The era of repurposed data: Exploring health via Data Science and new forms of data.

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by

Alec Edward Davies

November 2019

Acknowledgements

I am extremely grateful for the time, effort and tireless support my primary supervisor Dr Mark Green provided during my PhD studies. Mark enabled me to flourish during my PhD by believing in me and providing valuable advice and critical feedback over the past three years. Mark's door has always been open, and I truly appreciate the time he allocated to discussing research ideas, statistical methods and opportunities for events and personal development. Mark went far beyond expectations involving me in research outside of my PhD, enabling me to develop my experience in collaborative projects.

My appreciation extends to the support of my second supervisor Professor Alex Singleton. If it were not for Alex, I would have never pursued a PhD. Alex also gave a lot of time to me and I would like to thank him for his valuable advice and feedback. In particular, thanks for the support at conferences and for introducing me to your network at various events which proved key in enabling opportunities such as my visit to Toronto.

To all CDRC staff at both University of Liverpool and UCL thank you for the provision of data, infrastructure and administration which were vital in the first two papers of this thesis. I would like to thank all the members of the Geographic Data Science Lab whom I have met and worked with throughout the duration of my studies (especially Ellen, Hai and Kostas). I have learnt so much from the many discussions about Data Science in the last three years, so thanks for your advice and support.

I would like to thank Dr Michael Widener for allowing me to visit SAUSy Lab at University of Toronto, and for his support when I was a visiting researcher. His support went beyond what was expected and I appreciate this greatly.

I would also like to thank Dr Dean Riddlesden for allowing me to undertake a Data Science internship at Walgreens Boots Alliance, facilitating me to further develop my skillset and understanding. I echo these thanks for his advice on methodological approaches.

Without the support of my family I would not have been able to complete this thesis. Thanks for your relentless understanding, for believing in me and for allowing me to pursue my PhD. Mum, Dad, James, Rachel and Buster; I dedicate this work to you.

Finally, thanks to the Economic and Social Research council for the 1+3 Studentship, allowing me to gain a MSc in Geographic Data Science and pursue this PhD. This work was

supported by the Economic and Social Research Council [grant numbers ES/J500094/1, ES/L011840/1]. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. Thanks to the North West Doctoral Training Centre and in particular Hayley Meloy and Dr Julie McColl for the administrative support for RSTG forms, my internship and overseas visit to Toronto.

Ethical approval was granted for this research by the University of Liverpool's Research Ethics Committee (ref #752; ref #4610). The Canada Food Study was reviewed by and received ethics clearance through a University of Waterloo Research Ethics Committee (ORE# 21631). A full description of the Canada Food Study and methods can be found in the Canada Food Study technical report.

 The Canada Food study has been made possible through funding from the Public Health Agency of Canada. The views expressed herein do not necessarily represent the view of the Public Health Agency of Canada. Additional funding for the Canada Food Study has been provided by a PHAC–CIHR Chair in Applied Public Health, which supports David Hammond, staff and students at the University of Waterloo.
Further information regarding data sources, methods and software packages used can be found at: https://github.com/sgadavi3/nfd_self-medication.

Abstract

This thesis explores how new forms of data can help us better understand health outcomes and behaviours. The aims are to examine the contribution of new forms of data within health research; explore how they can be improved by geographical context; and identify applications of Data Science within health research.

Transaction level loyalty card data was acquired from a national high street retailer for England (2012-2014). Analysis explored this high dimensional data by examining the associations of features influencing the purchasing of self-medication products and predicting future purchasing. Results show new insights into self-medication behaviour, such as the difference in purchasing by sex for sun preps and the geographic variations in purchasing (e.g. North-South disparities). Clear seasonality was observed reflecting the climatic drivers of minor ailments. A scalable and accurate machine learning methodology is presented.

Data from the Canada Food Study (396 Canadians, aged 16-30, in five Canadian cities) was used to examine food preparation behaviours. A typology was built (k=10) which displayed potentially problematic food preparation behaviours (e.g. 5% were service food reliant). A measure of time weighted exposure to services (e.g. fast food) was created using respondents GPS trajectories. Findings included exposure to fast food being positively associated, whereas sport and leisure facilities negatively associated, with BMI. A positive relationship was found for the cluster of predominantly service food and BMI. The relationship between increased access to fast food and BMI is further highlighted.

Text summaries (abstracts) for every US obesity-related bill enacted 2001-2017 were used to investigate policy enactment. What is included in obesity policy abstracts, how enactment varies by state and time, and what influences this is considered. A text mining approach included measures of term usage (word and comparison clouds), rarity (TF-IDF), connectedness (Markov chain), and sentiment. Results displayed a childhood focus within policy abstracts (e.g. *school* and *physical activity* are prominent terms). The most populous states enacted the most bills with no clear geographic trend observed. New information was shown, such as the variance by presidential administration is seen (e.g. an initial spike observed during the Obama administration) although modelling suggests limited effects.

The thesis demonstrates how new forms of data offer unique opportunities for insights that traditional data sources are unable to consider. Geographic context provides further value to these data sources. Loyalty card records are shown as valuable in supplementing our understanding of self-medication behaviours, through efficient, cheap and objective purchasing data which could aid data driven population health surveillance and disease monitoring. Analysing obesity-related bill abstracts highlights variance in enactment across the US, bringing unprecedented context via the use of text data. Using sequence analysis and extending investigation beyond resident location via GPS trajectories facilitated new opportunities for understanding dietary behaviours. The findings and Data Science approaches demonstrate the usefulness of these new forms of data which present novel opportunities for greater understanding that is relevant for public health policy and planning.

Table of Contents

List of Figures

List of Tables

Chapter 1 : Introduction

1.1. Background

Health research is more than just the biomedical sciences. Revolutionary insights and contributions have come from outside the field of medicine. The ground-breaking insight that lung cancer is caused by smoking came from an epidemiologist (Sir Richard doll) not biologists or clinicians (Doll and Peto, 1981; The BMJ, 2005). William Farr is another example of contribution from outside the field, where his work in creating a national surveillance system catalysed disease observation and detection (Lilienfeld, 2007). Geographers and public health researchers attempt to understand health behaviours and have gathered and contributed a lot of innovation and knowledge extending beyond the work of biomedicine.

The nature of doing health research is constantly evolving. As Ballester, Michelozzi and Iñiguez (2003) describe, it is typical of public health studies to first consider Hippocrates. In ancient society Hippocrates fundamentally changed medicine from superstition to observation (Kleisiaris, Sfakianakis and Papathanasiou, 2014). 19th century developments are highlighted by the contributions of William Farr using an early example of causal inference for disease outbreak and the first national statistics based health surveillance system (Langmuir, 1976; Eyler, 2001; Lilienfeld, 2007), and John Snow who utilised an early form of GIS to investigate the 1854 London Cholera outbreak, finding the contamination cause by visualising the spatial pattern of infections (Drexler, 2014; Khoury and Ioannidis, 2014). Sir Richard Doll utilised questionnaires in his study of smoking and lung cancer (Doll and Peto, 1981; The BMJ, 2005).

The approaches of the past are, however, not necessarily the only future of research. Considering the examples of Hippocrates, Farr, Snow and Doll it is clear that data, methods and approaches are ever changing. A rapid recent evolution has occurred through new forms of data which are typically big data. For example, health sensors such as smart watches provide greater precision than all of the aforementioned pioneers could ever have imagined. Drexler (2014) also highlights how contemporary technologies (e.g. GIS software) would have reduced Snow's solution time considerably. At present research is at a crossroads where the promise of new forms of (big) data is starting to be achieved and the potential of this data is becoming realised through, as Raghupathi and Raghupathi (2014) highlight, applications of big data in health. New forms of data (and the associated data driven approaches) offer the opportunity to further develop health related research.

This thesis examines how new forms of data and advanced statistical processes (via Data Science and machine learning) can be applied within public health research that enable important health related research questions to be answered.

1.2. Overall research question and aims

The intent of this thesis is to answer the following research question:

> *To what extent can new forms of data help us better understand health outcomes or behaviours?*

Three interlinked aims have been selected to meet this intent:

1. *Examine the contribution of new forms of data to health research.*
2. *Explore how geographical context can supplement and improve the quality of information obtained from new forms of data.*
3. *Identify applications of machine learning that can be applied in health research.*

The first aim defines the overall rationale for this thesis and each of the research papers contained. By studying existing literature, it is possible to understand what new forms of data are and how they fit within the disciplines of Health Geography and Quantitative Geography. The current research climate and existing applications are examined enabling research ideas to be conceived. This in turn enables the selection of research questions, datasets and approaches for each of the quantitative chapters. The discussion in each quantitative chapter also details the specific contribution of each application presented. There is a critique of big data and new forms of data throughout which allows scrutiny of these data sources, and a critical personal reflection of their value from an applied point view within the conclusion.

The second aim considers the importance of geographic information within new forms of data. This will be achieved through three case studies: utilising residential location of customers in retail transaction data; linking time referenced GPS data with survey data; and exploring the dynamics of space within health policy. The exploration of space will primarily be exploratory and the statistical methods employed may not be spatially explicit, however the contribution of space is important as both Tobler's first law details (Tobler, 1970) and the pioneering work of John Snow highlight (Drexler, 2014; Khoury and Ioannidis, 2014).

The third aim will evaluate how machine learning can aid quantitative health (Geography) research. This is achieved by reviewing literature to understand the direction Quantitative

Geography is moving and examining existing applications. Justifications as to why advanced statistical and machine learning methods are appropriate and necessary, and the opportunities they bring, are included in the methods and discussion sections of each quantitative chapter.

To ensure the realisation of these aims a number of objectives are included. Table 1.1. highlights each objective, the aim it focusses upon and the chapter in which it is met.

Table 1.1. Thesis objectives

	Objective	Aim	Chapter
1	Conduct an in-depth review of existing literature.	All	2-6
2	Clearly define new forms of data and big data.	1	2
3	Understand how current new forms of health data being used.	1,2	2
4	Outline the promise new forms of data bring to health research.	1	2
5	Assess current applications of machine learning within health research.	3	2-6
6	Consider the limitations of new forms of data and a data driven approach.	All	2
7	Assess available new forms of data and perform exploratory analysis.	1,2	3-6
8	Select appropriate methods for the analysis of each dataset.	3	3-6
9	Apply and openly document statistical and machine learning models.	3	3-6
10	Analyse model performance using appropriate statistical tests.	3	3-6
11	Consider how results fit within the wider research environment.	2,3	3-6
12	Assess the opportunities each new form of data and the methods bring.	All	3-7
13	Outline possible further research.	All	7

1.3. Structure

Seven chapters are contained within this thesis. Chapter 2 provides an in-depth review of existing literature and discusses the need for, and contribution of, both new forms of data and a Data Science approach within health research. New forms of data and big data are introduced and defined, as well as detailing how they are produced, and how they are linked with other data. The evolution of Quantitative Geography is explored to understand where this approach fits within the discipline and current research environment. Exemplary examples of new forms of data are detailed. The promise of new forms of data (as well as notable applications) within Health Geography are then reviewed bringing an understanding of current methodological approaches and contributions to the field. The challenges of both new forms of data and big data, as well as spatial analysis, are detailed enabling an awareness of possible issues and limitations associated with such applications. Finally, research gaps outline the possibilities and opportunities of this thesis.

Chapter 3 is the first of four research papers included in this thesis and is the first of two chapters utilising loyalty card records. This chapter details how retail transactions linked with loyalty card records (~10 million customers between 2012-2014) acquired from a major high street retailer can be used in the study of self-medication. It focuses upon four medication groups (coughs and colds, hay fever, pain relief and sun preps) and aims to evaluate how this new form of data can be used to help inform our understanding of self-medication in England. An exploration of 50 socioeconomic and health accessibility features perceived (from literature) to impact upon self-medication is performed to understand how these features contribute in explaining self-medication purchasing behaviours. Contained is a scalable transferable machine learning methodology that could easily be applied in further applications. Results demonstrate the usefulness of loyalty card data for producing insights at the national level, and the data address issues that traditional studies are unable to consider and provide new information for groups that are not contained in other data.

Chapter 4 is the second of two papers utilising the same rarely available loyalty card record dataset to explore self-medication behaviours. This chapter considers the opportunities that new forms of data bring to population health surveillance by using objective purchasing behaviour (i.e. loyalty card records) to examine and then predict future self-medication purchasing (for coughs and colds and hay fever products). Analyses used more than 300 features (reduced during model building and optimization) and a state-of-the-art predictive algorithm to predict 17 months of purchasing based on a year of historical data. Examining feature importance then provided further context within these models from the influence of features. Results display new information often missing from self-medication products of how purchasing seasons vary from traditional (or known) ailment seasons and how the influence of features differ between the two product types and temporally. The quality, possibilities and promise of this data source for supplementing our understanding of health are considered as well as the presentation of a scalable, transferable methodology of prediction.

Chapter 5 has the objective of examining food preparation behaviours amongst young adults in Canada (396 individuals in five Canadian cities) by using survey data linked with food logs and GPS trajectories. The relationship between obesity and food preparation, exposure to food environment and demographic measures are also considered. The paper details the application of data mining via sequence analysis and clustering to build a typology of food preparation. This paper extends beyond static studies that only consider residential location-based exposure with the incorporation of GPS trajectories which allow a measure of time

weighted exposure to facilities that affect health (e.g. fast food). Regression models facilitate the examination of the outcomes of cluster membership and BMI. The results of this study highlight potentially problematic behaviours and substantial inequalities within food preparation of young adult Canadians, particularly in regard to access to fast food and BMI.

Chapter 6 is the final research paper contained and utilises the often-overlooked data type: text. Text abstracts of enacted obesity related policy for all US states (and Washington DC) 2001-2017 are used from Centers for Disease Control and Prevention. The chapter examines enacted obesity-related policy abstracts in the US by exploring what is included in obesity policy abstracts, and how policy enactment varies by state and over time. Analyses follow a text mining approach which includes word and comparison clouds, term frequency – inverse document frequency, a Markov chain, and sentiment analysis. Negative binomial mixed effect models were also used to explore factors (e.g. president, state political party, state income medians) that may influence the count of enacted obesity related bills. Unprecedented context is presented via the use of text mining, bringing new information for obesity-related policy with clear potential for expansion of this research further (and in particular internationally).

The final chapter (7) draws this thesis to a close by summarising the main findings as well as addressing how the research question, aims and objectives are met. The robustness of this thesis is ensured through the consideration and discussion of limitations. Future opportunities are highlighted which may advance the novel approaches and findings presented from using new forms of data and a Data Science approach and may address the limitations of this thesis.

1.4. Author contributions and outputs

Chapter 1: Alec Davies wrote this chapter.

Chapter 2: Alec Davies wrote this chapter. A summary was presented at the '*30th European Regional Science Associations Summer School'* in Lesvos Greece, 2017. A condensed and restructured version of parts of this chapter (co-authored with Mark Green) is published in the '*Routledge Handbook of Health Geography*'. Citation:

> Davies, A., & Green, M. (2018). Health Geography and the Big Data Revolution. In: Crooks, V., Andrews, G., & Pearce, J. *Routledge Handbook of Health Geography*. Routledge. p324-330.

Chapter 3: Alec Davies was the lead author; Mark Green and Alex Singleton co-authored. Davies, Green and Singleton conceived and designed the analysis. Davies planned, performed all data cleaning and analysis, and wrote the chapter. Green and Singleton contributed to revisions. A slightly adapted journal paper version of this chapter is published in the *'Machine learning in health and biomedicine'* special issue in the journal '*PLoS One*'. This research was presented at the *'American Association of Geographers Annual Meeting: New Orleans 2018'* conference. Citation:

> Davies, A., Green, M., & Singleton, A. (2018) Using machine learning to investigate self-medication purchasing in England via high street retailer loyalty card data. *PLoS ONE*. 13(11): p1-14.

Chapter 4: Alec Davies was the lead author; Mark Green, Dean Riddlesden and Alex Singleton co-authored. Davies, Green and Singleton conceived and designed the study. Davies planned, performed all data cleaning and analysis, and wrote the chapter. Green, Riddlesden and Singleton contributed revisions to the paper. A slightly adapted journal paper version of this chapter is published in *'Applied Marketing Analytics'*. This research was presented at the *'American Association of Geographers Annual Meeting: Washington DC 2019'* conference and in a colloquium session at the Department of Geography at the University of Toronto, Canada. Citation:

> Davies. A., Green, M., Riddlesden, D., & Singleton, A. (2020) Using loyalty card records and machine learning to understand how self-medication purchasing behaviours vary seasonally in England, 2012–2014. *Applied Marketing Analytics*. 5(4): p354-370.

Chapter 5: Alec Davies was the lead author; Michael Widener, Mark Green, Alex Singleton and David Hammond co-authored. Davies, Widener and Green conceived and designed the study. Davies planned, performed all data cleaning and analysis, and wrote the chapter. Widener, Green, Singleton and Hammond contributed revisions to the paper. This work was resultant from an ESRC Overseas Institutional Visit award (£2000) which funded a visit to the University of Toronto. The work included collaboration between the University of Liverpool and two Canadian universities (Universities of Toronto and Waterloo).

Chapter 6: Alec Davies was the lead author; Mark Green and Alex Singleton co-authored. Davies, Green and Singleton conceived and designed the study. Davies planned, performed all data cleaning and analysis, and wrote the chapter. Green and Singleton contributed revisions to the paper.

Chapter 7: Alec Davies wrote this chapter.

All co-authors have approved the inclusion of each paper within this thesis.

As well as the research feeding directly into the PhD, further outputs have been achieved during this PhD study. Publications include:

Daras, K., Davies, A., Green, M., & Singleton, A. (2018). Developing indicators for measuring health-related features of neighbourhoods. In: Longley, P., Cheshire, J., & Singleton, A. *Consumer Data Research*. UCL Press, London. p167-177.

Daras, K., Green, M., Davies, A., Barr, B., & Singleton, A. (2019) Open data on health-related neighbourhood features in Great Britain. *Scientific Data*. 6(107), p1-10.

Davies, A., Dolega, L. & Arribas-Bel, D. (2019) Buy online collect in-store: Exploring grocery click and collect using a national case study. *International Journal of Retail and Distribution Management*. 47(3), p278-291.

Green, M., Daras, K., Davies, A., Singleton, A., Barr, B. (2018) Developing an openly accessible multi-dimensional small area index of 'Access to Healthy Assets and Hazards' for Great Britain, 2016. *Health and Place,* 54(2018), p11-19.

Presentations include:

GISRUK 2017, University of Manchester: *How does competition affect grocery click and collect performance?*

CDRC Data Partner Forum, Saïd Business School, University of Oxford: *Where are healthy places?*

Additionally, datasets were contributed to for the Consumer Data Research Centre (CDRC); Access to Healthy Assets and Hazards (AHAH).

Chapter 2 : Literature review

This chapter focuses on new forms of data. First, new forms of data and big data are introduced, as well as considering why they are important and how they are produced. It outlines the promise presented from such data and why specifically health research should take note of and utilise these data. Following this, the quantitative subdisciplines of Geography are considered. This context allows understanding for how quantitative applications have evolved in Geography and includes new subdisciplines such as Geographic Data Science which are shaping how data are applied. The promise that new forms of data bring to Health Geography is then explored by considering what Health Geography is. Notable applications of new forms of data within the field are provided. Exploring these applications and existing approaches (e.g. exploratory analysis, data mining and predictive modelling) helped to shape the focus of this thesis and aided in the selection of appropriate methods used in the quantitative chapters. Limitations of both data and analysis are discussed which are important for ensuring that appropriate data and methodological approaches are used, as well as ensuring findings are accurate (e.g. models are interpreted at the correct scale to avoid individualistic and ecological fallacies). Research gaps are then detailed to outline exactly what this research may tackle. Finally, conclusions are drawn.

2.2. New forms of data

2.2.1. What are new forms of data?

In a 2001 keynote at the first European conference of GIS in public health, Löytönen (2001) presented how online services and the evolution of mobile technologies would allow accessible (authorisation and ethics pending) automated real-time geolocated data. This prediction came shortly after the internet was commercialised and years before the first mass market smartphones and wearable technologies. 18-years later, all new smartphones feature built in ‘health’ apps (as well as numerous third-party apps) that quantify a wide array of health phenomena (e.g. sleep tracking and heart rate monitoring) (Pentland, Reid and Heibeck, 2013). Smart wearables have further catalysed this data capture with devices such as smart watches and fitness bands affordable within the mass market.

New forms of data are non-traditional data sources that are collected for purposes other than research (e.g. loyalty card records). Connelly et al (2016) distinguish between made (data collected for specific research purposes) and found (data which may be valuable for researchers but was collected for non-research purposes – alternatively titled repurposed data) data. They are new in a sense that these data have only recently started becoming available to researchers. For example, social media (e.g., Facebook or Twitter) did not exist

10-15 years ago yet the data that they generate have increasingly been used by researchers to understand health-related behaviours (Khoury and Ioannidis, 2014).

As these data are routinely collected by companies (for non-research purposes), collection is typically less intrusive. Society is increasingly linked to a smartphone which has facilitated the monitoring and control of aspects of life. Massive quantities of both physical and social information (e.g. loyalty cards or health tracking) have resulted (Pentland, Reid and Heibeck, 2013). The digital traces attached to such devices have opened up new avenues for research that bring objectivity and ecological validity of everyday behaviour (Pentland, Reid and Heibeck, 2013).

Repurposed data have long been used in Geography. Unemployment data or credit information are common in geodemographics (Harris, Sleight and Webber, 2005; Singleton and Spielman, 2014), and environmental data (e.g. pollution data) have been applied in the study of exposure and ill health (Löytönen, 2014). Integrating such new forms of data has provided vital insights into the trends and identification of at-risk populations (Ginsberg *et al.*, 2009; Raghupathi and Raghupathi, 2014), however newer data such as energy performance certificates are generating further context to such issues. These new forms of data however must be scrutinized with use.

Within geographical research it is common for data to be aggregated to census-based geographies that offer universal scale and are widely familiar (Wise, Haining and Ma, 2001; Duque, Ramos and Suriñach, 2007). Data can easily be captured and aggregated to these geographies, which also allow for confidentially control where results are non-disclosive. Scale is fundamentally limited by the composition of data and licenses which means using census geographies allows for incorporating and linking many datasets. New challenges exist with new types of data (Brunsdon and Singleton, 2015). For example, Goodchild (2010) details how the increased availability of positioning systems (e.g. cell triangulation or GPS) have facilitated advances in real time measurement (e.g. Uber data) but bring new challenges which may include the variance in signal quality from measurement devices (e.g. cheap versus expensive smartphones); the need to extensively clean such data into usable formats (e.g. identifying trajectories from raw GPS data); and the ethics of being able to track individuals movement.

Despite extensive national coverage, traditional censuses are far from comprehensive and are limited in complexity (i.e. less than 50 questions) (Kitchin, 2014; Birkin, Clarke and Clarke, 2017). A decennial cycle means census data can often be out of date before full release. The

scale and information contained within the census is continually important, however applying new data brings opportunity to bridge gaps in data coverage (e.g. the index of multiple deprivation) (Deas et al., 2003). Supplementing new information allows analysis to continue to benefit healthcare planning and complement existing data sources during interim years where coverage is lacking.

2.2.2. Big data

The era of big data has resulted in dramatic changes in research (Miller, 2010; Boyd and Crawford, 2012). Greater computational resources are enabling the use of the vast quantities of data available (Miller, 2010; Kitchin, 2014; Singleton and Arribas-Bel, 2019). Accessing new forms of (big) data are opening up novel opportunities for research. These opportunities can be achieved through data driven methods (Miller and Goodchild, 2015). Researchers are seeking access to these large datasets which facilitate access to previously unobtainable insight (Boyd and Crawford, 2012; Mahrt and Scharkow, 2013). Application has however been limited by a lack of expertise (e.g. programming and big data architecture) that are necessary in the interrogation and extraction of insight from this data (Manovich, 2015). Carefully interpreting this data by utilising expertise can bring valuable insights (van Dijck, 2014).

A lack of processing power and high hardware costs have historically constrained our ability to use big data despite the availability of database management systems (i.e. querying, management and analysis software) (Fry and Sibley, 1976; Scholten and de Lepper, 1991; Longley et al., 2011). Developments in Data Science are enabling the extraction of usable information from data (Murdoch and Detsky, 2013) and now that computational architecture has improved to the necessary standards, analysing big data is increasingly possible (Herland, Khoshgoftaar and Wald, 2014).

2.2.3. Defining big data

There is no widely accepted definition of big data (Herland, Khoshgoftaar and Wald, 2014). Traditionally if data was too large for typical software to handle efficiently then it was referred to as big data; however, constant improvements in computing capacity mean this size is always increasing (Manovich, 2015). While size is inherently important (e.g. data contained within national healthcare systems are predicted to exceed billions of gigabytes (Andreu-Perez et al., 2015)) increased emphasis is being placed on our ability to utilise large data (Boyd and Crawford, 2012; Miller and Goodchild, 2015). Data production speeds have

increased due to automated collection bringing forth further defining characteristics which include speed and diversity (Andreu-Perez et al., 2015).

Kitchin (2014) identifies the "three V's" of big data research. *Volume* (or size) typically receives the majority of attention (Birkin, Clarke and Clarke, 2017). *Velocity* refers to data that is generated (and received) quickly, often real-time, which creates practical usage issues (Kitchin, 2014). For example, transaction data are generated as individuals' purchase products providing a continuous data production. Data are often received in a *variety* of heterogeneous forms, and often completely unstructured (e.g. text data), thereby placing emphasis on manipulating such data (Andreu-Perez et al., 2015).

Boyd and Crawford (2012) instead define big data as a combination of technology, analysis and mythology. They detail how technology enables computation, analysis allows interpretation, and mythology is the belief that big data offer greater insights. Such a universal definition is less technical and better aligned to the philosophy of doing big data research. Andreu-Perez et al (2015) also recommend including value into any definition. There are many small data sets available that can help to answer most research questions. Big data should complement or extend these analyses.

2.2.4. Big data creation

Data are generated in all aspects of daily routines ranging from purchasing food using loyalty and credit cards at a supermarket to visiting a doctor. Many of these data were being generated in the past at similar levels that would be considered big data today, however these data were not being stored or analysed. Large dataset creation has historically been constrained by the costs of imputation and processing (Miller, 2010). Fry and Sibley (1976) estimated the price of labour for an average dataset to be US$1000 in 1976. As such, we have seen national data sets (e.g. censuses) being limited in granularity of both space (through aggregation) and time (e.g. every 10 years) (Kitchin, 2014). Automated data collection is however rendering these issues obsolete.

Velocity, which represents the continuous production of data that has led to extensive detail (Kitchin, 2014), is facilitated via datafication (van Dijck, 2014; Baack, 2015). To reduce the costs of producing data companies are increasingly looking to repurpose automatedly collected data to supplement existing frameworks. This has resulted in an exponential increase in data and has accelerated the transition to datasets that contain more features than observations (Efron and Hastie, 2016). Increasingly a social norm, constant data collection via unobvious methods such as CCTV, smartphones and sensors has allowed new

information to be measured (e.g. interests via social media) (Kitchin, 2014; van Dijck, 2014). Despite the rate and size that data is produced, unstructured data is common. Data is rarely intuitively indexed and often lacks detailed metadata (Mahrt and Scharkow, 2013; Miller and Goodchild, 2015). Nonetheless expertise can overcome a lack of structure and software is increasingly enabling the use of unstructured data (e.g. GPS traces and computer vision).

The creation and utilisation of data has been catalysed by advances in computing and in particular cloudware (Kitchin, 2014). Companies such as Google offer products and services (e.g. Gmail) while retaining the right to mine user data (Andrejevic, 2007). This business model highlights the capabilities of further quantification of behaviour, as access to services are provided in return for user data (Andrejevic, 2007). This highlights why big data is being claimed "the holy grail of behavioural knowledge" (van Dijck, 2014, p199) and the "new oil" (Andreu-Perez et al., 2015, p1204). Datafication is however also linked to open source culture which allows the creation and usage of data away from the commercial "gold rush" (e.g. OpenStreetMap) (van Dijck, 2014; Baack, 2015).

2.2.5. Big data and health

The ability to obtain improved and further knowledge within the many facets of healthcare and public health research is facilitated through big data (Khoury and Ioannidis, 2014; Raghupathi and Raghupathi, 2014). Data driven approaches provide the means for the potential of big data (e.g. automated diagnosis) to be enabled (Herland, Khoshgoftaar and Wald, 2014). While traditional health data sources are rarely the largest in terms of size, the potential offered through data linkage to newer forms of data (e.g. linking survey data to purchasing behaviours through loyalty card records) could offer novel and exciting insights in health-related behaviours (Kayyali, Knott and Kuiken, 2013; Herland, Khoshgoftaar and Wald, 2014).

Medicine has been pioneering in the onus that it has long placed on both information and scientific evidence (Murdoch and Detsky, 2013). Despite this longstanding acknowledgement, healthcare delivery has been slow to utilise the rich information contained within its own infrastructure (Safran et al., 2007). There are significant opportunities for public health to study the success other fields have had in evolving to large dataset application (Kayyali, Knott and Kuiken, 2013; Weber, Mandl and Kohane, 2014), such as transaction data in Retail, or telescope data in Astronomy (Murdoch and Detsky, 2013).

2.2.6. Data linkage

The ability to combine many types of data, which measure a range of phenomena, into a comprehensive database can offer increased value via heterogeneity (Pentland, Reid and Heibeck, 2013; Andreu-Perez *et al.*, 2015). Both structured and unstructured data can be combined to bring further detail (Denny, 2012). While there has been good progress in linking administrative sources (e.g., health records as Pentland, Reid and Heibeck (2013) describe), there have been few examples using commercial data sources. This is due to factors such as difficulties in data linkage, or companies being concerned about potentially sharing competitive advantage. Initiatives such as the Consumer Data Research Centre have pioneered collaborations with industry, however ethical issues remain. As Boyd and Crawford (2012) argue, there is a fine line between accessibility and ethics, which means progress is slow.

2.3. The quantitative subdisciplines of Geography

Science is suggested to be entering a data driven fourth phase (Hey, Tansley and Tolle, 2009; Miller and Goodchild, 2015). A digital revolution is occurring in Geography, whereby digital devices are becoming essential in research, communication and dissemination (Ash, Kitchin and Leszczynski, 2018). Research has shifted to digital form and is finding wider audiences (Kitchin, 2014). Qualitative research is similarly affected as communication has shifted to the digital space (Ash, Kitchin and Leszczynski, 2018), such as social media, video communication and increased use of mobile devices (Lima and Musolesi, 2012).

Geography's interdisciplinary nature is deep routed (Singleton and Arribas-Bel, 2019). The quantitative revolution enabled new avenues of research to develop. Qualitative research is also increasingly embracing the quantities and richness of data and the digital research environment (Miller, 2010; Ash, Kitchin and Leszczynski, 2018). Evolving beyond the roots of Quantitative Geography, data driven subdisciplines have emerged including Geographic Information Science, Geocomputation and most recently Geographic Data Science.

2.3.1. Quantitative Geography

The quantitative revolution within Geography emerged from a need for rigorous spatially explicit theory and models to achieve objective analysis and research with spatial data (Marchand, 1974; Fotheringham, Brunsdon and Charlton, 2000). Evolving from mathematical roots, the aim is "to maximise knowledge on spatial processes with the minimum of error" (Fotheringham, Brunsdon and Charlton, 2000, p4). Spatial data may be present; however, often the methods employed are borrowed from other disciplines (e.g.

Statistics) and are therefore aspatial (Fotheringham, Brunsdon and Charlton, 2000; Wise, Haining and Ma, 2001; Murray, 2010).

At the most basic level Quantitative Geography incorporates Statistics (e.g. measures of central tendency or regression analysis) (Murray, 2010). Methods that were integral to the revolution are continually relevant (e.g. linear regression); however, Geographers were quick to modify these methods to enable spatial attributes to be recognised (e.g. geographically weighted regression) (Fotheringham, Brunsdon and Charlton, 2000). Advanced nonparametric methods (e.g. tree ensembles) have enabled greater measurement complexity and the inclusion of diverse data that bring improved efficiency and performance (Efron and Hastie, 2016).

2.3.2. Geographic Information Systems/ Science

Geographic Information Systems became popular in the 1980s (Goodchild, 2010). As a software for both exploratory analysis and spatial modelling, specialist spatial data manipulation software underpin Geographic Information Systems (Fotheringham, Brunsdon and Charlton, 2000; Murray, 2010; Longley et al., 2011; Haining, 2014). Emerging from developments primarily within computing, Geographic Information Systems have substantial importance to Quantitative Geography, with functionalities ranging from digitization to spatial interaction models (Murray, 2010). Geographic Information Systems can handle diverse data types (e.g. raster or vector data) and are powerful tools for geographic analysis (Fotheringham, Brunsdon and Charlton, 2000; Kistemann, Dangendorf and Schweikart, 2002).

Geographic Information Science instead refers to "the basic research field that seeks to redefine geographic concepts and their use in the context of [Geographic Information Systems]" (Mark, 2000, p48). Early Geographic Information Systems lacked integrated spatial theory because of a commercial focus (Fotheringham, Brunsdon and Charlton, 2000). Decades later computing architecture has shifted to the cloud and has enabled wider software accessibility (e.g. fully interactive web hosted software such as ArcGIS Online or Carto) as powerful workstations are no longer a requirement (Schuurman, 2009). Modern graphical user interfaces are much more intuitive and large tools have developed (e.g. PostGIS) meaning adoption has increased. Geographic visualisation reincarnates the value of images versus text, and is a useful tool in exploring and understanding spatial distributions; however, as visualisation limits detail, data cleaning is vital for visualisations to be accurate (Jacquez, 2014).

2.3.3. Geocomputation

Early Geographic Information Systems were limited in statistical complexity which catalysed the emergence of Geocomputation (Gahegan, 2012). Widespread data availability has created a need for Geographers to learn skills from Computer Science (e.g. programming) to draw geographic insight (Brunsdon and Singleton, 2015). Geocomputation is defined as "the art and science of solving complex spatial problems with computers" (Geocomputation.org, cited in Cheng, Haworth and Manley, 2012, p481). Geocomputation refreshed focus for geographical analysis by providing scalable tools for addressing complex issues (Gahegan, 2012). As a further paradigm of Quantitative Geography, Geocomputation borrows and compiles methods from multiple disciplines (e.g. classification and predictive modelling) (Cheng, Haworth and Manley, 2012). The increasingly digital world has allowed real-world phenomena to be captured and modelled with tools from Geocomputation (Brunsdon and Singleton, 2015). Blurred lines are present though, as it could be argued Geocomputation is just combining more computational resources with Geographic Information Science (Brunsdon and Singleton, 2015).

2.3.4. Geographic Data Science

The desire to utilise massive datasets has shaped modern data analytics (Miller, 2010; Kitchin, 2014; Efron and Hastie, 2016). These massive datasets lay outside the typical sources that social scientists utilise and are allowing new ways to measure phenomena (Singleton and Arribas-Bel, 2019). Traditional methods at the core of Quantitative Geography are continually attractive due to their ease of application and interpretability, however, Data Science techniques are necessary in order to utilise these new forms of (big) data (Cheng, Haworth and Manley, 2012).

Efron and Hastie (2016) recount how the discipline of Statistics was initially theory led due to computational resource limitations, however this barrier is no longer apparent allowing these methods to be utilised. Both the availability of data and the need for state-of-the-art prediction has seen machine learning improve predictive performance (Luo, 2016; Mullainathan and Spiess, 2017). Nonetheless both approaches have a place within quantitative research as "the question shall determine the method" (Elliott, 1999, p240).

A simplistic definition of Data Science is "the science of dealing with data" (Naur, 1974, p397). Naur (1974) notably suggests insights and interpretation of the results from Data Science application are reliant on collaboration with experts from other disciplines. Singleton and Arribas-Bel (2019) elaborate to detail how Data Science is a combination of

techniques and tools, but also a mindset. A key proficiency is the ability to generate insight (Cleveland, 2001; Provost and Fawcett, 2013). Data Scientists have long existed in the private sector (where focus is on monetization and competitive advantage) but have historically been known as data mining or machine learning specialists (Provost and Fawcett, 2013; Varian, 2014; Singleton and Arribas-Bel, 2019).

While big data is often seen as aspatial, underlying geographic dimensions are present providing important context (Arribas-Bel, 2014; Singleton and Arribas-Bel, 2019). For example, retail patronage decisions and the resultant transactions are inherently spatial and are impacted by further geographic factors such as weather. Geographic Data Science allows both geographic theory and Data Science methods to be fused in the pursuit of greater insight (Singleton and Arribas-Bel, 2019). As Singleton and Arribas-Bel (2019) highlight, there are many possible future opportunities applying Data Science methods in Geography. These could include predictive modelling, data mining, clustering, dimensionality reduction or sequence analysis.

2.3.5. Software

Geographic Information Systems offer the means for spatial analysis but are often limited in depth (Gahegan, 2012; Brunsdon and Comber, 2015). Although technological advances have brought considerable developments to software, the industry standard GIS (ESRI ArcInfo) is limited in scope (Haining, 2014). Computing is however developing at an incredible pace. Innovations such as cloud computing have created solid foundations for Geographic Information Systems, opening up analysis to wider audiences (Kistemann, Dangendorf and Schweikart, 2002).

There has been a rise in open source software that are free and constantly updated (e.g. Python, R and QGIS). Programming is increasingly necessary to utilise such data meaning statistical software has adopted the integration of spatial techniques, and spatial software has embedded programming (Jacquez, 2014; Brunsdon and Singleton, 2015). Software such as R and Python have enabled further application of machine learning (Mullainathan and Spiess, 2017). R for example evolved from the statistical language S, where statistical capabilities are deeply rooted (Brunsdon and Comber, 2015). Brunsdon and Singleton (2015) credit both R and Python for the development of Geocomputation. Access to data has also improved. Application programming interfaces (commonly *APIs*) allow for the easy download of data through querying.

Built computing is becoming more efficient through multicore software, parallel computing and GPU processing which have optimised performance. Physical computing power is however becoming less necessary. It is possible to pool computing resources in clusters using software such as h2o, docker and Kubernetes. Powerful virtual machines are also accessible on demand through services such as Amazon Web Services, Microsoft Azure and Google Cloud Platform. Extremely powerful virtual machines that are equivalent to physical machines that would cost many thousands can be rented for increasingly lower prices (a few pounds per hour). Research facilities are increasingly offering access to servers, clusters, and cloud providers to enable the processing of large data.

2.4. New forms of (health) data

Data Science has facilitated the use of the vast data available that comes in a variety of forms (i.e. structured and unstructured data) (Andreu-Perez et al., 2015). Healthcare is largely data driven, however traditionally data has been unstructured and in physical copy form (e.g. paper records) (Raghupathi and Raghupathi, 2014). Technological advances have led to easier data creation and storage. Patient records are increasingly being digitised (i.e. electronic health records) (Safran et al., 2007; Raghupathi and Raghupathi, 2014) and both open and secure data are increasingly available at various level of detail. For example, the NHS provides both open access data (e.g. prescription data at small area geographies) as well as secure data which are restrictive in storage conditions and statistical disclosure (e.g. hospital episode statistics).

Data is typically sourced from health infrastructure, however, new forms of data are increasingly becoming available (Herland, Khoshgoftaar and Wald, 2014) enabling enriched data driven possibilities (Pentland, Reid and Heibeck, 2013). Applications are well established in pandemic management where Data Science methods can provide information for response during disease outbreaks (Andreu-Perez *et al.*, 2015). The most prominent example is Google Flu Trends which predicts outbreaks of influenza from influenza-related internet searches (Dugas et al., 2012), although similar studies have used Twitter data to estimate the prevalence of influenza (Lampos, De Bie and Cristianini, 2010).

While the possibilities are wide ranging for the potential applications of big data in Health Geography, a few exemplary examples are focused on to illustrate feasible opportunities within the field.

2.4.1. Loyalty card data

Loyalty cards offer a lot of potential for health research. The original purpose of loyalty cards was to encourage brand loyalty through incentives (Sharp and Sharp, 1997; Mauri, 2003). Organisations, particularly the grocery sector, soon realised the potential to understand consumer behaviour and shape personalised shopping experiences (e.g. discount coupons and deciding which products to stock in stores). The importance of focusing on existing customers became apparent and communication networks were created (Hart *et al.*, 1999; Wright and Sparks, 1999). Loyalty card schemes (and more recently smartphone loyalty Apps) further developed with innovations in e-commerce. Electronic point of sale technology is now widely implemented (Byrom *et al.*, 2001) and new developments (e.g. contactless payments) have led to further digitisation of currency. Retailers can greater understand patronage via the information resultant of these technologies (Byrom et al., 2001; Felgate et al., 2012).

Commercial datasets provide novel opportunities for health research, however, to date there has been limited usage because of the difficulties associated with access, disclosure control and the protection of commercial advantage. Despite this, researchers are increasingly gaining access to these sources of information. Notable examples of the use of commercial datasets include:

- Silver et al. (2017) used point-of-sale data on purchasing behaviours in supermarkets (15.5 million transactions in 26 stores) to evaluate the impact of the introduction of a sugar sweetened beverage tax in Berkeley, California. Results suggested that a considerable decline in sales occurred where the tax was introduced. These findings from objective purchasing information were novel and important in understanding how successful a similar tax could be elsewhere, as modelled estimates and survey data are typically used. Loyalty card data would have strengthened the quality of the study allowing for a better understanding of changes in behaviours post-intervention, however they still demonstrate the usefulness and potential of consumer data.
- Nevalainen et al. (2018) used loyalty card records to investigate purchasing cycles of the most frequently purchased products in Finland. Findings identified problematic behaviour of high purchasing of beer, cigarettes and soft drinks. Although being limited to only one company in Finland, these findings brought new product exposure information, alongside new context into different shopper behaviours (e.g. personnel versus customers).
- Loyalty card records have been employed in a pioneering application of cancer surveillance. Flanagan et al. (2019) linked loyalty card records with diagnosis history to explore the opportunities of historical medication purchasing for

indicating early disease onset. Despite their study containing only small sample of participants, purchasing of related medication did occur prior to diagnosis demonstrating the potential of data driven health surveillance based off objective purchasing data (Flanagan *et al.*, 2019).

Despite the vast amount of new information packed within loyalty card and transaction level data, such data must be scrutinised before use as data provenance is typically unavailable and often quality is assumed (Wigan and Clarke, 2013). The majority of loyalty card schemes are free to join, and it is a customer's choice whether or not to join a scheme. Brand strength and offerings will strongly determine membership. From a research perspective there is no control on the sample. Membership will vary geographically with the presence of stores, brand awareness and demographics which can lead to bias within the data. Nevalainen *et al.* (2018) describes how store format (e.g. superstores versus local) may influence behaviour which could impact basket size or product selection. Differences in the characteristics of people who hold loyalty cards and those who do not must be acknowledged when analysing such data, particularly if findings may contribute to health policy. Nevalainen *et al.* (2018) detail how an additional step could be to validate findings against existing sources (e.g. against a health outcome survey or clinical data).

While they are an imperfect data source since not all individuals have loyalty cards where they shop (and individuals may have multiple cards) and the population may be biased (e.g. Waitrose shoppers are not representative of the UK population), they can provide objective data on behaviours not available previously. This is important given the under-reporting of behaviours within traditional self-reported data sources (Flegal, 1999; Nevalainen et al., 2018). For example, the majority of research exploring factors associated with over the counter medicine has utilised self-reported survey data which have shown to exhibit bias as individuals may not correctly recall usage (Green et al., 2016). Loyalty card datasets contain considerable detail of objective purchasing behaviour and provide opportunities beyond just retailer commercial advantage (Nevalainen et al., 2018). Despite the presence of representativeness, loyalty card data is a very useful supplement for health outcome research.

2.4.2. GPS data

GPS enabled devices (e.g. smartphones) are constantly producing information of movement (Pentland, Reid and Heibeck, 2013). GPS enabled devices collate information from a number of satellites to provide coordinates of their position. When combined with a timestamp, important contextual detail for the investigation of geographic patterns and the modelling of human behaviour is presented (Löytönen, 2014; Brunsdon and Singleton,

2015). Geolocation can be in real-time (via GPS) but can also occur post event (via combining transactions to store addresses).

The novelty of GPS data is the ability to move beyond residential location within research. GPS enabled mobile technologies detail where we move and where we consume, facilitating an additional focus upon activity spaces and the movement of people (Pentland, Reid and Heibeck, 2013; Widener *et al.*, 2018). Caryl et al. (2019) applied GPS to study children's exposure to tobacco retailers finding higher exposure (in terms of duration and the density of retailers) of children from deprived areas. Cell phone data providing GPS information have also been used in West Africa to track travel patterns and predict where Ebola might spread to after detecting an outbreak in a town or village, allowing public health officials to set up early preventative barriers for containing the spread of the disease (Wesolowski et al., 2014). These examples highlight the contribution of GPS data to study health outcomes.

GPS movement data is increasingly being collected and combined in research with other resources such as smartphone based questionnaires (Pentland, Reid and Heibeck, 2013). The provision of extensive GPS movement data, and the ability to perform data linkage to combine this with comprehensive information from other sources (e.g. food diaries) enables greater depth in behavioural research (Stopher and Greaves, 2007). The use of trajectory data from GPS devices in food behaviour research is increasing (e.g. Chaix *et al.* (2012); Scully *et al.* (2017;2019); Widener *et al.* (2018)). For example, Widener *et al.* (2018) utilised GPS data in the study of activity-based exposure for the creation of individual movement trajectories in the study of food environments. Findings suggested a negative association for exposure to fast food and grocery purchasing, whereas a positive association was found for immediate consumption and exposure to fast food (Widener et al., 2018).

2.4.3. Text data

Text data is a type of unstructured data (i.e. data of unstandardized length and format) (Delgado *et al.*, 2002), and is hypothesized to account for up to 80% of all business related information (Grimes, 2008). Text data is widely generated in health care ranging from clinicians notes to social media posts accounting for a considerable amount of information. Text mining involves the use of specific data mining techniques than can be applied on text data (Dörre, Gerstl and Seiffert, 1999; Delgado *et al.*, 2002). Text data is rich with information however its unstructured nature must be converted into a usable form (Raja *et al.*, 2008). Jensen, Jensen and Brunak (2012) explain how despite clear opportunities within health research text mining has been constrained by both ethics and a shortage in skill.

Electronic Health Records (a "byproduct of routine clinical care" (Denny, 2012, p1)) have been used in numerous studies investigating adverse drug reactions through text analysis (Warrer *et al.*, 2012). Methods vary in complexity from free text searching (basic) to natural language processing (complex) (Warrer *et al.*, 2012). For example Wang *et al.* (2009) studied the prevalence of adverse reactions to medication (e.g. skin rash, fatigue and hypertension) from text within discharge summaries. Utilising a text analysis approach in this context facilitated new and further knowledge to be obtained of the problematic side effects of medication.

Outside of traditional health text data (i.e. electronic health records) new forms of text data are increasingly available. Text data from search engines and social media have been found to be correlated with actual disease incidence (particularly in the study of influenza), highlighting a useful, cheap and fast way to monitor health outcomes (Wilson *et al.*, 2009; Corley *et al.*, 2010; Valdivia *et al.*, 2010). Lee, Agrawal and Choudhary (2013) extended this work to the real-time automated monitoring of both influenza and cancer in the US via Twitter, where geolocation allows this information to be visualised spatially.

Policy documents are a further example of a new form of text data, however, in the limited application to date research has consisted of exploratory research (e.g. Lopez-Zetina, Lee and Friis, 2006) and the impact of individual policies such as healthy eating and physical education (e.g. Cawley and Liu, 2008; Eyler et al., 2012; Lankford et al., 2013). Further applying text analysis poses opportunities for understanding the evolution of policy and decision making within a quantitative framework.

2.4.4. Further examples

New forms of (big) data have been used for further facets of health science. Wearable sensors (e.g. smart watches or heart rate monitors) have become mainstream within society, enabling the ability to accurately and regularly measure important health indicators (e.g. vital signs) and improving the information available to understand health outcomes (Heintzman, 2016; Xie *et al.*, 2018). This benefit extends to patients as these devices can be used to improve chronic disease management (e.g. type 1 diabetes) (Heintzman, 2016).

Mapping services such as Google Street View have also been mined as a new data source for collecting information about neighbourhoods such as aesthetics, location of food outlets and land use (Bethlehem et al., 2014). Fully automating the process has proved difficult but represents a useful research avenue, particularly with new developments in image processing, machine learning and computer vision. Google has also incorporated air quality

sensors onto their street view cars to collect data on levels of Nitric Oxide, Nitrogen Dioxide and Black Carbon (Tuxen-Bethman, 2017), and the data can be requested from Google for research purposes. There is also growing interest in using remote sensing data to develop desk-based audits of features of the physical and built environment (Charreire et al., 2014).

2.5. Health Geography and the promise of new forms of (big) data

Public health research has extended beyond the traditional focus on disease incidence and prevalence to consider further aspects such as healthcare and health behaviours (Rosenberg, 1998; Kearns and Moon, 2002). Health Geography has historically incorporated contemporary human geography themes (e.g. deprivation and wealth inequality) in a wide literature (Asthana et al., 2002) and has developed as a recognized subdiscipline (Rosenberg, 1998; Dummer, 2008).

Early health-related geographical enquiry was categorized into two separate facets of research (Parr, 2002; Dummer, 2008). Medical Geography typically relates to the study of disease incidence, the associated spatial variation and their causes (Parr, 2002; Dummer, 2008). Health Geography was born out of debates involving Medical Geographers and others (e.g. (Mayer and Meade, 1994; Kearns, 1995; Kearns and Moon, 2002)) where it was felt biomedical focus on the determinants of health never fully captured the true role of place on health and wellbeing. Kearns (1995) suggested a Health Geography was better aligned with Social Geography, which was followed by Rosenberg (1998) calling for an inclusive discipline as both subdisciplines suffered similar weaknesses and combining expertise could facilitate greater depth within research.

Extending beyond applied Medical Geography, Health Geography brought acknowledgement that sociodemographic characteristics contribute to health behaviours which enabled the concept of space and place to enrich the research environment (Kearns, 1995). Health research extended to incorporate wider measures of physical and social aspects of society beyond the traditional disease focus (Elliott, 1999). Examples such as the index of *Access to Healthy Assets and Hazards* (AHAH) (Green et al., 2018) combined expertise from both remits as well as from Geographic Data Science to produce valuable insights and a product that is useful to wide audience beyond Health Geography.

The relationship between health and place is a key concept to Health Geography (Crooks et al., 2018). Geography has an intrinsic value to health research due to the presence of an inherent spatial element in health outcomes (Scholten and de Lepper, 1991; Dummer, 2008).

The deep routed spatial element is concentrated on human-environment associations (Dummer, 2008). Being able to visualise disease distributions geographically was a significant contribution that Geography enabled (Kearns, 1995). The incorporation of space ranges beyond basic exploratory analysis and visualisation (i.e. mapping), to advanced theory and modelling.

2.6. Applications of new forms of data in Health Geography

Health Geographers are recognised for thinking appropriately, but also innovatively, around how space and place influence health (Elliott, 2018). These techniques are often borrowed from other disciplines reflecting the multidisciplinary nature of subject area; however, it is fundamental that the technique is suitable for the analysis (Elliott, 1999). This section considers categories of applications in Health Geography that range in methodological complexity (i.e. exploratory analysis to applied machine learning) highlighting the need for research along this complexity spectrum.

2.6.1. Exploratory analyses

Exploratory analysis within health research dates back as far as 1851 when John Snow used mapping to demonstrate the inherent spatial nature of cholera deaths (Drexler, 2014; Khoury and Ioannidis, 2014). Tobler's first law of geography (i.e. '*everything is related to everything else but near things are more related than distant things'* (Tobler, 1970, p236)), emphasizes this inherent spatial component which is important to consider within health research. Application has occurred in studying the spread of disease (e.g. John Snows Cholera map), accessibility to healthcare (e.g. the Index of Access to Healthy Assets and Hazards (Green et al., 2018), or understanding the influences of health behaviours (e.g. the relationship between exposure to tobacco and deprivation (Caryl *et al.*, 2019)).

The relationship between the environment and health-related outcomes are complex and may have multiple causes (Trinca, 2014). For example, there are many determinants of hay fever (e.g. multiple pollen species, allergies or pollution), meaning prediction is notoriously difficult (Davies and Smith, 1973; McInnes *et al.*, 2017; MetOffice, 2018a, 2018c). Spatial investigation can enable the identification of at-risk populations and contextual insight (Haining, 2014). Through the mapping of plant species (e.g. McInnes et al., 2017) knowledge can be added for explaining disease incidence.

A common application has been the development of composite indicators. Providing ranked measures of phenomena, composite indicators are acknowledged as tools that can provide

considerable context in the study of health phenomena and aid policy (OECD, 2008). Data availability has resulted in indices increasing in depth (Deas et al., 2003) as these measures have evolved from the Townsend score which contained only four variables, to indices which contain considerably more information (Jordan, Roderick and Martin, 2004). The 2015 Index of Multiple Deprivation is an example which includes non-traditional information (e.g. road traffic accidents) in the health domain (Smith et al., 2015). Green et al. (2018) instead specifically focused their research on health accessibility creating the Index of Accessibility to Healthy Assets and Hazards. Green and colleagues used retail outlet locations to compute distances for all postcodes (also aggregated to Local Super Output Area) in Great Britain to their nearest amenities such as fast food outlets, pubs or off-licenses. These national level metrics from diverse data sources enable a better understanding of the role of neighbourhood features on health and health-related behaviours without the need for extensive data manipulation.

2.6.2. Data mining

Extracting insight from data using statistical or machine learning methods is known as data mining (Hastie, Tibshirani and Friedman, 2009). Insights can be mined using traditional methods (e.g. linear regression or Ward's hierarchical clustering) or advanced machine learning algorithms (e.g. XGBoost or DBSCAN) (Efron and Hastie, 2016; Luo, 2016). The opportunities of data mining have been recognised with the numerous applications that already exist (Herland, Khoshgoftaar and Wald, 2014). Regression is widely applied, for example sociodemographic and health covariates have been used to infer prescription and self-medication behaviours (Green et al., 2016). In areas where usage has lacked there is a clear need for application (e.g. pharmacovigilance). Wilson, Thabane and Holbrook (2004) detail the need for data mining to detect adverse drug reactions which pose a significant drain on healthcare resources and could be mitigated with greater information.

Social media (e.g. Twitter) offers a novel source of information on human relationships and social interactions; however, application has been limited due to data complexity. Strategies to utilise this data have ranged from randomly sampling tweets (e.g. Eichstaedt et al. (2015)), to classifying the content of posts with machine learning (e.g. Nguyen et al., (2016)). A spatial dimension can be added by combining geotagged information provided in tweets. For example, Nguyen et al., (2016) used machine learning techniques to classify whether individuals were tweeting about fast food or high calorie/energy dense foods finding a positive correlation between the number of fast food outlets and the state level prevalence of obesity. Similar studies have been undertaken using Twitter to measure geographical

patterns in happiness (Gore, Diallo and Padilla, 2015), physical activity (Nguyen et al., 2016), diet (Nguyen et al., 2016; Widener et al., 2018), psychological distress (Eichstaedt et al., 2015) and ailments (Paul and Dredze, 2011). Despite these opportunities, geotagged tweets only represent a <1% subset, and those who geotag tweets are demographically different to those who do not (Sloan and Morgan, 2015).

Further possibilities come in the form of bespoke health geodemographic classifications. Widely applied in marketing (e.g. MOSAIC) geodemographic classifications consider neighbourhoods to carry insights of their residents (Harris, Sleight and Webber, 2005; Abbas, Ojo and Orange, 2009). "Geodemographics are small area classifications that provide summary indicators of the social, economic and demographic characteristics of neighbourhoods" (Adnan *et al.*, 2010, p283). Creation is however computationally intensive as big data creates pressure on clustering algorithms to group data (Adnan et al., 2010). Geodemographic tools pose the opportunity for many applications with public health (Petersen *et al.*, 2011). Health specific composite indices (e.g. the Index of Multiple Deprivation (Smith et al., 2015) and the Index of Access to Healthy Assets and Hazards (Green et al., 2018)) have proven successful in developing area level health measures and knowledge. Geodemographics offer further data mining opportunities to advance small area health knowledge beyond composite indicators. Population profiles substantiate detail within geodemographic classifications and highlight neighbourhood similarities (Abbas, Ojo and Orange, 2009). Applying Geodemographics within further analysis (e.g. predictive modelling) also brings opportunities for further contextual measures of health-related phenomena.

2.6.3. Predictive modelling

Predictive modelling is the process of forecasting outcomes, ranging from statistical analysis (e.g. linear regression) to complex machine learning algorithms (e.g. tree based ensemble methods and deep learning) (Kuhn, 2013; Efron and Hastie, 2016). It is regarded as a vital tool for healthcare as it facilitates the utilisation of big data (e.g. predicting diagnosis or Hospital admissions) (Luo, 2016). Within Geography predictive features (or covariates) are typically socioeconomic or demographic variables (e.g. Orueta *et al.* (2013); Green *et al.* (2016)), although research such as Orueta *et al.* (2013) have shown how further context specific measures can bring greater explanatory power.

Applying machine learning algorithms can bring superior performance (Efron and Hastie, 2016; Luo, 2016). These models provide very accurate and fast predictions and have the ability to handle nonparametric data. Application has, however, been limited in healthcare

due to the black box nature of these algorithms lacking interpretability (versus classical Statistics) as clinicians typically lack the Data Science understanding but also struggle to trust these complex methods (Luo, 2016).

Existing applications of predictive modelling within the field of health are varied. Dekker, Verkerk and Jongen (2000) detail how the vegetable industry has responded to increasing dietary consciousness by attempting to predict nutrient loss in farming to ensure produce meets sufficient nutrition content. Predictive modelling has been applied in disease prediction combining numerous data sources (Pentland, Reid and Heibeck, 2013; Herland, Khoshgoftaar and Wald, 2014; Andreu-Perez et al., 2015). An example is Google Flu Trends, which predicts outbreaks of flu from flu-related internet searches (Dugas et al., 2012). Twitter data has also been used to estimate the likelihood of food poisoning in New York (Sadilek, Brennan and Kautz, 2013). Sadilek and colleagues (2013) present a surveillance tool which aligns with official hygiene inspections. The application matches the locations of restaurants with geotagged tweets in real-time and suggests amount of sick visitors are a key predictor of food hygiene (Sadilek, Brennan and Kautz, 2013).

Predictive risk models are another example. These models identify high risk patients with the aim of better allocating healthcare resources (Panattoni et al., 2011; Bates et al., 2014). Orueta et al. (2013) predicted the expected cost of patients in Spain and their health care consumption in the next year, enabling efficient resource planning. Similarly patient medical histories (via electronic health records) have been used to predict strokes (Nwosu et al., 2019). Panattoni et al. (2011) suggest the NHS have been pioneering in their application of these models, however disclosure issues of patient level predictions can hamper the deployment of these effective planning tools as ethical considerations must be met.

Advanced models can bring more accurate predictions, however error is continually present (Kuhn, 2013). For example, classification problems can result in false positives and negatives. In a health care setting (e.g. does a patient have a disease) this error may result in the wrong or no treatment being given. Regression problems (e.g. predicting the number of people with a disease) can experience similar issues relating to over or underperformance. Google Flu Trends is the most prominent example of predictive error where predicting flu based on search engine behaviour resulted in massive over prediction (Butler, 2013; Lazer et al., 2014).

2.7. Challenges of new forms of data

2.7.1. Challenges of big data

While big data offer many possibilities, there are important limitations. Privacy concerns are increasingly important to the public, but they may be unaware that their data is being used in external research or being linked to other sources. Boyd and Crawford, (2012) describe how despite the opportunities presented, the ability to measure all aspects of life is a "manifestation of big brother" (Boyd and Crawford, 2012, p664). Society has changed because of big data (Lyon, 2014) where the transfer of information in return to access to digital services is now ingrained (van Dijck, 2014). This is problematic as data collection has in many applications evolved to become unconscious mass surveillance (Lyon, 2014). For example smart sensors have been used to measure passive Wi-Fi signals from phones to study changes in high street footfall (Soundararaj, Lugomer and Trasberg, 2019). This can also bring unintended consequences. Lyon (2014) highlights further privacy concerns that occur with long term or permanent data storage of health data that are further impacting society, describing how mental health records can cause restriction on the ability to travel or obtain visas.

As with all research, but particularly when using big data, ethics is important. Big data presents ethical concerns as data may contain potentially identifiable information. In the remit of health this is particularly sensitive. Many datasets are collected automatically for an initial purpose (e.g. transaction data is necessary for business finances), however the full use cases may not be initially apparent, meaning individuals may not have a transparent view of how their data may be used (Lyon, 2014). As health data increasingly moves into digital space new issues arise such as cybersecurity that did not exist with paper health records (Schukat *et al.*, 2016). Providing access to such data brings additional risk of improper use, for example Strandburg (2013) details a breach from the commercial sector where an individual improperly accessed data about a colleague. Within research, ethics are acknowledged by stringent approval processes that must be passed before research is approved to be conducted. Once passed, access to individual level data is rare without confidentially agreements and secure facilities. Such careful design and planning are key for mitigating risk and reducing the potential for unintended consequences (Schukat *et al.*, 2016). As data are typically anonymised and aggregated to larger scales where individuals are distanced and unidentifiable, the individual is often protected but applications of analysis are restricted instead to group level.

"Bigger data are not always better data" (Boyd and Crawford, 2012, p668). Moving away from thinking about the size of data and incorporating Boyd and Crawford's (2012) philosophy is more appropriate. We should not view a difference between big and small data; they are both data. Similarly, just because more data is available and can be easily linked, does not mean we should just do so, as data quality and compatibility issues (e.g. timing) will undoubtedly arise (Wigan and Clarke, 2013). Data should be judged on what it can add to a research question, as well as the quality that it offers (i.e. garbage in equals garbage out). Despite the increasing ability to use many features, every measure should still be justified. Rather than a big data revolution, we should acknowledge the all-data revolution where importance is placed on innovative analytics to better understand the world (Lazer *et al.*, 2014). Taking a (Geographic) Data Science approach avoids many of these issues and focuses on the methods necessary for analysis.

As Khoury and Ioannidis (2014) highlight, in order to avoid issues such as ecological fallacies we must be aware of the noise contained as "big error can plague big data" (Khoury and Ioannidis, 2014, p1054). The complexities found within big data (e.g. nonparametric or correlated features) violate the assumptions of traditional methods and require the use of newer methods (e.g. machine learning). Aggregation is also common which allows the main patterns to be observed, but it is also used because group behaviours are easier to model (in terms of speed and performance). Further difficulty comes as data provenance is often lacking (this may be intellectual property of a data provider, or alternatively just not disclosed) meaning that data quality is assumed to be good (Wigan and Clarke, 2013). In order to achieve insights from big data and contribute useful health knowledge, data quality must be considered and issues (e.g. the noise present within big data) must be addressed (Khoury and Ioannidis, 2014; Lyon, 2014). If an application opts to use big data then it should be unbiased, and results should be validated against some form of ground truth (e.g. comparing results to established health surveys) if available (Nevalainen *et al.*, 2018). This also places emphasis on ensuring the technique employed is carefully selected and suitable for the analysis (Elliott, 1999).

Large sample sizes are often used to improve predictions, however interpretability can be limited when applying black box algorithms, which as Luo (2016) describes, has limited applications of machine learning within clinical practice. As new forms of data are collected for purposes other than research, understanding data is vitally important (Boyd and Crawford, 2012), which can be difficult when metadata is typically lacking. Classic research is hypothesis driven where data collection is focused on answering a specific question. Repurposing data moves us away from this and raises issues of the potential presence of

unintended consequences that can influence behaviour. Examples of how big data can alter behaviour include:

- social media fundamentally aims to monetise behaviour and utilises complex marketing algorithms that filter what people see (van Dijck, 2014);
- customer relationship management promotions can alter behaviour based on criteria such as purchasing cycles and frequency, meaning behaviour that may not represent life 'as-is' (Andrejevic, 2007; Wigan and Clarke, 2013; van Dijck, 2014);
- wearable technologies provide users with health data that was previously unobtainable in real time (e.g. calorie tracking) and the software on these devices rewards increased and sustained participation (Schukat *et al.*, 2016).

It is not the case though that big data will suddenly improve health applications overnight. In 2013, Google Flu Trends overestimated doctor visits for influenza-associated conditions by twice what validated data suggested it should be (Butler, 2013). The same service also underestimated the H1N1 pandemic due to how people were searching for the condition (Cook et al., 2011). It is important not to overestimate the potential of big data by acknowledging the big data hubris; big data should supplement, rather than replace, traditional approaches (Lazer *et al.*, 2014). Understanding data samples and bias is a key consideration if new forms of data or big data are going to be used to inform policy or health in decision making, as despite the large sample sizes, specific groups of people may be unintentionally excluded from analysis. For these reasons it is important to stress the value part of any big data application.

2.7.2. Challenges of the spatial analysis of data

Spatial scale is important as aggregation can alter results and may cause ecological fallacies or modifiable areal unit problem (Fotheringham, Brunsdon and Charlton, 2000). Dependent on the scale (e.g. local or national) the measurement of health outcomes can vary greatly therefore geographic scale should be carefully selected (Dummer, 2008; Jacquez, 2014). Aggregation is common particularly as census geographies are widely accepted and familiar (Wise, Haining and Ma, 2001; Duque, Ramos and Suriñach, 2007). Both individual and ecological studies offer considerable insight to public health knowledge (Subramanian *et al.*, 2009; Idrovo, 2011).

Ecological fallacy (originally discovered by Robinson (1950) when studying illiteracy) is concerned with the difference in insight when the interpretation scale is different to the measurement scale (e.g. if assumptions are made for individuals when an outcome is aggregate) (Fotheringham, Brunsdon and Charlton, 2000). To add further complication the

reverse can be true (e.g. assumptions for the aggregate based on the individual), known as individualistic fallacy (Alker, 1969 in: Subramanian et al., 2009). The research question should define the scale; analysis and interpretation scale must match (Openshaw, 1984a). Ecological level research must be careful that findings are not presented at individual level and vice versa.

The modifiable areal unit problem refers to the aggregation of data to changeable zones (e.g. census geographies could change with urbanisation of a rural area) (Fotheringham, Brunsdon and Charlton, 2000). Aggregation determines this extent as "at the level of the individual field there is no spatial association" (Openshaw, 1984b, p3). Areal geographies are fundamental in the analysis and visualisation of outcomes; however, data are sensitive to their unit of measurement (e.g. population density varies considerably within census geographies) (Openshaw, 1984b; Fotheringham, Brunsdon and Charlton, 2000).

Despite aggregation levels being selected by researchers, increasingly the scale of measurement is becoming determined by consistent metrics, indicators and geodemographic data products (e.g. Lower Super Output Area or Local Authority District are typical UK scales). Interpretation must account for the modifiable nature of these geographies (Openshaw, 1984b). A possible solution is to study at an individual level (i.e. not aggregating) (Fotheringham, Brunsdon and Charlton, 2000) but the ability to visualise or model big data in this way is often inefficient. However, as area level observation is often important, varying spatial scale alternatively allows for the further and important context of outcomes at national, regional and local levels (Dummer, 2008; Subramanian *et al.*, 2009), allowing identification of at risk populations and efficient policy (Hay *et al*., 2005). These solutions are somewhat hampered by data disclosure issues and the need to preserve privacy. Alternatively, spatial models offer an acknowledgement of space (Openshaw, 1984b). These issues must be considered within study and ideally outcomes should be measured at the finest scale possible to minimise aggregation effects, however it is often not possible to meet these requirements.

2.8. Research gaps

"The application of big data to health care is inevitable" (Murdoch and Detsky, 2013, p1352) and is driven by the enormous associated costs of healthcare (Kayyali, Knott and Kuiken, 2013). Murdoch and Detsky (2013) detail how new insights, better information distribution, personalised medicine and self-care are enabled through big data and will improve the healthcare economy. Despite the presence of constraints (e.g. skill shortages, ethics or

computational power), this literature review has demonstrated the clear opportunities and need for the application of big and new forms of data within health.

This thesis falls under the remit of utilising extremely detailed data from various sources, linking them with existing datasets and applying a Data Science approach (i.e. data mining and machine learning methods) to derive new insights. These applications fall in areas where research is lacking but knowledge is necessary. The areas of self-medication and obesity are specifically focused upon due to the extensive drain these areas have upon healthcare (Heikkinen and Järvinen, 2003; Pillay et al., 2010). The opportunities presented come from both fine resolution information of individuals (e.g. over the counter product purchasing or food preparation diaries linked with GPS data) as well as policy related documents.

Fine resolution public health information is vital in determining at risk populations (Hay *et al.*, 2005). For example, data driven approaches have aided the identification of life-threatening complications (e.g. thoracic risk detection) (Andreu-Perez et al., 2015) and evidence based prescribing (Raghupathi and Raghupathi, 2014). Despite this, high data creation and capture costs associated with clinical data have limited the temporal coverage available (Andreu-Perez et al., 2015). A further limitation is that clinical data account for only those who make GP visits, whereas public health extends beyond this into the globally adopted hybridised practice of self-care.

Deregulated low strength medications used in the treatment of minor ailments are widely available. Health-literacy, emergent from the self-care movement, has developed amongst the general population where the use of over the counter medicines (in short term treatment) is high (Magruder, 2003). Existing research that has explored self-medication have utilised self-reported data from health surveys (e.g. Green *et al.*, 2016). Accessing big data in the new form of transactions linked with loyalty card records offer significant opportunities to bring new information to minor ailment prevalence research, and novel insights into self-medication behaviours. For example Flanagan et al. (2019) used a this data as a proof of concept for detecting ovarian cancer earlier. Opportunities are presented for both understanding the drivers of self-medication as well as predicting future purchasing.

Within obesity there are similar opportunities for further knowledge. Large amounts of information from dietary surveys are not new, however focus has typically been on consumption. Knowing where food is prepared is an important consideration of consumption as those who prepare their own food are more likely consume the recommended levels of nutrients (Larson et al., 2006). Eating out has also been found strongly associated with fast

food consumption where higher caloric consumption is consequential (Lachat *et al.*, 2012; An, 2016; Penney *et al.*, 2017). Exposure to such services, associated with low cost and convenience, is a key determinant of obesity (Tremblay and Willms, 2003; Burgoine *et al.*, 2014). The provision of extensive GPS movement data, and the ability to perform data linkage to combine this with comprehensive information from other sources (e.g. food diaries) provides considerable opportunity to build on existing research in this area (e.g. Chaix et al. (2012); Scully et al. (2017;2019); Widener et al. (2018)) and enables for greater depth in behavioural research (Stopher and Greaves, 2007).

There are also opportunities beyond the focus of people and their behaviours. Obesity policy has been considered in numerous papers that include exploratory research (e.g. Lopez-Zetina, Lee and Friis, 2006; Cawley and Liu, 2008; Eyler et al., 2012; Lankford et al., 2013), as well as modelling features that explain policy enactment (Boehmer et al., 2007; Hersey et al., 2010; Donaldson et al., 2015). This research however has often failed to account for the text contained as data. Few studies have analysed the text contents of bills (e.g. Cawley et al., 2008; Lankford et al., 2013), which are largely limited to qualitatively analysing and summarising texts. The opportunity of utilising text data poses great potential for understanding the drivers of obesity policy and how this has changed over time.

2.9. Conclusion

New forms of data have generated a lot of interest across many disciplines. While many challenges remain, a lot of promise is offered by repurposing data. The opportunities don't necessarily only relate to new data, but also the possibilities Data Science brings. Beyond data, this literature review has highlighted the value of a data driven (Geographic Data Science) approach which provides the means of exploring these new datasets. Through machine learning and data mining novel applications are possible and the opportunities and value of these datasets increases vastly. This thesis aims to provide exemplary examples of the possibilities available, addressing the research gaps that have been highlighted within this literature review.

Chapter 3 : Using machine learning to investigate self-medication purchasing in England via high street retailer loyalty card data.

This chapter is the first of four quantitative chapters presented and is the first of two applications using loyalty card records from a major high street retailer. This chapter focuses on how utilising this new form of data can enhance and extend knowledge beyond traditional datasets that are typically employed to study minor ailments. The chapter uses data that has coverage of approximately 20% of the adult population in England and focuses on four medication groups (coughs and colds, hay fever, pain relief and sun preps). As the number of characteristics suggested to impact self-medication is extensive, the machine learning methods employed allow the influence of many features to be compared. New information is presented at the national level with the inclusion of groups that are not contained within other data.

3.1. Introduction

The economic health-care burden of minor ailments (e.g. coughs and colds or sunburn) on the National Health Service is extensive (Pillay *et al.*, 2010). Self-care, a globally adopted movement, empowers patients to take control of their healthcare (World Health Organization, 2000a; Hughes, McElnay and Fleming, 2001; Foley *et al.*, 2015). Self-medication occurs via over the counter medicines used to treat minor ailments. Patients assume a greater health management responsibility as they diagnose and select suitable medical treatment, which can reduce the burden on health care providers; the process is typically hybridised with a combination of health care professionals and most recently online services such as WebMD (Hughes, McElnay and Fleming, 2001).

Traditionally over the counter products were weaker than medicines available through prescription, although stronger medication is increasingly becoming available at pharmacies via deregulation (Keen, 1994; Hughes, McElnay and Fleming, 2001); however, key differences of pack size and cost remain (Bradley and Bond, 1995; Morthorst *et al.*, 2018). Over the counter pharmaceuticals have purchase quantity restrictions and therefore are typically used as cheap short term treatments, whereas longer term treatment may require repeat prescriptions from GPs (Keen, 1994; Bradley and Bond, 1995; Hughes, McElnay and Fleming, 2001). Cost is influential for medication route as some population groups in England are prescription fee exempt (e.g. elderly, pregnant). The costs of prescribing cheap or weak medication has witnessed scrutiny with paracetamol highlighted as a high cost to the

National Health Service (NHS England, 2017). It is possible that social factors such as poverty or income could influence the likelihood of self-medication.

Despite the benefits to the health-care industry, self-care may result in mistreatment of health conditions which could have severe consequences and increased burden (Hughes, McElnay and Fleming, 2001). Consultation of ailments between patients and clinicians may be lacking within self-care dependence (Hughes, McElnay and Fleming, 2001; Foley *et al.*, 2015; Morthorst *et al.*, 2018). Delay of treatment or misdiagnosis, concurrent medication and unrelated medical conditions cause increased risk during self-medication (Bradley and Bond, 1995; Hughes, McElnay and Fleming, 2001). Side effects due to additional health complications (and other behaviours such as alcohol consumption) can be serious particularly if products are not correctly labelled or if patients are not medication literate (Montastruc *et al.*, 1997; Lee *et al.*, 2017; Morthorst *et al.*, 2018). Accidental and purposeful poisoning creates a considerable issue to the NHS with paracetamol related poisonings accounting for 16% of total poisonings (Morthorst *et al.*, 2018). Painkillers are most likely over the counter drugs to be abused (Wazaify *et al.*, 2005). Developing effective population surveillance systems to identify potential harms represents an important yet difficult venture.

The self-care movement has to an extent been fuelled by smart devices and fitness tracking which enable individuals to measure their own health and fitness (e.g. via heart rate monitors or smart watches) (Steinbrook, 2008). Health records are increasingly digitised (Raghupathi and Raghupathi, 2014). Data linkage across health care (e.g. centralised patient records or access to health data from smart devices) would allow practitioners greater awareness of patient medication to reduce the risk of side effects (Sivarajah *et al.*, 2017; Trifirò, Sultana and Bate, 2018). People also now have greater access to healthcare through digital services such WebMD and video appointments, meaning diagnosis is the most accessible it has ever been. As new forms of health data are becoming available, the ability to apply methods to deal with these data is important in allowing this data to mined and further insights created, showing an importance and relevance of applied big data research.

New forms of (big) data are non-traditional data sources collected for purposes other than research (e.g. loyalty card records, social media profiles, smart sensors) and are increasingly available to health researchers. One of these new forms of data, loyalty card records, offers interest to researchers and policy makers. Traditional research that has explored how self-medication behaviours differ throughout the population have only utilised self-reported data from health surveys (Green *et al.*, 2016). Self-reported data has been shown elsewhere to be affected by bias (Green *et al.*, 2016) and objective purchasing behaviours may offer one

solution for minimising such bias. Such data are often 'big' and cover national scales, compared to smaller health surveys that are often localised to smaller regions and therefore have less relevance to the national scale where public health policy decision making is often made. They also offer a less intrusive form of data collection since data are collected routinely by organisations. Real time purchase information for minor ailment medicines may be useful for improving surveillance systems (particularly through data linkage).

The aim of this study is to investigate how high street retailer data can help to inform our understanding of how individuals self-medicate in England.

3.2. Methods

3.2.1. Data

The outcome data explored in this study is transaction records linked to customer loyalty cards provided from a national high street retailer. The primary use of loyalty cards is to increase customer knowledge and thus strengthen retailer loyalty (Byrom *et al.*, 2001). When a customer purchases a product and provides a loyalty card their transaction is logged against their account in return for incentives and promotions. When customers register for a loyalty card, they are asked to provide additional details including age, gender and address.

Data were provided as individual transactions for ~300 categories of products. Upon accessing the data, this was cleaned from 15 million to 10 million customers by age and postcode. The cleaning process was required to account for unrealistic ages (e.g. greater than 100 years) or missing data (e.g. no age provided). The majority of this the reduction in sample size was via the removal of all non-England postcodes. This was due to differences in how prescribed medicines are funded between countries of the UK as well as the availability of predictors of which many are limited to England (e.g. Output Area Classification). Due to the sensitive nature of the data used, we are limited by the sample characteristics that can be reported. Despite the data being representative as it only contains loyalty records for one high street retailer, the large amount of information provided for self-medication purchasing enables detail at a much greater scale than previously available.

Transactions were aggregated by customer and product group to determine whether a customer purchased a product within the two-year period, April 2012 to 2014. The product groups were *coughs and colds* (e.g. cough suppressants, throat lozenges), *hay fever* (e.g. antihistamines), *pain relief* (e.g. paracetamol, ibuprofen) and *sun preps* (e.g. sun lotions). These categories were the lowest level and most detailed aggregation available. This allowed

for a comparison between over the counter medicines whilst maintaining as much detail as possible. Higher aggregations were provided in a hierarchy but using these would mean a loss of self-medication context (e.g. sun preps would be grouped as *toiletries*). This information was aggregated for Local Authority District and Lower Super Output Area using National Statistics Postcode Lookup (Office for National Statistics, 2016), and converted to the proportion of total customers per geography. Local Authority District (n = 326) was used as this is the lowest level allowed to publish data spatially by the data provider; Lower Super Output Area (n = 32844) was used in our analytical models to provide more detailed spatial resolution of our sociodemographic predictors.

We selected a diverse range of sociodemographic explanatory variables to explore how they related to self-medication purchasing patterns (shown table 3.1). These were selected based on previous research that has found that multiple aspects of an individual's social circumstances are associated with their likelihood of consuming self-medicines (Green *et al.*, 2016; Lee *et al.*, 2017). The objective was to utilise many sociodemographic variables as no single variable can best measure any social issue, as well as leveraging the machine learning approach that can handle a large number of features. Explanatory (variables) included Output Area Classification (Gale *et al.*, 2016), Rural Urban Classification (Bibby and Shepherd, 2004), the Index of Multiple Deprivation (Smith *et al.*, 2015) and the Index of Access to Healthy Assets and Hazards (Green *et al.*, 2018). Output Area Classification groups were used to measure population characteristics and were aggregated to Lower Super Output Area level using proportions of each group. Index of Multiple Deprivation score was used to account for deprivation. The Index of Access to Healthy Assets and Hazards was included as it comprises a range of health-related environmental measures such as air quality and accessibility to healthcare (Green *et al.*, 2018). As the methods selected (detailed later) are non-parametric, the models can handle similar or correlated measures meaning features such as the Index of Multiple Deprivation and Index of Access to Health Assets and Hazards can be used in the same model despite some overlap between environmental and access domains.

Table 3.1. Predictors included in machine learning models

Variables	Dataset
Age (median)	High Street Retailer loyalty card data via CDRC
AHAH (components: Dentists; Emergency Departments; Fast Food; Gambling; GP Practices; Greenspace 900m; Leisure; NO2; Off Licences; Pharmacies; PM10; Pubs; SO2; Tobacconists)	CDRC (Access to Healthy Assets and Hazards)
IMD Score	Gov.uk (Index of Multiple Deprivation 2015)
Rural Urban Classification (groups A1-E2)	Nomis (ONS Census Key Statistics)
Output Area Classification (groups 1a-8d)	ONS via CDRC (2011 Output Area Classification Geodata pack)

3.2.2. Statistical analyses

Exploratory analysis was performed to understand national level patterns using the Local Authority District level aggregated data. We calculated the overall distribution of purchasing each product stratified by gender to examine which medicines were most common. We then mapped overall purchasing patterns to explore how behaviours varied geographically.

A machine learning approach was applied to explore important sociodemographic characteristics of purchasing patterns for self-medication products. The rationale was the effectiveness and scalability of these statistical methods in capturing data complexity when utilising large data (Chen and Guestrin, 2016). Non-parametric modelling allows analysis of large numbers of observations and measures that require better predictive models in feasible timeframes; traditional models typically perform weaker in comparison (Efron and Hastie, 2016). Machine learning, in particular tree based models, are nonparametric in nature enabling the exploitation of data and are widely applied and highly effective particularly in ensemble methods (Chen and Guestrin, 2016). Various feature types as well as large feature and sample sizes can be utilised as each feature is treated separately.

Two regression tree ensemble methods were applied. The first, Random Forests, is a tree ensemble method that recursively partitions data in a greedy fashion – constantly improving (Efron and Hastie, 2016). Features are selected automatically from a sample which adds the 'randomness'. The method was selected as the baseline for model performance as it requires little hyperparameter tuning (i.e. the model hyperparameters are relatively optimal to begin with). The randomness prevents model overfitting, and the method is robust to noise as it selects strong complex learners with low bias (Kuhn and Johnson, 2013).

Boosting and in particular Extreme Gradient Boosting (XGBoost) was the second tree ensemble method selected. Boosting combines weak classifiers to produce an ensemble classifier with superior generalised misclassification of error (Kuhn and Johnson, 2013). An overall classifier with superior performance is determined by voting based off an ensemble of iteratively created weak learners; each new tree addresses the errors of its predecessors by reweighting misclassified points (Kuhn and Johnson, 2013; Efron and Hastie, 2016). Hyperparameter tuning enables an algorithm to be optimised and is fundamental to boosting as it can significantly improve predictive performance, however both searching for and tuning parameters bring greater computational complexity (Efron and Hastie, 2016). Hyperparameter tuning is strict towards overfitting. The key difference is Random Forests is focused on reducing variance, whereas XGBoost reduces bias to build a model. XGBoost is used as the method we are most interested in due to the increased performance that comes from hyperparameter tuning, whilst the parallel application allows greater computational complexity in shorter time frames.

The four self-medication product groups were used: *coughs and colds*, *hay fever*, *pain relief* and *sun preps.* Random Forests and XGBoost models for each product class were created. Data for each product contained n = 32844 records. These data were split into 70% training datasets (n = 22993). The remaining 30% (n = 9851) was used as holdout datasets (unseen test datasets) to assess model performance. The unit of analysis are Lower Super Output Areas (n = 32844). This is detailed in Table 3.2.

Random Forests have few hyperparameters to tune, hence the reputation for being a very accurate out of the box learning method. The column subsample (number of features) for each tree was 1/3, and the number of trees (rounds) was constrained to 500 as there was little gain by extending beyond this. The model was utilised with default settings from the randomForest R package (Liaw and Wiener, 2002).

Contrastingly as hyperparameter tuning is very important for optimising XGBoost models. Hyperparameters were found using an aggressive grid search to find the best combination of parameters within a range provided. The grid search included 10-fold cross-validation allowing for optimal hyperparameters to be found for each model (shown Table 3.2). Random Forests computation time was greater than XGBoost; XGBoost uses shallower tree depth and a parallel computing implementation. Model performance is analysed using performance metrics of R^2 and RMSE. Feature importance ranking is used to compare feature selection across model types, and partial dependence plots are used to explore the relationship between the most important features and the outcome variables of proportional product purchase. Despite machine learning algorithms witnessing performance increase,

context is often lost. Partial dependence plots are similar in function to coefficients in OLS regression, allowing for context to be retained (Greenwell, 2017). Partial dependence plots hold all variables constant within the model except the specified variable which is varied across its range. This allows interpretation of how the target variable changes as the specified variable changes, capturing correlations.

Analysis was performed using R (R Core Team, 2014). Random Forests were created in the randomForest R package (Liaw and Wiener, 2002), gradient boosting in the XGBoost R package (Chen *et al.*, 2018) and the data splits, hyper parameter search and model evaluation was performed using the caret R package (Kuhn, 2008). The 'pdp' r package (Greenwell, 2017) was used to explore the marginal effect of the top five ranked features and ggplot2 (Wickham, 2016) was used for visualisations.

3.3. Results

3.3.1. Overall purchasing behaviours

Figure 3.1. shows each of the product group proportion distributions by gender. Pain relief is shown to have the highest proportion of purchasing (median of 65.94%), whereas hay fever the lowest (median of 29.41%). One explanation for why hay fever has lower purchasing than the other products is that the associated condition does not affect the whole population. Pain Relief and coughs and colds (median 65.84% and 58.56%) both have high purchasing proportion due their high availability in England, in part related to how common they are as ailments (Morris, Cantrill and Weiss, 2001; Thielmann et al., 2018). Each of the products have similar distributions for both males and females, suggesting there isn't gender sensitivity within loyalty card customers for these product groups. Sun preps purchasing is the only product with a significant difference in the distribution, with proportions almost double for females (median 29.98 male, 47.01 female).

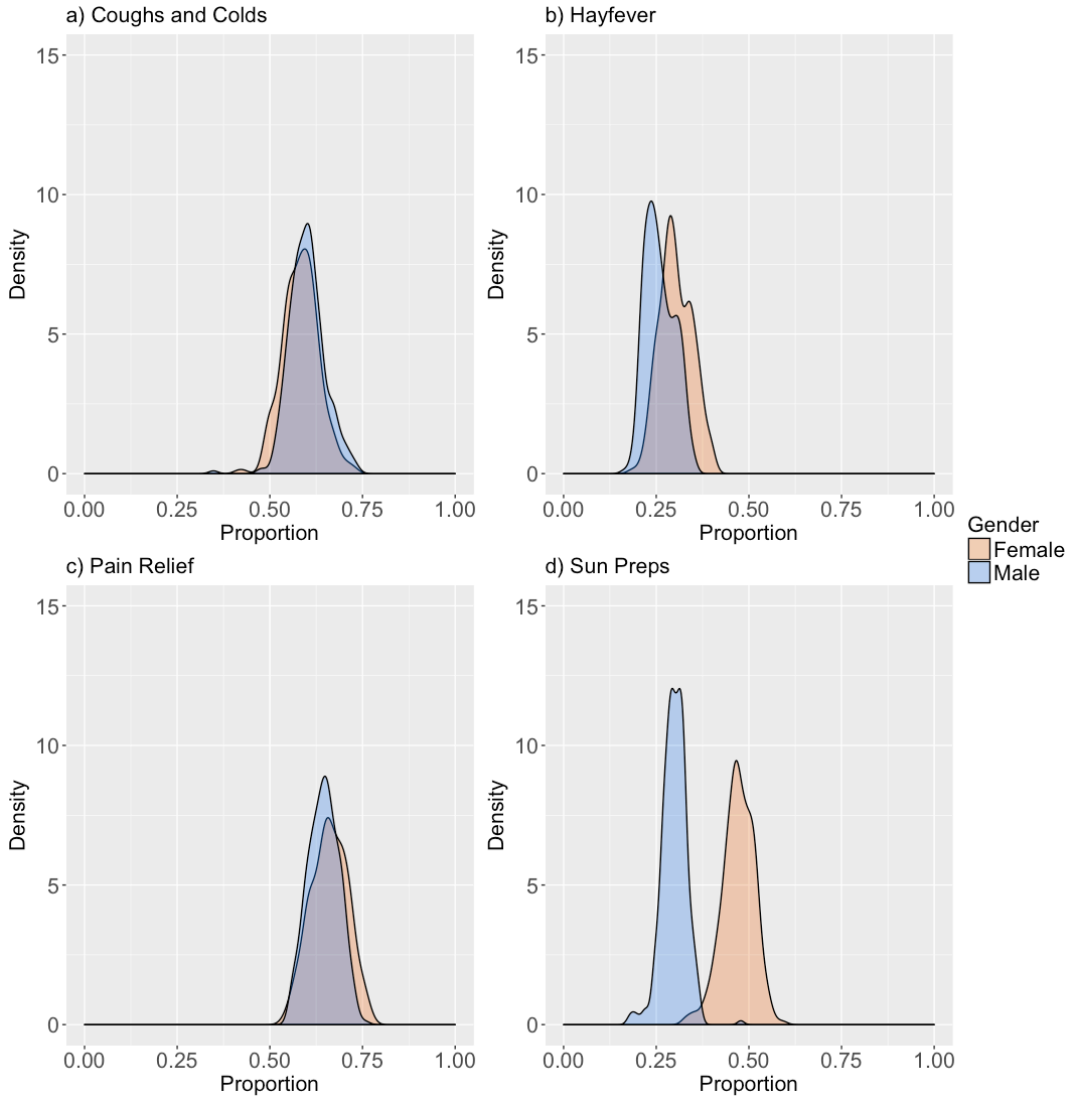

Figure 3.1. Proportion purchasing per Local Authority Level of self-medication products by gender

Figure 3.2. plots the geographic variation in purchasing of each product by quintiles at Local Authority District level. A consistent spatial pattern of higher purchasing in London and the South-East region is observed for each product bar sun preps. For coughs and colds, hay fever and pain relief there are distinct North-South differences with the North-West regions exhibiting lower purchasing. Sun preps exhibit a differing spatial pattern from the other medicines, with urban and central areas displaying higher proportion of sales compared to costal and rural areas (e.g. East Anglia and the South-East).

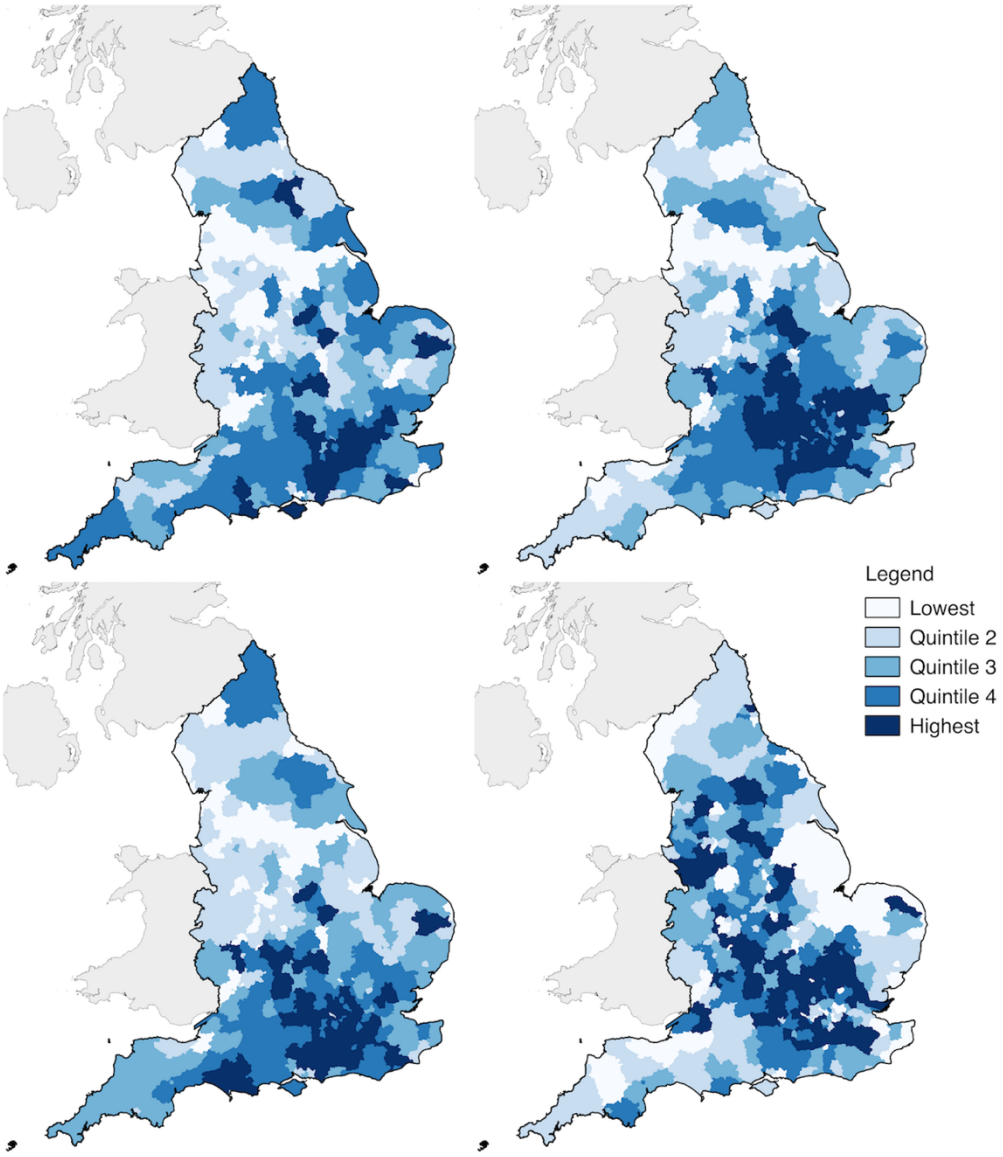

Figure 3.2. Proportion purchasing per Local Authority District of self-medication products (top left coughs and colds, top right hay fever, bottom left pain relief, bottom right sun preps)

3.3.2. Explaining sociodemographic correlates of purchasing behaviours

Table 3.2. shows the performance metrics of Root Mean Squared Error (RMSE) and R^2 for model. XGBoost performs better for both metrics except for coughs and colds where the performance is marginally worse (.002 worse for R^2, .0001 for RMSE). Sun preps has the best predictive performance; however, this product group exhibits the greatest variance between performance metrics with Random Forests performing .0231 worse with for R^2. Despite the poorest performance being for coughs and colds at .5010, there is good predictive performance across all our models. The difference in predictive ability shows that

of the variables included the variance is explained for some products more than others. Further variables may be included if the goal was solely predictive performance.

Table 3.2. Comparison of machine learning model performance

	Coughs and colds		Hay fever		Pain relief		Sun preps	
	Random Forests	XGBoost	Random Forests	XGBoost	Random Forests	XGBoost	Random Forest	XGBoost
Training sample size	70%	70%	70%	70%	70%	70%	70%	70%
Hyper-parameters								
Learning Rate		0.01		0.01		0.01		0.01
Gamma		0		0		0		0
Minimum child weight		1		1		1		1
Column subsample	.33	.7	.33	.7	.33	.7	.33	.7
Row subsample		.8		.8		.8		.8
Maximum depth		6		6		6		6
Rounds	500	5000	500	5000	500	5000	500	5000
Performance								
R2	.5030	.5010	.5881	.5993	.6010	.6063	.6148	.6379
RMSE	.0492	.0493	.0391	.0388	.0427	.0423	.0475	.0460
Run Time (minutes)	10	2	10	2	10	2	10	2

Learning rate = step size shrinkage used to make model conservative; Gamma = minimum loss reduction to make further partition; Minimum child weight = minimum instance weight needed in a child; Maximum depth = maximum depth of a tree (number of splits) (Chen and Guestrin, 2016); RMSE = Root Mean Squared Error

The purpose of this modelling is to investigate which sociodemographic factors are important for predicting purchasing patterns. We focus on the top five most important features from each model as these have the highest influence on overall model performance, with the remaining variables having less impact. The top five variables account for as much as 50% of loss reduction in the models. To visualise feature importance, we use Alluvial plots (an extension of Sankey diagrams) to show how ranks vary between models for each medicine. Figure 3.3. shows the ranks coloured by decile. The highest feature importance is stable for each medicine, showing similar features are consistently important for both methods. There is greater variability seen further down the variable rankings where variables have smaller effects.

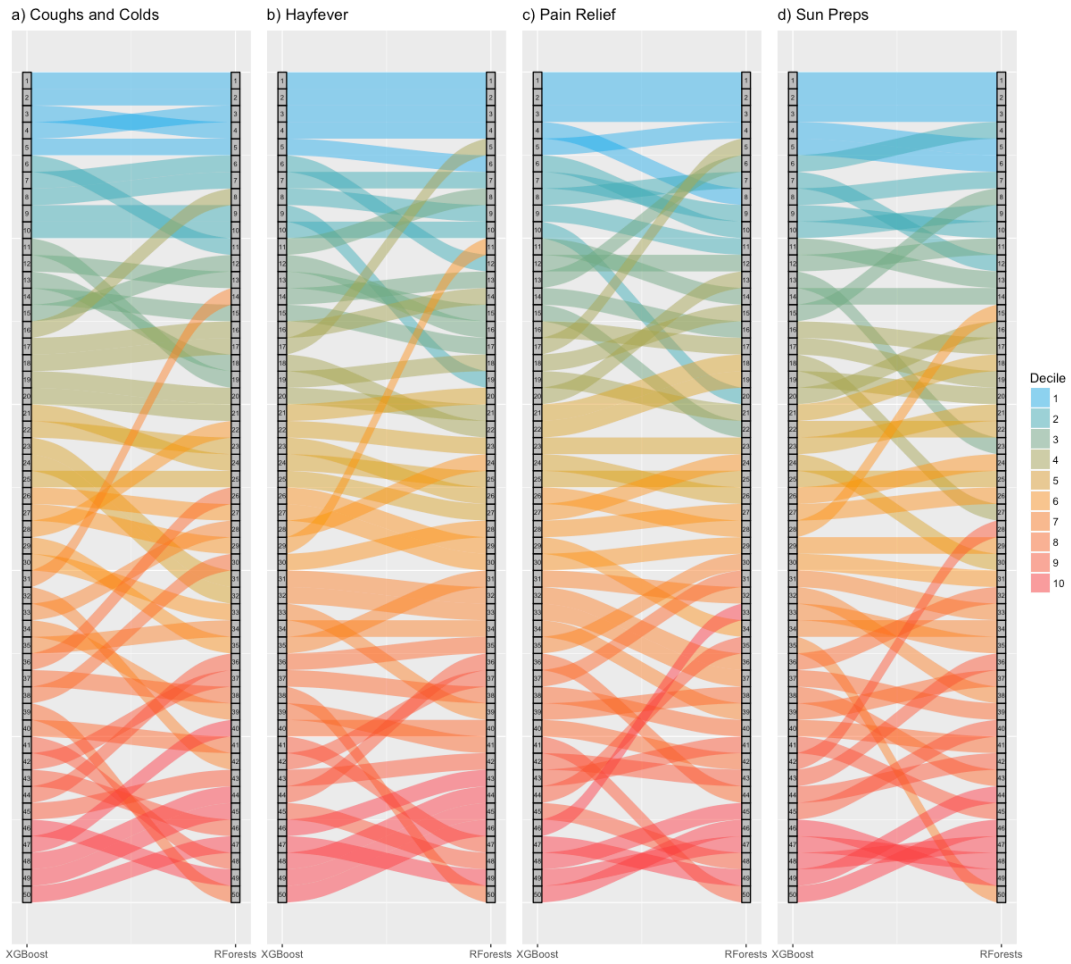

Figure 3.3. Rank comparison of feature importance

(note: 'Decile' refers to the decile of ranks from XGBoost)

Figure 3.4. presents the partial dependence plots for each model of the top five most important variables. A yhat value of 0 (y axis) represents the average proportion of customers. Positive values are interpreted as an increase, negative a decrease from the average value. There are three broad patterns observed. Socioeconomic features were stable and commonly high ranking in each model, particularly the Index of Multiple Deprivation score (9 of 20 occurrences). Areas that had higher Index of Multiple Deprivation scores were negatively associated with purchasing patterns. Air quality variables were also common (10 out of 20). Particulate matter (PM10) and nitrogen dioxide (NO_2) were both positively associated with purchasing patterns for coughs and colds, hay fever and pain relief. Sulphur dioxide (SO_2) was negatively associated with coughs and colds, and hay fever. Age only appeared in the top five once and was negatively associated to sun preps. Across all the models six features rank in the top ten. Eight features rank in the top ten for all models except sun preps, showing consistently important features across all product groups.

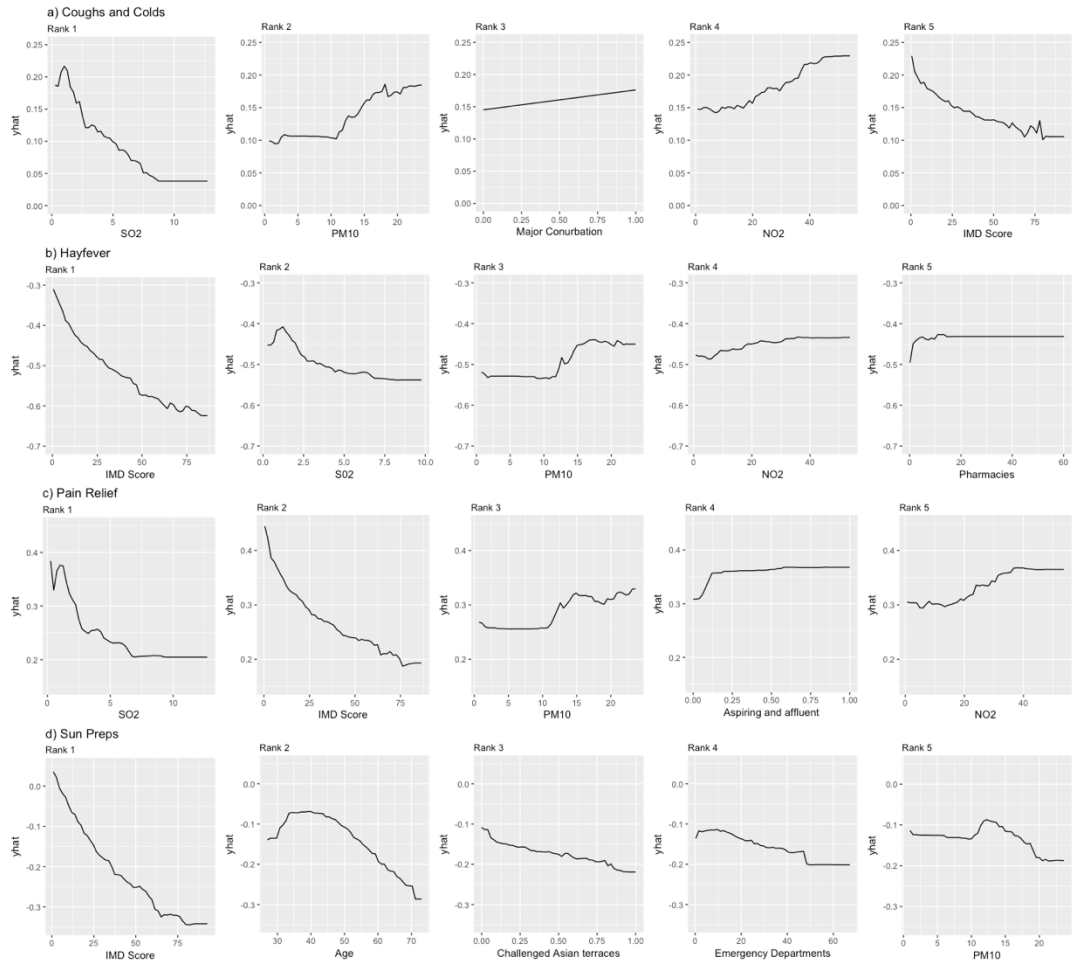

Figure 3.4. Partial dependence plots

(note: top five products from each XGBoost model)

3.4. Discussion

Loyalty card records from a national high street retailer have provided intriguing findings about self-medication patterns in England via novel application of machine learning. The large sample size of national level data on objective behaviours provides new context for customer behaviour of purchasing medicine, building on previous studies that have relied on small self-reported samples from specific regions that may be biased or less applicable to national-level decision making.

Our findings demonstrate that coughs and colds and pain relief medicines both have high proportions of purchasing, representing their common prevalence as minor ailments, with median proportions per Local Authority District above 55%. Sun preps were the least common medication purchased, particularly for males. There are numerous potential explanations for this. One explanation is that females are more likely to be responsible or

informed about the adverse effects of the sun, and therefore engage in protective measures (Miles *et al.*, 2005). Such differences may account for skin cancer rates being higher in males. Targeting males through loyalty card records may offer one approach for tackling such patterns. That being said, sun prep purchasing patterns are far lower than self-reported estimates from other surveys (e.g. Peacey et al. (2006)), which may represent their bias or that individuals purchase sun preps from other locations as well. Another possible explanation is that Sun Preps are solely preparatory whereas the other medicines can serve as response to immediate discomfort. This may influence people purchasing for their household and despite physically purchasing a product they may not actually consume the product, particularly in the instance of families. Surprisingly, we detect little difference between sex for the other medicines which contrasts to the wider literature demonstrating females having higher likelihood of consuming non-prescribed medicines (Figueiras, Caamano and Gestal-Otero, 2000; Green *et al.*, 2016).

We detect considerable geographical inequalities in purchasing patterns for each of our medications. A North-South divide is highlighted, with the distribution of purchasing patterns following the known distribution of socioeconomic measures and in particular poverty/deprivation (Smith *et al.*, 2015). This observation extends to southern population centres clearly highlighted as having the higher proportions of purchasing, and in particular the suburban surrounds of London. Our data offers potential for the geographic targeting of locations to increase self-medication behaviours. Sun preps once again differ in their distribution, with higher purchasing in urban and central regions of England. It is important to note that purchasing behaviours were lowest in coastal regions, which have been found to have higher UV radiation levels compared to inland locations (Kazantzidis *et al.*, 2015). These areas though are also characterised by older populations and given that purchasing behaviours for sun preps declined with age (figure 3.4.) this may also explain our findings. Given the difference in protective behaviours and risk of skin cancers, these represent important areas to target interventions.

Socioeconomic features were consistently shown to be associated with the purchasing of each medicine. Index of Multiple Deprivation score is consistently important in all models, exhibiting a negative association. For pain relief, aspiring and affluent Output Area Classification group has a positive spike between 0 and .1 with a slight positive correlation observed. Challenged Asian terraces Output Area Classification group are shown negatively correlated. The Output Area Classification pen portraits describes a group that exhibits high unemployment and overcrowding (Office for National Statistics, 2014). These findings follow previous research which has found positive associations between higher

socioeconomic status and over the counter medication usage (Figueiras, Caamano and Gestal-Otero, 2000; Green *et al.*, 2016). These associations link to income and education levels associated with such occupations. Individuals with higher levels of income have greater disposable resources that can be invested in purchasing self-medications. Increased educational attainment may also represent greater cognitive resources and therefore greater awareness towards understanding how or the need to self-treat ailments (Lee *et al.*, 2017). The socioeconomic findings, particularly Index of Multiple Deprivation score, show a correlation between deprivation and decreased proportion of purchasing over the counter products.

Age was identified as an important feature in the sun preps model. The partial dependence plot shows a negative association with age. Potential causes are that protection against the sun declines with age, with younger ages representing customers purchasing for dependent others (i.e. mothers protecting their children against sunburn), or lower compliance with medicine guidelines as age increases (Jarrett, Sharp and McLelland, 1993; Lowe *et al.*, 1995). Targeting older individuals who may be at risk of sunburn and skin cancer represents an important focus for policy makers.

Air quality was found to be an important contextual predictor of purchasing behaviours for all products other than sun preps (given there is little causal expectation of such a relationship for sun preps, this was expected). PM10 and NO2 are shown to be positively correlated with purchasing in the coughs and colds, hay fever and pain relief models. This relates to rates being higher in urban areas resultant of transport (Kukkonen *et al.*, 2001; Bealey *et al.*, 2007; Charpin and Caillaud, 2017; Green *et al.*, 2018). PM10 exhibiting high feature importance as well as a positive correlation with increased levels aligns with research that suggests risk of hay fever and also reduced lung function (possibly increasing susceptibility to respiratory issues such as coughs and colds) are associated to traffic-related air pollution (Charpin and Caillaud, 2017). SO2 distribution in the partial dependence plots is unconventional being negatively correlated to purchasing behaviours, then increasing and levelling off. SO2 is considered harmful at high concentrations, and such levels are often found in areas of intense industry which are typically not urban (DEFRA, 2017). Similar to pollution, major conurbations (Rural Urban Classification) exhibits a positive association for cough and colds, possibly linked to the ailments typically being viral.

There are several limitations to our study. The data agreement signed by the high street retailer means that sample characteristics must remain anonymous. This constrains our ability to report on how representative the data are, a necessary component of any research.

Despite the inclusion of 50 features, the study only utilises a select group of variables limiting the exploration to purely socioeconomic and environmental characteristics. Data linkage could identify further knowledge, such as Hospital Admission data or even open prescription data, although information could only be linked at aggregated geographic scale due to the scale that the data is available at. In this study, we consider only whether someone has purchased a product within the 2-year period. Involving seasonality could aid further understanding. This approach could see further data from weather stations involved to see if there are seasonal effect apparent. The limitation of not knowing who the individuals are purchasing for (i.e. themselves or significant others) means that the results are purely based on purchasing and demand side factors. We are also unaware of actual usage of products. Our analyses are also cross-sectional and are limited in their ability to draw inferences about relationships to sociodemographic variables. There are also ecological fallacies and inferences about how they apply towards understanding individual-level relationships that should be avoided.

3.5. Conclusion

This research utilises big data giving an understanding of large sample purchasing behaviour. The data contains close to 20% of the adult population in England, far larger than any previous self-medication study. The data driven approach using loyalty card data allows for actual purchasing behaviour captured within the data, allowing unprecedented context within data. This approach is a novel contribution to current self-care debate, hopefully allowing for further research expanding on the findings.

Chapter 4 : Using loyalty card records and machine learning to understand how self-medication purchasing behaviours vary seasonally in England, 2012-2014

This chapter uses the same loyalty card dataset from chapter 3 and builds on the opportunities suggested in the discussion. Two product groups are focused upon (*coughs and colds* and *hay fever*) due the known seasonality of the associated ailments. This chapter considers the opportunities for using new forms of data in population health surveillance and builds on existing approaches (e.g. the use of social media or search engine data) by providing objective purchasing information. As more than 300 features were originally available, a data driven methodology allowed the reduction of features to create accurate predictive models to predict 17-months of future self-medication product purchasing. Results offer new detail into the temporality of self-medication purchasing for these products and highlights the promise of both this data and approach in population health surveillance.

4.1. Introduction

Fine resolution public health information is vital for determining at-risk populations (Hay *et al.*, 2005). Data driven applications are improving health surveillance frameworks (increasingly in real-time) which have proven successful in pursuit of discovering such at-risk populations (Ginsberg *et al.*, 2009; Raghupathi and Raghupathi, 2014). Identification of potentially life threatening complications (e.g. thoracic aortic dissection) (Andreu-Perez *et al.*, 2015) and evidence-based prescribing (Raghupathi and Raghupathi, 2014) are possible when deploying a data driven approach to medicine. Despite this, clinical based data lack temporality and have high associated creation and collection costs (Andreu-Perez *et al.*, 2015).

Repurposing data from non-traditional sources (e.g. over the counter medicine transactions) are improving how we approach public health (Davies, Green and Singleton, 2018). These new forms of data are collected automatically (e.g. real-time transactions) and have allowed new approaches to healthcare surveillance through the utilisation of big data (Ginsberg *et al.*, 2009). Successful applications include using search engine data to predict influenza outbreaks (Google Flu Trends) (Cook *et al.*, 2011), social media (e.g. Twitter data) to track post-earthquake Cholera outbreaks in Haiti (St Louis and Zorlu, 2012), and loyalty card data to explore self-medication purchasing (Davies, Green and Singleton, 2018). These surrogate sources contribute superior speed and detail, providing a framework for fast estimates, inferences and early detection of disease (Magruder, 2003; Butler, 2013; Olson *et al.*, 2013;

Raghupathi and Raghupathi, 2014; Santillana *et al.*, 2014), and have been found correlated to actual disease data (Valdivia *et al.*, 2010).

Transactions linked with loyalty card data create a significant opportunity to improve knowledge regarding the prevalence and seasonality of minor ailment prevalence via over the counter medication purchasing information. These data do not only offer potential for merely monitoring prevalence but also in understanding the drivers of self-medication behaviours and predicting future behaviour. Self-medication offers a significant benefit to reducing the healthcare burden of minor ailments (Heikkinen and Järvinen, 2003; Pillay *et al.*, 2010), however as most of this information is held within industry, access is rare. Existing research has explored associations between primarily socioeconomic features and both prescription and over the counter medicines (e.g. Green *et al.* (2016)), where surveys are a common data source. Temporality within over the counter purchasing has been considered (e.g. Magruder (2003) and Magruder *et al.* (2004)) although applications have largely been exploratory and few applications have had access to loyalty information (e.g. Davies, Green and Singleton, 2018 or Nevalainen *et al.*, 2018). There is a research gap for utilising this real-time objective purchasing information for both understanding and predicting self-medication behaviours temporally.

Health-literacy, emergent from the self-care movement, has developed amongst the general population where over the counter medicine usage is high (Magruder, 2003). Sales of these medicines have been found highly correlated with physician records whilst reaching wider audiences than prescriptions (Magruder, 2003). Insights of purchasing behaviour are important for understanding the prevalence of over the counter medication which can infer the extent of ailments. Alternatively, this information could be used in a preventative framework to identify at-risk populations based on over-the-counter purchasing behaviours, which could aid clinicians in addressing issues such as self-medication dependence, misdiagnosis and concurrent medication (Bradley and Bond, 1995; Hughes, McElnay and Fleming, 2001). Accessing over the counter transaction data offers an opportunity for novel insights into self-medication behaviours, and the possibility of knowledge for future disease trends. The combination of transactions with anonymised loyalty information address the issues seen in other data (e.g. aggregation (Ginsberg *et al.*, 2009) or self-reporting bias (Green *et al.*, 2016)), allowing accurate information retention.

The aim of this study is to utilise loyalty card records to understand self-medication behaviours; explore how this varies over time and the drivers of these trends; and highlight opportunity for using them to predict future purchasing.

4.2. Methods

4.2.1. Data

We used anonymised transaction records linked to customer loyalty records from a national high street retailer, 2012 to 2014. Data are automatically collected and combined with loyalty accounts when a customer presents a loyalty card during transaction. Data contained anonymised individual level transactions for ~15 million customers grouped into ~300 categories. Data cleaning removed unrealistic (e.g. ages below 18 and above 100), missing values, and customers from outside of England. Data were constrained to England as prescription practices vary throughout the constituent countries of the UK.

We selected two outcomes – *hay fever* and *coughs and colds*. These minor ailments were chosen because they were identifiable within the high street retailer's hierarchical product categories. Both ailments are associated with commonly self-treated conditions and provide contrasting seasonal patterns. Other medicines categories were less distinct in the hierarchy are therefore excluded. We opted to use the finest level of detail available (lowest hierarchy groups) to avoid loss of context. We aggregated transactions to Lower Super Output Area Level which are administrative areas containing a mean population of ~1500 people (n = 32843, excluding the Isles of Scilly) (Office for National Statistics, 2016). Aggregated values were the proportion of customers purchasing each outcome per month.

A data driven approach was taken for the selection of explanatory variables (detailed later). We included any predictor available that had been demonstrated in the literature to be associated with self-medication or health behaviours, resulting in an initial count of ~300 predictors.

Environmental predictors of weather (Robinson *et al.*, 2017) and yearly pollution data (Brookes *et al.*, 2016) were aggregated from a national coverage raster grid (1x1km) to produce monthly LSOA averages. These data sources (CHESS and DEFRA) were selected as they are openly downloadable and useable for research, providing accurate modelled national coverage raster information. The outcomes are inherently seasonal therefore the influence of the weather and environment is an important consideration. Research suggests an environmental influence for these ailments (e.g. air quality and rhinitis) (Charpin and Caillaud, 2017).

Accessibility measures included predictors from the Index of Access to Healthy Assets and Hazards; a comprehensive data resource measuring contextual and geographical features related to health (e.g. air quality, green space) and overall index combining all measures (Green *et al.*, 2018). Air quality measures (e.g. SO2, PM10) are cited as causes of both ailments and have previously been identified as predictive features (Hajat *et al.*, 2001; Heikkinen and Järvinen, 2003; Davies, Green and Singleton, 2018). Individual measures of accessibility to pharmacies and GPs (from the Index of Access to Healthy Assets and Hazards) were included as a proxy for healthcare access. Physician diagnosis have previously been found correlated with over the counter medication sales (Magruder, 2003).

Socioeconomic status has previously been found to influence self-medication usage, where higher status has led to increased over the counter medication usage (Green *et al.*, 2016). The Index of Multiple Deprivation (Smith *et al.*, 2015) was used as a proxy for neighbourhood deprivation. The Output Area Classification (Gale *et al.*, 2016) was selected to measure the demographic characteristics of neighbourhoods (included as a proportion of Lower Super Output Area per group) and has been found a predictor of over the counter medication purchasing (Davies, Green and Singleton, 2018). Rural Urban Classification (Bibby and Shepherd, 2004) is utilised as a proxy for the effects of living environments, particularly as exposure (e.g. to viruses (PM2.5) and dust (PM10) (Charpin and Caillaud, 2017)) varies considerably within different environments. The sources of the socioeconomic variables can be seen in table 3.1.

Finally, we utilised information from the high street retailer data (including median age of loyalty card holders and previous sales). When predicting sales, historical purchasing features have been found important (Žylius, Simutis and Vaitkus, 2015). Previous month product and related product transactions were aggregated using the same method as the outcome for use as predictors (e.g. tissues, pain relief). Further product information including total product sales values were also included.

4.2.2. Statistical analyses

Machine learning models (e.g. tree ensembles) have been demonstrated to perform better than commonly used time-series methods (e.g. ARIMA) and are more flexible in dealing with large numbers of predictors (Adamowski *et al.*, 2012; Žylius, Simutis and Vaitkus, 2015; Pavlyshenko, 2016). Tree Ensembles are commonly applied in prediction and bring superior performance for complex nonparametric data (Chen and Guestrin, 2016). Extreme Gradient Boosting (XGBoost) is a scalable parallel implementation that combines weak learners to create superior leaners, using regularisation to minimise overfitting (Chen and

Guestrin, 2016). Possessing better efficiency and speed above other algorithms, XGBoost has been shown to outperform SVMs and Random Forests (Ogutu, Piepho and Schulz-Streeck, 2011).

A monthly forecasting approach is selected allowing a detailed temporal resolution whilst keeping model computation feasible. Training data contain a year's worth of monthly observations per Local Super Output Area. 10-fold cross validation and a 70-30 train-test data split were used in parameter tuning. Initial models were *static*, trained on month 1-12 (April 2012 to May 2013) and used to predict to month 13-27 (April 2013 to September 2014). A *dynamic* approach was then employed, retraining the model in a moving 12-month window producing a separate model per month to predict months 13-27. Dynamic retraining approaches have been observed to improve performance (Santillana *et al.*, 2014). The comparison of modelling allows evaluation of the accuracy of predicting 17 months in advance and scrutiny of how the models change with the inclusion of updated information (key for evaluating their potential in population health surveillance).

Initially ~300 predictors were available. Following a backward feature selection approach (including correlation, variance inflation factor and feature importance analysis) features were reduced to 40 for coughs and colds and 43 for hay fever. Performance increased with feature reduction suggesting more complexity is not necessarily better (Lazer *et al.*, 2014). Further engineered features included temporal information (month), and seasonality measures of typical seasons for coughs (Autumn to Winter (Heikkinen and Järvinen, 2003)) and hay fever (Spring to Summer (MetOffice, 2018c)) which improved performance. Hyperparameters and features were kept constant to aid comparison of models.

Analysis was performed in R (R Core Team, 2014). Modelling was performed in the XGBoost (Chen *et al.*, 2018), caret (Kuhn, 2008) and ALEPlot packages (Apley, 2018) and visualisations made using ggplot2 (Wickham, 2016).

4.3. Results

Monthly purchasing is more common for coughs and colds than hay fever medicines (1.7%-6.3%, and 0.5%-3.4%, respectively) (figure 4.1.). Monthly proportions are considerably smaller than the total proportion of customers purchasing products throughout the whole time period (58.6% and 29.4% respectively).

Seasonality is observed for both medicines however hay fever seasons are more clearly defined. Coughs and colds proportions rise through Winter peaking in December (6.3% in 2012 and 5.5% in 2013). A summer trough is observed with the lowest proportions in June to August (figure 4.1.). Contrastingly hay fever demonstrates a clear Autumn to Winter off-season. The highest proportions of hay fever are observed March to September (maximum July 2013 (3.3%) and June 2014 (3.4%)). Summer 2012 exhibits a lower peak at 2.5%, however, this was the coldest June for two decades (MetOffice, 2013). The interquartile range is greater for coughs and colds suggesting more variance nationally (possibly as hay fever is distinctly seasonal).

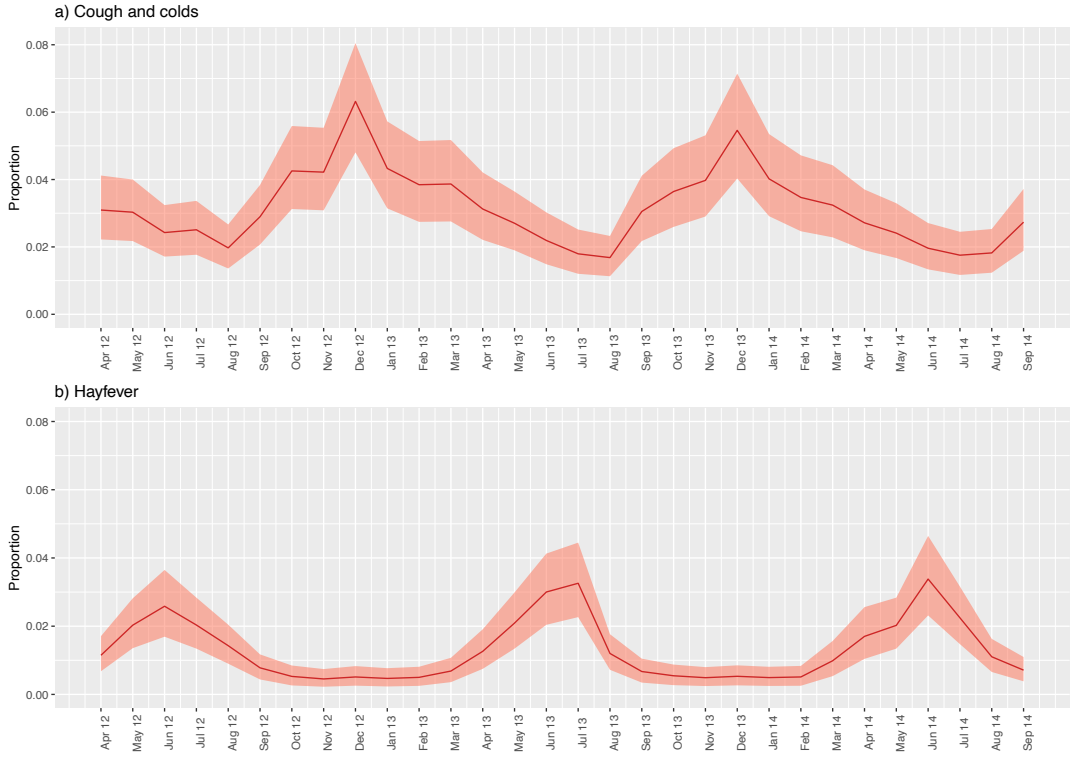

Figure 4.1. Median and interquartile range of proportion purchasing products per month

We next fit models to predict purchasing trends between May 2013 and September 2014. The static model consistently over predicted coughs and colds purchasing (figure 4.2.a). The predicted values for dynamic retraining find similar trends in the data, however, a time lag is observed where large changes occur (e.g. September and December 2013). Figure 4.1. showed that interquartile range increased with increased sales. R^2 values were highest where purchasing proportions are highest reflecting the benefits of greater variation in the training data. The range of R^2 value (0.5-0.7) outlines good performance. The worst performance is seen in August 2013 and 2014 where the lowest median proportions are found. Normalised

Root Mean Square Error (nRMSE) is consistent across the 17 months and follows a similar trend to the R^2 value. Model performance increased with dynamic retraining.

Modelling hay fever (see figures 4.2.d-f) results in trends similar to the training data. July 2013 sees the peak purchasing for 2013 whereas in 2012 and 2014 July witnesses declining purchasing showing that this approach fails to pick up yearly changes. This is likely reflecting the low variation in values across most months, limiting the model training performance. During the off-season, stable trends for hay fever mean predicted medians are closer to the actual values. However, model performance (e.g. R^2 value) is poorer during this period as well. The decrease in R^2 is however expected as this is influenced by the decrease of range within the data therefore explanation of the variance is reduced.

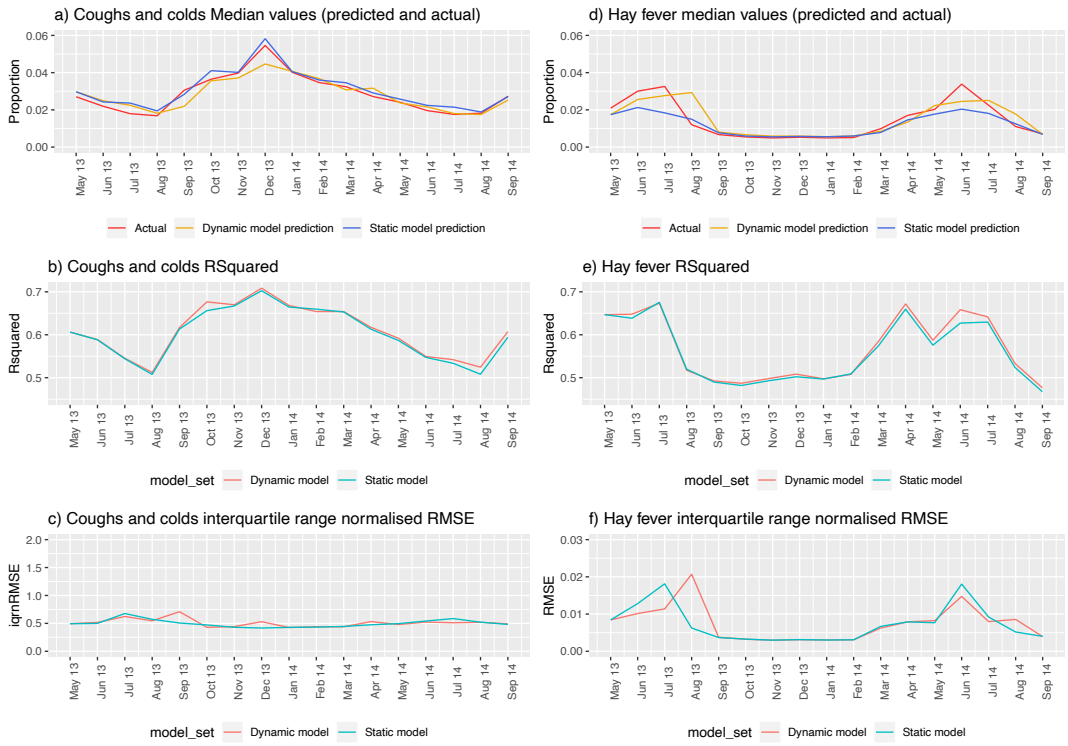

Figure 4.2. a) Coughs and colds median sales and predictions; b) Coughs and colds R2 performance; c) Coughs and colds interquartile range nRMSE; d) Hay fever median sales and predictions; e) Hay fever R2 performance; f) Hay fever interquartile range nRMSE

The dynamic modelling approach generally performs better than the static model, however a time lag occurs with the abrupt changes in sales (August 2013 and 2014 are over predicted). nRMSE is highest for the dynamic model at these time lags. The models struggle to predict the peaks; however, this is constrained by only the availability of one year of training data. A

greater coverage of historically data could improve predictive performance with seasons identified.

Exploring the feature importance across our models allows an evaluation of the predictors of self-medication (figure 4.3.). Only the top eight features (top 10%) are considered since they have the largest effect on the reduction of model error. For both categories, previous month product and related product purchasing are the most important features consistently. Month and buying season are important as temporal identification features. Distinct seasonality of hay fever purchasing is shown with buying season most important across seven months (figure 4.3.b). Temperature is also observed as consistently high ranking suggesting a climate influence. Sulphur dioxide pollution level is the only environmental predictor here for coughs and colds. No social predictors were observed as important here.

Comparably feature importance ranking is erratic for hay fever likely due to the greater seasonality of this product. Mean age of loyalty card holders is higher ranking suggesting this product group is sensitive to age. The largest number of changes in rank is seen for hay fever (August 2013 and 2014) corresponding with the highest nRMSE where purchasing medians decline for the off season.

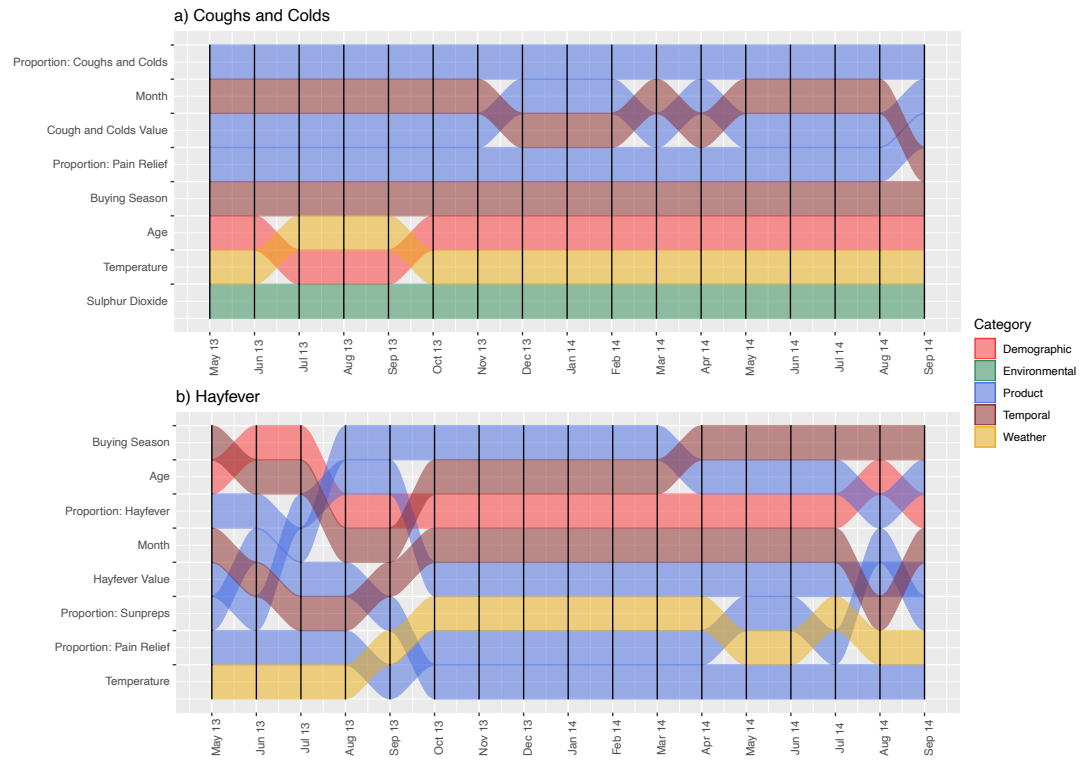

Figure 4.3. Feature importance rank change across models for eight most important features

In order to obtain further context from XGBoost models, Accumulated Local Effect plots (Apley, 2016) are used to understand associations between important features and the outcome in black box methods, particularly when correlation is present between predictors (Apley, 2016; Molnar, 2019). Accumulated Local Effect plots vary a feature across its range to consider its association with the outcome expressed as 'delta'. Again, only the top 10% of features are considered due to the highest reduction of error (and therefore influence) on the models.

Figure 4.4. shows ALE for coughs and colds. As expected, purchasing of product and related product features (cough and colds proportion, cough value and pain relief proportion) are all positively associated with an increase in delta. Similarly (as expected) buying season is positively associated with purchasing. Seasonality is observed within the feature 'month' in-line with typical seasons (Heikkinen and Järvinen, 2003). Positive delta is seen for Autumn to Winter, and negative for Spring to Summer. The largest increase is found where delta is highest, observed in December. This would suggest there is a large positive increase in proportion of customer purchasing in December. Age displays positive delta for ages between 40-60 for coughs and colds. Cold incidence rate is known to be "inversely proportional to age" (Heikkinen and Järvinen, 2003, p52), therefore it is likely that purchasing is for significant others (particularly children). Temperature is relatively static with small fluctuations from zero delta; however, delta is slightly elevated between 2.5-7.5°C. Coughs and colds are associated with a number of viruses that have varying seasonality which would likely explain the stability of temperature (Heikkinen and Järvinen, 2003).

Increased previous month product and related product features (hay fever proportion, hay fever value, sun preps proportion, pain relief proportion) are again associated positively for hay fever (figure 4.5.). Seasonal trends are observed, with buying season positively associated, and months Spring to Summer having large positive delta. Age, again as seen in coughs and colds, exhibits positive delta between 35-60 years old. It is possible these age ranges are purchasing for dependent others (i.e. parents purchasing for children), as decreasing and negative delta is viewed outside this range (Gray, Boardman and Symonds, 2011). For hay fever, positive delta is observed between temperature ranges 10-15°C and at 19°C, suggesting these temperatures increase sales. These ranges relate to optimal temperature ranges for trees (10-15°C), and 19°C is within the optimal range for grass species to release pollen (MetOffice, 2018a).

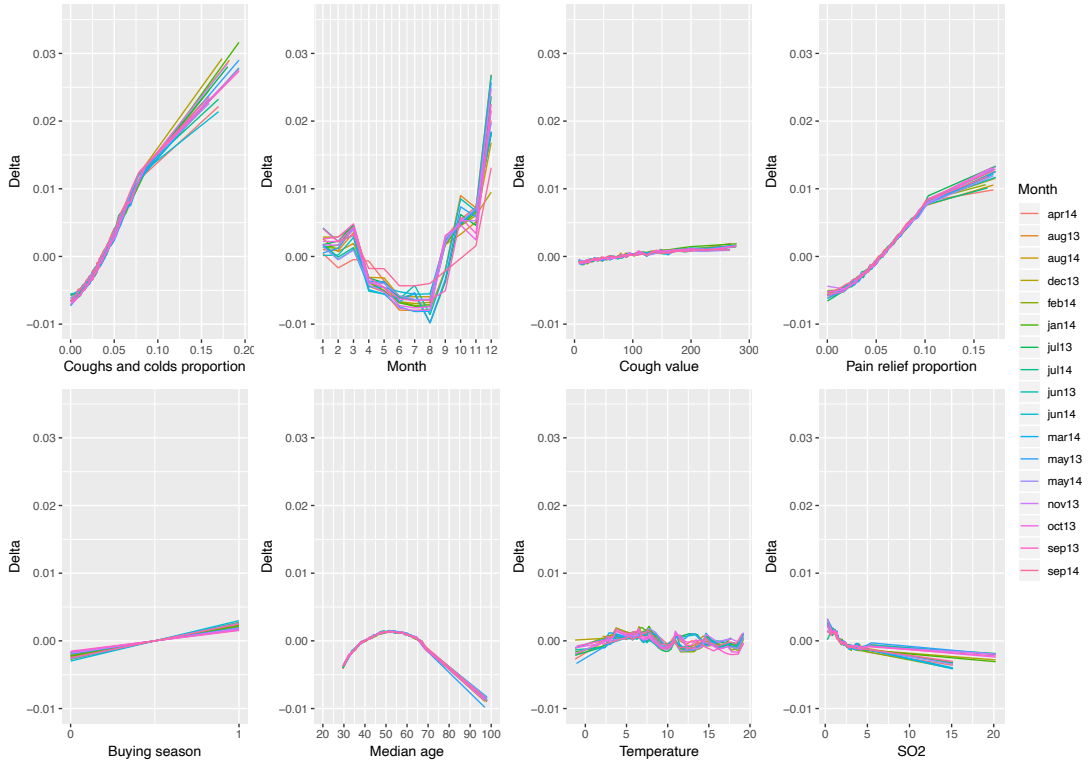

Figure 4.4. Accumulated local effects plot for eight most important features (coughs and colds)

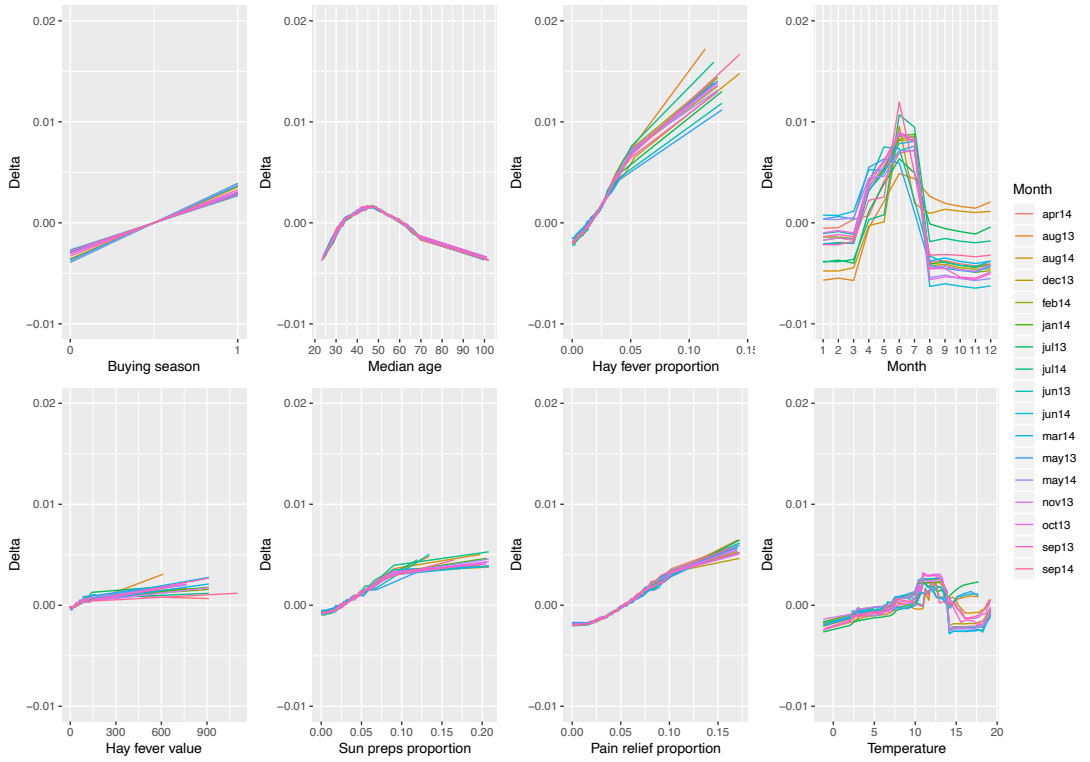

Figure 4.5. Accumulated local effects plot for eight most important features (hay fever)

4.4. Discussion

Using transaction level loyalty card data has provided valuable insights into the temporality of over the counter purchasing for the product groups considered. Distinct seasonality in purchasing was apparent with coughs and colds products more common in Winter and hay fever in Summer. Modelling trends in purchasing confirmed the importance of seasonality, as well as temperature and median age. We also found that our dynamically retrained modelling approach was in general better at predicting purchasing behaviours than a static approach. Our results demonstrate the potential of using such data for population health surveillance and forecasting.

Buying season is an important variable for both products but is ranked higher for hay fever than coughs and colds which is likely due to the more distinct purchasing season. This indicates positive influence (shown in ALE plots figures 4.4.-4.5.) of known coughs and colds season (Heikkinen and Järvinen, 2003) and pollen season for hay fever (MetOffice, 2018b). The observed epidemiological trend of the common cold "increases rapidly in autumn, remains fairly high through winter and decreases again in spring" (Heikkinen and Järvinen, 2003, p52). We find that over the counter coughs and colds purchasing observes the same initial increase and decline in Autumn and Winter, however we also observe an additional peak in Winter (shown Figure 4.1.). The large increase in December (and highest Delta (figure 4.4.)) may relate to a lag from Autumn (buying when needed), but possibly also preparatory purchasing for Winter, particularly as January and February have purchase decline. The less defined seasonality within purchasing (i.e. no clear buying or prolonged off-season) is likely attributable to the amount of associated viruses and respiratory ailments that have varying seasons (Heikkinen and Järvinen, 2003).

Forecasting hay fever is notoriously difficult as the season varies substantially year-on-year (Davies and Smith, 1973). Our approach offers new information of the buying season of hay fever products. We observe peak purchasing between June and July, which concurs with historically observed peaks in early June (Davies and Smith, 1973) and widely disseminated information to the public (MetOffice, 2018c). We observe a purchasing season from April to September, coinciding with seasonal temperatures which are likely to effect purchasing. Temperatures between 10 and 15°C have a positive delta (figure 4.5) and at 19°C increase is observed, relating to optimal pollen release temperatures (MetOffice, 2018a). The inclusion of a seasonality feature based on public advisory information (e.g. MetOffice (2018c)) increased model performance, highlighting influence on purchasing.

Environmental features were important reflecting the seasonality of products (increasing performance when included). Temperature is highly ranked for both products which contrasts to research suggesting weather does not bring performance improvement over historical information when predicting sales (Žylius, Simutis and Vaitkus, 2015). Temperature plays differing roles for our outcomes. For hay fever, it is a proxy variable that correlates to the production of pollen (although is directly driving that production hence indirectly influencing hay fever). In contrast, respiratory conditions (e.g. cold viruses and influenza) are influenced by colder weather (Heikkinen and Järvinen, 2003). We did not detect strong associations though for our other environment measures including air quality with only Sulphur Dioxide ranking in the top 10% of features for coughs and colds. This is despite evidence demonstrating that poor air quality is a determinant of both hay fever (e.g. PM10) and respiratory conditions associated with coughs and colds (Hajat *et al.*, 2001; Charpin and Caillaud, 2017).

We did not find any evidence in the importance of any social or demographic predictors. This was surprising since previous research has demonstrated the importance of social inequalities in self-prescribed medicine behaviours (i.e. lower socioeconomic status groups being less likely to self-medicate) (Green *et al.*, 2016). Despite this, the inclusion of these predictors brought model performance improvement highlighting some (albeit small) predictive importance. Median age of loyalty card holders in areas was found to be important. The result reflects that people aged 35-60 years had the highest proportion of medicine purchasing (and positive Delta in ALE plots), however this age range has been found to exhibit purchasing for dependent others or replenishing family medicine stock (Gray, Boardman and Symonds, 2011).

A number of limitations are present within this research. The limited time series data used (2012 to 2014) constrains historical training data to one year, limiting the quality of training and therefore predictions. The dynamic models show that retraining improves capturing trends, however nRMSE shows that where large differences occur compared with previous years the model performs poorly. Our model therefore presents more of a proof of concept for potential usage in a predictive surveillance model. Greater use of historical data could provide a feasible implementation for utilising over the counter product sales as an early indicator of disease trends, however there are many possible obstacles (e.g. ethical concerns and data linkage between multiple retailers and health records). Purchasing information does not equate to consumption of medicines and is a key limitation of these data. However, sales data have been shown to correlate to disease incidence rates highlighting value (Magruder *et al.*, 2004). Model performance was not perfect and would stress the need for utilising such

data alongside other (more traditional) data to fully understand trends in self-prescribed medications. Further involvement of environmental features such as pollen within the hay fever models (e.g. Ito *et al.*, (2015)) would likely bring performance gain however access to such data is limited. Interpretation of our results must be careful to avoid committing any ecological fallacies. Inferences of our results can only be made at Local Super Out Area level, limiting the application of our models. One opportunity to extend the model would be to explore the spatial patterns in purchasing over time and how they relate to disease outbreaks (e.g. Magruder (2003)). We also only focus on residence location and do not account for movement or spatial exposure (e.g. commuting) (Hanigan, Hall and Dear, 2006).

4.5. Conclusion

Presented are insights from a novel application of machine learning with new forms of data via a scalable Data Science approach for predicting trends in purchasing of self-medication. We build on previous over the counter medicine applications with the inclusion of loyalty card records (Magruder, 2003; Magruder *et al.*, 2004). The application could act as an early indicator of ailment incidence that could complement existing methods (e.g. Santillana *et al.*, (2014)), and may offer cheaper and more efficient means of data collection than existing disease surveillance systems that employ traditional health data (Ginsberg *et al.*, 2009).

Chapter 5 : Using Machine Learning to explore food preparation amongst young adults in Canada

While chapters 3 and 4 have shown how national scale transaction level loyalty card records can benefit public health knowledge, both studies are restricted to residential based location and do not consider the movement of individuals. The Canada Food Study presented an opportunity to extend focus to dynamic individual level exposure through GPS trajectories linked to a health survey. This dataset also highlights how big data does not necessarily relate purely to size (e.g. number of observations) but also can offer size in terms of width which in this dataset is thousands of features. The chapter considers the relationship between obesity and food preparation via the use of sequence analysis and clustering to create a typology of food preparation. The outcomes of cluster membership and then BMI are used in regression models for further understanding. This application identifies problematic food preparation behaviours in Canada and highlights a novel opportunity within obesity research.

5.1. Introduction

Global obesity rates have witnessed widespread increase coinciding with shifts in global food systems and changes in dietary practices. Resultant of prolonged energy imbalance, obesity is defined as the excessive build-up of body fat that may impair health (World Health Organization, 2000b; Katzmarzyk, 2002). Obesity rates in Canada increased from 5.6% in 1985 to 25.3% in 2017, with similar increases observed across all provinces (Katzmarzyk, 2002; Statistics Canada, 2019). Severity has risen and Body Mass Index is now more commonly exceeding 40 (i.e. class III obesity) causing a major issue to public health and its economy (e.g. increased incidence of obesity-related non-communicable diseases) (World Health Organization, 2000b; Kim and Basu, 2016; Lebenbaum *et al.*, 2018). Obesity threatens “both the overall health status of Canadians and the Canadian healthcare system” (Anis et al., 2010, p31), with 2021 estimates suggesting a CAN$9 billion direct cost, accounting for 2.5% of healthcare spending (Birmingham *et al.*, 1999; ObesityCanada, 2019).

Accessible, cheap, and extensively promoted ‘junk’ food is considered a key contributor to obesity increase (Hill and Peters, 1998; World Health Organization, 2000b; Guthrie, Lin and Frazao, 2002; Barlow, McKee and Stuckler, 2018; Fong *et al.*, 2019). To date research has commonly focused on understanding associations between obesity and the built environment, food contents from dietary recalls, and physical activity (Hill and Peters, 1998; Tremblay and Willms, 2003; Kestens *et al.*, 2012; Satija *et al.*, 2015; Crowe *et al.*, 2018).

While a multitude of studies examine these individual relationships, few have incorporated this information together, particularly linking nutritional and health surveys along with built environment data (e.g. exposure to food outlets (Casey *et al.*, 2012; Burgoine *et al.*, 2014)).

Childhood obesity has dramatically increased since 1980, bringing long-term risk, consequences, and severe public health concerns (World Health Organization, 2000b; Johnson-Taylor and Everhart, 2006; Abarca-Gómez *et al.*, 2017). Globally, child and adolescent obesity was 4.5 times higher in 2016 (18%) versus 1975 (4%), and 5% higher versus adults (13%) (World Health Organization, 2018). Despite research commonly focusing upon adolescence, there is a lack of understanding about important factors that may impact obesity rates among young adults, including preparation habits, food-related behaviours, and environmental exposures (Johnson-Taylor and Everhart, 2006; Larson *et al.*, 2006). With youth obesity more prevalent than ever before and young adults (particularly men) more likely to consume fast food (Duffey *et al.*, 2009; Lachat *et al.*, 2012), there is an urgent need for research that explores dietary behaviours during the transition to adulthood (ages 16-30).

Examining where and by whom food is prepared brings important context to nutrition studies. Findings from Larson *et al.* (2006) suggest those who prepare their own food are more likely to consume the recommended level of nutrients. An inverse relationship between consuming food sourced out of home and obesity is widely cited, associating service food (e.g. fast food) with poor nutrient intake (vs prepared at home) and weight gain (Duffey *et al.*, 2009; Smith *et al.*, 2009; Powell, Nguyen and Han, 2012; Penney *et al.*, 2017). Exposure to such services, associated with low cost and convenience, is a key determinant of obesity (Hill and Peters, 1998; Tremblay and Willms, 2003; Burgoine *et al.*, 2014). In contrast, skipping meals (particularly breakfast) has been linked to grazing behaviour, negatively impacting nutrition and hampering cognitive function (Benton and Parker, 1998; Waterhouse *et al.*, 2005; DeJong *et al.*, 2009). Skipping breakfast has been suggested as an associated cause of obesity (DeJong *et al.*, 2009). To our knowledge, there have been no examinations of how different sequential meal behaviour patterns (including origin, location, and meals skipped) is affected by exposure to food environments, and how this subsequently links to BMI.

To fill the aforementioned gap in the literature, our study uses data from the 2016 Canada Food Study. Launched as a part of a larger multi-country effort to explore links between diet and health, the Canada Food Study combines comprehensive assessments of diet and dietary practices over time with high resolution urban built environment exposure data. To achieve

this, the Canada Food Study has utilised the ubiquity of GPS-enabled smartphones to capture daily spatial activity patterns. The use of trajectory data from GPS devices in food behaviour research is increasing (e.g. Chaix *et al.* (2012); Scully *et al.* (2017;2019); Widener *et al.* (2018)). The provision of extensive GPS movement data, and the ability to perform data linkage to combine this with comprehensive information from other sources (e.g. food diaries) enables for greater depth in behavioural research (Stopher and Greaves, 2007).

Methods of sequence analysis and data mining make it possible to derive interpretable typologies of behaviour (Gabadinho *et al.*, 2011). Commonly utilised in career path analysis, these methods focus on complete events sequences (Abbott and Tsay, 2000), and provide analysis opportunities of the sequential food preparation data collected in the Canada Food Study. We examine food preparatory behaviours by creating a typology of preparation sequences. Doing this, alongside an assessment of time weighted GPS trajectory data, while also accounting for important covariates, will allow a novel exploration of the relationship between obesity, food preparation, consumption, and exposure to food environments.

5.2. Methods

5.2.1. Data

The dataset used (the 2016 Canada Food Study) is described in detail by Hammond, (2017), Hammond, White and Reid, (2017) and Widener *et al.*, 2018). Summarising Widener and colleagues' description, respondents, aged 16-30, were recruited in five Canadian cities (Edmonton, Halifax, Montreal, Toronto and Vancouver) via in-person intercept sampling (from a sample of stratified sites). Eligibility is determined by residence, age, internet and device access, and no prior Canada Food Study enrolment. Consent was provided prior to completing the study. Of the 3000 young adults in the initial survey, 1568 were invited for a smartphone sub-study and 686 participated. Respondents must have expressed interest for the follow-up study and had an eligible smartphone. Using a GPS-enabled smartphone and the *CFSMobile* app, high resolution trajectory data was collected which included a temporal component. Participants received a CAN$2 recruitment cash incentive, a CAN$20 Interac e-transfer after survey completion and an additional CAN$25 Interac e-transfer after GPS study completion.

After excluding respondents due to incomplete participation (e.g. failing to complete all separate sections of the study, travelling outside Canada during the study or significant amounts of 'refuse to answer'), 396 individuals were used in our analysis. Summary statistics of the overall sample are as follows. 35% were of the sample were (assigned the

sex at birth of) male and 65% female. The median age was 21. In terms of ethnicity, 52% of the sample identified as White, 4% Aboriginal, 4% Black, 10% Chinese, 7% South Asian and 23% other. 86% of the sample had completed a minimum education level of a high school diploma or equivalent. The median BMI of the sample was 23.4.

5.2.2. Food preparation sequence typology

The Canada Food Study includes food preparation recall questions over a one-week period. Respondents provided the location and who prepared food for three daily meals (breakfast, lunch, dinner) for seven consecutive days. Sequences were standardised by order (beginning Monday breakfast). Possible answers were:

- Home, by you (including minimal prep),
- Home, by someone else (e.g. family/ partner),
- Out of home service (e.g. restaurant/ takeout),
- Someone else's home,
- Did not eat.

This data contained 21 columns of data for each respondent, with the aforementioned five possible answers. In its raw form this data is not particularly useful nor easy to comprehendible by the human eye. Sequence analysis enabled this information to be mined providing insights from methods such as clustering. While the data is novel, it is the sequence analysis that adds value to this data. Food preparation sequences are mined in order to understand how food preparation varies through a week, and to explore whether there are typical behaviours within our sample. Considering these data as sequences provides the opportunity to easily interpret behaviours at a population level. Clustering enables this analysis to extend to consider whether behaviours such as skipping breakfast or eating out on the weekend exist and to understand the characteristics of the groups that exhibit these behaviours. Understanding these behaviours as an outcome and exploring the characteristics of individuals who exhibit them are key to be able to develop knowledge and policy within obesity and nutrition. Analysis was performed in R (R Core Team, 2014).

Sequence analysis

Sequence analysis allows the visualisation and data mining of sequential data. Following a methodology similar to McVicar and Anyadike-Danes (2002), sequence analysis was performed using TraMineR (R package) (Gabadinho et al., 2011). The sequences are made up of 21 meals from Monday breakfast to Sunday Dinner. To analyse these sequences optimal matching is necessary to examine the similarity of sequences. This allows the creation of a cost matrix of substitutions, allowing sequences (e.g. *Did not eat – Home, by*

you – Home, by you and *Did not eat – Home, by you – Home someone else*) to be judged similar, despite not matching exactly (Dijkstra and Taris, 1995). The Hamming (Hamming, 1950) method for optimal matching was used as equal weighting is given to position. As dissimilarity is determined by the number of positions that do not match, the approach is strict to time warp (e.g. *Out of home service – Home, by you – Home, by you* and *Home, by you – Out of home service – Home someone else* have no positional matches and are therefore different) (Studer and Ritschard, 2016).

Clustering

Both Ward's method of hierarchical clustering (Ward, 1963) and K-medoids (Park and Jun, 2009) were tested to cluster the food preparation sequences. Ward's method is agglomerative, categorising data into a tree of clustered groups (Liao, 2005). K-medoids arbitrarily categorises data into clustered groups and iteratively minimizes an objective function until within cluster sum of distances between medoids is minimised (Hartigan and Wong, 1979; Liao, 2005; Park and Jun, 2009). Clusters were created for a range of k (2-10) to allow a parsimonious solution to be identified. Ward's method (k=10) was selected via qualitative exploration of representative sequences and state frequency plots, elbow plots and the gap statistic. A need for a high number of splits was highlighted (e.g. for low splits respondents with exclusively Out of home service and Home, by you were grouped together). The gap statistic (Tibshirani, Walther and Hastie, 2001) and qualitative exploration suggested k=10 optimal. Ward's method was selected over K-medoids since the latter produced less optimal results (e.g. multiple Home, by you clusters were produced only varying at one state). Clustering was performed using the cluster (Maechler *et al.*, 2018) and kmed (Budiaji, 2019) R packages, with model metrics explored using the factoextra package (Alboukadel Kassambara and Mundt, 2017).

5.2.3 Time weighted exposure

Mobility patterns are theorised to be consistent, with high probabilities of returning to familiar locations (Lin and Hsu, 2014; An, 2016). GPS movement data initiates the generation of time weighted average daily exposure to food and health-based facilities (e.g. fast food, sport and leisure). Using the itinerum-trip-breaker algorithm (SAUSy-Lab, 2019), cleaned representations of activity spaces are extracted for each participant. As SAUSy-Lab (2019) describe, the algorithm removes erroneous points (i.e. similar points to temporal neighbours and large jumps indicating no signal), segments data and then performs location detection using a time-weighted kernel density estimation, generating coordinates with a temporal component indicating activity occurrence. A parameter of ten minutes was

determined by the authors as a reasonable threshold to determine whether a person has remained in a location for long enough to be considered a place of note. The algorithm also calculates travel time between locations.

Multiple buffers were generated around each place of note, with distances intended to account for direct sight (250m) and an approximate 15-minute walk (1km). Business register data from Statistics Canada was used to quantify obesity-relevant businesses in the built environment (e.g. food retail and sports facilities). North American Industry Classification System (NAICS) codes were used to classify businesses by count, aggregated at the Dissemination Area level (Statistics Canada (Business Register Division), 2017). The proportion of the Dissemination Area's area (Statistics Canada, 2017) intersecting each activity location buffer was calculated and used to weight Dissemination Area business count for that site. For example, if a Dissemination Area contained ten fast food retailers and the buffer covered 50% of the area, the location would be considered exposed to five fast food retailers. A daily total exposure to the built environment measurement per participant is calculated by:

- Multiplying the number of relevant businesses within activity space buffers by the time spent in each buffer
- Summing the total weighted counts across all activity space buffers.

These daily exposures are averaged across the number of days GPS data was provided per individual, creating a dynamic movement based average exposure measurement.

5.2.4. Statistical analyses

Summary statistics, state frequencies, and representative sequences were calculated to provide descriptive context for our typology.

As personal contexts drive where food is sourced and consumed (Widener *et al.*, 2018), multinomial logistic regression was used to examine whether demographic and built environment features explained an individual engaging in our clusters of food preparation behaviours. Covariates included:

- Age - linked to increased likelihood of consuming Out of home service;
- Sex and race - associated to food-preparatory behaviours (Larson *et al.*, 2006);
- Household size - found impacting likelihood of eating Out of home service (Datar, 2017); and
- Time weighted exposure to facilities (fast-food, restaurants, sport and leisure).

Linear regression was used to examine BMI as a function of our sequence clusters, exposure, and demographic features. Exploratory features include:

- Age - associated with differences in obesity rates;
- Sex - obesity is typically higher for women (World Health Organization, 2000b);
- Ethnicity - risk of complications from obesity vary with ethnicity as fat is stored differently (World Health Organization, 2000b; Katzmarzyk, 2002; Gittelsohn *et al.*, 2018);
- Household size - found inversely related to BMI (Datar, 2017); and
- Time weighted exposure to facilities (fast-food, restaurants, sport and leisure) - caloric access is considered a key influencer of obesity (Tremblay and Willms, 2003).

Regression models were performed using base R (R Core Team, 2014) and the nnet R package (Venables and Ripley, 2002) and regression summary tables were created using the stargazer package (Hlavac, 2015).

5.3. Results

5.3.1. Descriptive analysis

Looking at all sequences from the sample, Home, by you is the most frequent preparation at breakfast (.6), however, Did not eat is notably prevalent at breakfast (.2) versus other mealtimes (shown figure 5.1.a). Out of home service accounts for larger frequency for lunch and dinner, increasing at the weekend. Evening meals are rarely missed, with Home, by you again displaying the largest frequency (.45). Someone else's home increases towards and is highest at weekends. Entropy for each state (figure 5.1.b) is lowest at breakfast, suggesting lacking diversity of food preparation behaviours compared to lunch/dinner. A small increase in entropy is observed chronologically, likely due to higher variance at weekend.

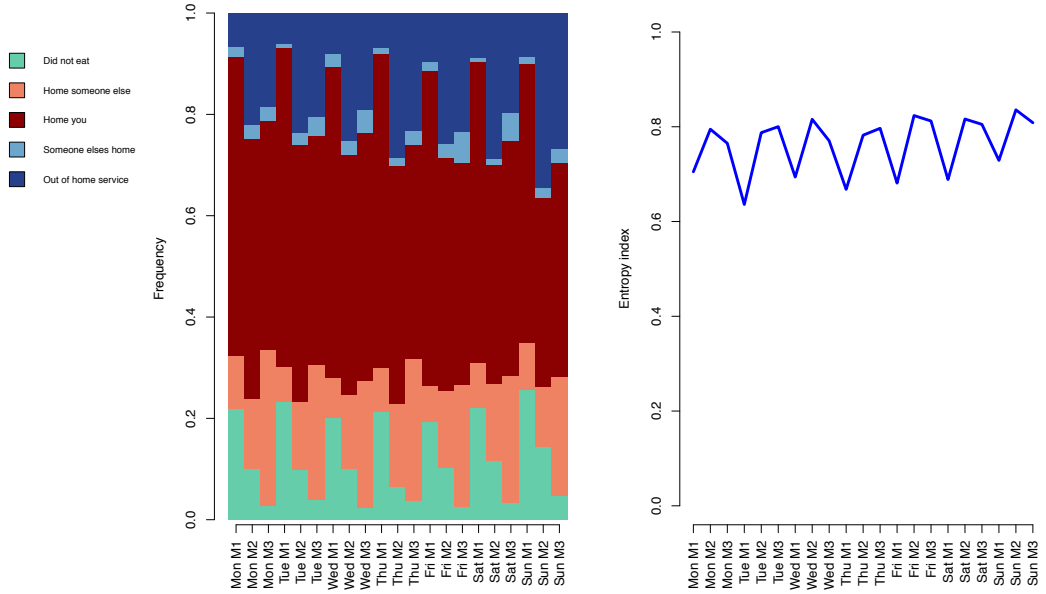

Figure 5.1. a) State frequency (left); b) state entropy (a measure of uncertainty) (right).

Note: the X axis details the meal number of the day (e.g. M1 is meal 1).

5.3.2. Exploratory analysis of typology

Figure 5.2. provides a graphic breakdown of our ten clusters, containing a vast amount of information. Further documentation of these plots can be found in the TraMineR R package (Gabadinho et al., 2011) documentation, however, the following description should aid interpretation of Figure 5.2.. Part a. displays representative sequences that make up each cluster. The Y axis shows the number of representative sequences that are contained within each cluster, as well as the cluster size. The symbols (e.g. ▫) distinguish between sequences on the Y axis and also on the discrepency and the mean distance measures (above each set of representative sequences). The X axis is split into 21 blocks, each block represents a meal in order (i.e. position one is Monday breakfast or meal nine is Wednesday dinner). For example, cluster 9 has one representative sequence (all meals are prepared at home by someone else). As these sequences are representative, figure 5.2.b. instead shows the frequency of each food preparation type within each of cluster. Again using the example of cluster 9, one representative sequence is shown in part a., whereas part b. shows that from Wednesday onwards, out of home service food actually increases in frequency. This information links back to part a. with the discrepency being three and the mean distance four for this cluster.

Ten clusters are found within the food preparation sequences. While the largest cluster contains 160 people, the smallest cluster contains 11. Despite the cluster size being

unbalanced, using a smaller number of clusters meant that distinct sequences were dispersed through other clusters losing context (highlighted in figure 5.2. within clusters 7, 9 and 10). Ten clusters were kept to ensure this context was not lost. Within our ten homogenous clusters, there are five prominent behaviours (figure 5.2.a-b); Home, by you (clusters 1, 4, 6 and 8), Home, by someone else (clusters 2 and 9), mixed Home, by you and Home, by someone else (clusters 3 and 5), Out of home service (cluster 7) and Did not eat (cluster 10). Distinct subsequences are found such as clusters missing breakfast (1, 2, 6, 7 and 10), and increased weekend Out of home service (clusters 1, 5, 8).

Of the Home, by you prominent clusters, cluster 4 (n=160) is the largest with Home, by you across all meals. Cluster 6 (n=29) is similar, however breakfast is typically missed and a higher BMI is witnessed (.7 increase). Both cluster 4 and 6 infrequently include Out of home service. Here the highest ages are found (cluster 4=23.2, cluster 6=23.3). Fast food exposure is high at 250m and 1km buffers, despite low Out of home service bevhaviours.

Cluster 1 (n=26) sees inclusion of Did not eat and Out of home service (particularly Out of home service at weekends), while cluster 8 (n=24) sees only higher Out of home service. Participants in cluster 1 typically skip breakfast, and have a BMI 1.1 above average. Cluster 1 sees frequent Out of home service, despite low exposure to fast food. Cluster 8 is younger (20.8 years), however BMI is .7 above average and travel time is 7.3 minutes above average, possibly indicative of more Out of home service. Cluster 2 (n=34), displays similarities with cluster 1 (indentical BMI), however Home, by someone else replaces HY. Did not eat breakfast is common. These sequences frequent Out of home service, despite below average exposure to fast food.

Those in cluster 9 (n=16) have one representative sequence: all meals are Home, by someone else except Sunday dinner (HY). Travel time and fast food exposure allign with average.

Home, by you/ Home, by someone else mix is found in clusters 3 (n=59) and 5 (n=16). Cluster 3 has static behaviour of breakfast (Home, by you), lunch (Home, by you/ Home, by someone else mix) and Dinner (Home, by someone else). Age is low (19.9), household size is high (median=4), and exposure to fast food is low for this group. Constrastingly, cluster 5 displays no clear pattern and sporadic Out of home service inclusion. Above average BMI is possibly influenced by the lack of food preparation structure, alongside higher fast food exposure, compared with cluster 3. There is some Out of home service at weekends for both groups.

Cluster 7 (n=29) has majority Out of home service. Two representative sequences are observed: all Out of home service, or all Out of home service but Did not eat breakfast. Age (20.3) is low for this cluster, along with household size (2.5). Fast food exposure is low at 250m (.08), increasing at 1km.

Cluster 10 (n=11) exhibits isolated behaviour with Did not eat dominant (Breakfast is typically skipped). This group has the lowest median age (18.7), however median BMI is above average. Travel time is the largest observed (87.2 minutes), and fast food exposure is low.

Table 5.1. Summary statistics of clusters

Cluster	n	Age (mean)	BMI (median)	Travel time (mean mins)	Household size (median)	Fast Food 250m (median)	Fast Food 1000m (median)
1	26	21.5	23.4	84.5	3	.26	2.96
2	34	21.2	23.4	87.3	3.5	.11	2.71
3	59	19.9	21.9	80.0	4	0.12	2.83
4	160	23.2	22.3	77.7	2	.22	4.40
5	16	22.1	22.8	71.0	3	.21	2.26
6	29	23.3	23.0	62.0	2	.32	4.95
7	22	20.3	22.5	66.5	2.5	.08	3.69
8	24	20.8	21.5	85.3	3	.15	2.91
9	16	21.1	23.0	80.3	3.5	.20	3.21
10	11	18.7	22.9	87.2	3	.07	2.44
All	397	21.9	22.3	78.0	3	0.18	3.77

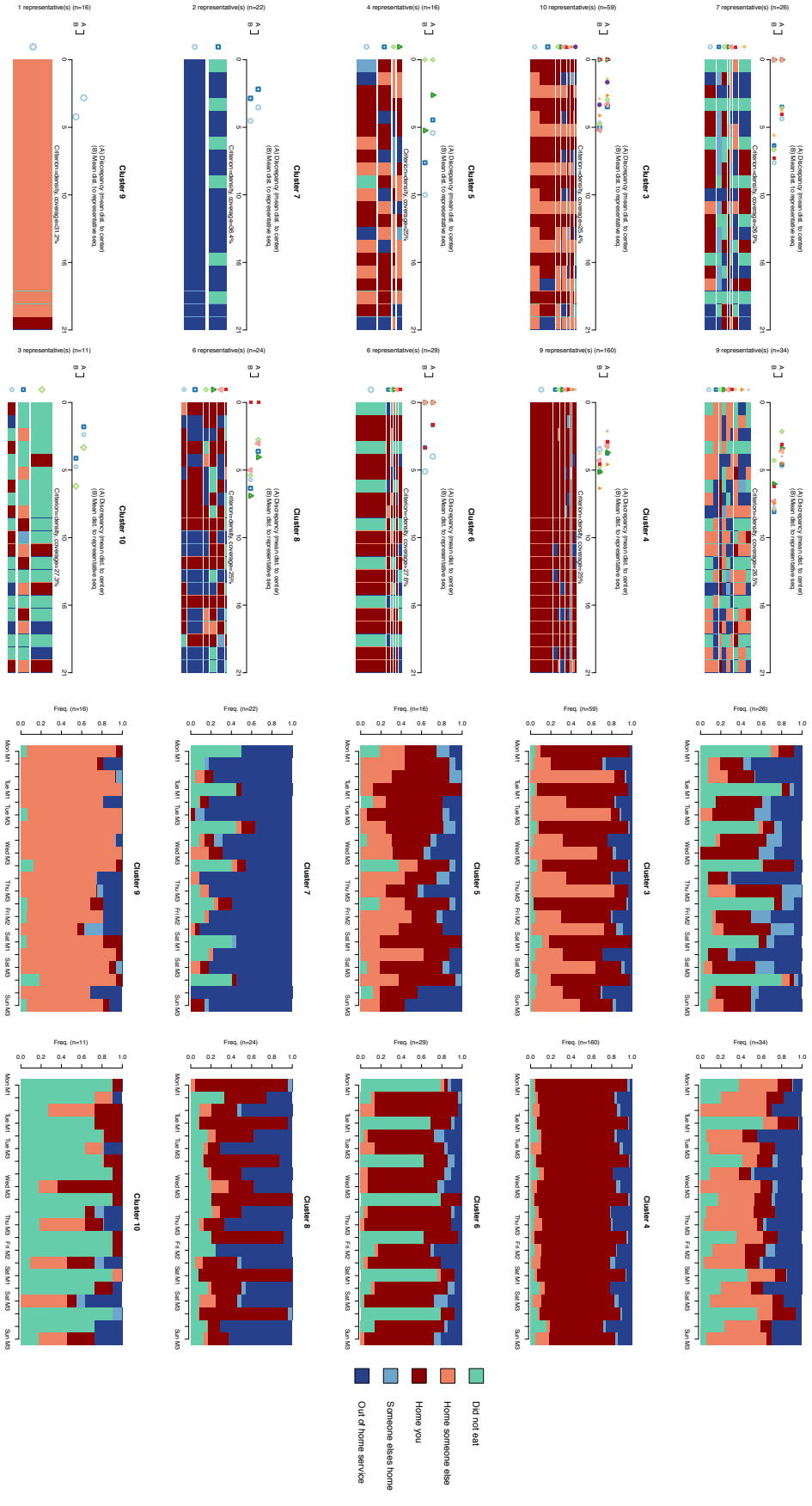

Figure 5.2. a) Representative sequences for typology b) state frequencies for typology

5.3.3. Explaining food preparation behaviours

Cluster 4 (predominantly Home you) is the reference category for our outcome as this is largest cluster which represents consistent at home preparation behaviour. Covariate reference categories are sex = female, location = Toronto, and race = white.

Age is significant for all clusters ($p<.001$). A one-unit increase decreases the odds of being in each cluster (except cluster 6 displaying a positive relationship). Sex is significant for all groups except clusters 8 and 10 ($p<.001$); all odds of cluster membership are greater than one (i.e. males were more commonly associated with clusters 1-3, 5-7 and 9). Cluster 7 displays the largest coefficient (OR = 8.97) suggesting being male increases the likelihood of predominantly Out of home service. Similarly, a large odds ratio (3.60) was observed for Home, by someone else (cluster 9) versus Home, by you (cluster 4).

Few associations were observed between locations. Significant odds ($p<.001$) are found for cluster 9, meaning the odds of Home, by someone else versus Home, by you is higher for Montreal versus Toronto. Significance ($p=.002$) is seen for cluster 5 with higher odds for a mix of Home, by you / Home, by someone else / Out of home service vs Home, by you for individuals in Vancouver versus Toronto. All other associations were non-significant.

Significant odds were found for participants identifying as Chinese and clusters lacking consistent home preparation (1-2, 7, 9 and 10). The highest odds (OR = 17.71) for participants identifying as Chinese was cluster 9 (which has higher occurrence of Home, by someone else), which is similar for participants identifying as South Asian (OR = 10.67) and race=Other (16.72). Chinese and South Asian exhibit the only higher odds of being in cluster 7 (Out of home service majority).

Access to leisure facilities have non-significant odds for all clusters, however 250m access to fast food or restaurants was found to have varying effects. For example, cluster 7 (Out of home service dominant) exhibits non-significant odds for fast food, but a significant per unit increase of 8.8% is found for restaurants ($p=.003$).

All odds are significant ($p<.001$) for travel time. The highest odds (1.012) are found for cluster 10 (the most sporadic eating behaviour vs cluster 4). The odds are lowest (.982) for cluster 6, despite this group typically missing breakfast.

Household size is significant for all clusters. Cluster 9 (predominantly Home, by someone else) has the highest odds for an increase in household size whereas cluster 7 (Out of home service) and cluster 6 (Home, by you - most similar to 4) exhibit the lowest odds.

Table 5.2. Multinomial logistic regression (exponentiated odds ratios)

	Dependent Variable:								
	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 5**	**Cluster 6**	**Cluster 7**	**Cluster 8**	**Cluster 9**	**Cluster 10**
Age	0.900***	0.894***	0.798***	0.940***	1.027***	0.793***	0.817***	0.890***	0.660***
	(0.065)	(0.061)	(0.055)	(0.075)	(0.058)	(0.083)	(0.075)	(0.082)	(0.150)
Sex Male	1.731***	2.866***	2.207***	1.702**	1.890***	8.970***	1.059	3.600***	0.195
	(0.482)	(0.443)	(0.371)	(0.596)	(0.455)	(0.565)	(0.549)	(0.643)	(1.174)
Location	0.346	0.460	0.841	0.916	1.160	0.298	0.787	0.950	0.207
Edmonton	(0.874)	(0.674)	(0.512)	(1.004)	(0.648)	(0.926)	(0.747)	(1.000)	(1.250)
Location	0.721	0.348	0.358	0.534	0.323	0.589	0.351	0.450	0.397
Halifax	(0.657)	(0.647)	(0.522)	(0.930)	(0.706)	(0.698)	(0.741)	(1.294)	(0.983)
Location	0.851	0.486	0.594	1.787	1.110	0.322	1.287	3.140***	0.523
Montreal	(0.802)	(0.897)	(0.663)	(0.923)	(0.657)	(0.963)	(0.953)	(0.909)	(1.352)
Location	0.559	0.685	0.731	2.696**	0.227	0.219	1.004	1.876*	0.379
Vancouver	(0.688)	(0.633)	(0.547)	(0.828)	(0.862)	(0.936)	(0.715)	(0.900)	(1.113)
Race	0.696	1.684**	0.256	0.862	1.075	0.00002	1.375	0.001	0.00004*
Aboriginal	(1.133)	(0.801)	(1.128)	(1.170)	(1.156)	(0.00001)	(0.892)	(0.0005)	(0.00002)
Race	0.559	0.403	0.438	0.526	0.861	0.577	0.00001	0.0003	2.629*
Black	(1.177)	(1.192)	(0.821)	(1.215)	(0.926)	(1.359)	(0.00003)	(0.0005)	(1.329)
Race Chinese	4.868***	4.637***	1.168	0.00001	1.351	6.675***	1.105	17.709***	4.165**
	(0.710)	(0.685)	(0.677)	(0.00001)	(0.794)	(0.865)	(0.910)	(1.344)	(1.416)
Race South	1.022	1.385	0.437	0.00001	1.075	3.795***	0.00001	10.673***	0.00000
Asian	(0.923)	(0.767)	(0.778)	(0.00003)	(0.902)	(0.833)	(0.00002)	(1.368)	(0.00001)
Race Other	0.697	0.667	0.749	0.544	0.548	0.867	0.863	16.724***	2.509**
	(0.660)	(0.622)	(0.449)	(0.703)	(0.614)	(0.721)	(0.632)	(1.189)	(0.892)
Sport Leisure	0.583	0.239	0.319	0.888	0.209	0.758	0.221	0.278	1.588
	(0.786)	(1.091)	(0.677)	(0.743)	(1.089)	(0.976)	(1.107)	(1.195)	(1.528)
Fast food	1.288*	1.231	2.525***	0.514	1.301*	0.743	5.410***	6.128***	1.570
	(0.569)	(0.635)	(0.421)	(0.801)	(0.595)	(0.683)	(0.595)	(0.775)	(1.281)
Restaurants	0.987**	0.750	0.796***	1.348***	0.844**	1.088**	0.151	0.304	0.107
	(0.312)	(0.403)	(0.256)	(0.358)	(0.300)	(0.364)	(0.825)	(0.586)	(1.354)
Travel time	1.005***	1.003***	1.002***	0.993***	0.982***	0.991***	1.005***	1.001***	1.012***
(mean)	(0.005)	(0.005)	(0.005)	(0.009)	(0.008)	(0.008)	(0.005)	(0.007)	(0.009)
Household size	1.311***	1.594***	1.561***	1.361***	0.860***	0.838***	0.923***	1.723***	1.350***
	(0.158)	(0.143)	(0.121)	(0.189)	(0.173)	(0.215)	(0.185)	(0.193)	(0.228)
Constant	0.599	0.640	13.040***	0.329	0.883	36.299***	20.389***	0.017	155.753***
	(1.685)	(1.557)	(1.330)	(2.082)	(1.574)	(1.993)	(1.817)	(2.419)	(3.321)
AIC	1556.076	1556.076	1556.076	1556.076	1556.076	1556.076	1556.076	1556.076	1556.076
Sample size	396	396	396	396	396	396	396	396	396

Note: Exponentiated odds ratios, standard errors and p-values (*p<0.05; **p<0.01; ***p<0.001) are displayed

5.3.4. Examining whether good preparation is associated to body weight

Age is positively associated with BMI (.140, p=.05). Sex and household size are non-significant. Aboriginal race was positively associated with a 2.386 BMI increase (p=.05) whereas all other categories demonstrating no association. All cities exhibit increased BMI compared with the reference (Toronto) however Montreal is the only significant coefficient (2.174, p=.012).

Fast food exposure (250m) was positively associated with BMI increase (1.329, p=.019), whereas leisure exposure (250m) was inversely related to BMI (-1.639, p=.025). Exposure to restaurants and travel time are non-significant. The typology suggests individuals within cluster 7 may have higher BMI than individuals in cluster 4, indicating increased BMI for those with food predominantly prepared Out of home service, however the level of significance was inconclusive (p=.093). No other cluster is found significant.

Table 5.3. Linear Regression

	Dependent variable: Body Mass Index
Age	0.140* (0.071)
Sex Male	0.482 (0.535)
Location Edmonton	1.296 (0.796)
Location Halifax	0.961 (0.742)
Location Montreal	2.174* (0.858)
Location Vancouver	0.863 (0.740)
Race Aboriginal	2.386* (1.213)
Race Black	0.572 (1.204)
Race Chinese	-1.325 (0.888)
Race South Asian	0.032 (1.035)
Race Other	-0.029 (0.652)
Cluster 1	0.290 (1.024)
Cluster 2	1.254 (0.946)
Cluster 3	-0.219 (0.786)
Cluster 5	0.286 (1.249)
Cluster 6	0.029 (0.977)
Cluster 7	1.916 (1.137)
Cluster 8	-1.618 (1.063)
Cluster 9	1.493 (1.316)
Cluster 10	1.602 (1.513)
Sport Leisure	-1.639* (0.728)
Fast Food	1.329* (0.562)
Restaurants	-0.204 (0.351)
Travel Time (mean)	0.0002 (0.006)
Household size	-0.101 (0.173)
Constant	19.253*** (1.914)
Observations	396
R^2	0.095
Adjusted R^2	0.034
Residual Std. Error	4.701 (df = 370)
F Statistic	1.560* (df = 25; 370)

Note: Beta coefficients, p values (*p<0.05; **p<0.01; ***p<0.001) and standard errors are displayed

5.4. Discussion

Utilising food diaries alongside GPS movement data allows for unprecedented context when analysing food preparatory behaviours. Little, if any research, has explored how food preparation behaviours vary sequentially, how environmental context may relate to this, and how these important factors link to health outcomes like BMI. We build on previous dietary behaviour studies with the novel use of sequences, contributing new information to the national and global concern of improving diet (An, 2016).

We demonstrate that potentially problematic behaviours exist with food preparatory behaviours, as 5% of our sample predominantly have food prepared out of home service. Eating out is strongly associated with fast food consumption, and higher caloric consumption is consequential (Lachat *et al.*, 2012; An, 2016; Penney *et al.*, 2017). Furthermore, concerning sequences are observed with increased Out of home service at weekends (clusters 1, 5 and 8) as it is suggested that consumers do not offer the same scrutiny to ingredients compared with Home, by you preparation (Guthrie, Lin and Frazao, 2002).

In contrast, 30% of our study population fall in representative sequences missing meals (Did not eat breakfast). Research has shown skipping meals can negatively impact daily nutrition, is linked to obesity, and can hamper some cognitive function (Benton and Parker, 1998; Waterhouse *et al.*, 2005; DeJong *et al.*, 2009). Skipping breakfast also promotes unfavourable grazing behaviours (Waterhouse *et al.*, 2005). We find clusters regularly missing breakfast exhibit higher median BMI.

The highest median ages tend to be in Home, by you clusters (4 and 6). This is consistent with research suggesting younger adults are more likely to consume Out of home service (Larson *et al.*, 2006). Significant associations of younger age and Home, by someone else clusters possibly capture younger individuals living with family. In our data, however, a positive association between increased age and BMI is exhibited aligning with global trends (World Health Organization, 2018). This polarity could be suggestive of increased wealth (from the transition from education to work) influencing portion size and therefore caloric intake (Lachat *et al.*, 2012).

We detect higher odds of males belonging to clusters lacking Home you, consistent with the notion that young adult males are more likely to consume fast food, and a greater likelihood of young adult females being invloved in food preparation (Larson *et al.*, 2006; Duffey *et al.*, 2009). Non-significant effects are found for sex predicting BMI. Despite obesity rates

typically being higher for women, this finding aligns with obesity prevalence predictions in Canada which suggest similar rates in men and women (World Health Organization, 2000b; Gotay *et al.*, 2013).

Household size is suggestive of Out of home service cluster membership, corresponding with smaller household sizes being linked to consuming food away from home (Datar, 2017). Literature associates larger families with lower BMI, however we find no significant effects (Datar, 2017). Household size has been cited as a determinant to skipping meals, whereas we do not observe any obvious relationship (DeJong *et al.*, 2009; Datar, 2017).

We observe varied effects by race. Significant effects for participants who identify as Chinese and cluster 7 and 9 membership show a reliance on others for food preparation. When predicting BMI, participants who identify as Aboriginal, a group with a greater risk of obesity, is the only race displaying a significant positive coefficient (Katzmarzyk, 2002; Gittelsohn *et al.*, 2018). Evidence of varied preparation by race, and associations with BMI, show an important determination in food behaviours and obesity when considering at-risk populations.

The theory that widespread access to caloric foods is linked to the consumption of such foods (Tremblay and Willms, 2003) is not convincingly found in this analysis. Summary statistics show that exposure is low for cluster 7 (predominantly Out of home service) and high for clusters with little Out of home service (4 and 6). Moreover, log odds are non-significant for clusters with high Out of home service, and fast food exposure (5 and 7). Restaurants are however significantly positively associated suggesting that Out of home service may not necessarily be influenced purely by fast food outlets. Contrastingly, clusters (1, 3, 8) with sporadic inclusion of Out of home service exhibit significant positive log odds for time weighted fast food exposure, suggesting complex exposure interactions. In relation to BMI, exposure measures concur with expectations that access to calorie dense food is a key influencer of obesity (Tremblay and Willms, 2003). Access to restaurants is non-significant, possibly suggesting a difference in quality versus fast food (Lachat *et al.*, 2012; Faber *et al.*, 2013; Penney *et al.*, 2017). As expected, access to sport and leisure facilities are negatively associated with BMI increase.

Cluster 7 (predominantly Out of home service) exhibits the highest significance of any cluster. Showing a positive association with BMI, this agrees with the wider research of a positive relationship between Out of home service and obesity (Hill and Peters, 1998; Tremblay and Willms, 2003; Penney *et al.*, 2017). This finding is consistent with research

suggesting involvement in food preparation increases the likelihood of meeting nutrient guidelines, however the significance level is low (p=.093) (Larson *et al.*, 2006).

There are numerous limitations in our study. Despite wide application, food surveys are limited by cost and time, substantiated with our lack of complete data (Biro *et al.*, 2002). Eligibility, a reduction of invites for the additional GPS study, out of country travel, data completeness and data linkage issues meant only subsample could be used in our study. Similarly, the available covariates were restricted – e.g. snack data had no temporal or quantity metadata and therefore was omitted. Socioeconomic status is also not considered which may influence both access and consumption choice, and has been associated with higher (but also different types of) away from home energy consumption (DeJong *et al.*, 2009; Lachat *et al.*, 2012; Penney *et al.*, 2017). Furthermore, dietary recall surveys feature measurement error, under-reporting and self-selection bias (Faber *et al.*, 2013; Satija *et al.*, 2015; Crowe *et al.*, 2018), however some error is assumed for every dietary recording method (Biro *et al.*, 2002).

In this study data preparation groups include generic features (e.g. Out of home service). Despite a general literature consensus that Out of home service is negative based on the premise that fast food is the main source, quality varies considerably between fast-food (large, cheap, low-quality) and sit-down restaurants (expensive, specifically sourced ingredients) (Lachat *et al.*, 2012; Faber *et al.*, 2013; Penney *et al.*, 2017). Similarly, home preparation infers healthiness, despite including food that may require little preparation (e.g. frozen meals). We also fail to consider ease, speed, price, or the number of people food is being prepared (all potential influencers of food choice) (Waterhouse *et al.*, 2005). Nonetheless, utilising preparation instead of food consumption is likely to reduce reporting bias as exact foods are withdrawn.

Diet is exposure and time dependent therefore an individual's behaviour can dramatically change (Widener and Shannon, 2014). Although GPS data are typically restricted to one week (e.g. Chaix et al. (2012); Scully et al. (2017)) our findings cannot be considered a ground truth for all Canadians aged 16-30 (Satija *et al.*, 2015). GPS data are widely accepted as supplementary data and bring clear value, but limitations include unstable signals, data processing and transfer, and variance in device quality (Shen and Stopher, 2014). Other spatial issues could also affect results. For example, we attribute environmental exposure to a 250m buffer around activity spaces in our models, accounting for direct sight. Other distances could be relevant, based on the type of facility, and issues like cumulative exposure are not accounted for.

Despite these limitations, there are real opportunities for using the presented methods with datasets in other regions and across other populations. The increase in high resolution location information, along with data from time use and dietary diaries provide exciting possibilities to conduct new studies that seek to merge data sources to provide a more holistic view of food-related behaviours and subsequent health outcomes.

5.5. Conclusion

The increase in obesity, high associated costs to Canadian healthcare, and predicted further caloric availability means understanding food preparatory behaviours for young adults in Canada is of great importance (Birmingham *et al.*, 1999; Barlow, McKee and Stuckler, 2018). Population surveillance is deemed necessary to explore the obesity epidemic in Canada (Katzmarzyk, 2002). Unfortunately, quantifying food preparation and consumption is extremely difficult and like most nutritional studies the only viable data collection is via survey. Our study is not nationally representative (An, 2016), however we are able to highlight problematic behaviours among young urban adults in North America. Our approach is novel in dietary behaviour research, and contributes information to the obesity epidemic in Canada, with a methodology that is transferable and scalable provided data availability which hopefully will be used to inform policy and change social norms.

Chapter 6 : An application of text mining for understanding the evolution of US obesity related policy (2001-2017)

Chapters 3, 4 and 5 have examined new forms of data that either have a residential location or a dynamic movement people focus, however, there is a lot of available data that extend beyond individuals and groups of people. This chapter instead highlights opportunities for using text data for further insight into Public Health. Text summaries (abstracts) of enacted obesity related bills are used to explore the focus of obesity policy, and how this varies geographically and temporally. Following a text mining approach, a range of methods are employed from word clouds to sentiment analysis to explore these data. The influences of obesity related policy enactment are also modelled using negative binomial mixed effect models. This chapter highlights the knowledge and context that is obtainable within another non-traditional data source that impacts upon health behaviours and exhibits clear benefit and potential for further application.

6.1. Introduction

US Obesity prevalence is one of the highest internationally, reflecting issues such as average calorie consumption per capita exceeding double recommended levels (Abelson and Kennedy, 2004; Hales *et al.*, 2017). Adult obesity trends have increased from 30.5% in 2001 to 39.6% in 2017 (Hales *et al.*, 2017). Morbid obesity, the most severe form of obesity has a current prevalence of 7.7% (National Institute of Diabetes and Digestive and Kidney Diseases, 2019) which may present further issues for future healthcare resources (World Health Organization, 2000b; Sturm, 2007; US Department for Health and Human Services, 2010; Kim and Basu, 2016). Driving obesity increases are significant changes to dietary patterns and environments (Hill and Peters, 1998). Nutrition environments (increased exposure and access to cheap high calorie food and drink) alongside physical inactivity (increased sedentary lifestyles) are cited as key determinants (Hill and Peters, 1998; World Health Organization, 2000b; Guthrie, Lin and Frazao, 2002; Sturm, 2007; US Department for Health and Human Services, 2010; Fong *et al.*, 2019).

The significance of the high prevalence of obesity has seen considerable policy effort aimed at tackling and reversing trends, however success has been limited. The 2001 Surgeon-General's *Call To Action* sought collaboration to improve (i.e. reduce and prevent) the obesity situation in the US (US Department for Health and Human Services, 2001); however, despite providing guidelines, response to the call has been argued as weak (Abelson and Kennedy, 2004). The proceeding 2010 Surgeon-Generals vision stated that

"the prevalence of obesity, obesity-related diseases, and premature death remains too high" (US Department for Health and Human Services, 2010, p1). At the most basic level individuals need to make healthier choices (e.g. exercise, reduce sugar intake) (US Department for Health and Human Services, 2010). Individual freedom of choice cannot be regulated; however, state policy and regulations can act to improve environments (McKinnon *et al.*, 2009). With the determinants of obesity complex and numerous, the solutions will need to cover many different stakeholders (e.g. individuals, private sector and communities) and many facets require action (e.g. agriculture, advertising and community programmes) (McKinnon *et al.*, 2009; Gortmaker *et al.*, 2011). In both Surgeon-General reports (US Department for Health and Human Services, 2001, 2010) the numerous opportunities for legislation within society are clearly outlined, ranging from education (e.g. food service and curriculums) to the environment (e.g. advertising and parkland).

Understanding how features influence the enactment of policies enables greater depth in the knowledge of successful obesity response. Current research includes exploratory research (e.g. understanding the relationship the built environment and obesity) (Lopez-Zetina, Lee and Friis, 2006); evaluating the impact of individual policies (e.g. healthy eating or physical education) (Cawley and Liu, 2008; Eyler *et al.*, 2012; Lankford *et al.*, 2013); critical analysis and frameworks for future polices (e.g. coordinating food environment interventions) (McKinnon *et al.*, 2009; Sacks, Swinburn and Lawrence, 2009; Gortmaker *et al.*, 2011); and exploring the drivers of bill enactment (i.e. bills passing through legislation), for example, to understand obesity prevention legislation (Boehmer *et al.*, 2007; Hersey *et al.*, 2010; Donaldson *et al.*, 2015). The majority of these studies routinely focus on themes such as childhood obesity, addressing taxes, exposure and media image (e.g. school environments account for a considerable quantity of obesity bills) (Boehmer *et al.*, 2007; Cawley and Liu, 2008; Frieden, Dietz and Collins, 2010; Lankford *et al.*, 2013). A lack of focus on adults is noticeable, and there is a need for further examination of why particular bills are enacted versus others (Donaldson *et al.*, 2015).

Previous research has found bill topic and the semantics of policy phrasing important for understanding obesity policy trends. Soda tax and nutrition labelling are notable strategies that are increasingly prominent in legislation (Eyler *et al.*, 2012; Lankford *et al.*, 2013; Donaldson *et al.*, 2015). The language used in bills has also proven important. Specific emphasis within bills has progressed over time (e.g. increased detail for the management of vending machines) (Lankford *et al.*, 2013). Support for bills is also influenced by perceptions of the cause of obesity (i.e. whether the onus of obesity is an individual's fault) (Barry *et al.*, 2009; Donaldson *et al.*, 2015; Joslyn and Haider-Markel, 2019). Decision

makers are likely to avoid the possibility of creating feelings of discrimination by not supporting bills that contrast with their constituents' beliefs (Barry *et al.*, 2009; Donaldson *et al.*, 2015; Joslyn and Haider-Markel, 2019).

Research has typically not considered how phrasing might explain bill enactment rates or how the phrasing has evolved over time. Few studies have analysed the contents of bills (e.g. Cawley and Liu (2008); Lankford *et al.* (2013); Donaldson *et al.* (2015)), and studies which have are largely limited to qualitative analyses. Related studies have primarily focused on enactment rates of bills (particularly childhood obesity related) as an outcome (Donaldson *et al.*, 2015) and typically fail to consider bills texts as data. These approaches are restricted by short study durations and the ability to analyse and quantify increasing quantities of data (e.g. measuring trends of words terms over time). Utilising methods such as text analysis and the mining of large repurposed unstructured data (i.e. obesity policy abstracts), brings significant opportunities for improving the understanding of what influences obesity-related policy enactment. Taking advantage of this new form of data allows for novel exploration and contribution to knowledge focusing on the characteristics of policy abstracts; how bill enactment varies geographically; and what drives policy enactment over time.

The aim of this study is to investigate how utilising text data in the form of obesity-related policy abstracts can help inform our understanding of obesity-related bill enactment in the US.

6.2. Methods

6.2.1. Data

Data were obtained from the Centers for Disease Control and Prevention (CDC) who collate information on health-related legislation (Centers for Disease Control and Prevention, 2019b). An earlier and less comprehensive version of these data have been utilised in other studies (Hersey *et al.*, 2010; Lankford *et al.*, 2013). The dataset contains information of every bill enacted from 2001-2017 for all 50 states and Washington DC, the state and year bills were enacted in, the topic of a bill, and a text summary (abstract). We focused only on policy where the topic was defined as nutrition (n=1699), physical activity (n=1212) or obesity (n=1609). Full bills lack structure (e.g. considerable amounts of cross-referencing between bills and structure that would require removal before analysis) and would require individual searching and downloading from an archive (n=4520), therefore abstracts were instead utilised in this study. The term count of abstracts ranged from 5 (e.g. provision 5491 - Alaska 2003) to 281 (provision 6472 – Arkansas).

As obesity-related law is largely devolved from federal law with a state level legislative focus, we correspondingly focus analysis here at the state level (Boehmer *et al.*, 2007). Previous studies have used enactment rate as the outcome (Boehmer *et al.*, 2007; Hersey *et al.*, 2010; Donaldson *et al.*, 2015), however our dataset only provided full coverage for enacted policies. We therefore use enactment count per state per year as our outcome, calculated for all enacted policies, as well as stratifying by health topic (nutrition, obesity and physical activity).

Covariates were derived from literature and included measures of health, political majority and demographics and were combined by year and state. Our health measures included the percentage of individuals with obesity (defined as body mass index of 30+) and the percentage of people who exercised in the last month, which were both derived from the CDCs Behavioural Risk Factor Surveillance System (BRFSS) (Centers for Disease Control and Prevention, 2019a). The influence of obesity on health has varied in studies. Obesity prevalence has been found inversely associated with overall obesity law enactment at state level (Marlow, 2014), and obesity rates have been found unpredictive of obesity legislation in short duration studies (i.e. 3-year data) (Eyler *et al.*, 2012; Donaldson *et al.*, 2015). We explore the impact of these measures over a much wider time frame.

The effects of state majority political party varied in previous enactment research. Changes in state majority party has been found influential of law enactment (Cawley and Liu, 2008), however state political party association has been found irrelevant within the context of obesity prevention policy (2011-13) (Donaldson *et al.*, 2015). We use state political majority in the last election, as well as president at the time as our political measures (The U.S. National Archives and Records Administration, 2019).

Our socioeconomic predictors included yearly median income estimates from the US Census Bureau's annual social and economic supplements of their current population survey (US Census Bureau, 2019). Whilst previous studies have included a measure for income (e.g. Eyler *et al.* (2012); Marlow (2014)) limited effects have been observed (e.g. on state level obesity related legislation enactment) (Marlow, 2014); our study extends to incorporate a much wider time frame.

6.2.2. Statistical analyses

A text mining approach was applied in order to extract insights from the policy abstracts. Text mining involves the use of specific data mining techniques that can be applied on text

data (Dörre, Gerstl and Seiffert, 1999; Delgado *et al.*, 2002). Text data is rich with information, however, its unstructured nature must be converted into a usable form (Raja *et al.*, 2008). For each abstract stop words (e.g. and, or, to) and numbers were removed, and words were stemmed to their root to remove duplication via tense (e.g. *schools* is stemmed to *school*). The ability to analyse unstructured data extends beyond basic analytics enabling deeper understanding and insight (Dörre, Gerstl and Seiffert, 1999; Delgado *et al.*, 2002; Raja *et al.*, 2008).

Five text mining methods were applied ranging from simple visualisations to more complex measures of word rarity and sentiment. Each method provides a specific measure relating to the structure of the text and they are used to attempt to understand the contents of obesity bill abstracts. First, word clouds are a visual method of displaying the frequency of term occurrence within documents which provide an overview of words and bigrams that are prevalent within the documents (Cidell, 2010). Within a word cloud, the larger the size of the text, the more frequently a term is used. Comparison clouds instead enable the difference in frequency of term usage to be quantified across groups of texts (Kopp, 2019). Colour highlights the group that terms are used most by, and the text size indicates the size of difference. These two methods are the most basic visualisation, but both highlight summary characteristics of the policy abstracts, enabling initial exploratory understanding of their content.

Term frequency – inverse document frequency (TF-IDF) is a measure of word rarity within texts (Leskovec, Rajaraman and Ullman, 2014; Silge and Robinson, 2017). We use TF-IDF to expand beyond the basic summary statistics displayed within word clouds and explore term usage in greater depth in the form of word rarity. The intent is to distinguish differences between policy topics by using TF-IDF to highlight the presence of specific and possibly specialist terms (Leskovec, Rajaraman and Ullman, 2014). To summarise the explanations of the TF-IDF calculation by Leskovec, Rajaraman and Ullman (2014), Silge and Robinson (2017) and Hamdaoui (2019), for each word term frequency is first calculated by dividing the number of times a word is used in a document by the overall count of words in that document. Inverse document frequency is then calculated by dividing the total number of documents in a group by the number of documents that contain the word and taking the log of this result. Finally, the term frequency is multiplied by the inverse document frequency to create TF-IDF, where larger values signify greater rarity.

Considering how words are used in sequence is also important for understanding text structure and key phrases. Exploring bigrams (pairs of words) instead of singular words is

one way to obtain this information, but we are also interested in combinations of words that are commonly used in sequence. We use Markov chains to extend the analysis and our understanding from word or bigram scores to wider network of connections. Markov chains, an implementation of graph theory, produce a network of nodes for terms commonly used in sequence; connecting lines with arrows detail the direction, and the transparency level emphasises the strength, of connection (Kiss, 1968; Silge and Robinson, 2017). From the words that make up a collection of documents, a probability is assigned for each bigram from each initial word, enabling a network to be created with probability thresholds for inclusion (Kiss, 1968; Silge and Robinson, 2017). For example, if we use the term *body* as our initial word of a bigram and the second term of *mass*. This bigram (*body mass*) has a large number of connections and therefore a high probability of *mass* to be used after the term *body*, therefore the two terms are connected with an arrow from *body* to *mass*. *Mass* is also commonly connected to *index* therefore a further connecting arrow links *mass* to *index* highlighting *body mass index*.

Finally, sentiment analysis allows the quantification of underlying opinions with documents to be mined and extracted (Liu, 2012; Silge and Robinson, 2017). We apply sentiment analysis to explore the presence of opinion with the abstracts, particularly focusing on whether this changes over time, whether this changes during transition between presidential administrations, and whether this aligns with enactment counts. This may indicate how the text contents of bills may relate to enactment. When applying this analysis, sentiments are assigned to every word with a document. Lexicons contain these measures of sentiment, which score words by category (e.g. positive or negative) and can be combined to create an overall sentiment score of a text, enabling comparison with similar documents (Silge and Robinson, 2017). The Bing lexicon for example contains a large dictionary of positive and negative terms (approximately 7000) however there is a greater proportion of positive words (Hu and Liu, 2004; Silge and Robinson, 2017).

To explore the influences of obesity related policy enactment negative binomial mixed effects models were applied. These models enable the use of this longitudinal data and has previously been applied in this context for shorter duration studies (Eyler *et al.*, 2012; Donaldson *et al.*, 2015). The outcome is enactment count per state. The fixed effects are political (president and state political party), health (percentage population with BMI 30-98.9 and percentage population who exercised in the last month) and demographic (median income). The random effect is year.

Analysis was performed in R (R Core Team, 2014) using the wordcloud (Fellows, 2018), tidytext (Silge and Robinson, 2016), ggplot2 (Wickham, 2016), igraph (Csardi G, 2006), ggraph (Pedersen, 2018), tidyverse (Wickham, 2017), tmap (Tennekes, 2018) and lme4 (Bates *et al.*, 2015) packages. The regression summary table was produced using the stargazer package (Hlavac, 2015).

6.3. Results

6.3.1. What is included in US obesity-related policy abstracts?

Figure 6.1 provides a visual quantification of term usage within enacted obesity related bill summaries. Term prevalence across all enacted policy abstracts is visualised by individual term (figure 6.1.a) and combination of terms (i.e. bigrams, figure 6.1.b). Within figures 6.1.a-b. the size and colour gradient of the terms details their frequency of use (i.e. small light words are used less frequently than large dark words, with large dark words being the most common). The most common individual term is *school*, whereas the most common pair of words (bigram) is *physical activity* (*physic activ*). We see that *school* is present in bigrams (e.g. *public school*, *school district*, *school program*). This would suggest a focus of obesity related policies on children. *Physical* (*physic*) and *activity* (*activ*) are prevalent as individual words, however as a pair these words are much more frequent. We also see themes such as *task force*, *development* (*develop*) and *tax* are also common within policy abstract text.

Comparison clouds allow the variance by theme to be observed (figure 6.1.c-d). Here the frequency is again shown by size but instead relates to the difference in usage between topics. Colour instead highlights the policy theme (e.g. Nutrition is blue) and the size indicates the topic the term is used most in. For example, *food* is used more within Nutrition bill abstracts, whereas *physic* within Physical Activity abstracts. The most common term *school* (and similarly *education*) is the most prominent in all topics, however *physical activity* is more frequent for bigrams because of the frequency of use in physical activity policy abstracts. *Food* has the largest variance of frequency for any term (under nutrition). Nutrition policy abstracts has higher variance of usage for agriculture and nutrition terms, whereas for physical activity policy abstracts *education*, *transport* and *development* appear as specific keywords. Comparably *education* is common across all topics, however *transport* is only common for physical activity. Despite obesity bill abstracts having the highest average word count per abstract (51; nutrition is 46 and physical activity is 48), there is much less specific frequency variance of terms. Nonetheless obesity sees a focus on *weight gain*, *prevention* and *obesity*.

Bigrams (figure 6.1.b) show that *physical activity* is frequent for all topics (and is the most frequent bigram if stratified by topic), however the highest variance of usage comes within physical activity policy abstracts (figure 6.1.d). Bigrams with comparably higher use for physical activity policy abstracts are again themed around education and transport (e.g. *transit oriented* or *health education*). Prevalence of agriculture specific themes are found in nutrition policy abstracts (e.g. *farmers market* and *agricultural products*). Obesity abstracts lacked bigrams that are comparably used more than other topics, suggesting difficulty in distinguishing obesity bill abstracts compared to the other topics.

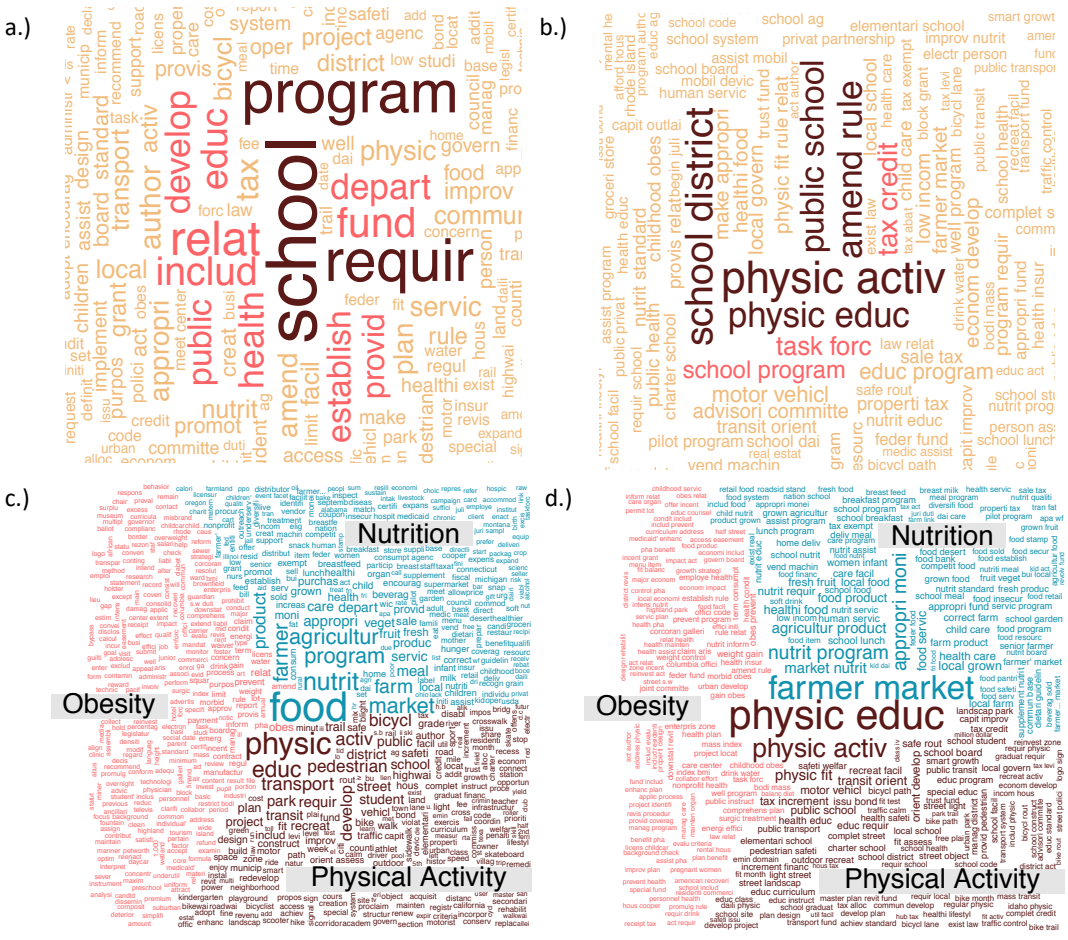

Figure 6.1. Abstract summary term statistics for a) word cloud of term frequency; b) bigram cloud of frequency for combination of words; c) word comparison cloud of frequency usage; d) bigram comparison cloud of frequency for combination of words

Figure 6.2. develops the information from figure 6.1. by instead focusing on word rarity by focusing on terms that are specifically used by each separate topic. Term frequency - inverse document frequency was calculated for both terms and bigrams by health topic (figure 6.2.). The y axis displays the word or bigram, and the x axis is the rarity score. Only the highest scoring words and bigrams are including for each topic. For example, for Nutrition terms *breastfe* (*breastfeeding* stemmed) has the highest score and therefore is the rarest term. This

enables unique themes to be identified; for example, nutrition policy abstracts see some focus around *breastfeeding*, whereas obesity policy abstracts see a surgery focus, and physical activity policy abstracts include government programmes (e.g. *FAPE* – Free Appropriate Public Education). Bigrams are noticeably different; nutrition sees numerous uses of *farm*, obesity sees direct mention of obesity (e.g. *morbid obesity*, *weight gain*) as well food delivery, whereas physical activity sees the involvement of incentives (e.g. *bike month*).

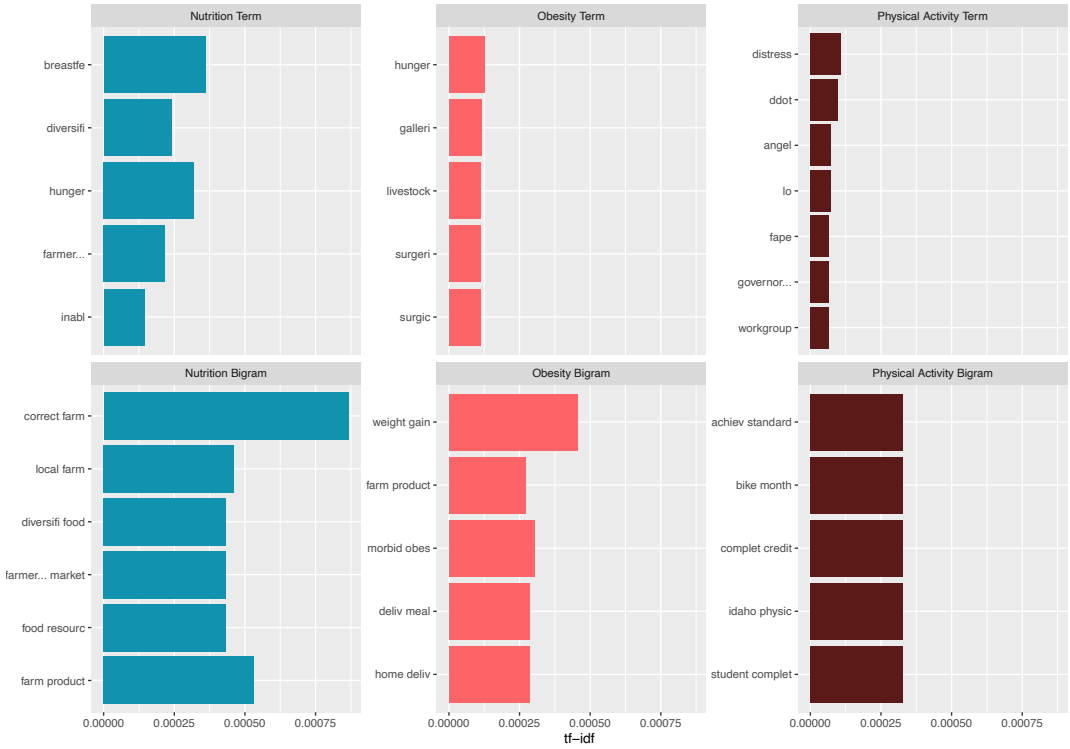

Figure 6.2. Visualisation of the rarity of words and bigrams (combinations of words) for each obesity topic (using term frequency - inverse document frequency)

Figure 6.3. extends the text analysis to focus on at connections of words. This allows us to not only focus on frequency, but also word parings and their connections to highlight key words or phrases that are frequently used in sequence using a Markov chain. Figure 6.3. visualises word pairings with a frequency greater than 50 times across all abstracts. The words are represented by nodes and arrows show the direction of connection (e.g. *rules* follows *amends*, so in text this would read *amends rules*); the darker the arrow the greater the frequency. Key connector words are highlighted e.g. *Physical* which is connected to a number of terms: *activity*, *fitness*, *education*. Similarly, *education* follows *health*, *physical*, and *nutrition*. *School* pro/precedes terms, explaining the high prevalence within the word cloud however low for bigram word cloud (e.g. *school district*, *school lunch*, *public school*). Contrastingly some individual phrases are also observed with no other connections (e.g.

childhood obesity or *vending machines*). Multiple connections are also observed (e.g. *amends rules relating*, *healthy food products/service*, *body mass index*, and *public transit-oriented development*). These results suggest the presence of regular buzzwords within policy summaries highlighting themes that are consistent in enacted bills. It may be that these themes are of key interest to policy, and therefore the high prevalence may relate policy addressing these issues are enacted. Knowing this information may also help decide future policy themes by understanding the contents of successful policy abstracts.

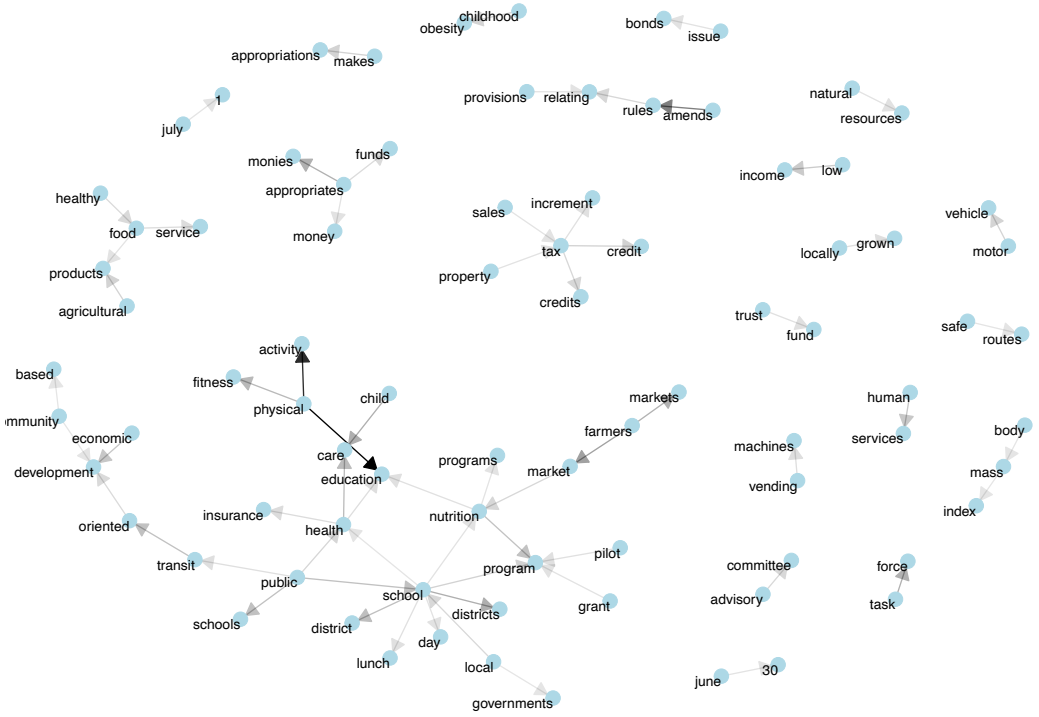

Figure 6.3. Visualisation of connecting words appearing more than 50 times (using a Markov Chain)

6.3.2. How does obesity-related policy vary by state?

Between 2001 and 2017, 4520 obesity-related bills were enacted. The mean number of policies enacted per state was 89, with a range 12-274. The lowest overall count came from South Dakota (12), Nebraska (21), Kansas (23), Wyoming (24) and Montana (26). The highest enacted counts were exhibited in the most populous states of California (most populous; 274), Texas (second most populous; 240), and Illinois (fourth most populous; 224). States with significantly smaller populations (e.g. Rhode Island and Arkansas) still enacted a higher count of bills (143 and 152 respectively).

Stratifying by health topic, 1699 bills were enacted for nutrition, 1609 obesity and 1212 physical activity. The lowest state mean was for physical activity (24), with higher means for

nutrition (33) and obesity (32). The largest range was seen for physical activity bills, boasting the both the highest (California = 98) and lowest (South Dakota = 1) counts by health topic. Nutrition and Obesity have similar ranges, with South Dakota again the lowest (5 and 6), however Illinois had the highest count for nutrition (90) and Texas for obesity (98). California, Texas and Illinois, boasted the highest counts for all three topics.

Figure 6.4.a. shows how counts vary geographically. The lowest overall count came from states in bordering between the Mountain and Northern Midwestern regions (South Dakota, Nebraska, Kansas, Wyoming and Montana). Texas and California (most populous and largest by area) behave differently from all neighbouring states, with significantly higher enactment counts. Similarly, Illinois differs from its neighbours, with Wisconsin and Indiana exhibit some of the lowest overall counts. With the exception of New Hampshire, states in New England and Middle Atlantic are above average.

Varied emphasis on health topics is observed (Figure 6.4.b-d). At regional level this variance is most noticeable in the North-eastern states (Middle Atlantic and New England). Here New York exhibited a lower count for overall obesity-related bills, however had the fourth highest number of nutrition bills (a 45% share), whereas the neighbouring Pennsylvania has high nutrition enactment for the area and low enactment for other topics. The Mountain and Midwestern states enactment is better aligned and consistent. At state level examples of differing bill focus (or topic priorities) are highlighted by Texas exhibits a low focus (24.6%) on nutrition, whereas in Oregon 49.4% of policies enacted are nutrition-based bills.

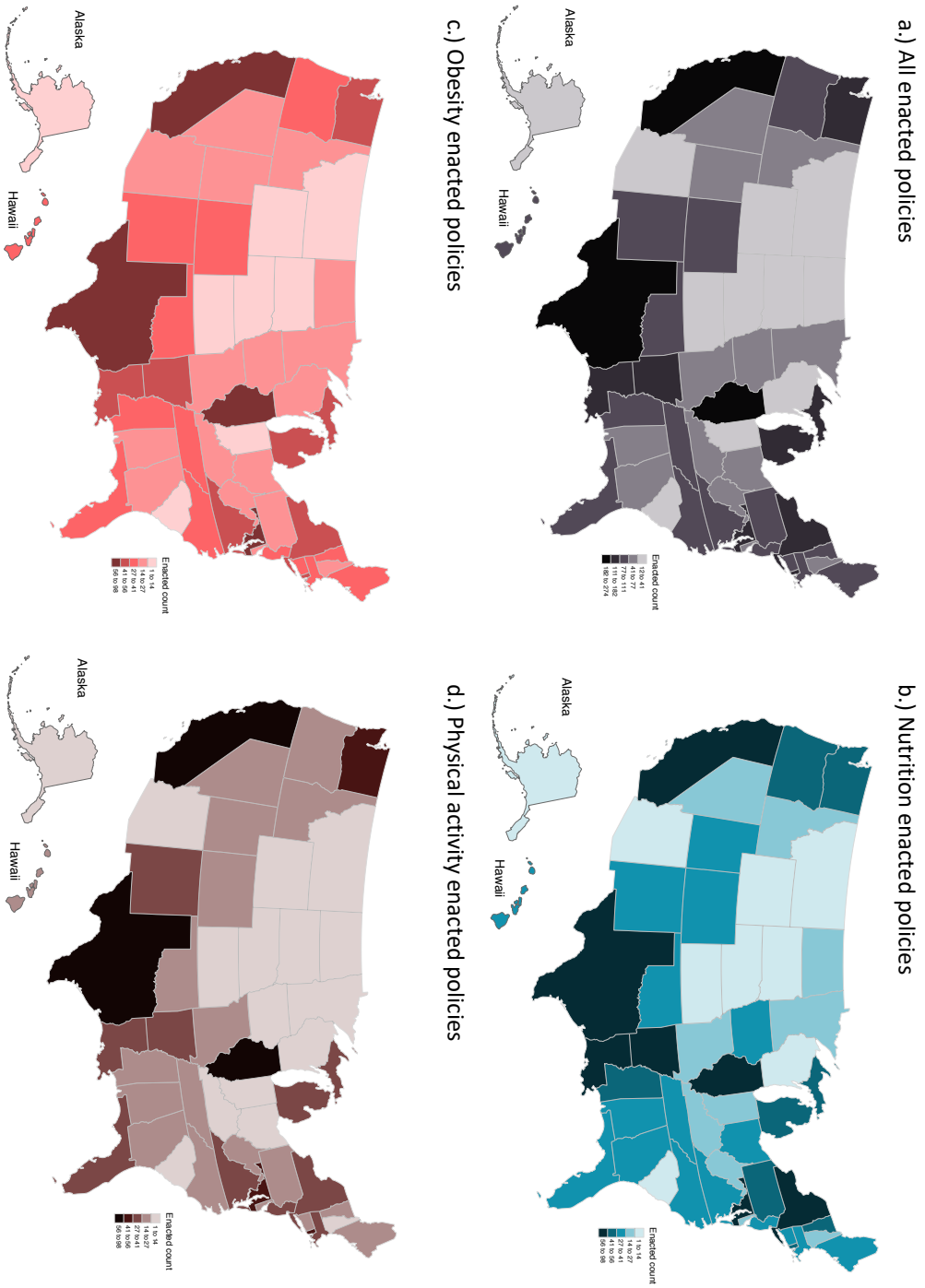

Figure 6.4. Policy enactment count by state a) all policy, b) nutrition, c) obesity, d) physical activity.

(note: boundary polygons were obtained from the spData R package (Bivand, Nowosad and Lovelace, 2019)).

6.3.3. How has obesity-related policy changed over time?

In the 17-year time frame of this data, there has been three presidents: Presidents Bush (Republican), Obama (Democratic) and Trump (Republican). 64% of enacted policies were in the Democratic Obama administration. The same topics are favoured in the Bush and Obama Administrations, with Nutrition (38% and 37%) and Obesity (34% and 37%) having the largest shares of enacted bills, and Physical activity the lowest (28% and 26%). Contrastingly however, during the first term of the Trump administration 71% of bills enacted were Nutrition, whereas 8% were obesity and 20% physical activity, showing an apparent shift in focus. At state level, Democrat states had the largest amount of bills enacted (61.6%). There was little difference by topic with nutrition between political party.

Figure 6.5. plots trends in summary statistics of policy abstract contents across the three political periods. Line plots show overall state averages, with smoothed conditional means used to ease interpretation. We stratify enactment count by political party of a state to examine if trends vary between Republican and Democratic states. We also use the measures of average sentiment (percentage of words that have positive sentiment) and average wordcount, to quantify the structure of policy abstracts. Wordcount is used to measure text structure which ranges from 5 to 281 words. We track this over time to explore how variation changes in relation to political periods (i.e. is more detail included in the final year before presidential change to push bills through).

Average enacted count remained fairly stable during the Bush administration, before a slight increase in the final few years. Early promise was seen in the Obama administration with enactment counts far higher than the previous administration (reaching 18 in 2010). Average count however drops considerably to less than four in 2013 and for the remainder of this administration. The first year of the Trump administration continues the low enactment (average less than one enactment per state). Comparing state political parties', Democratic states were consistently higher over the period than compared to Republican states. The gap between Democrat and Republican states was widest during the first half of the Obama administration, although trends were similar.

For the measures of text structure, we see a gradual increase in average word count, peaking in 2011 (60) during the Obama administration. Average word count then falls to similar levels as seen pre 2010, before a second rise in 2017 (60 during the Trump administration). Average word count is largely similar by political party, before separating post 2013 with declining Republican length. The percentage of words that have positive sentiment are held consistent between two and three throughout the majority of the 17-year period, however an

increase is witnessed to 3.5 in 2017 (Trump administration first year). The sentiment of Democratic states is predominantly slightly more positive however the difference is very small (typically one more positive word on average).

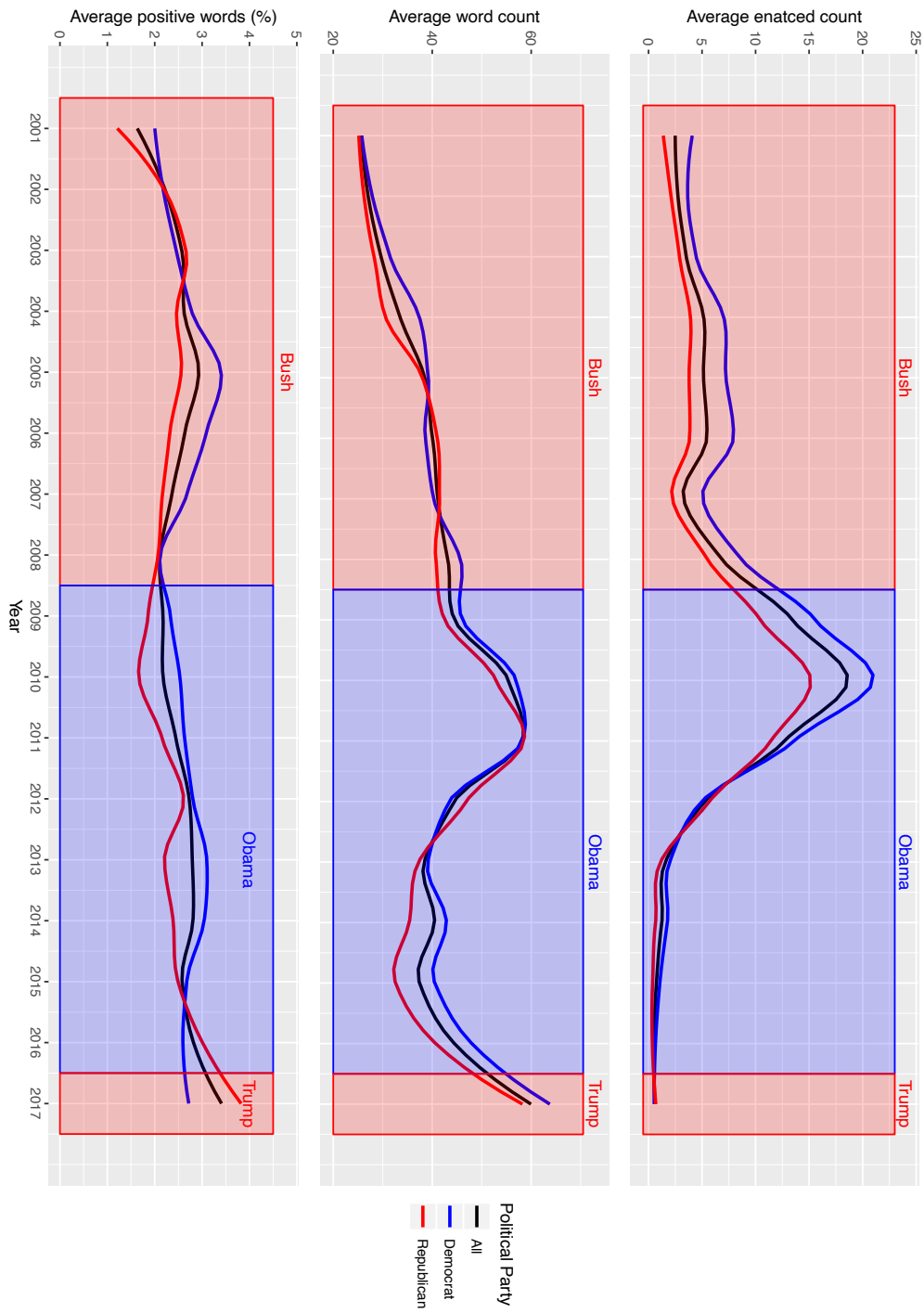

Figure 6.5. Smoothed conditional yearly state means for count of policies enacted (top), average abstract word count (middle) and percentage of words positive (bottom)

Individual word clouds for Democrat and Republican state bills are largely the same (results not shown), as well as being very similar to figure 6.1.a, suggesting little difference in term

frequency by party. Despite this similarity a comparison cloud highlights the differences in term use (figure 6.6.). We detect differences in term use and political party majority. Democrat states displayed more specific usage of transport and monetary related terms (e.g. *transport*, *transit*, *funds*, and *development*). Contrastingly, Republican state bills see higher focus on the keyword *physical*, as well as *activity*, *education* and *nutrition*.

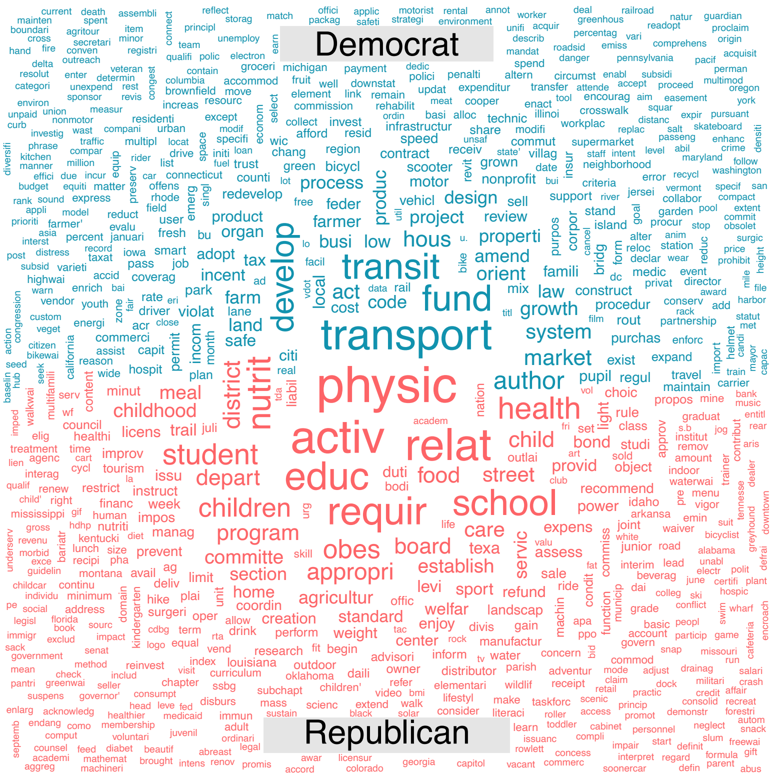

Figure 6.6. Comparison cloud of frequency of terms by political party

6.3.4. What influences obesity-related policy?

Negative binomial mixed effects models are used to explore the influences of obesity-related policy enactment. Our outcome is the count of policies enacted (per state, per year).

For all obesity-related policy, we found no significant effect for policy enactment during President Obama's or Trump's administration (in comparison to the Bush administration). Republican states were significantly negatively associated with an increase in enacted bills (IRR=.604 $p<.001$). For our health measures, state level obesity prevalence (IRR=.959, $p=.045$) and physical activity in the last month (IRR=.918, $p<.001$) are also negatively associated with an increase in enacted bills. Median income of a state was not associated to the number of policies enacted, suggesting little difference in policy behaviours between rich and poorer states. Our random effect, year, accounts for 52% of the variance explained. The coefficient of determination suggested fixed affects account for 15.3% of variation, increasing to 59.1% including random effects.

Each heath topic is also explored individually. For each theme, Obama administration is non-significant. While the Trump administration had non-significant effects for nutrition policy, we detect negative associations for obesity (IRR=.020, p=.025) and physical activity (IRR=.083, p=.023) compared to our reference category Bush administration. Republican states are negatively associated with an increase in enacted bills for nutrition (IRR=.518, p<.001), obesity (IRR=.642, p<.001) and physical activity (IRR=.626, p<.001).

State level obesity prevalence is only found significant in the topic of physical activity with a negative association (IRR=.953, p=.045). State level physical activity in the last month is negatively associated with an increase in enactment count across all topic models (IRR= .925 and .926, p<.001 for both). Median income ($1000s) is non-significant for nutrition, whereas it is found positively associated with enacted count for obesity (IRR=1.02, p=.028) and physical activity (1.018, p=.050).

The random effects show that temporal trends account for a 44% of variation for nutrition policy and 42% for physical activity policy. A much higher proportion of variation is explained by year for obesity (65%). The Marginal R2 is similar across our topic models (15%-19%). Conditional R2 however noticeably increases for Obesity (71%), with 18% more variance being explained by combined effects, highlighting the temporal influence being greater for obesity policy.

Table 6.1. Results of negative binomial mixed effects regression models analysing state obesity related policy enactment (n= 866)

Variable	*Dependent variable:* Count of Enacted policies for:			
	Model 1 All	**Model 2 Nutrition**	**Model 3 Obesity**	**Model 4 Physical Activity**
Fixed effects				
President Bush	Reference	Reference	Reference	Reference
President Obama	1.015 (.977)	1.230 (.624)	.492 (.348)	1.069 (.890)
President Trump	.121 (.051)	.238 (.115)	.020 (.025)	.083 (.023)
Democratic state	Reference	Reference	Reference	Reference
Republican state	.604 (<.001)	.518 (<.001)	.642 (<.001)	.626 (<.001)
Obesity	.959 (.045)	.972 (.138)	.979 (.377)	.953 (.045)
Physical activity	.918 (<.001)	.925 (<.001)	.926 (<.001)	.926 (.001)
Median Income ($1000s)	1.015 (.060)	1.004 (.588)	1.020 (.028)	1.018 (.050)
(Intercept)	4.228 (<.001)	1.746 (.063)	1.457 (.477)	1.173 (.640)
Random effects				
σ^2	.88	.81	1.16	1.11
τ_{00} **Year**	.94	.64	2.12	.81
Intraclass Correlation Coefficient	.52	.44	.65	.42
Marginal R2	.153	.153	.188	.183
Conditional R2	.591	.526	.713	.527
AIC	4046.445	2851.357	2459.582	2442.782

Note: Incident rate ratio (p-value).

6.4. Discussion

Using data on US obesity-related policy and applying a Data Science approach has helped to identify valuable insights into what is included in obesity-related policy abstracts, and how this has changed between 2001 and 2017. Our study has several notable strengths. We analyse one of the longest time frames of obesity-related policy during a period of considerable political change. We extend previous research through utilising the unstructured text information within policy abstracts to examine the evolution of policy content. It provides a detailed investigation of trends in the language used in policy and moves beyond prior research that does not consider text data when exploring obesity bill enactment (e.g. Boehmer *et al.* 2008 and Donaldson *et al.* 2015). Our results demonstrate the

potential of utilising policy text summaries when exploring policy that can be applied within multiple contexts for understanding responses to obesity.

We demonstrate that keywords (frequency of use) within obesity related policy abstracts are primarily related to education and physical activity. Childhood obesity is a prevalent theme within policy and often aims to target the development of behaviours early in life (US Department for Health and Human Services, 2010). The higher usage of terms relating to childhood policy found is possibly indicative of lower adult obesity focus in other studies (e.g. obesity prevention policy) (Donaldson *et al.*, 2015). We also see themes such as *task force*, *development* (develop) and *tax* are also common within our policy abstracts, which are often favoured approaches due to feasibility of implementation (Boehmer *et al.*, 2007; Donaldson *et al.*, 2015).

Despite overall key words primarily relating to education and physical activity, we detect unique focus within topics (e.g. nutrition policy focusing on agriculture and physical activity infrastructure). This prevalence aligns with research suggesting themes such as healthier food provision are successful policy interventions for nutrition, whereas for physical activity themes such as transport infrastructure achieves greater impact (Mayne, Auchincloss and Michael, 2015). Obesity policy abstracts comparably lacked specific term usage compared to other topics, further highlighted by low TF-IDF rarity scores (except *weight gain*). Extending analysis to include word networks displays the presence common buzzwords through prevalent term linkage (e.g. *farmers markets* or *body mass index*).

Regional clusters (particularly in the Southern region) exhibiting high obesity prevalence have been displayed spatially in the US (Agovino, Crociata and Sacco, 2019). Our analysis suggests that population appears to play a greater role than space for enactment count, with the most populous states enacting the highest amounts of bills. Low counts appear to be located in the Midwestern region, however beyond this there is no clear spatial pattern.

Changes in state political party have been found to be influential of policy enactment. We find Democratic states enacted more policies in the 17-year period (61.6%) aligning with the party's political views of promoting government intervention (e.g. the provision of healthcare) (Blendon *et al.*, 2008). Modelling enactment count found Republican states significantly negatively associated with an increase in enacted bills for all topics. These findings align with previous research which has suggested that health and nutrition laws are less likely to pass within Republican states (Boehmer *et al.*, 2008; Cawley and Liu, 2008; Marlow, 2014). Notably however we detect a higher usage of the term *nutrition* within

republican policy abstracts where nutrition laws are suggested as less likely to be pursued (Cawley and Liu, 2008), and although count is higher, the share of policies enacted from each topic remains consistent across all three topics. Despite this, nutrition accounts for the highest amount of obesity related policy enactment (38% Democrat vs 37% Republican) suggesting it receives more intervention through policy than obesity and physical activity.

Visualising change over time displayed variance by presidential administration. The largest change is seen when the Obama administration was elected with a change in the national political party appearing to increase policy enactment. Despite this, our mixed effects model suggests no significant effect of the Obama administration for any topic. We however detect negative associations for obesity and physical activity policy enactment for Trump administration vs our reference category (Bush administration), indicative of the differences seen in enactment shares compared with other administrations (e.g. Obesity policy 26% lower for Trump vs Bush administration). 2017 is however the only year in the Trump administration was in office in this dataset. It is typically acknowledged that there is a shift in focus in re-election years, as term limits (which have been found predictive of enactment rates) can constrain long term policy (Eyler *et al.*, 2012). In contrast we find that average enactment counts generally align with previous years within the same term, and that the first year in term is low for both the Bush and Trump administration.

Modelling suggests that our health covariates have varied effects. We find obesity prevalence negatively associated with overall and physical activity policy enactment; however, the effect has low significance. No significant effect is found for nutrition and obesity policy. These contrasting effects appear to situate amongst varied influence of obesity seen within literature, highlighting the need for further exploration (Eyler *et al.*, 2012; Marlow, 2014; Donaldson *et al.*, 2015). Physical activity in the last month is however negatively associated with enactment across all topics. This suggests that as state level exercise participation increases, enactment count drops across all obesity related policy.

Using median income as a proxy for poverty, we find no significant effects for overall or nutrition policy, suggesting little difference between rich and poorer states. In contrast, we observed a positive association between obesity and physical activity, suggesting richer states enact more of these bills. We consider year as our random effect which accounts approximately half of the variation explained by each model. Studies with short durations note the lack of longitudinal data as a key limitation (Hersey *et al.*, 2010). We find when combined with fixed effects the coefficient of determination increases considerably, showing the noticeable influence of time and highlighting that our modelling approach is appropriate.

There are a number of limitations to this study. When examining keywords, we only consider terms and bigrams, which could be further extended to n-grams. Similarly, the amount of context derived from documents could be extended, however our research is exploratory in this area. We are limited by the data only containing full records for bills which passed (i.e. were enacted). The CDC stopped including information for introduced, vetoed and dead bills within this dataset from 2016 (Centers for Disease Control and Prevention, 2019b) and therefore we cannot derive enactment rates as we only have complete information for enacted bills. We also do not account for further potential influences of behaviour. There are many features beyond our measurement such as time spent focusing on re-election, that have previously been found predictive (Eyler *et al.*, 2012). We only consider bill abstracts and do not quantify the introduction of taxes (e.g. soda tax) which have been favoured over the introduction of restrictive bills (Thow *et al.*, 2010; Eyler *et al.*, 2012; Donaldson *et al.*, 2015). Quantifying the impact of taxes is difficult (e.g. farmers markets are funded by soda tax offsetting potential impacts) (Donaldson *et al.*, 2015). Despite bringing new context by focusing on topic as well as overall obesity related policy, we do not examine policy in further scope whereas individual bill themes are increasingly favoured due to feasibility (e.g. task forces) (Boehmer *et al.*, 2007; Donaldson *et al.*, 2015). Knock on effects are also apparent, where bills are passed in one state this leads to adoption in further states (e.g. vending machine bills in California) (Boehmer *et al.*, 2007). Funding has also been found associated with obesity law enactment (e.g. CDC funding in 2005 in Kentucky) (Hersey *et al.*, 2010).

6.5. Conclusion

This research utilised a data driven approach giving an understanding of obesity related policy enactment. The data contains 17 years' worth of all obesity related bills and the inclusion of text abstracts bring far more scope than any previous obesity policy study. The data driven approach allows text to be mined as data, bringing unprecedented context. We present a novel contribution to obesity policy knowledge, hopefully allowing further research to expand upon our approach and findings.

Chapter 7 : Conclusions

This chapter concludes this thesis by providing a summary of the findings, limitations and possible future extensions of this research. Research findings detail how each aim, objective and the overall research question are addressed. Limitations are considered to ensure robustness as well as the correct interpretation of this work. Future opportunities are highlighted, which may address the limitations and enable a continuation of the novel findings and approaches presented. Finally, the concluding statement draws this thesis to a close.

7.1. Research findings

This thesis displays several notable strengths of new forms of data in the emerging fourth phase of Science (Hey, Tansley and Tolle, 2009). The lack of applications within the wider research environment (due to data access, computational resources and methodological approach) is addressed by providing four novel studies involving new forms of data. The opportunities of these surrogate sources is clearly demonstrated within both the literature review and each of the quantitative applications contained, highlighting the greater depth, accuracy and new information obtainable within public health research (Hay *et al.*, 2005; Dummer, 2008). This section translates how these core findings relate to and answer my overall research question, aims and objectives.

7.1.1. Aim 1: Examine the contribution of new forms of data to health research.

In order to achieve this aim, an in-depth literature review was conducted (objective 1). New forms of data and big data were defined (objective 2), and those currently being used (objective 3) and available (objective 7) were considered. The promise of new forms of data in health research was evaluated (objective 4) with limitations explored helping shape the applications presented (objective 6). In each quantitative chapter the new forms of data available were discussed and exploratory analysis performed (objective 7). The contribution of each new form of data was considered throughout (objective 12).

Evident gaps within public health knowledge appeared when reviewing existing literature. Knowledge of self-medicating individuals beyond self-reported surveys is limited despite data (e.g. open prescription) being available. In the study of dietary behaviour, research is typically nutrition or accessibility focused; food preparation has seen little attention. Within policy enactment research there is a lack of understanding as to why bills are enacted. This thesis demonstrates how the use of new forms of data can benefit public health research as standalone (e.g. loyalty card records) and supplementary sources (e.g. GPS data and

surveys). The opportunities for repurposing existing data as well as acknowledging that data can come in a variety of forms (e.g. sequences or text data) is shown.

Loyalty card records contribute new information of the exposure of self-medication products. Pain relief and coughs and colds products were purchased by up to 75% of customers per LSOA, displaying wide exposure. Sun prep products contrasting displayed a low customer reach and clear differences by sex were observed (e.g. males purchasing less). New detail in the seasonality of minor ailments showed further detail beyond observed epidemiological trends (e.g. cough and cold purchasing in December) (Heikkinen and Järvinen, 2003). This data is significantly larger than other self-medication studies with the inclusion of loyalty records which brings objective purchasing information extending beyond predominantly self-reported information (e.g. (Green *et al.*, 2016)).

Viewing food diaries as sequences allows for behaviours to be mined enabling greater value with the resultant insights. The typology created highlights problematic behaviours of young adult Canadians for meal preparation (e.g. 30% of the sample do not eat breakfast). While current applications of linking surveys and GPS information do exist, their focus is predominantly exposure based (e.g. Chaix *et al.* (2012); Scully *et al.* (2017;2019); Widener *et al.* (2018)). Linking food preparation sequences and GPS based exposure, and considering how these factors link to health outcomes (like BMI) brings new information to the national and global concern of improving diet (An, 2016).

Text is acknowledged as a data source and highlights further detail of the determinants of bill proposal success. The prevalent policy theme of childhood obesity (US Department for Health and Human Services, 2010) is observed with the presence of key words (e.g. *education or physical activity*). Themes relating to more feasible intervention (e.g. *task force* or *tax*) were common (Boehmer *et al.*, 2007; Donaldson *et al.*, 2015). Including and acknowledging unstructured text information extends beyond studies in bill enactment (e.g. Boehmer *et al.* 2008 and Donaldson *et al.* 2015) highlighting opportunities for more detailed investigation into notable public health issues (e.g. obesity).

7.1.2. Aim 2: Explore how geographical context can supplement and improve the quality of information obtained from new forms of data.

Health outcomes are intrinsically spatial and therefore geographic context is an important consideration and contribution (Dummer, 2008). The wider research area of quantitative geography and the evolutions within the field are first considered within chapter 2 (objective

1), which details current applications of health data (objective 3) and the limitations present (e.g. ecological fallacy) (objective 6). The importance of spatial context is also seen in each quantitative chapter using exploratory analysis (objective 7).

The visualisation of enactment count by state enables the conveyance of information of variations in enactment by policy topic, as well as the similarities between neighbouring states in the US. As regional clusters have previously been found for obesity prevalence (e.g. (Agovino, Crociata and Sacco, 2019)) it is important to consider how policy may be influenced spatially; however, visualisations suggest population size appears to have a greater impact than space as the most populous states enact the most bills (e.g. California and Texas). Visualising enactment spatially facilitated greater depth in the extraction of insight.

By combining loyalty card records with residential location a North-South divide was displayed highlighting geographical inequalities as purchasing patterns appearing to follow existing deprivation metrics (e.g. Smith *et al.* (2015)). Problematically low purchasing of sun prep products was observed in coastal regions where UV radiation is highest (Kazantzidis *et al.*, 2015). In this form, geographic context yields valuable detail and has potential for determining locations to target policy, action or intervention.

Creating a time-weighted measure of exposure using the GPS traces allowed an extension of knowledge beyond resident based location. Results displayed no convincing presence of correlation between access to calorie-dense foods and using these establishments. Exposure was low for clusters predominantly eating out and high for clusters not. This increased detail of movement data enables a further and more accurate understanding of the complexities in the relationship between unhealthy eating and exposure to food environments.

The contribution of geographical information within statistical (or machine learning) modelling is also shown. Aggregated air quality measures are found as an important contextual predictor of purchasing. PM10, which is associated with respiratory issues (Charpin and Caillaud, 2017), exhibited positive correlations with hay fever purchasing. When predicting monthly purchasing, aggregated temperature ranked highly. Temperature correlates to the production of pollen (MetOffice, 2018a), whereas respiratory conditions are influenced by colder temperatures (Heikkinen and Järvinen, 2003). Exposure measures showed access to calorie dense food does influence BMI (Tremblay and Willms, 2003). State level sociodemographic measures were found as influential features of obesity-related policy enactment. A positive association was observed between obesity bill enactment and

median income suggesting richer states enact more bills. The geographic context within these measures highlight its importance within research for further understanding health outcomes.

7.1.3. Aim 3: Identify applications of machine learning that can be applied in health research.

The high street retailer dataset presented is considerably larger than historically available minor ailment data (i.e. the transactions of 10 million customers). Despite data from the Canada Food Study only containing 396 individuals, more than a thousand contextual variables were contained from the combination of food diaries and GPS traces. The unstructured nature of the text summaries presented different complexities of size as contained were more than 4000 unstructured individual policy summaries. Further information (e.g. socioeconomic features) is similarly abundant and easily linked at census geographies. This size (of both observations and predictors) meant the selection of appropriate modelling techniques was important to maximise the opportunities these datasets present (in terms of computing resources and time). By assessing current applications (objectives 3 and 5) appropriate modelling techniques were selected (objective 8).

Machine learning was often presented as the most effective approach. Forecasting hay fever season is notoriously difficult (Davies and Smith, 1973). The state-of-the-art predictive algorithm employed enabled high performance on a limited historical times series. Further interpretation of these models is made possible using Partial Dependence and Accumulated Local Effects analysis; for example, hay fever purchasing was highest at optimal pollen release temperatures (MetOffice, 2018a). This dataset may offer cheaper and more efficient means of data collection than existing disease surveillance systems (Ginsberg *et al.*, 2009) (objective 12). The methodology employed could be applied as an early indicator of ailment incidence complementing existing methods (e.g. Santillana *et al.*, 2014) (objective 9).

The use of sequence analysis within food preparation study is innovative and provides further context for understanding dietary behaviours via the creation of an easily interpretable typology of food preparation. The typology demonstrates 5% of the sample is out of home service reliant which is problematic as eating out is strongly associated with fast food and higher caloric consumption (Lachat *et al.*, 2012; An, 2016; Penney *et al.*, 2017). The data mining approach is again scalable with a clear methodology presented (objective 9).

Despite the limited acknowledgement of text as data, the semantics of policy phrasing have previously been found important. For instance, soda tax and nutrition labelling are increasingly prominent in legislation (Eyler *et al.*, 2012; Lankford *et al.*, 2013; Donaldson *et al.*, 2015). Similarly, specific term usage is known (e.g. increasingly detailed physical activity programmes (Lankford *et al.*, 2013)). The methodological approach highlights the opportunities for text as data, bringing further detail of bill policy abstract keywords and how semantics change over time. This scalable and transferable methodology can enable research to extend in terms of depth, providing greater understanding of bill enactment (objective 9).

Each of the models presented in this thesis were statistically tested in order to analyse model performance (objective 10). Various stages of performance evaluation were passed (i.e. method selection, parameter tuning and performance metrics). The results were discussed and considered against existing research in each respective area (objective 11); an important step in evaluating the opportunities of each modelling approach (objective 12).

7.1.4. Overall research question: To what extent can new forms of data can help us better understand health outcomes or behaviours?

The aforementioned aims have proven important in answering the overall research question of this thesis. The two health themes studied are responsible for the use of considerable amounts of healthcare resources. Minor ailments have been scrutinised due to their extensive drain on the NHS (Pillay *et al.*, 2010; NHS England, 2017). Obesity represents a significant issue globally which is highlighted by the considerable amounts of targeted policy (Abelson and Kennedy, 2004) (e.g. US Department for Health and Human Services (2001; 2010)). The new forms of data used were vital in enabling these applications and have provided novel insights.

This thesis extends beyond the reach of many studies limited to traditional datasets. New information within minor ailment research is presented for a group that typically cannot be considered due to a lack of available data. The inclusion of sequence analysis and a typology of food preparation sequences, and the subsequent linking with GPS trajectories, is also a new approach for obesity-related research. One of the longest time frames of obesity-related policy during a period of considerable political change is studied. The use of unstructured text information and text analysis is an evolution in the study of obesity policy.

The value of new forms of data is clearly displayed by the new facets of research they enable. As repurposed data are generally cheaper and more efficiently collected the opportunities for application are considerable (Ginsberg *et al.*, 2009). Results demonstrate the potential of using such data within population health surveillance, forecasting, and the understanding of health behaviours; all of which are important for preventative frameworks (Bradley and Bond, 1995; Hughes, McElnay and Fleming, 2001). The inclusion of objective information (e.g. purchasing records, GPS movement) and focus on outcomes that are less sensitive to reporting biases (e.g. food source instead of actual meal) have applicability in national-level decision making. The applications may act as early indicators or complement existing methods (e.g. Santillana *et al.*, 2014).

Important also is the data driven approach undertaken to handle the quantities and complexities of these data which facilitates the extraction of unprecedented context. A clearly detailed, transferable and scalable methodology is presented in each quantitative chapter enabling others to understand and dissect the approaches used (objective 9). While new forms of data offer benefit, Data Science tools contribute significantly to the ability of using these data, enabling their potential to be achieved.

7.2. Limitations

While this thesis has focused on the novelty, relevance and importance of new forms of data and a Data Science approach within the study of health outcomes, there are several limitations.

While new forms of data and big data are increasingly desired by and available to researchers, the sensitive nature of individual level data create important and necessary barriers to access (Boyd and Crawford, 2012; Mahrt and Scharkow, 2013). Ethical approval and approval from the data provider alongside specific data usage and storage conditions are required. The sensitivity of individual level information means disclosive sample characteristics often must remain anonymous, constraining the ability to fully report the representativeness of the data. The rigorous application processes and very specific usage terms somewhat hamper the reproducibility of science; however, these procedures are necessary when utilising highly sensitive data.

Although data is linked from a number of sources within each quantitative chapter, the linkage process is simple and restricted to census geographies. The scale of aggregation was primarily determined by disclosure requirements and the scale of data that is being linked

(e.g. deprivation indexes or area level business counts); however, the aggregation levels were kept as detailed as possible (Subramanian et al., 2009). Data linkage could though occur at finer scales. Flanagan *et al.* (2019) were able to link data at the individual level, combining diagnosis histories with individual purchasing behaviour. While this would bring a greater level of granularity, Flanagan *et al.* (2019) had only 11 participants in their study. Joining information at the individual level would pose significant ethical concerns, particularly at the national scale.

Despite the regularity of aggregation to widely accepted and familiar scales (e.g. census geographies) (Wise, Haining and Ma, 2001; Duque, Ramos and Suriñach, 2007; Ginsberg et al., 2009), the potential presence of ecological fallacy and modifiable areal unit problem must be acknowledged (Fotheringham, Brunsdon and Charlton, 2000). The findings presented in this thesis can only be applied to the level of aggregation used (Openshaw, 1984a). The loyalty card analysis is residential location focused and therefore movement or spatial exposure is neglected (Hanigan, Hall and Dear, 2006). Using GPS data combats this limitation; however, the sample sizes and time frame are considerably restricted. While it is common for GPS data to be restricted to one week (e.g. Chaix et al. (2012); Scully et al. (2017)), and for studies to confine results to aggregate measures (e.g. census geographies), the findings cannot be considered as ground truth for all groups (Satija *et al.*, 2015).

The findings presented in these analyses are association-based and not casual in their inferences. The methods employed (e.g. regression) enable prediction and inference, however, they are fundamentally based on correlations between outcomes and predictors (Freedman, 1997). Tree ensemble methods (e.g. Random Forests and XGBoost) are based regression trees which "were developed for the purposes of prediction and classification, not causal inference" (Gass *et al.*, 2014, p6). Interpretation of the results should therefore be made carefully. Conclusive causes of health outcomes are difficult to obtain (Hill, 1965; Lucas and McMichael, 2005) as "the *cause* of illness may be immediate and direct, [or] it may be remote and indirect" (Hill, 1965, p295). Despite this, criteria for causation (as outlined by Hill (1965)) such as coherence, strength and consistency are displayed within this thesis and elude to clear conclusions (Freedman, 1997). The results presented are important but cannot be deemed causal.

Despite the presentation of new insights into facets of health that have rarely (if at all) been studied in this way, the results are representative and focus on demand side factors. For example, in chapters 3 and 4 it is vital that interpretation acknowledges that the data only contains purchasing information and is limited to approximately 20% of adults in England.

The insights only describe the behaviours of those within this sample. There is no context as to why or for whom the products are being purchased (in the case of significant others) or when or why these products are being consumed. While this limits the potential insights derived from analyses, sales data has been found correlated with physician records suggesting that they still have value (Magruder *et al.*, 2004).

These datasets are also limited in longitudinal scale. The high street retailer dataset is only available from 2012-2014 constraining the performance ability of modelling. Greater availability of historical data could improve the findings with more detail. The Canada Food Study data employed is limited by completeness of data (constrained by cost and time (Biro *et al.*, 2002)). GPS traces and preparation sequences are limited to one week. Eligibility, a reduction of invites for the GPS study, out of country travel, data completeness and data linkage issues meant only a subsample could be used in our study.

Reflecting on my personal experiences of performing applied data science and big data research, there is a lot of hype and excitement surrounding new forms of data. However, this size brings considerable limitations in the usability and speed in which analysis can be performed and insights gained. Such data requires specialist skills in data querying in order to convert raw data into useable forms. While the datasets employed within this thesis are huge in their raw form, it is necessary for dimensions to be reduced or data to be aggregated in order to make this interpretable and useable. For example, the loyalty card dataset was reduced from transaction level information to aggregations by both time and spatial scale, meaning that by the time data is fed into models the size had considerably reduced. Infrastructure adds additional complexity as in order to be able to use some of these data sets specific infrastructure is necessary. Whether that be secure facilities or access to big compute, there are further requirements that are necessary which can limit even accessing the data in the first place.

The methods employed (and necessary) to use this data are advanced and require both advanced understanding (e.g. machine learning models) and further skillsets that are heavily reliant on programming. Knowledge of relational databases and SQL is necessary to initially be able to query big relational data, whereas text mining instead moves analysis away from tabular data and into the realms of unstructured data. Proficiency in the software R (R Core Team, 2014) proved key for me to be able to apply the necessary machine learning approaches on the data. Method specific knowledge was also important, such as parameter tuning being vital when applying tree ensemble methods in order to achieve improved predictive performance. Specialist understanding is also necessary to be able to interpret

results. Combining data cleaning, modelling and interpretation means than researchers must play the role of data engineer, data scientist and data visualisation analyst, along with having expertise in their domain. These skillsets are difficult to acquire and are rare in unison meaning that the ability of researchers to actually apply new forms of data and big data are limited. When considering using such data researchers must ensure they have the necessary proficiency.

There are many opportunities for research to include new forms of data and big data, however, the limitations have shown the many facets to how it is only supplementary. There are many limitations of the data that must be considered, as well as ethics, infrastructure requirements and specialist skills necessary to use such data. These data are also untested in the long term, compared with data (such as censuses) which have a stable history. A key stumbling block is also the lack of data provenance, metadata and documentation that is provided with these datasets which constrains sample understanding and use. The excitement that surrounds big data such not taint existing and established data sources, as big data should only be considered supplementary to these.

7.3. Future opportunities

The research presented in this thesis highlights how new forms of data can be used to explore public health outcomes. There are several opportunities of extending this research, which may address the limitations of this research (objective 13).

Loyalty card data could be linked with actual health outcome data to understand how medicine purchasing relates to how individuals manage health conditions. Researchers such as Flanagan *et al.* (2019) have demonstrated the possibilities for this by linking cancer diagnosis and consumer data. Linking NHS prescription data would highlight areas of self-medication or prescription reliance. Alternatively, data could be combined with Hospital episodes statistics data from the NHS which would allow further investigation into issues such as paracetamol poisoning (Wazaify *et al.*, 2005; Morthorst *et al.*, 2018). Due to the sensitivity of the high street retailer data, hospital episodes statistics and individual level survey data, such applications would have severe ethical and confidentiality concerns that would need to be addressed before interesting multidisciplinary projects could be pursued.

Sequence analysis of food preparation and time weight exposure could be extended to multiple years of the Canada food study to understand whether increase in age changes dietary behaviour. Exposure measures could focus specifically by day instead of a weekly

aggregated measure (provided data availability). The use of GPS data could evolve to attempt to understand why exposure matters. For example, identifying the specific moments of the day individuals visit restaurants or fast food outlets, and whether such behaviours are associated with dietary behaviours or obesity. Complex effects of exposure are noted; however, the focus does not extend beyond associations of preparation and exposure. Alternatively, further application could link data from loyalty programmes or purchasing receipts, allowing understanding of what foods are eaten or the nutritional value of each meal. Greater detail on the measures and outcomes in this investigation will help produce more accurate models.

The application of text mining highlighted new opportunities for the study of health outcomes. There is clear opportunity for text mining to be applied beyond obesity-related policy abstracts and could be applied to further texts. Public health reports are one example, where understanding could be contributed of how perceptions of health outcomes have changed over time, and whether there are underlying drivers within the phrasing of health-related reports. There is a considerable opportunity to use text mining internationally. As demonstrated by Bautin, Vijayarenu and Skiena (2008), text analysis methods (e.g. sentiment analysis) can be applied to multiple languages allowing international research. Exploring policy responses to health outcomes (e.g. obesity) on an international scale would enable an understanding of effective policy response and may highlight the most appropriate or successful global approaches.

7.4. Concluding statement

The value of accessing and applying new forms of data in public health research is clearly shown in this thesis. Whether applications be in exploratory research investigation, or more specific in public health surveillance, there are clear opportunities and value for unlocking these new forms of data. As both minor ailments and obesity pose such extensive population level issues both nationally and globally, the insights and knowledge presented within the quantitative chapters of this thesis have clear relevance and impact in understanding these issues. The (Geographic) Data Science approach necessary in applying these datasets has been clearly documented and could be applied and transferred to similar public health themes. The findings may be used in public health planning and policy and will hopefully help to catalyse future research.

References

Abarca-Gómez, L. *et al*. (2017) 'Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016', *The Lancet*, 390(10113), pp. 2627–2642.

Abbas, J., Ojo, A. and Orange, S. (2009) 'Geodemographics - a tool for health intelligence?', *Public Health*, 123(1), pp. 35–39.

Abbott, A. and Tsay, A. (2000) 'Sequence analysis and optimal matching methods in sociology: Review and prospect', *Sociological Methods and Research*, 29(1), pp. 3–33.

Abelson, P. and Kennedy, D. (2004) 'Editorial: The Obesity Epidemic', *Science*, 304(5676), p. 1413.

Adamowski, J. *et al*. (2012) 'Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada', *Water Resources Research*, 48(1), pp. 1–14.

Adnan, M. *et al*. (2010) 'Towards real-time geodemographics: Clustering algorithm performance for large multidimensional spatial databases', *Transactions in GIS*, 14(3), pp. 283–297.

Agovino, M., Crociata, A. and Sacco, P. L. (2019) 'Proximity effects in obesity rates in the US: A Spatial Markov Chains approach', *Social Science and Medicine*, 220(January 2019), pp. 301–311.

Alboukadel Kassambara and Mundt, F. (2017) *factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.5*. Available at: https://cran.r-project.org/package=factoextra.

An, R. (2016) 'Weekend-weekday differences in diet among U.S. adults, 2003–2012', *Annals of Epidemiology*, 26(1), pp. 57–65.

Andrejevic, M. (2007) 'Surveillance in the Digital Enclosure', *The Communication Review*, 10(4), pp. 295–317.

Andreu-Perez, J. *et al*. (2015) 'Big Data for Health', *IEEE Journal of Biomedical and Health Informatics*, 19(4), pp. 1193–1208.

Anis, A. H. *et al*. (2010) 'Obesity and overweight in Canada: An updated cost-of-illness

study', *Obesity Reviews*, 11(1), pp. 31–40.

Apley, D. (2018) 'ALEPlot: Accumulated Local Effects (ALE) Plots and Partial Dependence (PD) Plots. R package version 1.1.' Available at: https://cran.r-project.org/package=ALEPlot.

Apley, D. W. (2016) 'Visualizing the effects of predictor variables in black box supervised learning models.', *arXiv*, 1612.08468, pp. 1–44.

Arribas-Bel, D. (2014) 'Accidental, open and everywhere: Emerging data sources for the understanding of cities', *Applied Geography*, 49(May 2014), pp. 45–53.

Ash, J., Kitchin, R. and Leszczynski, A. (2018) 'Digital turn, digital geographies?', *Progress in Human Geography*, 42(1), pp. 25–43.

Asthana, S. *et al*. (2002) 'Themes in British health geography at the end of the century: A review of published research 1998-2000', *Social Science and Medicine*, 55(1), pp. 167–173.

Baack, S. (2015) 'Datafication and empowerment: How the open data movement re-articulates notions of democracy, participation, and journalism', *Big Data & Society*, 2(2), pp. 1–11.

Ballester, F., Michelozzi, P. and Iñiguez, C. (2003) 'Weather, climate, and public health', *Journal of Epidemiology and Community Health*, 57(10), pp. 759–760.

Barlow, P., McKee, M. and Stuckler, D. (2018) 'The Impact of U.S. Free Trade Agreements on Calorie Availability and Obesity: A Natural Experiment in Canada', *American Journal of Preventive Medicine*, 54(5), pp. 637–643.

Barry, C. L. *et al*. (2009) 'Obesity metaphors: How beliefs about the causes of obesity affect support for public policy', *The Milbank Quarterly*, 87(1), pp. 7–47.

Bates, D. *et al*. (2015) 'Fitting Linear Mixed-Effects Models Using lme4.', *Journal of Statistical Software*, 67(1), pp. 1–48.

Bates, D. W. *et al*. (2014) 'Big data in health care: Using analytics to identify and manage high-risk and high-cost patients', *Health Affairs*, 33(7), pp. 1123–1131.

Bautin, M., Vijayarenu, L. and Skiena, S. (2008) 'International sentiment analysis for news and blogs', in *ICWSM 2008 - Proceedings of the 2nd International Conference on Weblogs and Social Media*, pp. 19–26.

Bealey, W. J. *et al*. (2007) 'Estimating the reduction of urban PM10 concentrations by trees within an environmental information system for planners', *Journal of Environmental Management*, 85(1), pp. 44–58.

Benton, D. and Parker, P. Y. (1998) 'Breakfast, blood glucose, and cognition', *American Journal of Clinical Nutrition*, 67(4), pp. 772–778.

Bethlehem, J. R. *et al*. (2014) 'The SPOTLIGHT virtual audit tool: A valid and reliable tool to assess obesogenic characteristics of the built environment', *International Journal of Health Geographics*, 13(1), pp. 1–8.

Bibby, P. and Shepherd, J. (2004) *Developing a New Classification of Urban and Rural Areas for Policy Purposes – the Methodology*, *National Statistics*. DEFRA, Stationery Office, London.

Birkin, M., Clarke, G. and Clarke, M. (2017) *Retail location planning in an era of multi-channel growth*. Edited by M. Birkin, G. Clarke, and M. Clarke. New York: Routledge.

Birmingham, C. L. *et al*. (1999) 'The cost of obesity in Canada', *CMAJ*, 160(4), pp. 483–488.

Biro, G. *et al*. (2002) 'Selection of methodology to assess food intake', *European Journal of Clinical Nutrition*, 56(2), pp. 25–32.

Bivand, R., Nowosad, J. and Lovelace, R. (2019) *spData: Datasets for Spatial Analysis*. *R package version 0.3.0*. Available at: https://cran.r-project.org/package=spData.

Blendon, R. J. *et al*. (2008) 'Health care in the 2008 presidential primaries', *New England Journal of Medicine*, 358(4), pp. 414–422.

Boehmer, T. K. *et al*. (2007) 'Patterns of childhood obesity prevention legislation in the United States.', *Preventing chronic disease*, 4(3), pp. 333–340.

Boehmer, T. K. *et al*. (2008) 'Preventing Childhood Obesity Through State Policy. Predictors of Bill Enactment', *American Journal of Preventive Medicine*, 34(4), pp. 333–340.

Boyd, D. and Crawford, K. (2012) 'Critical Questions for Big Data', *Information, Communication & Society*, 15(5), pp. 662–679.

Bradley, C. P. and Bond, C. (1995) 'Increasing the number of drugs available over the

counter: Arguments for and against', *British Journal of General Practice*, 45(399), pp. 553–556.

Brookes, D. *et al*. (2016) *Technical report on UK supplementary assessment under The Air Quality Directive (2008/50/EC), The Air Quality Framework Directive (96/62/EC) and Fourth Daughter Directive (2004/107/EC) for 2014*. London.

Brunsdon, C. and Comber, L. (2015) 'Introduction', in Brunsdon, C. and Comber, L. (eds) *An introduction to R for spatial analysis and mapping*. London: Sage, pp. 1–9.

Brunsdon, C. and Singleton, A. (2015) 'Preface', in Brunsdon, C. and Singleton, A. (eds) *Geocomputation: A Practical Primer*. London: Sage, pp. xiii–xiv.

Budiaji, W. (2019) *kmed: Distance-Based k-Medoids. R package version 0.2.0*.

Burgoine, T. *et al*. (2014) 'Associations between exposure to takeaway food outlets, takeaway food consumption, and body weight in Cambridgeshire, UK: Population based, cross sectional study', *BMJ*, 348(2014), pp. 1–10.

Butler, D. (2013) 'When Google got flu wrong', *Nature*, 494(7436), pp. 155–156.

Byrom, J. *et al*. (2001) 'Exploring the geographical dimension in loyalty card data', *Marketing Intelligence & Planning*, 19(3), pp. 162–170.

Caryl, F. *et al*. (2019) 'Socioeconomic inequalities in children's exposure to tobacco retailing based on individual-level GPS data in Scotland', *Tobacco Control*, 2019, pp. 1–7.

Casey, R. *et al*. (2012) 'Spatial accessibility to physical activity facilities and to food outlets and overweight in French youth', *International Journal of Obesity*, 36(7), pp. 914–919.

Cawley, J. and Liu, F. (2008) 'Correlates of state legislative action to prevent childhood obesity', *Obesity*, 16(1), pp. 162–167.

Centers for Disease Control and Prevention (2019a) *Behavioral Risk Factor Surveillance System Survey Data., Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2001-2017*.

Centers for Disease Control and Prevention (2019b) *CDC Nutrition, Physical Activity, and Obesity - Legislation, cdc.gov*. Available at: https://chronicdata.cdc.gov/Nutrition-Physical-Activity-and-Obesity/CDC-Nutrition-Physical-Activity-and-Obesity-Legisl/nxst-x9p4 (Accessed: 3 July 2019).

Chaix, B. *et al*. (2012) 'An interactive mapping tool to assess individual mobility patterns in neighborhood studies', *American Journal of Preventive Medicine*, 43(4), pp. 440–450.

Charpin, D. and Caillaud, D. (2017) 'Air pollution and the nose in chronic respiratory disorders', in Bachert, C., Bourdin, A., and Chanez, P. (eds) *The Nose and Sinuses in Respiratory Disorders: ERS Monograph*. European Respiratory Society, Sheffield, pp. 162–176.

Charreire, H. *et al*. (2014) 'Using remote sensing to define environmental characteristics related to physical activity and dietary behaviours: A systematic review (the SPOTLIGHT project)', *Health and Place*, 25(January 2014), pp. 1–9.

Chen, T. *et al*. (2018) *xgboost: Extreme Gradient Boosting. R package version 0.6.4.1*.

Chen, T. and Guestrin, C. (2016) 'XGBoost: A Scalable Tree Boosting System', in *KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, 13th–17th August*, pp. 785–794.

Cheng, T., Haworth, J. and Manley, E. (2012) 'Advances in geocomputation (1996-2011)', *Computers, Environment and Urban Systems*, 36(6), pp. 481–487.

Cidell, J. (2010) 'Content clouds as exploratory qualitative data analysis', *Area*, 42(4), pp. 514–523.

Cleveland, W. S. (2001) 'Data Science: An Action Plan for Expanding the Technical Areas of the field of statistics', *International Statistical Review*, 69(1), pp. 21–26.

Connelly, R. *et al*. (2016) 'The role of administrative data in the big data revolution in social science research', *Social Science Research*, 59(September 2016), pp. 1–12.

Cook, S. *et al*. (2011) 'Assessing Google Flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic', *PLoS ONE*, 6(8), pp. 1–8.

Corley, C. D. *et al*. (2010) 'Text and structural data mining of influenza mentions in web and social media', *International Journal of Environmental Research and Public Health*, 7(2), pp. 596–615.

Crooks, V. *et al*. (2018) 'Introducing the routledge handbook of health geography', in Crooks, V., Andrews, G., and Pearce, J. (eds) *Routledge handbook of Health Geography*. New York, NY, pp. 1–7.

Crowe, M. *et al*. (2018) 'Data Mapping From Food Diaries to Augment the Amount and Frequency of Foods Measured Using Short Food Questionnaires', *Frontiers in Nutrition*, 5(2018), pp. 1–9.

Csardi G, N. T. (2006) 'The igraph software package for complex network research, InterJournal, Complex Systems 1695. http://igraph.org'.

Datar, A. (2017) 'The more the heavier? Family size and childhood obesity in the U.S', *Social Science and Medicine*, 180(May 2017), pp. 143–151.

Davies, A., Green, M. A. and Singleton, A. D. (2018) 'Using machine learning to investigate self- medication purchasing in England via high street retailer loyalty card data', *PloS ONE*, 13(11), pp. 1–14.

Davies, R. R. and Smith, L. P. (1973) 'Forecasting the start and severity of the hay fever season', *Clinical & Experimental Allergy*, 3(3), pp. 263–267.

Deas, I. *et al*. (2003) 'Measuring neighbourhood deprivation: A critique of the Index of Multiple Deprivation', *Environment and Planning C: Government and Policy*, 21(6), pp. 883–903.

DEFRA (2017) 'Air Pollution in the UK 2016', *Annual Report 2016 Issue 2*, (September), p. 131. Available at: https://uk-air.defra.gov.uk/library/annualreport/viewonline?year=2016_issue_2.

DeJong, C. S. *et al*. (2009) 'Environmental and cognitive correlates of adolescent breakfast consumption', *Preventive Medicine*, 48(4), pp. 372–377.

Dekker, M., Verkerk, R. and Jongen, W. M. F. (2000) 'Predictive modeling of health aspects in the food production chain: a case study on glucosinolates in cabbage', *Trends in Food Science and Technology*, 11(4–5), pp. 174–181.

Delgado, M. *et al*. (2002) 'Mining Text Data: Special Features and Patterns', in *Pattern Detection and Discovery*. Berlin: Springer, pp. 140–153.

Denny, J. C. (2012) 'Mining Electronic Health Records in the Genomics Era', *PLoS Computational Biology*, 8(12), pp. 1–15.

van Dijck, J. (2014) 'Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology', *Surveillance and Society*, 12(2), pp. 197–208.

Dijkstra, W. and Taris, T. (1995) 'Measuring the Agreement between Sequences', *Sociological Methods & Research*, 24(2), pp. 214–231.

Doll, R. and Peto, R. (1981) 'The causes of cancer: Quantitative estimates of avoidable risks of cancer in the united states today', *JNCI: Journal of the National Cancer Institute*, 66(6), pp. 1192–308.

Donaldson, E. A. *et al*. (2015) 'Patterns and predictors of state adult obesity prevention legislation enactment in US states: 2010-2013', *Preventive Medicine*, 74(2015), pp. 117–122.

Dörre, J., Gerstl, P. and Seiffert, R. (1999) 'Text Mining: Finding Nuggets in Mountains of Textual Data', in *KDD '99 Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 398–401.

Drexler, M. (2014) *Big data's big visionary*. Available at: https://www.hsph.harvard.edu/magazine/magazine_article/big-datas-big-visionary/ (Accessed: 2 August 2019).

Duffey, K. J. *et al*. (2009) 'Regular Consumption from Fast Food Establishments Relative to Other Restaurants Is Differentially Associated with Metabolic Outcomes in Young Adults', *The Journal of Nutrition*, 139(11), pp. 2113–2118.

Dugas, A. F. *et al*. (2012) 'Google Flu Trends: Correlation with emergency department influenza rates and crowding metrics', *Clinical Infectious Diseases*, 54(4), pp. 463–469.

Dummer, T. J. B. (2008) 'Health geography: Supporting public health policy and planning', *CMAJ*, 178(9), pp. 1177–1180.

Duque, J. C., Ramos, R. and Suriñach, J. (2007) 'Supervised regionalization methods: A survey', *International Regional Science Review*, 30(3), pp. 195–220.

Efron, B. and Hastie, T. (2016) *Computer age statistical inference: algorithms, evidence, and data science*. Edited by B. Efron and T. Hastie. New York: Cambridge University Press.

Eichstaedt, J. C. *et al*. (2015) 'Psychological Language on Twitter Predicts County-Level Heart Disease Mortality', *Psychological Science*, 26(2), pp. 159–169.

Elliott, S. J. (1999) 'And the question shall determine the method', *Professional Geographer*, 51(2), pp. 240–243.

Elliott, S. J. (2018) '50 years of medical health geography(ies) of health and wellbeing', *Social Science and Medicine*, 196(C), pp. 206–208.

Eyler, A. A. *et al*. (2012) 'Patterns and predictors of enactment of state childhood obesity legislation in the united states: 2006-2009', *American Journal of Public Health*, 102(12), pp. 2294–2302.

Eyler, J. M. (2001) 'The changing assessments of John Snow's and William Farr's cholera studies', *Sozial- und Praventivmedizin*, 46(4), pp. 225–232.

Faber, M. *et al*. (2013) 'Presentation and interpretation of food intake data: Factors affecting comparability across studies', *Nutrition*, 29(11–12), pp. 1286–1292.

Felgate, M. *et al*. (2012) 'Using Supermarket Loyalty Card Data to Analyse the Impact of Promotions', *International Journal of Market Research*, 54(2), pp. 221–240.

Fellows, I. (2018) 'wordcloud: Word Clouds. R package version 2.6.' Available at: https://cran.r-project.org/package=wordcloud.

Figueiras, A., Caamano, F. and Gestal-Otero, J. J. (2000) 'Sociodemographic factors related to self-medication in Spain', *European Journal of Epidemiology*, 16(1), pp. 19–26.

Flanagan, J. M. *et al*. (2019) 'Self-Care Behaviors of Ovarian Cancer Patients Before Their Diagnosis: Proof-of-Concept Study', *JMIR Cancer*., 5(1), pp. 1–19.

Flegal, K. M. (1999) 'Evaluating epidemiologic evidence of the effects of food and nutrient exposures', *American Journal of Clinical Nutrition*, 69(6), pp. 1339-1344.

Foley, M. *et al*. (2015) 'The availability of over-the-counter codeine medicines across the European Union', *Public Health*, 129(11), pp. 1465–70.

Fong, M. *et al*. (2019) 'Modelling the Association between Core and Discretionary Energy Intake in Adults with and without Obesity', *Nutrients*, 11(3), pp. 1–13.

Fotheringham, S., Brunsdon, C. and Charlton, M. (2000) *Quantitative Geography: Perspectives on Spatial Data Analysis*. Edited by S. Fotheringham, C. Brunsdon, and M. Charlton. London: Sage.

Freedman, D. (1997) 'From Association to Causation via Regression', *Advances in Applied Mathematics*, 18(1), pp. 59–110.

Frieden, T. R., Dietz, W. and Collins, J. (2010) 'Reducing childhood obesity through policy change: Acting now to prevent obesity', *Health Affairs*, 29(3), pp. 357–363.

Fry, J. and Sibley, E. (1976) 'Evolution of data-base management systems', *ACM Computing Surveys*, 8(1), pp. 7–42.

Gabadinho, A. *et al*. (2011) 'Analyzing and Visualizing State Sequences in R with TraMineR', *Journal of Statistical Software*, 40(4), pp. 1–37.

Gahegan, M. (2012) *What is geocomputation? A history and outline*. Available at: http://www.geocomputation.org/what.html (Accessed: 31 July 2019).

Gale, C. G. *et al*. (2016) 'Creating the 2011 area classification for output areas (2011 OAC)', *Journal of Spatial Information Science*, 2016(12), pp. 1–27.

Gass, K. *et al*. (2014) 'Classification and regression trees for epidemiologic research: An air pollution example', *Environmental Health*, 13(1), pp. 1–10.

Ginsberg, J. *et al*. (2009) 'Detecting influenza epidemics using search engine query data', *Nature*, 457(7232), pp. 1012–1014.

Gittelsohn, J. *et al*. (2018) 'Specific Patterns of Food Consumption and Preparation Are Associated with Diabetes and Obesity in a Native Canadian Community', *The Journal of Nutrition*, 128(3), pp. 541–547.

Goodchild, M. F. (2010) 'Twenty years of progress: GIScience in 2010', *Journal of Spatial Information Science*, 2010(1), pp. 3–20.

Gore, R. J., Diallo, S. and Padilla, J. (2015) 'You are what you tweet: Connecting the geographic variation in America's obesity rate to twitter content', *PLoS ONE*, 10(9), pp. 1–16.

Gortmaker, S. L. *et al*. (2011) 'Changing the future of obesity: science, policy, and action', *The Lancet*, 378(9793), pp. 838–847.

Gotay, C. C. *et al*. (2013) 'Updating the Canadian obesity maps: An epidemic in progress', *Canadian Journal of Public Health*, 104(1), pp. 64–68.

Gray, N. J., Boardman, H. F. and Symonds, B. S. (2011) 'Information sources used by parents buying non-prescription medicines in pharmacies for preschool children', *International Journal of Clinical Pharmacy*, 33(5), pp. 842–848.

Green, M. A. *et al*. (2016) 'Investigation of social, demographic and health variations in the usage of prescribed and over-the-counter medicines within a large cohort (South Yorkshire, UK)', *BMJ Open*, 6(9), pp. 1–9.

Green, M. A. *et al*. (2018) 'Developing an openly accessible multi-dimensional small area index of "Access to Healthy Assets and Hazards" for Great Britain, 2016', *Health & Place*. Elsevier Ltd, 54(November), pp. 11–19.

Greenwell, B. M. (2017) 'pdp: An R Package for Constructing Partial Dependence Plots', *The R Journal*, 9(1), pp. 421–436.

Grimes, S. (2008) *Unstructured data and the 80 percent rule*. Available at: http://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/ (Accessed: 3 July 2019).

Guthrie, J. F., Lin, B. H. and Frazao, E. (2002) 'Role of food prepared away from home in the American diet, 1977-78 versus 1994-96: Changes and consequences', *Journal of Nutrition Education and Behavior*, 34(3), pp. 140–150.

Haining, R. (2014) 'Spatial statistics and the Analysis of Health Data', in Gatrrell, A. and Loytonen, M. (eds) *GIS and Health: GISDATA 6*. 6th edn. London: CRC Press, pp. 29–47.

Hajat, S. *et al*. (2001) 'Association between air pollution and daily consultations with general practitioners for allergic rhinitis in London, United Kingdom', *American Journal of Epidemiology*, 153(7), pp. 704–14.

Hales, C. M. *et al*. (2017) 'Prevalence of Obesity Among Adults and Youth: United States, 2015-2016.', *NCHS data brief*, 288(2017), pp. 1–8.

Hamdaoui, Y. (2019) *TF(Term Frequency)-IDF(Inverse Document Frequency) from scratch in python.*, *Towards Data Science*. Available at: https://towardsdatascience.com/tf-term-frequency-idf-inverse-document-frequency-from-scratch-in-python-6c2b61b78558 (Accessed: 29 February 2020).

Hamming, R. W. (1950) 'Error Detecting and Error Correcting Codes', *The Bell System Technical Journal*, 29(2), pp. 147–160.

Hammond, D. (2017) *2016 Canada Food Study User Guide and Codebook.*, *Waterloo, ON: University of Waterloo*.

Hammond, D., White, C. and Reid, J. (2017) *2016 Canada Food Study: Technical Report*,

Waterloo, ON: University of Waterloo.

Hanigan, I., Hall, G. and Dear, K. B. G. (2006) 'A comparison of methods for calculating population exposure estimates of daily weather for health research', *International Journal of Health Geographics*, 5(1), pp. 1–16.

Harris, R., Sleight, P. and Webber, R. (2005) 'Introducing Geodemographics', in *Geodemographics, GIS and Neighbourhood Targeting*. 7th edn. London: Wiley and sons, pp. 1–27.

Hart, S. *et al*. (1999) 'Are Loyalty Schemes a Manifestation of Relationship Marketing?', *Journal of Marketing Management*, 15(6), pp. 541–562.

Hartigan, J. A. and Wong, M. A. (1979) 'Algorithm AS 136: A k-means clustering algorithm', *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1), pp. 100–108.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) 'Introduction', in Hastie, T., Tibshirani, R., and Friedman, J. (eds) *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. New York: Springer, pp. 1–9.

Hay, S. I. *et al*. (2005) 'The accuracy of human population maps for public health application', *Tropical Medicine and International Health*, 10(10), pp. 1073–1086.

Heikkinen, T. and Järvinen, A. (2003) 'The common cold', *Lancet*, 361(9351), pp. 51–59.

Heintzman, N. D. (2016) 'A Digital Ecosystem of Diabetes Data and Technology: Services, Systems, and Tools Enabled by Wearables, Sensors, and Apps', *Journal of Diabetes Science and Technology*, 10(1), pp. 35–41.

Herland, M., Khoshgoftaar, T. M. and Wald, R. (2014) 'A review of data mining using big data in health informatics', *Journal of Big Data*, 1(1), pp. 1–35.

Hersey, J. *et al*. (2010) 'The Association between Funding for Statewide Programs and Enactment of Obesity Legislation', *Journal of Nutrition Education and Behavior*, 42(1), pp. 51–56.

Hey, T., Tansley, S. and Tolle, K. (2009) *The Fourth Paradigm: Data-Intensive Scientific Discovery*, *Data-Intensive Scientific Discovery*. Redmond: Microsft Research.

Hill, A. B. (1965) 'The environment and disease: association or causation?', *Journal of the*

Royal Society of Medicine, 58(1965), pp. 295–300.

Hill, J. O. and Peters, J. C. (1998) 'Environmental contributions to the obesity epidemic', *Science*, 280(5368), pp. 1371–1374.

Hlavac, M. (2015) *stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.*

Hu, M. and Liu, B. (2004) 'Mining Opinion Features in Customer Reviews', in *AAAI (Vol. 4, No. 4, pp. 755-760).*

Hughes, C. M., McElnay, J. C. and Fleming, G. F. (2001) 'Benefits and Risks of Self Medication', *Drug Safety*, 24(14), pp. 1027–1037.

Idrovo, A. J. (2011) 'Three criteria for ecological fallacy', *Environmental Health Perspectives*, 119(8), p. 332.

Ito, K. *et al*. (2015) 'The associations between daily spring pollen counts, over-the-counter allergy medication sales, and asthma syndrome emergency department visits in New York City, 2002-2012', *Environmental Health*, 14(1), pp. 1–12.

Jacquez, G. M. (2014) 'GIS as an Enabling Technology', in Gatrrell, A. and Loytonen, M. (eds) *GIS and Health: GISDATA 6*. London: CRC Press, pp. 17–28.

Jarrett, P., Sharp, C. and McLelland, J. (1993) 'Protection of children by their mothers against sunburn', *BMJ*, 306(6890), p. 1448.

Jensen, P. B., Jensen, L. J. and Brunak, S. (2012) 'Mining electronic health records: Towards better research applications and clinical care', *Nature Reviews Genetics*, 13(6), pp. 395–405.

Johnson-Taylor, W. L. and Everhart, J. E. (2006) 'Modifiable environmental and behavioral determinants of overweight among children and adolescents: Report of a workshop', *Obesity*, 14(6), pp. 929–966.

Jordan, H., Roderick, P. and Martin, D. (2004) 'The Index of Multiple Deprivation 2000 and accessibility effects on health', *Journal of Epidemiology and Community Health*, 58(3), pp. 250–257.

Joslyn, M. R. and Haider-Markel, D. P. (2019) 'Perceived causes of obesity, emotions, and attitudes about Discrimination Policy', *Social Science and Medicine*, 223(February 2019), pp. 97–103.

Katzmarzyk, P. T. (2002) 'The Canadian obesity epidemic, 1985-1998', *CMAJ*, 166(8), pp. 1039–1040.

Kayyali, B., Knott, D. and Kuiken, S. Van (2013) 'The big-data revolution in US health care : Accelerating value and innovation', *Mc Kinsey & Company*, 2(8), pp. 1–13.

Kazantzidis, A. *et al*. (2015) 'A modeling approach to determine how much UV radiation is available across the UK and Ireland for health risk and benefit studies', *Photochemical & Photobiological Sciences*, 14(6), pp. 1073–1081.

Kearns, R. A. (1995) 'Medical geography: Making space for difference', *Progress in Human Geography*, 19(2), pp. 251–259.

Kearns, R. and Moon, G. (2002) 'From medical to health geography: Novelty, place and theory after a decade of change', *Progress in Human Geography*, 26(5), pp. 605–625.

Keen, P. J. (1994) 'POM to P: useful opportunity or unacceptable risk?', *Journal of the Royal Society of Medicine*, 87(7), p. 422. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1294659/?page=2.

Kestens, Y. *et al*. (2012) 'Association between activity space exposure to food establishments and individual risk of overweight', *PLoS ONE*, 7(8), pp. 1–13.

Khoury, M. J. and Ioannidis, J. P. A. (2014) 'Big data meets public health', *Science*, 346(6213), pp. 1054–1055.

Kim, D. D. and Basu, A. (2016) 'Estimating the Medical Care Costs of Obesity in the United States: Systematic Review, Meta-Analysis, and Empirical Analysis', *Value in Health*, 19(5), pp. 602–613.

Kiss, G. R. (1968) 'Words, associations, and networks', *Journal of Verbal Learning and Verbal Behavior*, 7(4), pp. 707–713.

Kistemann, T., Dangendorf, F. and Schweikart, J. (2002) 'New perspectives on the use of Geographical Information Systems (GIS) in environmental health sciences', *International Journal of Hygiene and Environmental Health*, 205(3), pp. 169–181.

Kitchin, R. (2014) 'Big Data, new epistemologies and paradigm shifts', *Big Data & Society*, 1(1), pp. 1–12.

Kleisiaris, C. F., Sfakianakis, C. and Papathanasiou, I. V. (2014) 'Health care practices in

ancient Greece: The hippocratic ideal', *Journal of Medical Ethics and History of Medicine*, 7(6), pp. 1–5.

Kopp, B. (2019) *Creating a Word Cloud in R*, *RPubs*. Available at: https://rpubs.com/brandonkopp/creating-word-clouds-in-r (Accessed: 29 February 2020).

Kuhn, M. (2008) 'Building Predictive Models in R Using the caret Package', *Journal Of Statistical Software*, 28(5), pp. 1–26.

Kuhn, M. (2013) 'Introduction', in Kuhn, M. and Johnson, K. (eds) *Applied Predictive Modeling*. New York: Springer, pp. 1–16.

Kuhn, M. and Johnson, K. (2013) 'Regression Trees and Rule-Based Models', in Kuhn, M. and Johnson, K. (eds) *Applied Predictive Modeling*. New York, NY: Springer New York, pp. 173–220.

Kukkonen, J. *et al*. (2001) 'A semi-empirical model for urban PM10 concentrations, and its evaluation against data from an urban measurement network', *Atmospheric Environment*, 35(26), pp. 4433–4442.

Lachat, C. *et al*. (2012) 'Eating out of home and its association with dietary intake: a systematic review of the evidence', *Obesity Reviews*, 13(4), pp. 329–346.

Lampos, V., De Bie, T. and Cristianini, N. (2010) 'Flu detector-tracking epidemics on twitter', in Balcázar J.L., Bonchi F., Gionis A., S. M. (ed.) *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2010. Lecture Notes in Computer Science, vol 6323*. Berlin: Springer, pp. 599–602.

Langmuir, A. D. (1976) 'William Farr: Founder of modern concepts of surveillance', *International Journal of Epidemiology*, 5(1), pp. 13–18.

Lankford, T. *et al*. (2013) 'Analysis of state obesity legislation from 2001 to 2010', *Journal of Public Health Management and Practice*, 19(3), pp. 114–118.

Larson, N. I. *et al*. (2006) 'Food Preparation by Young Adults Is Associated with Better Diet Quality', *Journal of the American Dietetic Association*, 106(12), pp. 2001–2007.

Lazer, D. *et al*. (2014) 'The parable of google flu: Traps in big data analysis', *Science*, 343(6176), pp. 1203–1205.

Lebenbaum, M. *et al*. (2018) 'Trends in obesity and multimorbidity in Canada', *Preventive*

Medicine, 116(November 2018), pp. 173–179.

Lee, C. H. *et al*. (2017) 'Inappropriate self-medication among adolescents and its association with lower medication literacy and substance use', *PLoS ONE*, 12(12), pp. 1–14.

Lee, K., Agrawal, A. and Choudhary, A. (2013) 'Real-Time disease surveillance using twitter data: demonstration on flu and cancer', in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Chicago: ACM, pp. 1474–1477.

Leskovec, J., Rajaraman, A. and Ullman, J. D. (2014) 'Mining of massive datasets', in *Mining of Massive Datasets*. Second Edi. Cambridge: Cambridge University Press.

Liao, T. (2005) 'Clustering of time series data - A survey', *Pattern Recognition*, 38(11), pp. 1857–1874.

Liaw, A. and Wiener, M. (2002) 'Classification Regression by randomForest.', *R News*, 2(3), pp. 18–22.

Lilienfeld, D. (2007) 'Celebration: William Farr (1807 1883) an appreciation on the 200th anniversary of his birth', *International Journal of Epidemiology*, 36(5), pp. 985–987.

Lima, A. and Musolesi, M. (2012) 'Spatial dissemination metrics for location-based social networks', in *UbiComp'12 - Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pp. 972–979.

Lin, M. and Hsu, W. J. (2014) 'Mining GPS data for mobility patterns: A survey', *Pervasive and Mobile Computing*, 12(2014), pp. 1–16.

Liu, B. (2012) 'Sentiment Analysis: A Fascinating Problem', in Liu, B. (ed.) *Sentiment Analysis and Opinion Mining*. Cambridge: Cambridge University Press, pp. 7–14.

Longley, P. *et al*. (2011) *Geographic information science and systems*. Third Edit. Hoboken, New Jersey: John Wiley & Sons.

Lopez-Zetina, J., Lee, H. and Friis, R. (2006) 'The link between obesity and the built environment. Evidence from an ecological analysis of obesity and vehicle miles of travel in California', *Health and Place*, 12(4), pp. 656–64.

Lowe, C. *et al*. (1995) 'Effects of self-medication programme on knowledge of drugs and compliance with treatment in elderly patients.', *BMJ*, 310(6989), pp. 1229–1231.

Löytönen, M. (2001) 'Mobile devices and gis in health - new opportunities and threats.', in *Geographic Information Sciences in Public Health – 2001*. Sheffield, UK, p. 2.

Löytönen, M. (2014) 'GIS, Time Geography and Health', in Gatrrell, A. and Loytonen, M. (eds) *GIS and Health: GISDATA 6*. 6th edn. London: CRC Press, pp. 97–110.

Lucas, R. M. and McMichael, A. J. (2005) 'Association or causation: Evaluating links between "environment and disease"', *Bulletin of the World Health Organization*, 83(10), pp. 792–795.

Luo, G. (2016) 'Automatically explaining machine learning prediction results: A demonstration on type 2 diabetes risk prediction', *Health Information Science and Systems*, 4(1), pp. 1–9.

Lyon, D. (2014) 'Surveillance, Snowden, and Big Data: Capacities, consequences, critique', *Big Data and Society*, 1(2), pp. 1–13.

Maechler, M. *et al*. (2018) 'cluster: Cluster Analysis Basics and Extensions. R package version 2.0.7-1.'

Magruder, S. F. (2003) 'Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of human disease', *Johns Hopkins APL Technical Digest*, 24(4), pp. 349–353.

Magruder, S. F. *et al*. (2004) 'Progress in understanding and using over-the-counter pharmaceuticals for syndromic surveillance', *Morbidity and mortality weekly report*, 53(Suppl.), pp. 117–122.

Mahrt, M. and Scharkow, M. (2013) 'The Value of Big Data in Digital Media Research', *Journal of Broadcasting and Electronic Media*, 57(1), pp. 20–33.

Manovich, L. (2015) 'Trending: The Promises and the Challenges of Big Social Data', in Gold, M. K. (ed.) *Debates in the Digital Humanities*. University of Minnesota Press, pp. 460–475.

Marchand, B. (1974) 'Quantitative Geography: Revolution or Counter-Revolution?', *Geoforum*, 5(1), pp. 15–23.

Mark, D. M. (2000) 'Geographic information science: Critical issues in an emerging cross-disciplinary research domain', *URISA Journal*, 12(1), pp. 45–54.

Marlow, M. L. (2014) 'Determinants of state laws addressing obesity', *Applied Economics Letters*, 21(2), pp. 84–89.

Mauri, C. (2003) 'Card loyalty. A new emerging issue in grocery retailing', *Journal of Retailing and Consumer Services*, 10(2003), pp. 13–25.

Mayer, J. D. and Meade, M. S. (1994) 'A reformed medical geography reconsidered', *Professional Geographer*, 46(1), pp. 103-106.

Mayne, S. L., Auchincloss, A. H. and Michael, Y. L. (2015) 'Impact of policy and built environment changes on obesity-related outcomes: A systematic review of naturally occurring experiments', *Obesity Reviews*, 16(5), pp. 362–375.

McInnes, R. N. *et al*. (2017) 'Mapping allergenic pollen vegetation in UK to study environmental exposure and human health', *Science of the Total Environment*, 599–600(December 2017), pp. 483–499. Available at: http://dx.doi.org/10.1016/j.scitotenv.2017.04.136.

McKinnon, R. A. *et al*. (2009) 'Considerations for an Obesity Policy Research Agenda', *American Journal of Preventive Medicine*, 36(4), pp. 351–357.

McVicar, D. and Anyadike-Danes, M. (2002) 'Predicting successful and unsuccessful transitions from school to work by using sequence methods', *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 165(2), pp. 317–334.

MetOffice (2013) *Annual 2012*. Available at: https://www.metoffice.gov.uk/climate/uk/summaries/2012/annual (Accessed: 7 January 2019).

MetOffice (2018a) *How does the weather affect hay fever?* Available at: https://www.metoffice.gov.uk/health/public/pollen-forecast/how-does-the-weather-affect-hay-fever (Accessed: 28 February 2019).

MetOffice (2018b) *Pollen Forecast*. Available at: https://www.metoffice.gov.uk/health/public/pollen-forecast (Accessed: 11 January 2019).

MetOffice (2018c) *When is hay fever season in the UK?* Available at: https://www.metoffice.gov.uk/health/public/pollen-forecast/when-is-hayfever-season (Accessed: 17 December 2018).

Miles, A. *et al*. (2005) 'SunSmart? Skin cancer knowledge and preventive behaviour in a

British population representative sample', *Health Education Research*, 20(5), pp. 579–585.

Miller, H. J. (2010) 'The data avalanche is here. Shouldn't we be digging?', *Journal of Regional Science*, 50(1), pp. 181–201.

Miller, H. J. and Goodchild, M. F. (2015) 'Data-driven geography', *GeoJournal*, 80(4), pp. 449–461.

Molnar, C. (2019) *Accumulated Local Effects (ALE) Plot*, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Molnar. Available at: https://christophm.github.io/interpretable-ml-book/ale.html (Accessed: 28 February 2019).

Montastruc, J. L. *et al*. (1997) '[Pharmacovigilance of self-medication].', *Therapie*, 52(2), pp. 105–10.

Morris, C. J., Cantrill, J. A. and Weiss, M. C. (2001) 'GPs' attitudes to minor ailments', *Family Practice*, 18(6), pp. 581–585.

Morthorst, B. R. *et al*. (2018) 'Availability of Paracetamol Sold Over the Counter in Europe: A Descriptive Cross-Sectional International Survey of Pack Size Restriction', *Basic and Clinical Pharmacology and Toxicology*, 122(6), pp. 643–649.

Mullainathan, S. and Spiess, J. (2017) 'Machine Learning: An Applied Econometric Approach', *Journal of Economic Perspectives*, 31(2), pp. 87–106.

Murdoch, T. B. and Detsky, A. S. (2013) 'The inevitable application of big data to health care', *JAMA*, 309(13), pp. 1351-1352.

Murray, A. T. (2010) 'Quantitative geography', *Journal of Regional Science*, 50(1), pp. 143–163.

National Institute of Diabetes and Digestive and Kidney Diseases (2019) *Overweight & Obesity Statistics*, *niddk.nih.gov*. Available at: https://www.niddk.nih.gov/health-information/health-statistics/overweight-obesity (Accessed: 23 August 2019).

Naur, P. (1974) 'Concise Survey of Computer Methods.', *Sweden: Studentlitteratur*.

Nevalainen, J. *et al*. (2018) 'Large-scale loyalty card data in health research', *Digital Health*, 4(10), pp. 1–10.

Nguyen, Q. C. *et al*. (2016) 'Building a National Neighborhood Dataset From Geotagged

Twitter Data for Indicators of Happiness, Diet, and Physical Activity', *JMIR Public Health and Surveillance*, 2(2), pp. 1–16.

NHS England (2017) 'Items which should not be routinely prescribed in primary care: A Consultation on guidance for CCGs', *NHS England Gateway*, (2), pp. 1–48. Available at: https://www.engage.england.nhs.uk/consultation/items-routinely-prescribed/.

Nwosu, C. S. *et al*. (2019) 'Predicting Stroke from Electronic Health Records', in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–4.

ObesityCanada (2019) *Obesity in Canada*. Available at: https://obesitycanada.ca/obesity-in-canada/ (Accessed: 11 April 2019).

OECD (2008) *Handbook on Constructing Composite Indicators: Methodology and User Guide*. OECD publishing.

Office for National Statistics (2014) 'Pen Portraits for 2011 Area Classification for Output Areas', 1(April), p. 2014.

Office for National Statistics (2016) *National Statistics Postcode Lookup. Contains public sector information licensed under the open government license v3*. Available at: https://data.gov.uk/dataset/5d97ecf0-be29-4a13-8afb-377679f7bc99/national-statistics-postcode-lookup-may-2016.

Ogutu, J. O., Piepho, H. P. and Schulz-Streeck, T. (2011) 'A comparison of random forests, boosting and support vector machines for genomic selection', *BMC Proceedings*, 5(Suppl.), pp. 1–5.

Olson, D. R. *et al*. (2013) 'Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales', *PLoS Computational Biology*, 9(10), pp. 1–11.

Openshaw, S. (1984a) 'Ecological Fallacies and the Analysis of Areal Census Data', *Environment and Planning A: Economy and Space*, 16(1), pp. 17–31.

Openshaw, S. (1984b) *The modifiable areal unit problem*, *CATMOG 38*. Norwich: GeoBooks.

Orueta, J. F. *et al*. (2013) 'Predictive risk modelling in the Spanish population: A cross-sectional study', *BMC Health Services Research*, 13(269), pp. 1–9.

Panattoni, L. E. *et al*. (2011) 'Predictive risk modelling in health: Options for New Zealand and Australia', *Australian Health Review*, 35(1), pp. 45–51.

Park, H. S. and Jun, C. H. (2009) 'A simple and fast algorithm for K-medoids clustering', *Expert Systems with Applications*, 36(2), pp. 3336–3341.

Parr, H. (2002) 'Medical geography: Diagnosing the body in medical and health geography, 1999-2000', *Progress in Human Geography*, 26(2), pp. 240–251.

Paul, M. J. and Dredze, M. (2011) 'You are what you Tweet: Analyzing Twitter for public health.', in *Fifth International AAAI Conference on Weblogs and Social Media*, pp. 265–272.

Pavlyshenko, B. M. (2016) 'Linear, machine learning and probabilistic approaches for time series analysis', in *Proceedings of the 2016 IEEE 1st International Conference on Data Stream Mining and Processing (DSMP), Lviv, 23rd–27th August*. IEEE, pp. 377–81.

Peacey, V. *et al*. (2006) 'Ten-year changes in sun protection behaviors and beliefs of young adults in 13 European countries', *Preventive Medicine*, 43(6), pp. 460–465.

Pedersen, T. L. (2018) 'ggraph: An Implementation of Grammar of Graphics for Graphs and Networks. R package version 1.0.2.'

Penney, T. L. *et al*. (2017) 'Utilization of Away-From-Home Food Establishments, Dietary Approaches to Stop Hypertension Dietary Pattern, and Obesity', *American Journal of Preventive Medicine*, 53(5), pp. 155–163.

Pentland, A., Reid, T. and Heibeck, T. (2013) *Big data and health: Revolutionizing medicine and public health.*, *Report of the Big Data and Health Working Group*.

Petersen, J. *et al*. (2011) 'Geodemographics as a tool for targeting neighbourhoods in public health campaigns', *Journal of Geographical Systems*, 13(2), pp. 173–192.

Pillay, N. *et al*. (2010) 'The Economic Burden of Minor Ailments on the National Health Service (NHS) In the UK', *Self Care*, 1(3), pp. 105–116.

Powell, L. M., Nguyen, B. T. and Han, E. (2012) 'Energy intake from restaurants: Demographics and socioeconomics, 2003-2008', *American Journal of Preventive Medicine*, 43(5), pp. 498–504.

Provost, F. and Fawcett, T. (2013) 'Data Science and its Relationship to Big Data and Data-Driven Decision Making', *Big Data*, 1(1), pp. 51–9.

R Core Team (2014) ‘R: A language and environment for statistical computing.’, *R Foundation for Statistical Computing, Vienna, Austria*. Available at: https://www.r-project.org/.

Raghupathi, W. and Raghupathi, V. (2014) ‘Big data analytics in healthcare: promise and potential’, *Health Information Science and Systems*, 2(1), pp. 1–10.

Raja, U. *et al*. (2008) ‘Text mining in healthcare. Applications and opportunities.’, *Journal of Healthcare Information Management*, 22(3), pp. 52–56.

Robinson, A. H. (1950) ‘Ecological correlation and the behaviour of individuals’, *American Sociological Review*, 15(3), pp. 351–357.

Robinson, E. L. *et al*. (2017) ‘Climate Hydrology and Ecology Research Support System Meteorology Dataset for Great Britain (1961–2015) [CHESS-met] v1.2.’, *NERC Environmental Information Data Centre*. Available at: https://doi.org/10.5285/b745e7b1-626c-4ccc-ac27-56582e77b900.

Rosenberg, M. W. (1998) ‘Medical or health geography? Populations, peoples and places’, *International Journal of Population Geography*, 4(3), pp. 211–226.

Sacks, G., Swinburn, B. and Lawrence, M. (2009) ‘Obesity Policy Action framework and analysis grids for a comprehensive policy approach to reducing obesity’, *Obesity Reviews*, 10(1), pp. 76–86.

Sadilek, A., Brennan, S. and Kautz, H. (2013) ‘nEmesis: Which Restaurants Should You Avoid Today?’, in *First AAAI Conference on Human Computation and Crowdsourcing*., pp. 138–146.

Safran, C. *et al*. (2007) ‘Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper’, *Journal of the American Medical Informatics Association*, 41(1), pp. 1–9.

Santillana, M. *et al*. (2014) ‘What can digital disease detection learn from (an external revision to) Google flu trends?’, *American Journal of Preventive Medicine*, 47(3), pp. 341–347.

Satija, A. *et al*. (2015) ‘Understanding Nutritional Epidemiology and Its Role in Policy’, *Advances in Nutrition*, 6(1), pp. 5–18.

SAUSy-Lab (2019) *itinerum-trip-breaker*. Available at: https://github.com/SAUSy-

Lab/itinerum-trip-breaker (Accessed: 13 April 2019).

Scholten, H. J. and de Lepper, M. J. C. (1991) 'The benefits of the application of Geographical Information Systems in public and environmental health', *World Health Statistics quarterly. Rapport Trimestriel de Statistiques Sanitaires Mondiales*, 44(3), pp. 1–21.

Schukat, M. *et al*. (2016) 'Unintended Consequences of Wearable Sensor Use in Healthcare. Contribution of the IMIA Wearable Sensors in Healthcare WG', *IMIA Yearbook of Medical Informatics*, 25(01), pp. 73–86.

Schuurman, N. (2009) 'The new Brave New World: Geography, GIS, and the emergence of ubiquitous mapping and data', *Environment and Planning D: Society and Space*, 27(4), pp. 571–580.

Scully, J. Y. *et al*. (2017) 'GPS or travel diary: Comparing spatial and temporal characteristics of visits to fast food restaurants and supermarkets', *PLoS ONE*, 12(4), pp. 1–13.

Scully, J. Y. *et al*. (2019) 'A Time-Based Objective Measure of Exposure to the Food Environment', *International Journal of Environmental Research and Public Health*, 16(7), pp. 1–14.

Sharp, B. and Sharp, A. (1997) 'Loyalty programs and their impact on repeat-purchase loyalty patterns', *International Journal of Research in Marketing*, 14(5), pp. 473–86.

Shen, L. and Stopher, P. R. (2014) 'Review of GPS Travel Survey and GPS Data-Processing Methods', *Transport Reviews*, 34(3), pp. 316–334.

Silge, J. and Robinson, D. (2016) 'tidytext: Text Mining and Analysis Using Tidy Data Principles in R', *The Journal of Open Source Software*, 1(3), pp. 1–3.

Silge, J. and Robinson, D. (2017) *Text Mining with R: A tidy approach*. First Edit. Sepastopol, CA: O'Reilly.

Silver, L. D. *et al*. (2017) 'Changes in prices, sales, consumer spending, and beverage consumption one year after a tax on sugar-sweetened beverages in Berkeley, California, US: A before-and-after study', *PLoS Medicine*, 14(4), pp. 1–19.

Singleton, A. D. and Arribas-Bel, D. (2019) 'Geographic Data Science', *Geographical Analysis*, pp. 1–15.

Singleton, A. D. and Spielman, S. E. (2014) 'The Past, Present, and Future of Geodemographic Research in the United States and United Kingdom', *Professional Geographer*, 66(4), pp. 558–67.

Sivarajah, U. *et al*. (2017) 'Critical analysis of Big Data challenges and analytical methods', *Journal of Business Research*. The Authors, 70(2017), pp. 263–286.

Sloan, L. and Morgan, J. (2015) 'Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter', *PLoS ONE*, 10(11), pp. 1–15.

Smith, K. J. *et al*. (2009) 'Takeaway food consumption and its associations with diet quality and abdominal obesity: A cross-sectional study of young adults', *International Journal of Behavioral Nutrition and Physical Activity*, 6(1), pp. 1–13.

Smith, T. *et al*. (2015) 'The English Indices of Deprivation 2015'. London: Department for Communities and Local Government, pp. 1–123.

Soundararaj, B., Lugomer, K. and Trasberg, T. (2019) *Towards a Better Understanding of Footfall*, *cdrc.ac.uk*. Available at: https://www.cdrc.ac.uk/wp-content/uploads/2019/07/FF_Project-Update_July2019_1.3.pdf (Accessed: 6 March 2020).

St Louis, C. and Zorlu, G. (2012) 'Can Twitter predict disease outbreaks?', *BMJ*, 344(e2353), pp. 1–3.

Statistics Canada (2017) *Dissemination Area Boundary File, 2016 Census*. *Statistics Canada Catalogue no*. *92-169-X*. Available at: https://www150.statcan.gc.ca/n1/en/catalogue/92-169-X (Accessed: 14 March 2019).

Statistics Canada (2019) *Canadian Community Health Survey – Annual component (CCHS)*. Available at: https://www.statcan.gc.ca/eng/survey/household/3226 (Accessed: 3 July 2019).

Statistics Canada (Business Register Division) (2017) *Canadian Business Patterns; Dissemination Area (DA) Level [custom tabulation]*. Available at: https://doi.org/10.5683/SP/FLLHOV Scholars Portal Dataverse; V2 (Accessed: 12 April 2019).

Steinbrook, R. (2008) 'Personally controlled online health data - the next big thing in medical care?', *New England Journal of Medicine*, 358(16), p. 1653.

Stopher, P. R. and Greaves, S. P. (2007) 'Household travel surveys: Where are we going?',

Transportation Research Part A: Policy and Practice, 41(5), pp. 367–381.

Strandburg, K. J. (2014) 'Monitoring, Datafication, and Consent: Legal Approaches to Privacy in the Big Data Context', in Lane, J, Stodden, V, Bender, S, Nissenbaum, H. (ed.) *Privacy, Big Data, and the Public Good: Frameworks for Engagement*. Cambridge: University Press, pp. 5–43.

Studer, M. and Ritschard, G. (2016) 'What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures', *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 179(2), pp. 481–511.

Sturm, R. (2007) 'Increases in morbid obesity in the USA: 2000-2005', *Public Health*, 121(7), pp. 492–496.

Subramanian, S. V. *et al*. (2009) 'Revisiting Robinson: The perils of individualistic and ecologic fallacy', *International Journal of Epidemiology*, 38(2), pp. 342–360.

Tennekes, M. (2018) 'tmap: Thematic Maps in R.', *Journal of Statistical Software*, 84(6), pp. 1–39.

The BMJ (2005) *Sir Richard Doll*, *bmj.com*. Available at: https://www.bmj.com/content/suppl/2005/07/28/331.7511.295.DC1 (Accessed: 12 September 2019).

The U.S. National Archives and Records Administration (2019) *Electoral College results*. Available at: https://www.archives.gov/electoral-college/results (Accessed: 3 July 2019).

Thielmann, A. *et al*. (2018) 'Self-care for common colds: A European multicenter survey on the role of subjective discomfort and knowledge about the self-limited course - The COCO study', *PLoS ONE*, 13(4), pp. 1–11.

Thow, A. M. *et al*. (2010) 'The effect of fiscal policy on diet, obesity and chronic disease: a systematic review', *Bulletin of the World Health Organization*, 88(2010), pp. 609–614.

Tibshirani, R., Walther, G. and Hastie, T. (2001) 'Estimating the number of clusters in a data set via the gap statistic', *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 63(2), pp. 411–423.

Tobler, W. R. (1970) 'A Computer Movie Simulating Urban Growth in the Detroit Region', *Economic Geography*, 46(Sup1), pp. 234–240.

Tremblay, M. S. and Willms, J. D. (2003) 'Is the Canadian childhood obesity epidemic related to physical inactivity?', *International Journal of Obesity*, 27(9), pp. 1100–5.

Trifirò, G., Sultana, J. and Bate, A. (2018) 'From Big Data to Smart Data for Pharmacovigilance: The Role of Healthcare Databases and Other Emerging Sources', *Drug Safety*, 41(2), pp. 143–149.

Trinca, S. (2014) 'GIS Applications for Environment and Health in Italy', in Gatrrell, A. and Loytonen, M. (eds) *GIS and Health: GISDATA 6*. 6th edn. CRC Press, pp. 113–123.

Tuxen-Bethman, K. (2017) *Let's clear the air: Mapping our environment for our health*, *The Keyword*. Available at: https://blog.google/products/maps/lets-clear-air-mapping-our-environment-our-health/. (Accessed: 20 July 2017).

US Census Bureau (2019) *Table H-8. Median Household Income by State*. Available at: https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-income-households.html (Accessed: 8 July 2019).

US Department for Health and Human Services (2001) *The Surgeon-General's Call to Action to Prevent an Increase in Overweight and Obesity 2001*, *The Surgeon General's call to action to prevent and decrease overweight and obesity*. Rockville, MA: U.S. Department of Health and Human Services, Public Health Service, Office of the Surgeon General.

US Department for Health and Human Services (2010) *The Surgeon General's Vision for a Healthy and Fit Nation 2010*. Rockville; MD: Department of Health and Human Services, Office of the Surgeon General.

Valdivia, A. *et al*. (2010) 'Monitoring influenza activity in Europe with Google Flu Trends: Comparison with the findings of sentinel physician networks - results for 2009-10', *Eurosurveillance*, 15(29), pp. 1–6.

Varian, H. R. (2014) 'Big Data: New Tricks for Econometrics', *Journal of Economic Perspectives*, 28(2), pp. 3–28.

Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S*. Fourth Edi. New York: Springer.

Wang, X. *et al*. (2009) 'Active Computerized Pharmacovigilance Using Natural Language Processing, Statistics, and Electronic Health Records: A Feasibility Study', *Journal of the American Medical Informatics Association*, 16(3), pp. 328–337.

Ward, J. H. (1963) 'Hierarchical Grouping to Optimize an Objective Function', *Journal of the American Statistical Association*, 58(301), pp. 236–244.

Warrer, P. *et al*. (2012) 'Using text-mining techniques in electronic patient records to identify ADRs from medicine use', *British Journal of Clinical Pharmacology*, 73(5), pp. 674–684.

Waterhouse, J. *et al*. (2005) 'Chronobiology and meal times: internal and external factors', *British Journal of Nutrition*, 77(S1), pp. 29–38.

Wazaify, M. *et al*. (2005) 'Societal perspectives on over-the-counter (OTC) medicines', *Family Practice*, 22(2), pp. 170–176.

Weber, G. M., Mandl, K. D. and Kohane, I. S. (2014) 'Finding the missing link for big biomedical data', *JAMA*, 311(24), pp. 2479–2480.

Wesolowski, A. *et al*. (2014) 'Commentary: Containing the Ebola Outbreak - the Potential and Challenge of Mobile Network Data', *PLoS Currents Outbreaks*, 6(2014), pp. 1–21. Available at: http://currents.plos.org/outbreaks/index.html%3Fp=42561.html (Accessed: 2 August 2019).

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H. (2017) 'tidyverse: Easily Install and Load the "Tidyverse". R package version 1.2.1.'

Widener, M. J. *et al*. (2018) 'Activity space-based measures of the food environment and their relationships to food purchasing behaviours for young urban adults in Canada', *Public health nutrition*, 21(11), pp. 2103–16.

Widener, M. J. and Shannon, J. (2014) 'When are food deserts? Integrating time into research on food accessibility', *Health and Place*, 30(November 2014), pp. 1–3.

Wigan, M. R. and Clarke, R. (2013) 'Big data's big unintended consequences', *Computer*, 46(6), pp. 46–53.

Wilson, A. M., Thabane, L. and Holbrook, A. (2004) 'Application of data mining techniques in pharmacovigilance', *British Journal of Clinical Pharmacology*, 57(2), pp. 127–134.

Wilson, N. *et al*. (2009) 'Interpreting Google flu trends data for pandemic H1N1 influenza:

the New Zealand experience.', *Eurosurveillance*, 15(44), pp. 1–3.

Wise, S., Haining, R. and Ma, J. (2001) 'Providing spatial statistical data analysis functionality for the GIS user: The SAGE project', *International Journal of Geographical Information Science*, 15(3), pp. 239–254.

World Health Organization (2000a) *Guidelines for the regulatory assessment of medicinal products for use in self-medication*. (No. WHO/EDM/QSM/00.1). Geneva: World Health Organization.

World Health Organization (2000b) *Obesity: Preventing and managing the global epidemic*. (No. 894). World Health Organization.

World Health Organization (2018) *Obesity and overweight*. Available at: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight (Accessed: 7 May 2019).

Wright, C. and Sparks, L. (1999) 'Loyalty saturation in retailing: Exploring the end of retail loyalty cards?', *International Journal of Retail & Distribution Management*, 27(10), pp. 429–440.

Xie, J. *et al*. (2018) 'Evaluating the validity of current mainstream wearable devices in fitness tracking under various physical activities: Comparative study', *JMIR mHealth and uHealth*, 6(4), pp. 1–13.

Žylius, G., Simutis, R. and Vaitkus, V. (2015) 'Evaluation of computational intelligence techniques for daily product sales forecasting', *International Journal of Computing*, 14(3), pp. 157–64.