) to take part in an online study. It is hoped that the results of this survey will offer an insight into the usefulness of a Drug Prescription dashboard for use by citizens to explore drug prescribing trends in Northern Ireland. Participants attitude to citizen science will also be surveyed.

Citizen science is defined as a form of open collaboration where members of the public participate in the scientific process to address real-world problems in ways that include identifying research questions, collecting, and analysing data, interpreting results, making new discoveries, developing technologies and applications, and solving complex problems.

**If you would like to take part, please select the proceed link at the bottom of the page. For your information a short video illustrating the operation of the dashboard can be found here.**

There is no time limit on use of the dashboard, but I would ask you to complete the exit survey before leaving. No personally identifiable data is used or stored.

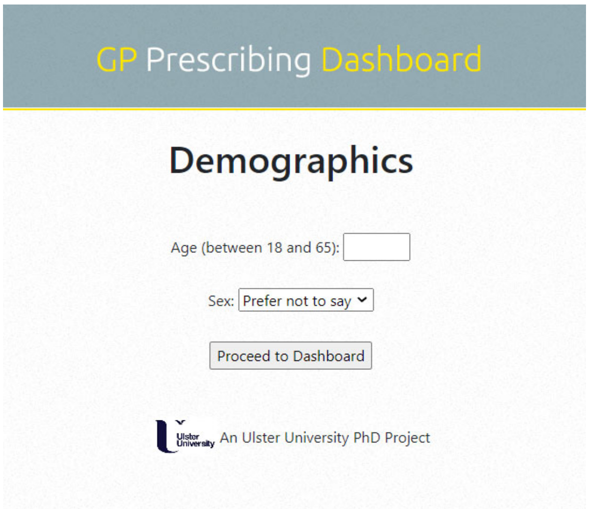
Your participation is greatly appreciated. If you are not eligible to take part, thank you for your time. If you have any further questions about the study, please contact:
PhD Researcher: Fred Booth (booth-f@ulster.ac.uk)

Consent

I understand that by proceeding, I confirm that I am aged 18-65, have read and understood the above information and have asked and received answers to any questions I may have. I understand that no personally identifiable data will be collected and I give my consent for the collection of usage data from the dashboard and my responses to survey questions to be stored for the duration of the survey.

An Ulster University PhD Project

FIGURE J.1: GP prescribing dashboard homepage



GP Prescribing Dashboard

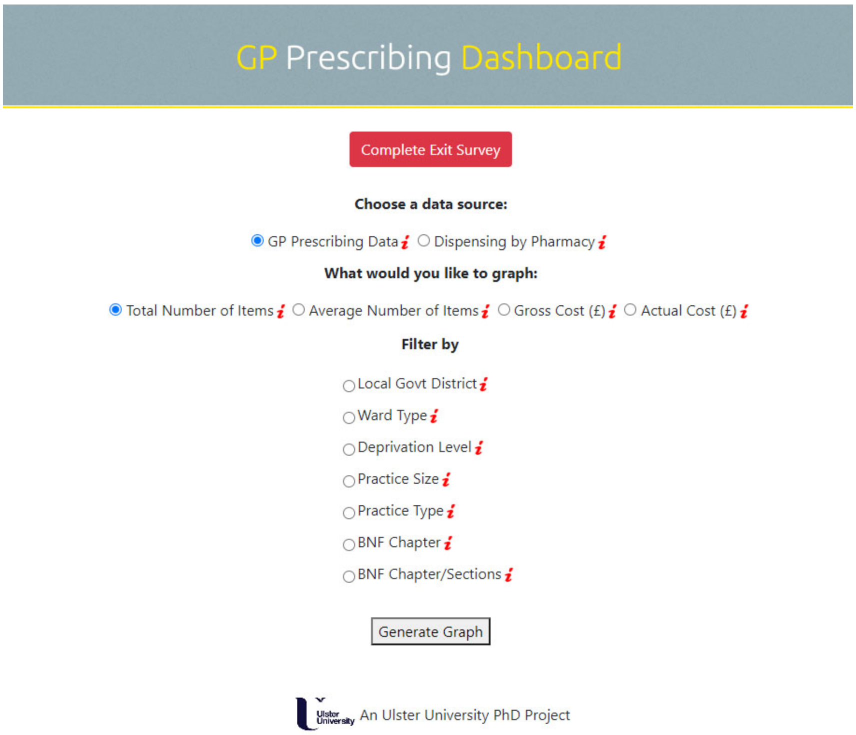
Demographics

Age (between 18 and 65):

Sex: Prefer not to say

An Ulster University PhD Project

FIGURE J.2: GP prescribing dashboard demographics page



GP Prescribing Dashboard

Complete Exit Survey

**Choose a data source:**

● GP Prescribing Data ℹ ○ Dispensing by Pharmacy ℹ

**What would you like to graph:**

● Total Number of Items ℹ ○ Average Number of Items ℹ ○ Gross Cost (£) ℹ ○ Actual Cost (£) ℹ

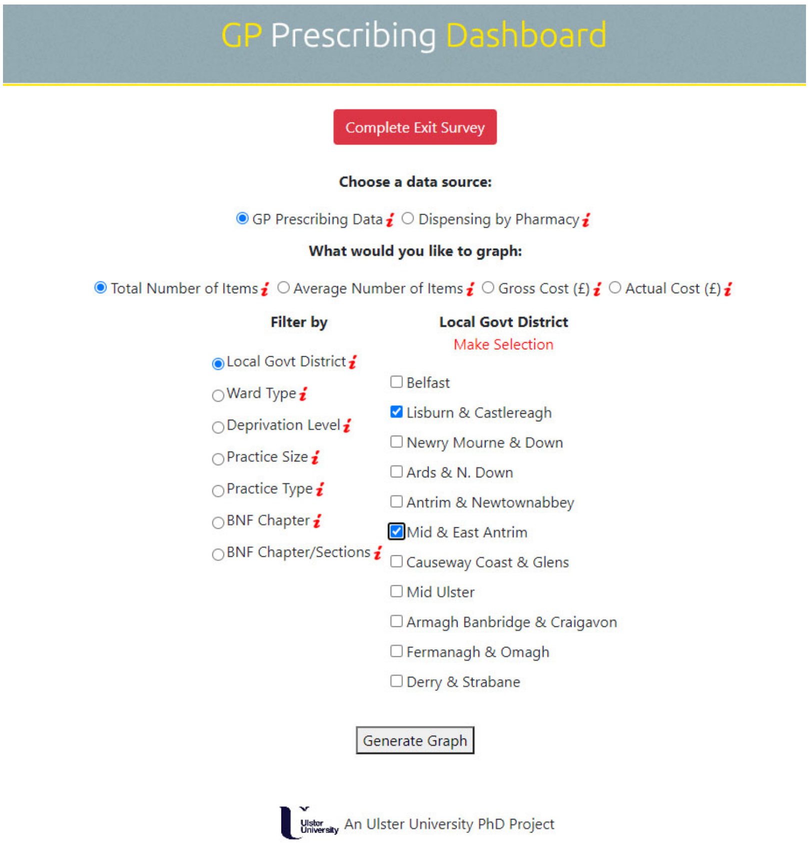
**Filter by**

○ Local Govt District ℹ
○ Ward Type ℹ
○ Deprivation Level ℹ
○ Practice Size ℹ
○ Practice Type ℹ
○ BNF Chapter ℹ
○ BNF Chapter/Sections ℹ

Generate Graph

Ulster University An Ulster University PhD Project

FIGURE J.3: GP prescribing dashboard interface



GP Prescribing Dashboard

Complete Exit Survey

**Choose a data source:**

● GP Prescribing Data ℹ ○ Dispensing by Pharmacy ℹ

**What would you like to graph:**

● Total Number of Items ℹ ○ Average Number of Items ℹ ○ Gross Cost (£) ℹ ○ Actual Cost (£) ℹ

**Filter by**

● Local Govt District ℹ
○ Ward Type ℹ
○ Deprivation Level ℹ
○ Practice Size ℹ
○ Practice Type ℹ
○ BNF Chapter ℹ
○ BNF Chapter/Sections ℹ

**Local Govt District**

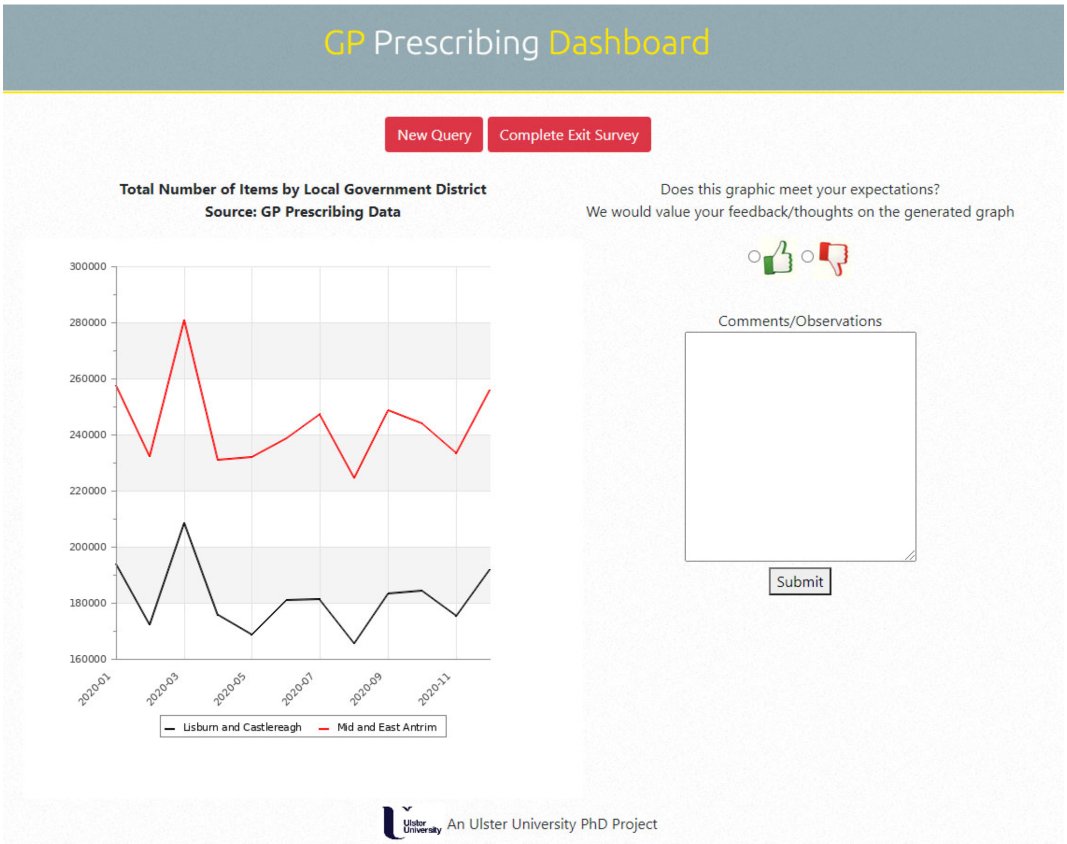
Make Selection

□ Belfast
☑ Lisburn & Castlereagh
□ Newry Mourne & Down
□ Ards & N. Down
□ Antrim & Newtownabbey
☑ Mid & East Antrim
□ Causeway Coast & Glens
□ Mid Ulster
□ Armagh Banbridge & Craigavon
□ Fermanagh & Omagh
□ Derry & Strabane

Generate Graph

Ulster University An Ulster University PhD Project

FIGURE J.4: GP prescribing dashboard interface with filter selected

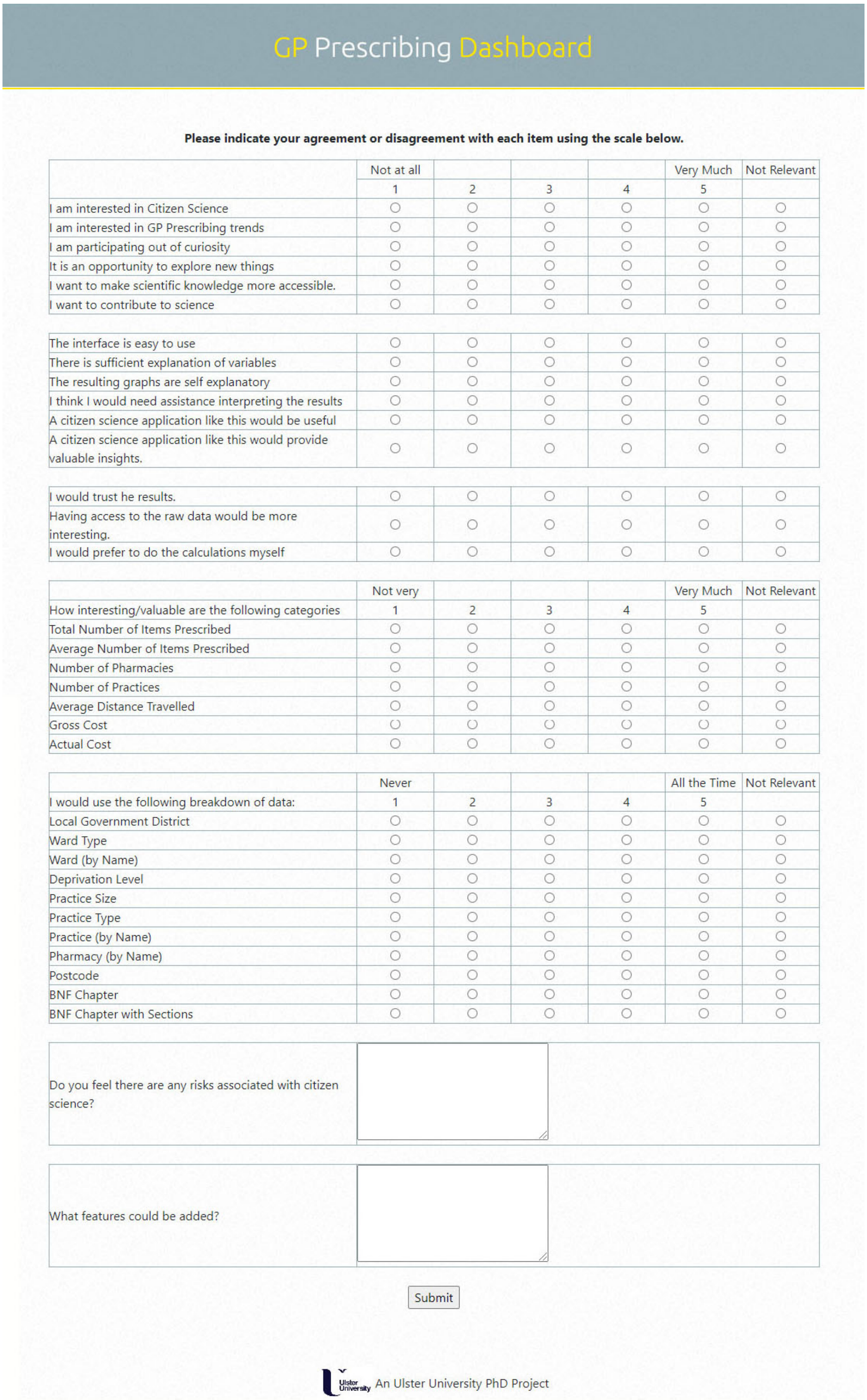


**Total Number of Items by Local Government District**
**Source: GP Prescribing Data**

Does this graphic meet your expectations?
We would value your feedback/thoughts on the generated graph

An Ulster University PhD Project

FIGURE J.5: GP prescribing dashboard result page



# GP Prescribing Dashboard

**Please indicate your agreement or disagreement with each item using the scale below.**

| | Not at all | | | | Very Much | Not Relevant |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | |
| I am interested in Citizen Science | ○ | ○ | ○ | ○ | ○ | ○ |
| I am interested in GP Prescribing trends | ○ | ○ | ○ | ○ | ○ | ○ |
| I am participating out of curiosity | ○ | ○ | ○ | ○ | ○ | ○ |
| It is an opportunity to explore new things | ○ | ○ | ○ | ○ | ○ | ○ |
| I want to make scientific knowledge more accessible. | ○ | ○ | ○ | ○ | ○ | ○ |
| I want to contribute to science | ○ | ○ | ○ | ○ | ○ | ○ |

| | | | | | | |
|---|---|---|---|---|---|---|
| The interface is easy to use | ○ | ○ | ○ | ○ | ○ | ○ |
| There is sufficient explanation of variables | ○ | ○ | ○ | ○ | ○ | ○ |
| The resulting graphs are self explanatory | ○ | ○ | ○ | ○ | ○ | ○ |
| I think I would need assistance interpreting the results | ○ | ○ | ○ | ○ | ○ | ○ |
| A citizen science application like this would be useful | ○ | ○ | ○ | ○ | ○ | ○ |
| A citizen science application like this would provide valuable insights. | ○ | ○ | ○ | ○ | ○ | ○ |

| | | | | | | |
|---|---|---|---|---|---|---|
| I would trust he results. | ○ | ○ | ○ | ○ | ○ | ○ |
| Having access to the raw data would be more interesting. | ○ | ○ | ○ | ○ | ○ | ○ |
| I would prefer to do the calculations myself | ○ | ○ | ○ | ○ | ○ | ○ |

| | Not very | | | | Very Much | Not Relevant |
|---|---|---|---|---|---|---|
| How interesting/valuable are the following categories | 1 | 2 | 3 | 4 | 5 | |
| Total Number of Items Prescribed | ○ | ○ | ○ | ○ | ○ | ○ |
| Average Number of Items Prescribed | ○ | ○ | ○ | ○ | ○ | ○ |
| Number of Pharmacies | ○ | ○ | ○ | ○ | ○ | ○ |
| Number of Practices | ○ | ○ | ○ | ○ | ○ | ○ |
| Average Distance Travelled | ○ | ○ | ○ | ○ | ○ | ○ |
| Gross Cost | ○ | ○ | ○ | ○ | ○ | ○ |
| Actual Cost | ○ | ○ | ○ | ○ | ○ | ○ |

| | Never | | | | All the Time | Not Relevant |
|---|---|---|---|---|---|---|
| I would use the following breakdown of data: | 1 | 2 | 3 | 4 | 5 | |
| Local Government District | ○ | ○ | ○ | ○ | ○ | ○ |
| Ward Type | ○ | ○ | ○ | ○ | ○ | ○ |
| Ward (by Name) | ○ | ○ | ○ | ○ | ○ | ○ |
| Deprivation Level | ○ | ○ | ○ | ○ | ○ | ○ |
| Practice Size | ○ | ○ | ○ | ○ | ○ | ○ |
| Practice Type | ○ | ○ | ○ | ○ | ○ | ○ |
| Practice (by Name) | ○ | ○ | ○ | ○ | ○ | ○ |
| Pharmacy (by Name) | ○ | ○ | ○ | ○ | ○ | ○ |
| Postcode | ○ | ○ | ○ | ○ | ○ | ○ |
| BNF Chapter | ○ | ○ | ○ | ○ | ○ | ○ |
| BNF Chapter with Sections | ○ | ○ | ○ | ○ | ○ | ○ |

Do you feel there are any risks associated with citizen science?

What features could be added?

FIGURE J.6: GP prescribing dashboard survey page

# Appendix K

# Algorithm for transforming JSON object to lines on graph

**Algorithm 5** Algorithm for transforming JSON object to lines on graph

```
$json \leftarrow jsonobject$
$obj \leftarrow decode(json)$
source = obj.source
filter = obj.filter
Variables = obj.variables
graph = obj.graph
if variables is null then
    sql = “SELECT “+graph+” FROM “+source+” GROUPBY datetime”
    Retrieve data from database
    Apply data to draw line on graph
else if length of variables equals 1 then
    sql = “SELECT “+graph+” FROM “+source+”WHERE ”+filter = variable[0]+” GROUPBY datetime”
    Retrieve data from database
    Apply data to draw line on graph
else if length of variables > 1 then
    for Each variable in variables do
        sql = “SELECT “+graph+” FROM “+source+”WHERE ”+filter = variable+” GROUPBY datetime”
        Retrieve data from database
        Apply data to draw line on graph
    end for
end if
```

## Appendix L

# Comments posted by participants regarding resulting graphs

TABLE L.1: Comments posted by participants relating to graphs produced by dashboard

| Age | Sex | Comment |
|---|---|---|
| 31 | F | the line colour for Derry and Belfast are too similar. Maybe add pop-up text of the place name when scrolling over the line, would be useful. |
| 25 | M | Too many zeros on y axis |
| 39 | M | Yes, the graphic meets my expectations. Just one minor comment - the y-axis perhaps needs a comma separator for numbers greater than 999 i.e. "3,750,000" (proposed) versus "3750000" (current). |
| 40 | M | This looks great. I will suggest that both horizontal and vertical axis are label to make the graph more descriptive. |
| 49 | M | The Y axis should be formatted 9,999,999 would be helpful |
| 42 | M | No option to pick either total or avg when picking districts |
| 36 | F | The graph needs axis labels to aid understanding and perhaps more user friendly explanatory text around what it is displaying. |
| 36 | F | Same comments as previous graph - it needs axis labels more user friendly explanatory text. |
| 22 | F | Could it be possible to make interactive graphs, where you can zoom in to reduce the axis range? I glanced at this graph quickly and assumed the decline in item number in 2020-08 was at zero. |
| 44 | X | It would be nice to be able to isolate/view individual data points |
| 55 | M | Selected LGD should be displayed in the graph. Include day of week beside dates. |

Continued on next page

**Table L.1 – continued from previous page**

| Age | Sex | Comment |
|---|---|---|
| 55 | M | What does average mean? Average per day, per GP, per pharmacy? |
| 55 | M | It would be helpful to be able to compare queries side by side, e.g. rural vs urban |
| 20 | F | Clear, concise. However labels on either axis of the graph would prove useful. |
| 44 | F | Can this be broken down into subgroups eg analgesia, antiepileptic medications? |
| 44 | F | A combined graph with deprivation levels and BNF subtypes would be useful so that population trends can be seen. This could give an early indication of evolving trends eg in drug misuse. |
| 45 | F | Maybe a seperator comma on the Y axes units, to help make sense of the number more quickly. |
| 45 | F | Not clear what the average refers to - average number per GP? Per person? |
| 45 | F | Separator commas on Y axes would be helpful to make sense of numbers. |
| 50 | M | can see the lockdown trends, panic in march 2020 |
| 50 | M | based on last query seems clear pattern |
| 50 | M | similar trends |
| 50 | M | trends all seem similar, clearly data changes. My dad was a pharmacist, so I new average prescriptions per month as I used to get paid £10 to stamp them |
| 36 | F | I guess. Not sure what I am looking at or the point |
| 22 | F | I thought it would be fairly consistent from month to month. |
| 44 | F | Clear |
| 23 | M | Different colours used in plotting make the graph clear and easily legible. |
| 23 | M | Plot colours used make it easily legible and clear. |
| 65 | F | just tried to leave comment but got kicked out |
| 22 | F | Would maybe be good to see potential explanations for peaks and troughs in data e.g. 2020-03 spike due to COVID-19 cases rising etc. |
| 44 | X | no data labels |
| 44 | X | Not sure what is represented here |

# Appendix M

# Exit survey comments

TABLE M.1: Comments posted in response to the question "Do you feel there are any risks associated with citizen science?"

| Age | Sex | Comment |
|---|---|---|
| 56 | M | Can change behaviours, e.g. patients demanding something from their GP when they have seen the trends elsewhere. |
| 31 | F | Generalised findings can be used to misrepresent service level issues. For example, prescribing trends are higher in Belfast because it has a bigger population, if this is not explicitly stated, citizens may perceive this as Belfast is more reactive in prescribing medications compared to other areas for example. Likewise, general prescription trends do not indicate differences in levels of prescriptions for pain medication, treatment of psychopathology, chronic or acute illness etc. That information would let citizens see if there is a difference in prescription trends for issues relating to their personal health concerns increasing their investment in the process. |
| 36 | F | I am pro-Citizen Science but feel it is important to be very careful about citizen involvement to ensure issues around lack of understanding or misinterpretation are well-managed in the process. |
| 50 | M | I see potential for gaining better insights and trends |
| 22 | F | Jumping to conclusions, the general public may not know how to interpret data in a systematic, unbiased way and take from it what they want to see. |
| 23 | M | Making sure enough people have assessed/analysed the data to obtain reliable results. |
| 54 | M | Mis-interpretation of data... |
| 65 | F | Misunderstanding/uninformed interpretation and spread of misinformation/disinformation as a result |
| 63 | M | no |

Continued on next page

**Table M.1 – continued from previous page**

| Age | Sex | Comment |
|---|---|---|
| 40 | M | no |
| 40 | F | no |
| 27 | F | No |
| 18 | M | No |
| 44 | F | No |
| 33 | M | No |
| 28 | F | no as all data would not be able to link to a specific person |
| 44 | M | No, everyone should have access to this type of information. Whether they use it or not is personal preference. |
| 23 | F | Not sure -never understood that it existed |
| 59 | M | Not when anonymised |
| 25 | M | Only misinterpretation |
| 27 | M | Potentially if the citizen scientists are unable to discuss their analysis with experts and then draw the wrong conclusions. On the whole though I believe it to be very beneficial - more eyes on data means more viewpoints and perspectives that can lead to more understanding. Citizen science is also effective in getting more people interested in science which is always a good thing! |
| 20 | F | Sometimes |
| 39 | M | Unsure. |
| 45 | F | Yes - a good understanding on what the data is based on, how it is collected etc. |

TABLE M.2: Comments posted in response to the question "What features could be added?"

| Age | Sex | Comment |
|---|---|---|
| 50 | M | Explanations of the data sets and trends |
| 31 | F | All graphs need to be labeled, not all citizens are used to reviewing graphs or visual representations of data. Graphs should be interactive, if you scroll over them you should see the number, place name etc. |
| 56 | M | Classifications by drug |
| 25 | M | More dynamic UI and faster graph rendering |
| 39 | M | Comma separator for amounts on the y-axis greater than £999 i.e. "3,750,000". |
| 54 | M | Dashboard could be more explanatory/visual |
| 36 | F | More labels on graphs, more explanatory text. |

Continued on next page

**Table M.2 – continued from previous page**

| Age | Sex | Comment |
|---|---|---|
| 20 | F | This surgery was very well made, easy to understand, and gathered keen interest in the field |
| 44 | M | More friendly UI. Better descriptions of variables. Variables that give more useful information such as a breakdown of what is being prescribed by category for example. Costs per category etc. Let people know where the money is being spent and possibly give insight into where the major problems are in healthcare. |
| 40 | F | I would like to see graph labels on axis and access to the raw data would be interesting but maybe not necessary. Overall it is a very interesting and well thought out process. |
| 23 | F | More description of the results graph. Needs axis labels and graph title needs to explain the results better |
| 59 | M | Breakdown by age group |
| 22 | F | Interactive graphs i.e. ability to adjust axis ranges or compare over a longer period of time, annotations of potential reasons for changes in a graph. I also think the language explaining the parameters to make a graph should be in lay terms, the average person would not be able to understand many of them. |
| 23 | F | An explanation with the graphs - x & y headers at least. More information about pharmacies/gp practices local to participants |
| 18 | M | A guide to interpret data |
| 44 | F | Alerts to prescription requests among the local population, this may provide and early warning to prescribers as to trends in misused drugs eg cyclizine. |
| 27 | M | Manual re-scaling of graphs might be helpful. Having the variables named directly on the plots would make them easier to use at a glance |
| 23 | M | A Select/Deselect All option. Optional linear/logarithmic plot scales. Date inputs and/or ranges, where the data allows (e.g. user input years/date ranges, or multi-year plots, or 12-month plots). |
| 33 | M | More graphs! |
| 45 | F | It would be nice not to have to look up each graph separately, and have it more along the lines of a Power BI dashboard. Thank you. |

# References

Adeley, J E A N M (2006). "e-Prescribing , Efficiency , Quality : Lessons from the Computerization of UK Family Practice". In: pp. 470–475. DOI: 10.1197/jamia.M2041.Introduction.

Agrawal, Ankit, Sanchit Misra, Ramanathan Narayanan, Lalith Polepeddi, and Alok Choudhary (2012). "Lung cancer survival prediction using ensemble data mining on SEER data". In: *Scientific Programming* 20.1, pp. 29–42. ISSN: 10589244. DOI: 10.3233/SPR-2012-0335.

Andreu-Perez, Javier, Carmen C.Y. Poon, Robert D. Merrifield, Stephen T.C. Wong, and Guang Zhong Yang (2015). "Big Data for Health". In: *IEEE Journal of Biomedical and Health Informatics* 19.4, pp. 1193–1208. ISSN: 21682194. DOI: 10.1109/JBHI.2015.2450362.

Ankerst, Mihael, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander (June 1999). "OPTICS: Ordering Points to Identify the Clustering Structure". In: *SIGMOD Rec.* 28.2, 49–60. ISSN: 0163-5808. DOI: 10.1145/304181.304187. URL: https://doi.org/10.1145/304181.304187.

Armitage, Richard (2020). "in England during". In: *The Lancet Psychiatry* 8.2, e3. ISSN: 2215-0366. DOI: 10.1016/S2215-0366(20)30530-7. URL: http://dx.doi.org/10.1016/S2215-0366(20)30530-7.

Bahri, Safa, Nesrine Zoghlami, Mourad Abed, and João Manuel R S Tavares (2019). "BIG DATA for Healthcare : A Survey". In: 7, pp. 7397–7408. DOI: 10.1109/ACCESS.2018.2889180.

Baldwin, Timothy, Paul Cook, Marco Lui, Andrew Mackinlay, and Li Wang (2013). "How Noisy Social Media Text , How Diffrnt Social Media Sources ?" In: *Proc. IJCNLP 2013* October, pp. 356–364.

Bandyopadhyay, Sunayan, Julian Wolfson, David M. Vock, Gabriela Vazquez-Benitez, Gediminas Adomavicius, Mohamed Elidrisi, Paul E. Johnson, and Patrick J. O'Connor (2015). "Data mining for censored time-to-event data: a Bayesian network model for predicting cardiovascular risk from electronic health record data". In: *Data Mining and Knowledge Discovery* 29.4, pp. 1033–1069. ISSN: 1573756X. DOI: 10.1007/s10618-014-0386-6. arXiv: 1404.2189.

Barakat, Nahla, Andrew P. Bradley, and Mohamed Nabil H. Barakat (2010). "Intelligible support vector machines for diagnosis of diabetes mellitus". In: *IEEE Transactions on Information Technology in Biomedicine* 14.4, pp. 1114–1120. ISSN: 10897771. DOI: 10.1109/TITB.2009.2039485.

Bender, Ralf and Stefan Lange (2001). “Adjusting for multiple testing—when and how?” In: *Journal of Clinical Epidemiology* 54.4, pp. 343–349. ISSN: 0895-4356. DOI: https://doi.org/10.1016/S0895-4356(00)00314-0. URL: https://www.sciencedirect.com/science/article/pii/S0895435600003140.

Bhardwaj, Ashutosh (2020). *Silhouette Coefficient. This is my first medium story, so... | by Ashutosh Bhardwaj | Towards Data Science*. https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c.

Biernacki, C., G. Celeux, and G. Govaert (2000). “Assessing a mixture model for clustering with the integrated completed likelihood”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.7, pp. 719–725. DOI: 10.1109/34.865189.

Bland, J Martin and Douglas G Altman (1995). “Multiple significance tests: the Bonferroni method”. In: *BMJ* 310.6973, p. 170. ISSN: 0959-8138. DOI: 10.1136/bmj.310.6973.170. eprint: https://www.bmj.com/content/310/6973/170.full.pdf. URL: https://www.bmj.com/content/310/6973/170.

Bohm, Klaus, Anett Mehler-Bicher, and Dennis Fenchel (2011). “GeoVisualAnalytics in the public health sector”. In: *ICSDM 2011 - Proceedings 2011 IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services*, pp. 291–294. DOI: 10.1109/ICSDM.2011.5969049.

Bonney, Rick, Tina B Phillips, Heidi L Ballard, and Jody W Enck (Jan. 2016). “Can citizen science enhance public understanding of science?” en. In: *Public Underst. Sci.* 25.1, pp. 2–16.

Boopathy, P (Aug. 2021). *Let’s Understand All About Data Wrangling! - Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2021/08/lets-understand-all-about-data-wrangling/. (Accessed on 01/12/2021).

Booth, FG (Apr. 2022). *Analytics, Visualisation and Machine Learning of General Practitioner Prescribing using Open Health Data*. DOI: 10.5281/zenodo.6409927. URL: https://doi.org/10.5281/zenodo.6409927.

Booth, Frederick G, Maurice D. Mulvenna, Raymond R. Bond, Kieran McGlade, Brian Cleland, Debbie Rankin, Jonathan Wallace, and Michaela Black (2021a). “COVID-19 and lockdown: The highs and lows of general practitioner prescribing”. In: *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp. 1–4. DOI: 10.1109/BHI50953.2021.9508575.

Booth, Frederick G, Maurice D. Mulvenna, Raymond R. Bond, Kieran McGlade, and Debbie Rankin (2020a). “Examining the Effect of Deprivation on Prescribing Behaviours in Northern Ireland”. In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1897–1900. DOI: 10.1109/BIBM49941.2020.9313132.

Booth, Frederick G., Maurice D. Mulvenna, Raymond R. Bond, Kieran McGlade, Debbie Rankin, and Jonathan Wallace (2020b). “Examining the Effect of General Practitioner Practice Size on Prescribing Behaviours in Northern Ireland”. In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2705–2708. DOI: 10.1109/BIBM49941.2020.9313570.

Booth, Frederick G., Raymond R Bond, Maurice D Mulvenna, Brian Cleland, Kieran McGlade, Debbie Rankin, Jonathan Wallace, and Michaela Black (2021b). "Discovering and comparing types of general practitioner practices using geolocational features and prescribing behaviours by means of K-means clustering". In: *Scientific Reports* 11.1, pp. 1–15. ISSN: 2045-2322. DOI: 10.1038/s41598-021-97716-3. URL: https://doi.org/10.1038/s41598-021-97716-3.

Branova, Bojan (2020). "Daily Monitoring of Emotional Responses to the Coronavirus Pandemic in Serbia : A Citizen Science Approach". In: 11.August. DOI: 10.3389/fpsyg.2020.02133.

Braun-Fahrländer, C, M Gassner, L Grize, U Neu, FH Sennhauser, HS Varonier, JC Vuille, and B Wüthrich (1999). "Prevalence of hay fever and allergic sensitization in farmer's children and their peers living in the same rural community. SCARPOL team. Swiss Study on Childhood Allergy and Respiratory Symptoms with Respect to Air Pollution". In: *Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology* 29.1, 28—34. ISSN: 0954-7894. DOI: 10.1046/j.1365-2222.1999.00479.x. URL: https://doi.org/10.1046/j.1365-2222.1999.00479.x.

Broder, Andrei, Lluis Garcia-Pueyo, Vanja Josifovski, Sergei Vassilvitskii, and Srihari Venkatesan (2014). "Scalable k-means by ranked retrieval". In: *WSDM 2014 - Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, pp. 233–242. DOI: 10.1145/2556195.2556260.

Bucholc, Magda, Maurice O Kane, Siobhan Ashe, and Kongfatt Wong-lin (2018). "Prescriptive variability of drugs by general practitioners". In: pp. 1–13.

Buczak, Anna L, Benjamin Baugher, Erhan Guven, Linda Moniz, Steven M. Babin, and Jean-Paul Chretien (2016). "Prediction of Peaks of Seasonal Influenza in Military Health-Care Data". In: *Biomedical Engineering and Computational Biology* 7s2, BECB.S36277. ISSN: 1179-5972. DOI: 10.4137/becb.s36277.

California State University (2022). *Tests of Statistical Significance*. https://home.csulb.edu/~msaintg/ppa696/696stsig.htm.

Carter, Mary, Sarah Chapman, and Margaret C. Watson (2021). "Multiplicity and complexity: A qualitative exploration of influences on prescribing in UK general practice". In: *BMJ Open* 11.1, pp. 1–10. ISSN: 20446055. DOI: 10.1136/bmjopen-2020-041460.

Carús Candás, Juan Luis, Víctor Peláez, Gloria López, Miguel Ángel Fernández, Eduardo Álvarez, and Gabriel Díaz (2014). "An automatic data mining method to detect abnormal human behaviour using physical activity measurements". In: *Pervasive and Mobile Computing* 15, pp. 228–241. ISSN: 15741192. DOI: 10.1016/j.pmcj.2014.09.007. URL: http://dx.doi.org/10.1016/j.pmcj.2014.09.007.

Carvalho, Luiz F.M., Carlos H.C. Teixeira, Wagner Meira, Martin Ester, Osvaldo Carvalho, and Maria Helena Brandao (2017). "Provider-Consumer Anomaly Detection for Healthcare Systems". In: *Proceedings - 2017 IEEE International Conference on Healthcare Informatics, ICHI 2017*, pp. 229–238. DOI: 10.1109/ICHI.2017.75.

CFFR (2016). “Affordable Housing Working Group: Issues Paper”. In: *Nips* January, p. 14. ISSN: <null>. URL: `https://papers.nips.cc/paper/2092-on-spectral-clustering-analysis-and-an-algorithm.pdf\%0Ahhttp://www.treasury.gov.au/ConsultationsandReviews/Consultations/2016/CFFR-Affordable-Housing-Working-Group`.

Chang, Chun Lang (2007). “A study of applying data mining to early intervention for developmentally-delayed children”. In: *Expert Systems with Applications* 33.2, pp. 407–412. ISSN: 09574174. DOI: `10.1016/j.eswa.2006.05.007`.

Chen, SY., Z. Feng, and X. Yi (2017). *A general introduction to adjustment for multiple comparisons - PMC*. `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5506159/`.

Chignard, S (2013). *A brief history of Open Data*. `http://www.paristechreview.com/2013/03/29/brief-history-open-data/`. (Accessed on 08/26/2021).

Chong, Siow Ann, Edimansyah Abdin, Janhavi Ajit Vaingankar, Derrick Heng, Cathy Sherbourne, Mabel Yap, Yee Wei Lim, Hwee Bee Wong, Bonnie Ghosh-Dastidhar, Kian Woon Kwok, and Mythily Subramaniam (2012). “A population-based survey of mental disorders in Singapore”. In: *Annals of the Academy of Medicine Singapore* 41.2, pp. 49–66. ISSN: 03044602.

Cleland, Brian, Jonathan Wallace, Raymond Bond, Michaela Black, Maurice Mulvenna, Deborah Rankin, Austin Tanney, Magee Campus, and Northern Ireland (2018). “Insights into Antidepressant Prescribing Using Open Health Data ”. In: *Big Data Research* 12, pp. 41–48. ISSN: 2214-5796. DOI: `10.1016/j.bdr.2018.02.002`. URL: `https://doi.org/10.1016/j.bdr.2018.02.002`.

Cunningham, Niall and Ian Gregory (2014). “Hard to miss, easy to blame? Peacelines, interfaces and political deaths in Belfast during the Troubles”. In: *Political Geography* 40, pp. 64–78. ISSN: 09626298. DOI: `10.1016/j.polgeo.2014.02.004`. URL: `http://dx.doi.org/10.1016/j.polgeo.2014.02.004`.

Curtis, Helen J and Ben Goldacre (2018). “OpenPrescribing : normalised data and software tool to research trends in English NHS primary care prescribing 1998 – 2016”. In: pp. 1–10. DOI: `10.1136/bmjopen-2017-019921`.

Dash, Manoranjan and Huan Liu (2000). “Feature Selection for Clustering”. In: *Knowledge Discovery and Data Mining. Current Issues and New Applications*. Ed. by Takao Terano, Huan Liu, and Arbee L. P. Chen. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 110–121. ISBN: 978-3-540-45571-4.

DataNovia (2018a). *5 Amazing Types of Clustering Methods You Should Know - Datanovia*. `https://www.datanovia.com/en/blog/types-of-clustering-methods-overview-and-quick-start-r-code/`.

— (2018b). *DBSCAN: Density-Based Clustering Essentials - Datanovia*. `https://www.datanovia.com/en/lessons/dbscan-density-based-clustering-essentials/`.

Delen, Dursun (2009). “Analysis of cancer data: A data mining approach”. In: *Expert Systems* 26.1, pp. 100–112. ISSN: 02664720. DOI: `10.1111/j.1468-0394.2008.00480.x`.

Den Broeder, Lea, Jeroen Devilee, Hans Van Oers, A. Jantine Schuit, and Annemarie Wagemakers (2018). "Citizen Science for public health". In: *Health promotion international* 33.3, pp. 505–514. ISSN: 14602245. DOI: 10.1093/heapro/daw086.

Dey, D (Jan. 2021). *Calinski-Harabasz Index – Cluster Validity indices | Set 3 - GeeksforGeeks*. https://www.geeksforgeeks.org/calinski-harabasz-index-cluster-validity-indices-set-3/.

Drury, Ethan (2020). "Antibiotics Unearthed : Antibiotic Discovery and Citizen Science December 2020". In: December.

E. Youssef, Ahmed (2014). "A Framework for Secure Healthcare Systems Based on Big Data Analytics in Mobile Cloud Computing Environments". In: *The International Journal of Ambient Systems and Applications* 2.2, pp. 1–11. ISSN: 23216344. DOI: 10.5121/ijasa.2014.2201.

Eccles, Abi, Michael Hopper, Amadea Turk, and Helen Atherton (2019). "Patient use of an online triage platform :" in: May, pp. 336–344.

Eriksson, Robert, Thomas Werge, Lars Juhl Jensen, and Søren Brunak (2014). "Dose-specific adverse drug reaction identification in electronic patient records: Temporal data mining in an inpatient psychiatric population". In: *Drug Safety* 37.4, pp. 237–247. ISSN: 11791942. DOI: 10.1007/s40264-014-0145-z.

Ester, M, H P Kriegel, J Sander, and Xu Xiaowei (Dec. 1996). "A density-based algorithm for discovering clusters in large spatial databases with noise". In: URL: https://www.osti.gov/biblio/421283.

Fan, Brandon, Weiguo Fan, Carly Smith, and Harold Skip (2020). "Adverse drug event detection and extraction from open data : A deep learning approach". In: *Information Processing and Management* 57.1, p. 102131. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2019.102131. URL: https://doi.org/10.1016/j.ipm.2019.102131.

Fernandez, J. (2020). *The statistical analysis t-test explained for beginners and experts | by Javier Fernandez | Towards Data Science*. https://towardsdatascience.com/the-statistical-analysis-t-test-explained-for-beginners-and-experts-fd0e358bbb62. (Accessed on 10/01/2021).

Frazer, John Scott and Glenn Ross Frazer (2020). "GP prescribing in Northern Ireland by deprivation index : retrospective analysis". In: pp. 1–9. DOI: 10.1136/fmch-2020-000376.

French, Declan (2009). "Residential segregation and health in Northern Ireland". In: *Health and Place* 15.3, pp. 888–896. ISSN: 13538292. DOI: 10.1016/j.healthplace.2009.02.012.

Frey, B.J. and Dueck D. (2007). *Clustering by Passing Messages Between Data Points*. https://www.science.org/lookup/doi/10.1126/science.1136800. (Accessed on 09/09/2021).

Gebka, Elisabeth, Antoine Clarinval, and Anthony Simonofski (2019). "Generating Value with Open Government Data : Beyond the Programmer". In: pp. 0–1.

Goldacre B.and MacKenna, B., H.J. Curtis, R. Croker, and A.J. Walker (2021). *Trends in antidepressant prescribing in England - The Lancet Psychiatry*. https://www.thelancet.com/journals/lanpsy/article/PIIS2215-0366(21)00081-X/fulltext. (Accessed on 09/27/2021).

Goldacre, Ben and Brian Mackenna (2020). “The NHS deserves better use of hospital medicines data Ben Goldacre and Brian MacKenna argue that hospital medicines data has huge potential to improve”. In: pp. 1–5. DOI: 10.1136/bmj.m2607.

Google Developers (2022). *k-Means Advantages and Disadvantages | Machine Learning | Google Developers*. https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages.

Grandview Research (2021). *Healthcare Analytics Market Size & Growth Report, 2020-2027*. https://www.grandviewresearch.com/industry-analysis/healthcare-analytics-market. (Accessed on 08/26/2021).

Grootendorst, Maarten (2019). *Validating your Machine Learning Model | by Maarten Grootendorst | Towards Data Science*. https://towardsdatascience.com/validating-your-machine-learning-model-25b4c8643fb7.

Gujral, Garima (2020). “ARTIFICIAL INTELLIGENCE ( AI ) AND DATA SCIENCE FOR DEVELOPING INTELLIGENT HEALTH INFORMATICS SYSTEMS”. In: January.

Haklay, Mordechai (Muki), Daniel Dörler, Florian Heigl, Marina Manzoni, Susanne Hecker, and Katrin Vohland (2021). “What Is Citizen Science? The Challenges of Definition”. In: *The Science of Citizen Science*. Ed. by Katrin Vohland, Anne Land-Zandstra, Luigi Ceccaroni, Rob Lemmens, Josep Perelló, Marisa Ponti, Roeland Samson, and Katherin Wagenknecht. Cham: Springer International Publishing, pp. 13–33. ISBN: 978-3-030-58278-4. DOI: 10.1007/978-3-030-58278-4_2. URL: https://doi.org/10.1007/978-3-030-58278-4_2.

Han, S. M., G. Greenfield, A. Majeed, and B. Hayhoe (Nov. 2020). “Impact of Remote Consultations on Antibiotic Prescribing in Primary Health Care: Systematic Review”. In: *J Med Internet Res* 22.11, e23482.

Hanley, J.P., E. Jackson, L.A. Morrissey, D.M. Rizzo, B.L. Sprague, I.N. Sarkar, and F.E. Carr (2015). *Geospatial and Temporal Analysis of Thyroid Cancer Incidence in a Rural Population | Thyroid*. https://www.liebertpub.com/doi/epub/10.1089/thy.2015.0039. URL: https://doi.org/10.1089/thy.2015.0039.

Hao, Bibo, Wen Sun, Yiqin Yu, and Guotong Xie (2017). “Developing Healthcare Data Analytics APPs with Open Data Science Tools”. In: *Studies in health technology and informatics* 235, 176—180. ISSN: 0926-9630. URL: http://europepmc.org/abstract/MED/28423778.

Harpaz, Rave, Herbert S. Chase, and Carol Friedman (2010). “Mining multi-item drug adverse effect associations in spontaneous reporting systems”. In: *BMC Bioinformatics* 11.SUPPL. 9, pp. 5–12. ISSN: 14712105. DOI: 10.1186/1471-2105-11-S9-S7.

Harpaz, Rave, Santiago Vilar, William DuMouchel, Hojjat Salmasian, Krystl Haerian, Nigam H. Shah, Herbert S. Chase, and Carol Friedman (2013). "Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions". In: *Journal of the American Medical Informatics Association* 20.3, pp. 413–419. ISSN: 10675027. DOI: 10.1136/amiajnl-2012-000930.

Hochberg, Yosef (Dec. 1988). "A sharper Bonferroni procedure for multiple tests of significance". In: *Biometrika* 75.4, pp. 800–802. ISSN: 0006-3444. DOI: 10.1093/biomet/75.4.800. eprint: https://academic.oup.com/biomet/article-pdf/75/4/800/1170595/75-4-800.pdf. URL: https://doi.org/10.1093/biomet/75.4.800.

Hogg, R., D. Ritchie, B. de Kok, C. Wood, and G. Huby (Apr. 2013). "Parenting support for families with young children - a public health, user-focused study undertaken in a semi-rural area of Scotland". In: *J Clin Nurs* 22.7-8, pp. 1140–1150.

Holm, Sture (1979). "A Simple Sequentially Rejective Multiple Test Procedure". In: *Scandinavian Journal of Statistics* 6.2, pp. 65–70. ISSN: 03036898, 14679469. URL: http://www.jstor.org/stable/4615733 (visited on 09/12/2022).

Hsu, I Ching and Chun Cheng (2020). "Integrating machine learning and open data into social Chatbot for filtering information rumor". In: *Journal of Ambient Intelligence and Humanized Computing* 0123456789. ISSN: 1868-5145. DOI: 10.1007/s12652-020-02119-3. URL: https://doi.org/10.1007/s12652-020-02119-3.

Huang, Yue, Paul McCullagh, Norman Black, and Roy Harper (2007). "Feature selection and classification model construction on type 2 diabetic patients' data". In: *Artificial Intelligence in Medicine* 41.3, pp. 251–262. ISSN: 09333657. DOI: 10.1016/j.artmed.2007.07.002.

Iliashenko, Oksana, Zilia Bikkulova, and Alissa Dubgorn (2019). "Opportunities and challenges of artificial intelligence in healthcare". In: *E3S Web of Conferences* 110. ISSN: 22671242. DOI: 10.1051/e3sconf/201911002028.

Imran, Sohail, Tariq Mahmood, Ahsan Morshed, and Timos Sellis (2021). "Big Data Analytics in Healthcare — A Systematic Literature Review and Roadmap for Practical Implementation". In: 8.1, pp. 1–22.

Jain, Priyanshu (2020). *Unsupervised Machine Learning: Validation Techniques - Guavus - Go Decisively*. https://www.guavus.com/technical-blog/unsupervised-machine-learning-validation-techniques/.

Jiang, Fei, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang (2017). "Artificial intelligence in healthcare: Past, present and future". In: *Stroke and Vascular Neurology* 2.4, pp. 230–243. ISSN: 20598696. DOI: 10.1136/svn-2017-000101.

Jin, Huidong, Jin Chen, Hongxing He, Graham J. Williams, Chris Kelman, and Christine M. O'Keefe (2008). "Mining unexpected temporal associations: Applications in detecting adverse drug reactions". In: *IEEE Transactions on Information Technology in Biomedicine* 12.4, pp. 488–500. ISSN: 10897771. DOI: 10.1109/TITB.2007.900808.

Jones, G. and B. Bhanu (July 1999). "Recognition of Articulated and Occluded Objects". In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 24.07, pp. 603–613. ISSN: 1939-3539. DOI: 10.1109/34.777371.

Karaolis, Minas A., Joseph A. Moutiris, Demetra Hadjipanayi, and Constantinos S. Pattichis (2010). "Assessment of the risk factors of coronary heart events based on data mining with decision trees". In: *IEEE Transactions on Information Technology in Biomedicine* 14.3, pp. 559–566. ISSN: 10897771. DOI: 10.1109/TITB.2009.2038906.

Kelly, E and G Stoye (2014). "Does GP Practice Size Matter ? GP Practice Size and the Quality of Primary Care". In:

Keusch, Florian (2015). *Why do people participate in Web surveys? Applying survey participation theory to Internet survey data collection*. Vol. 65. 3. Springer Berlin Heidelberg, pp. 183–216. ISBN: 1130101401. DOI: 10.1007/s11301-014-0111-y. URL: http://dx.doi.org/10.1007/s11301-014-0111-y.

Khan, Atif, John A Doucette, Robin Cohen, and Daniel J Lizotte (2012). "Integrating Machine Learning into a Medical Decision Support System to Address the Problem of Missing Patient Data". In: pp. 12–15. DOI: 10.1109/ICMLA.2012.82.

Krumholz M., Harlan (2014). "Big Data And New Knowledge In Medicine: The Thinking, Training, And Tools Needed For A Learning Health System". In: *Health affairs* 33.7, pp. 1163–1170. ISSN: 0278-2715. DOI: 10.1377/hlthaff.2014.0053. URL: http://search.ebscohost.com/login.aspx?direct=true{\&}db=jlh{\&}AN=2012640987{\&}site=ehost-live.

Lee, Wonsung, Gene Yi, Dain Jung, Minki Kim, and Il Chul Moon (2014). "Network analysis approach to study hospitals' prescription patterns focused on the impact of new healthcare policy". In: *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics* 2014-January.January, pp. 2643–2650. ISSN: 1062922X. DOI: 10.1109/smc.2014.6974326.

Li, Peiyi, Xiaoyu Liu, Elizabeth Mason, Guangyu Hu, Weimin Li, Mohammad S Jalali, and Yongzhao Zhou (2020). "How telemedicine integrated into COVID-19 strategies : case China ' s anti- from a National Referral Center". In: pp. 1–4. DOI: 10.1136/bmjhci-2020-100164.

MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observations*. https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings%20of%20the%20Fifth%20Berkeley%20Symposium%20on%20Mathematical%20Statistics%20and%20Probability,%20Volume%201:%20Statistics/chapter/Some%20methods%20for%20classification%20and%20analysis%20of%20multivariate%20observations/bsmsp/1200512992. (Accessed on 09/09/2021).

Maguire, A, I Douglas, L Smeeth, and M Thompson (2007). "Determinants of cholesterol and triglycerides recording in patients treated with lipid lowering therapy in UK primary care". In: *Pharmacoepidemiology and drug safety* 16.January, pp. 228–228. ISSN: 1053-8569. DOI: 10.1002/pds.

Maillard, Jean-yves, Sally F Bloom, Patrice Courvalin, Sabiha Y Essack, Sumanth Gandra, Charles P Gerba, Joseph R Rubino Ba, and Elizabeth A Scott (2020). "American Journal of Infection Control Reducing antibiotic prescribing and addressing the global problem of antibiotic resistance by targeted hygiene in the home and everyday life settings : A position paper". In: 48, pp. 1090–1099. DOI: `10.1016/j.ajic.2020.04.011`.

Malcolm, William, Ronald A Seaton, Gail Haddock, Linsey Baxter, Sarah Thirlwell, Polly Russell, Lesley Cooper, Anne Thomson, and Jacqueline Sneddon (Dec. 2020). "Impact of the COVID-19 pandemic on community antibiotic prescribing in Scotland". In: *JAC-Antimicrobial Resistance* 2.4. dlaa105. ISSN: 2632-1823. DOI: `10.1093/jacamr/dlaa105`. eprint: `https://academic.oup.com/jacamr/article-pdf/2/4/dlaa105/38275343/dlaa105.pdf`. URL: `https://doi.org/10.1093/jacamr/dlaa105`.

Masters, Karen, Eun Young Oh, Joe Cox, Brooke Simmons, Chris Lintott, Gary Graham, Anita Greenhill, and Kate Holmes (2016). "Science learning via participation in online citizen science". In: *Journal of Science Communication* 15.3. ISSN: 18242049. DOI: `10.22323/2.15030207`. arXiv: `1601.05973`.

MathWorks (2022). *Calinski-Harabasz criterion clustering evaluation object - MATLAB*. `https://www.mathworks.com/help/stats/clustering.evaluation.calinskiharabaszevaluation.html#:~:text=The%20Calinski%2DHarabasz%20criterion%20is,highest%20Calinski%2DHarabasz%20index%20value.`.

Maverick, JB (2021). *What Assumptions Are Made When Conducting a T-Test?* `https://www.investopedia.com/ask/answers/073115/what-assumptions-are-made-when-conducting-ttest.asp#:~:text=The%20common%20assumptions%20made%20when,of%20variance%20in%20standard%20deviation.`.

McLeod, S (2019). *Likert Scale Definition, Examples and Analysis | Simply Psychology*. `https://www.simplypsychology.org/likert-scale.html`. (Accessed on 10/18/2021).

Moher, David, Alessandro Liberati, Jennifer Tetzlaff, Douglas G. Altman, Doug Altman, Gerd Antes, David Atkins, Virginia Barbour, Nick Barrowman, Jesse A. Berlin, Jocalyn Clark, Mike Clarke, Deborah Cook, Roberto D'Amico, Jonathan J. Deeks, P. J. Devereaux, Kay Dickersin, Matthias Egger, Edzard Ernst, Peter C. Gøtzsche, Jeremy Grimshaw, Gordon Guyatt, Julian Higgins, John P.A. Ioannidis, Jos Kleijnen, Tom Lang, Nicola Magrini, David McNamee, Lorenzo Moja, Cynthia Mulrow, Maryann Napoli, Andy Oxman, Bá Pham, Drummond Rennie, Margaret Sampson, Kenneth F. Schulz, Paul G. Shekelle, David Tovey, and Peter Tugwell (2009). "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement". In: *PLoS Medicine* 6.7. ISSN: 15491277. DOI: `10.1371/journal.pmed.1000097`.

Müllner, Daniel (2011). "Modern hierarchical, agglomerative clustering algorithms". In: 1973, pp. 1–29. arXiv: `1109.2378`. URL: `http://arxiv.org/abs/1109.2378`.

NHS, England (2019). *Bronchodilators - NHS*. https://www.nhs.uk/conditions/bronchodilators/#:~:text=Bronchodilators%20are%20a%20type%20of,by%20inflammation%20of%20the%20airways.

NI Department for Infrastructure (2021). *The Travel Survey for Northern Ireland Headline Report 2017-2019 has been published today | Department for Infrastructure*. https://www.infrastructure-ni.gov.uk/news/travel-survey-northern-ireland-headline-report-2017-2019-has-been-published-today. (Accessed on 09/30/2021).

Nimmagadda, Shastri L. and Heinz V. Dreher (2014). "On robust methodologies for managing public health care systems". In: *International Journal of Environmental Research and Public Health* 11.1, pp. 1106–1140. ISSN: 16617827. DOI: 10.3390/ijerph110101106.

Olsson, Thomas (2020). "Challenges and Opportunities in Open Data Collaboration – a focus group study". In: pp. 205–212. DOI: 10.1109/SEAA51224.2020.00044.

O'Reilly, D. and M. Stevenson (2003). "Mental health in Northern Ireland: Have "the Troubles" made it worse?" In: *Journal of Epidemiology and Community Health* 57.7, pp. 488–492. ISSN: 0143005X. DOI: 10.1136/jech.57.7.488.

Panagiotakopoulos, Theodor Chris, Dimitrios Panagiotis Lyras, Miltos Livaditis, Kyriakos N. Sgarbas, George C. Anastassopoulos, and Dimitrios K. Lymberopoulos (2010). "A Contextual Data Mining Approach Toward Assisting the Treatment of Anxiety Disorders". In: *IEEE Transactions on Information Technology in Biomedicine* 14.3, pp. 567–581. DOI: 10.1109/TITB.2009.2038905.

Pezzotti, Patrizio, Piero Borgia, Gabriella Guasticchi, Francesco Chini, and Letizia Orzella (2011). "Can we use the pharmacy data to estimate the prevalence of chronic conditions? a comparison of multiple data sources". In: *BMC Public Health* 11.1. DOI: 10.1186/1471-2458-11-688.

Phillips-Wren, Gloria, Phoebe Sharkey, and Sydney Morss Dy (2008). "Mining lung cancer patient data to assess healthcare resource utilization". In: *Expert Systems with Applications* 35.4, pp. 1611–1619. ISSN: 09574174. DOI: 10.1016/j.eswa.2007.08.076.

Post, Andrew R., Tahsin Kurc, Sharath Cholleti, Jingjing Gao, Xia Lin, William Bornstein, Dedra Cantrell, David Levine, Sam Hohmann, and Joel H. Saltz (2013). "The Analytic Information Warehouse (AIW): A platform for analytics using electronic health record data". In: *Journal of Biomedical Informatics* 46.3, pp. 410–424. ISSN: 15320464. DOI: 10.1016/j.jbi.2013.01.005. URL: http://dx.doi.org/10.1016/j.jbi.2013.01.005.

PracticeIndex (2017). *Appointments to patients ratio: A complicated matter*. https://practiceindex.co.uk/gp/blog/appointments-patients-ratio-complicated-matter/#:~:text=Based%20on%20a%20widely%20accepted,a%20day%20over%20five%20days..

Prokosch, HU and T Ganslandt (2009). "Perspectives for medical informatics. Reusing the electronic medical record for clinical research". In: *Methods of information in* *medicine* 48.1, 38—44. ISSN: 0026-1270. DOI: `10.3414/me9132`. URL: `https://doi.org/10.3414/ME9132`.

Public Health England (2019). *Dependence on prescription medicines linked to deprivation - GOV.UK*. `https://www.gov.uk/government/news/dependence-on-prescription-medicines-linked-to-deprivation`. (Accessed on 09/26/2021).

PyShark (2022). *Dunn Index for K-Means Clustering Evaluation | Python-bloggers*. `https://python-bloggers.com/2022/03/dunn-index-for-k-means-clustering-evaluation/`.

Rao, A. Ravishankar, Subrata Garai, Daniel Clarke, and Soumyabrata Dey (2018). "A system for exploring big data: An iterative k-means searchlight for outlier detection on open health data". In: *Proceedings of the International Joint Conference on Neural Networks* 2018-July, pp. 1–8. DOI: `10.1109/IJCNN.2018.8489448`.

Razavian, Narges, Saul Blecker, Ann Marie Schmidt, Aaron Smith-Mclallen, Somesh Nigam, and David Sontag (2015). "Population-level prediction of type 2 diabetes from claims data and analysis of risk factors". In: *Big Data* 3.4, pp. 277–287. ISSN: 2167647X. DOI: `10.1089/big.2015.0020`.

Reuters (2020). *False claim: patients with respiratory conditions can receive 'rescue packs' from their doctor | Reuters*. `https://www.reuters.com/article/uk-factcheck-rescue-packs-respiratory/false-claim-patients-with-respiratory-conditions-can-receive-rescue-packs-from-their-doctor-idUSKBN21D2UV?edition-redirect=uk`. (Accessed on 09/27/2021).

Rezaei-darzi, Ehsan, Parinaz Mehdipour Id, Mariachiara Di Cesare, Farshad Farzadfar Id, Shadi Rahimzadeh, Lisa Nissen, and Alireza Ahmadvand (2021). "Evaluating equality in prescribing Novel Oral Anticoagulants ( NOACs ) in England : The protocol of a Bayesian small area analysis". In: pp. 1–14. DOI: `10.1371/journal.pone.0246253`. URL: `http://dx.doi.org/10.1371/journal.pone.0246253`.

Rich, Eliot, David F. Andersen, William Augustine, Felipe Cronemberger, Katrina Hull, Luis Luna-Reyes, Roderick Macdonald, Mahdi Najafabadi, Smita Sharma, Carson Tao, James P. Houghton, Jack Homer, and Xu Jianping (2015). "An experimental platform for interpreting open-source health data though integration with dynamic disease models and geoplots". In: *2015 17th International Conference on E-Health Networking, Application and Services, HealthCom 2015*, pp. 97–101. DOI: `10.1109/HealthCom.2015.7454480`.

Richard, Zehang, Li Id, Evaline Xie, Forrest W Crawford Id, Joshua L Warren, Kathryn Mcconnell Id, J Tyler Copple, Tyler Johnson Id, and Gregg S Gonsalves Id (2019). "Suspected heroin-related overdoses incidents in Cincinnati , Ohio : A spatiotemporal analysis". In: pp. 1–15.

Roberts, Adam P. (2020). "Swab and Send: A citizen science, antibiotic discovery project". In: *Future Science OA* 6.6, pp. 10–12. ISSN: 20565623. DOI: `10.2144/fsoa-2020-0053`.

Roelfsema, Chris, Ruth Thurstan, Maria Beger, Christine Dudgeon, Jennifer Loder, Eva Kovacs, Michele Gallo, Jason Flower, K-le Gomez Cabrera, Juan Ortiz, Alexandra Lea, and Diana Kleine (2016). “A Citizen Science Approach : A Detailed Ecological Assessment of Subtropical Reefs at Point Lookout , Australia”. In: pp. 1–20. DOI: 10.1371/journal.pone.0163407.

Royal College of General Practitioners (2019). *Ghost patients 'nothing sinister' – and the insinuation GPs are complicit in fraud is 'shocking', says RCGP*. https://www.rcgp.org.uk/about-us/news/2019/june/ghost-patients-nothing-sinister-and-the-insinuation-gps-are-complicit-in-fraud-is-shocking-says-rcgp.aspx. (Accessed on 09/10/2021).

Sakaeda, Toshiyuki, Kaori Kadoyama, and Yasushi Okuno (2011a). “Statin-associated muscular and renal adverse events: Data mining of the public version of the FDA adverse event reporting system”. In: *PLoS ONE* 6.12, pp. 1–5. ISSN: 19326203. DOI: 10.1371/journal.pone.0028124.

Sakaeda, Toshiyuki, Kaori Kadoyama, Akiko Tamon, and Yasushi Okuno (2011b). “Data mining of the public version of the FDA Adverse Event Reporting System, AERS: Colistin-associated adverse events”. In: *Japanese Journal of Chemotherapy* 59.6, pp. 610–613. ISSN: 13407007.

Santos, R. S., S. M.F. Malheiros, S. Cavalheiro, and J. M.Parente de Oliveira (2013). “A data mining system for providing analytical information on brain tumors to public health decision makers”. In: *Computer Methods and Programs in Biomedicine* 109.3, pp. 269–282. ISSN: 01692607. DOI: 10.1016/j.cmpb.2012.10.010. URL: http://dx.doi.org/10.1016/j.cmpb.2012.10.010.

Scherl, Marcus (2010). “Benchmarking of Cluster Indices”. In: URL: http://epub.ub.uni-muenchen.de/12797/1/DA_Scherl.pdf.

Sculley, D. (2010). “Web-Scale k-Means Clustering”. In: *Proceedings of the 19th International Conference on World Wide Web*. WWW ’10. New York, NY, USA: Association for Computing Machinery, 1177–1178. ISBN: 9781605587998. DOI: 10.1145/1772690.1772862. URL: https://doi.org/10.1145/1772690.1772862.

Senior, Martyn L., Huw Williams, and Gary Higgs (2003). “Morbidity, deprivation and drug prescribing: Factors affecting variations in prescribing between doctors’ practices”. In: *Health and Place* 9.4, pp. 281–289. ISSN: 13538292. DOI: 10.1016/S1353-8292(02)00061-8.

Sharma, A (Jan. 2021). *What is Data Wrangling? Its Tools & 6 Steps of Wrangling | FavTutor*. https://favtutor.com/blogs/data-wrangling. (Accessed on 01/12/2021).

Shin, Su-Jin, Je-Yong Oh, Sungrae Park, Minki Kim, and Il-Chul Moon (2015). “Hierarchical Prescription Pattern Analysis with Symptom Labels”. In: *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pp. 178–187. DOI: 10.1109/ICDMW.2015.138. URL: http://ieeexplore.ieee.org/document/7395669/.

Simpao, Allan F, Luis M Ahumada, Jorge A Gálvez, and Mohamed A Rehman (Apr. 2014). “A review of analytics and clinical informatics in health care”. In: *Journal* *of medical systems* 38.4, p. 45. ISSN: 0148-5598. DOI: `10.1007/s10916-014-0045-x`. URL: `https://doi.org/10.1007/s10916-014-0045-x`.

Simpson, Robert, Kevin R. Page, and David De Roure (2014). "Zooniverse: Observing the World's Largest Citizen Science Platform". In: *Proceedings of the 23rd International Conference on World Wide Web*. WWW '14 Companion. New York, NY, USA: Association for Computing Machinery, 1049–1054. ISBN: 9781450327459. DOI: `10.1145/2567948.2579215`. URL: `https://doi.org/10.1145/2567948.2579215`.

Statistics How To (2021). *T Test (Student's T-Test): Definition and Examples - Statistics How To*. `https://www.statisticshowto.com/probability-and-statistics/t-test/`. (Accessed on 10/01/2021).

Statistics Solutions (2021). *Pearson's Correlation Coefficient - Statistics Solutions*. `https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/pearsons-correlation-coefficient/`. (Accessed on 10/01/2021).

Stobierski, T (Jan. 2021). *Data Wrangling: What It Is & Why It's Important*. `https://online.hbs.edu/blog/post/data-wrangling`. (Accessed on 01/12/2021).

Sunyer, J., C. Spix, P. Quénel, A. Ponce-de León, A. Pönka, T. Barumandzadeh, G. Touloumi, L. Bacharova, B. Wojtyniak, J. Vonk, L. Bisanti, J. Schwartz, and K. Katsouyanni (1997). "Urban air pollution and emergency admissions for asthma in four European cities: The APHEA project". In: *Thorax* 52.9, pp. 760–765. ISSN: 00406376. DOI: `10.1136/thx.52.9.760`.

Temiz, Serdar (2018). "Open data and innovation adoption: lessons from Sweden". PhD thesis. KTH Royal Institute of Technology.

The Pacer Blog (2021). *An easy, beginner training plan to walk from couch to 5k - The Pacer Blog: Walking, Health and Fitness*. `https://blog.mypacer.com/2019/10/25/easy-beginner-training-plan-to-walk-from-couch-to-5k/`. (Accessed on 09/30/2021).

Tsipouras, Markos G, Student Member, Themis P Exarchos, Student Member, Dimitrios I Fotiadis, Senior Member, Anna P Kotsia, Konstantinos V Vakalis, Katerina K Naka, and Lampros K Michalis (2008). "Automated Diagnosis of Coronary Artery Disease Based on Data Mining and Fuzzy Modeling". In: *IEEE Transactions on Information Technology in Biomedicine* 12.4, pp. 447–458. ISSN: 1089-7771.

University of Illinois (2021). *What is Medical Informatics? | Health Informatics Online Masters*. `https://healthinformatics.uic.edu/blog/what-is-medical-informatics/`. (Accessed on 08/26/2021).

Vigo, Markel, Lamiece Hassan, William Vance, Caroline Jay, Andrew Brass, and Sheena Cruickshank (2018). "Britain Breathing: Using the experience sampling method to collect the seasonal allergy symptoms of a country". In: *Journal of the American Medical Informatics Association* 25.1, pp. 88–92. ISSN: 1527974X. DOI: `10.1093/jamia/ocx148`.

Wang, Jingqi, William Christopher Mathews, Huy Anh Pham, Hua Xu, Yaoyun Zhang, and San Diego (2020). "Opioid2FHIR : A system for extracting FHIR- compatible opioid prescriptions from clinical text". In: pp. 1748–1751.

Whitley, Elise and Jonathan Ball (2002). "Statistics review 5: Comparison of means". In: *Critical Care* 6.5, pp. 424–428. ISSN: 13648535. DOI: 10.1186/cc1548.

Wiggins, Andrea and John Wilbanks (2019). "The Rise of Citizen Science in Health and Biomedical Research". In: *The American Journal of Bioethics* 19.8, pp. 3–14. ISSN: 1526-5161. DOI: 10.1080/15265161.2019.1619859. URL: https://doi.org/10.1080/15265161.2019.1619859.

Wikipedia (Aug. 2021). *Whitening transformation - Wikipedia*. https://en.wikipedia.org/wiki/Whitening_transformation.

Yang, Wan Shiou and San Yih Hwang (2006). "A process-mining framework for the detection of healthcare fraud and abuse". In: *Expert Systems with Applications* 31.1, pp. 56–68. ISSN: 09574174. DOI: 10.1016/j.eswa.2005.09.003.

Yousef, Maria Mohammad (2021). "Pr ep rin t no t p ee ev Pr ep rin t no t p ee". In: 13.2.

Yu, Kun Hsing, Andrew L. Beam, and Isaac S. Kohane (2018). "Artificial intelligence in healthcare". In: *Nature Biomedical Engineering* 2.10, pp. 719–731. ISSN: 2157846X. DOI: 10.1038/s41551-018-0305-z. URL: http://dx.doi.org/10.1038/s41551-018-0305-z.

Zhang, Hongxiang and Lizhen Wang (2018). "An information-Theoretic outlier detection method for prescription data". In: *2017 3rd IEEE International Conference on Computer and Communications, ICCC 2017* 2018-Janua, pp. 2361–2365. DOI: 10.1109/CompComm.2017.8322957.

Zhang, Tian, Raghu Ramakrishnan, and Miron Livny (1996). "BIRCH: An Efficient Data Clustering Method for Very Large Databases". In: *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*. SIGMOD '96. New York, NY, USA: Association for Computing Machinery, 103–114. ISBN: 0897917944. DOI: 10.1145/233269.233324. URL: https://doi.org/10.1145/233269.233324.