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

Alec Davies*, Mark A. Green, Dean Riddlesden, Alex D. Singleton

Geographic Data Science Lab, Department of Geography and Planning, University of Liverpool

Current Address: Geographic Data Science Lab, Department of Geography & Planning, University of Liverpool, Liverpool, United Kingdom

*Corresponding Author

Email: a.e.davies@liverpool.ac.uk

Abstract

This paper examines objective purchasing information for inherently seasonal self-medication product groups using transaction-level loyalty card records. Predictive models are applied to predict future monthly self-medication purchasing. Analyses are undertaken at the lower super output area level, allowing the exploration of ~300 retail, social, demographic and environmental predictors of purchasing. The study uses a tree ensemble predictive algorithm, applying XGBoost using one year of historical training data to predict future purchase patterns. The study compares static and dynamic retraining approaches. Feature importance rank comparison and accumulated local effects plots are used to ascertain insights of the influence of different features. Clear purchasing seasonality is observed for both outcomes, reflecting the climatic drivers of the associated minor ailments. Although dynamic models perform best, where previous year behaviour differs greatly, predictions had higher error rates. Important features are consistent across models (e.g. previous sales, temperature, seasonality). Feature importance ranking had the greatest difference where seasons changed. Accumulated local effects plots highlight specific ranges of predictors influencing self-medication purchasing. Loyalty card records offer promise for monitoring the prevalence of minor ailments and reveal insights about the seasonality and drivers of over-the-counter medicine purchasing in England.

Keywords

Self-medication; over-the-counter medicines; minor ailments; machine learning; tree ensembles.

Introduction

Fine resolution public health information is vital in determining at-risk populations (Hay *et al.*, 2005). Data driven applications are improving health surveillance frameworks (increasingly in real-time) and have proven to help in the discovery of specific 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 allow 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 ailments by leveraging information regarding the purchase of over-the-counter medication. These data do not only offer potential for merely monitoring prevalence but also aid 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). This paper seeks to address this gap in the research into real-time objective purchasing information in terms of 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.

This study utilises loyalty card records to (1) understand self-medication behaviours; (2) explore how behaviours vary over time and the drivers of these trends; and (3) highlight opportunities for using such records to predict future purchasing.

Methods

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. The 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. Transactions were aggregated 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) was 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.

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.

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 than other algorithms, XGBoost has been shown to outperform SVMs and Random Forests (Ogutu, Piepho and Schulz-Streeck, 2011).

A monthly forecasting approach was selected allowing a detailed temporal resolution whilst keeping model computation feasible. Training data contained 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 (vital 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).

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 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 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).

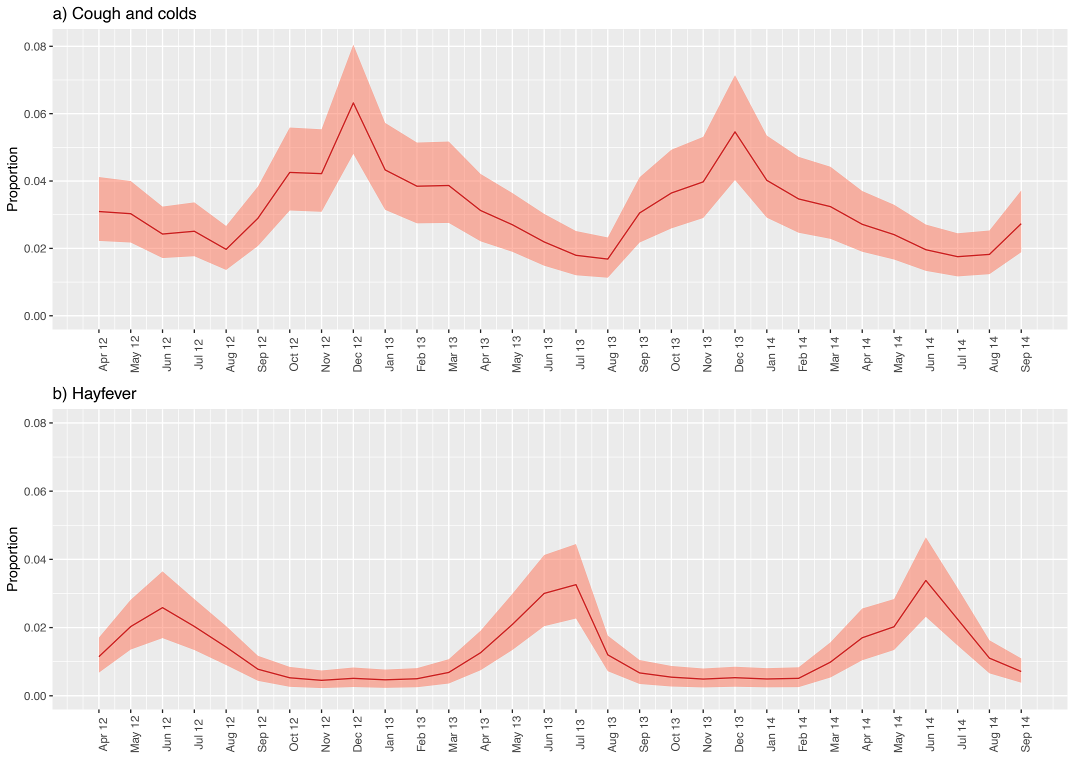

Figure 1. Median and interquartile range of proportion purchasing products per month

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 2a). 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 1 shows that interquartile range increased with increased sales. R^2 values were highest where purchasing proportions are highest (see figure 1) 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 2d-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. 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.

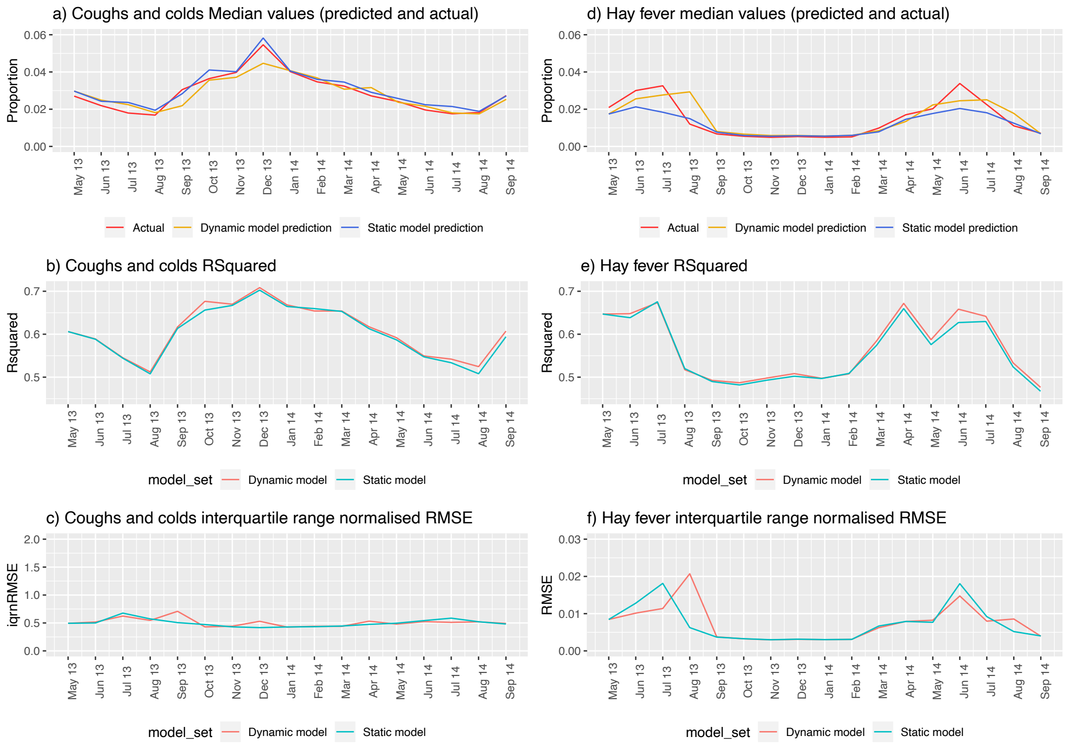

Figure 2. a) Coughs and colds median sales and predictions; b) Coughs and colds R2 performance; c) Coughs and colds interquartile range nRMSE; d) Hay fever median sales and predictions; e) Hay fever R2 performance; f) Hay fever interquartile range nRMSE

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 1 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 3). Only the top 8 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 7 months (figure 3b). 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.

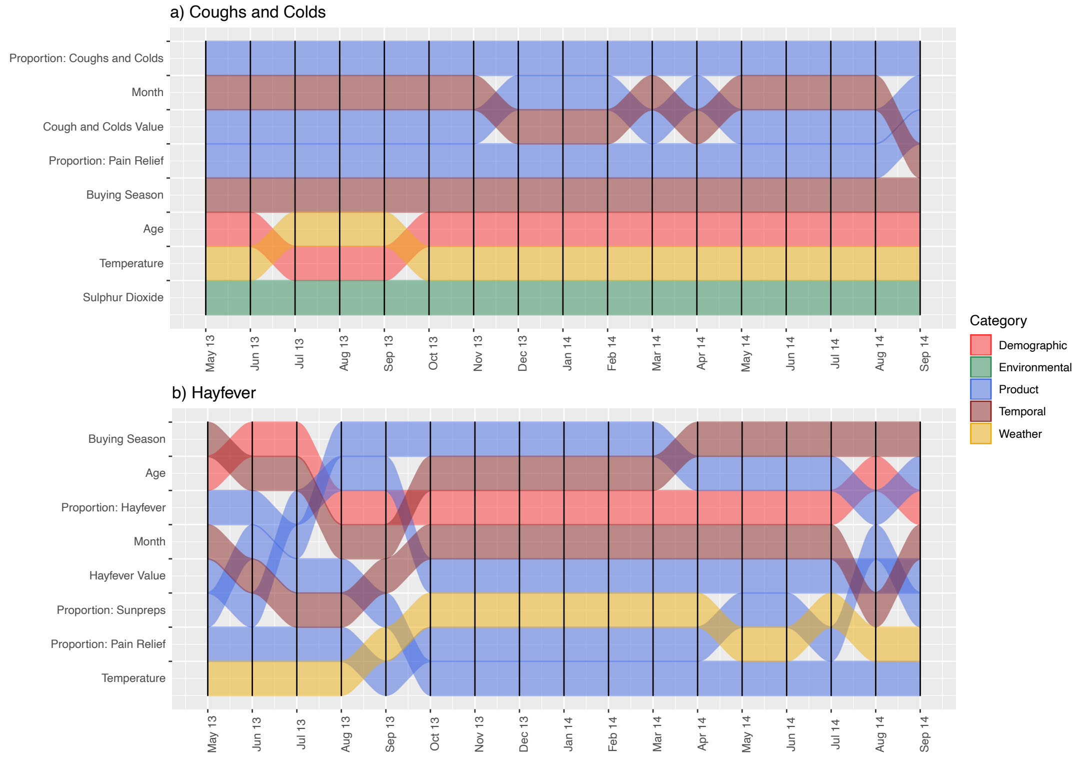

Figure 3. Feature importance rank change across models for 8 most important features

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 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 0 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 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, a 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).

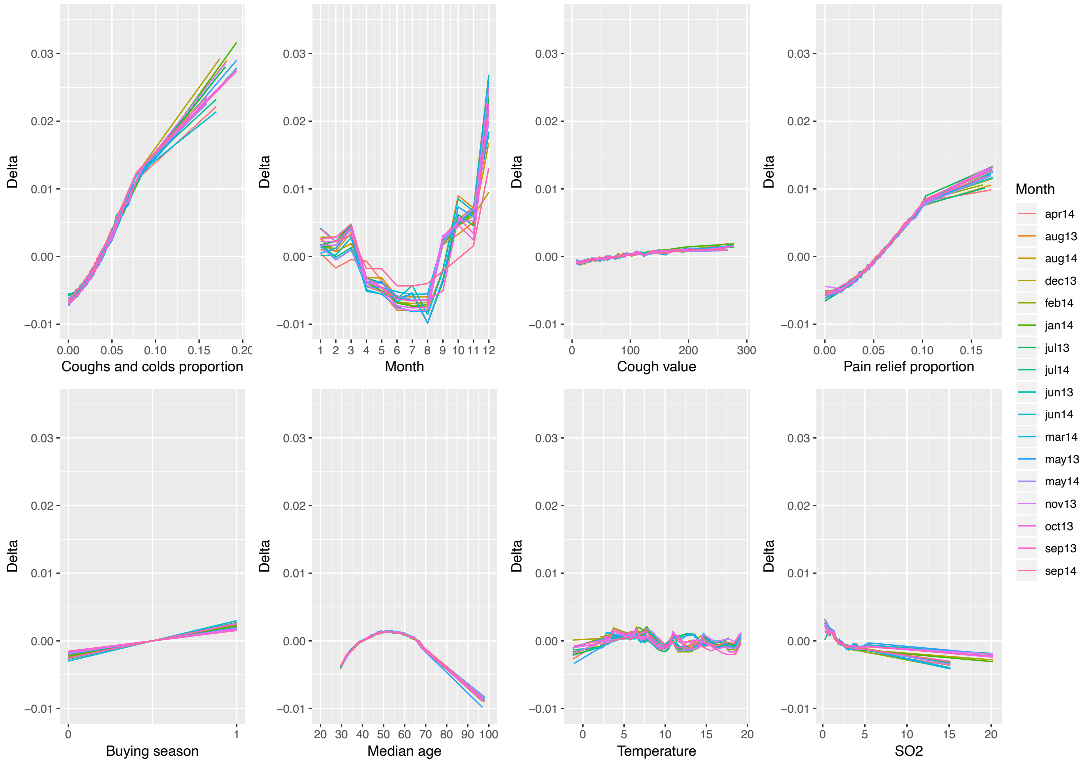

Figure 4. Accumulated local effects plot for 8 most important features (coughs and colds)

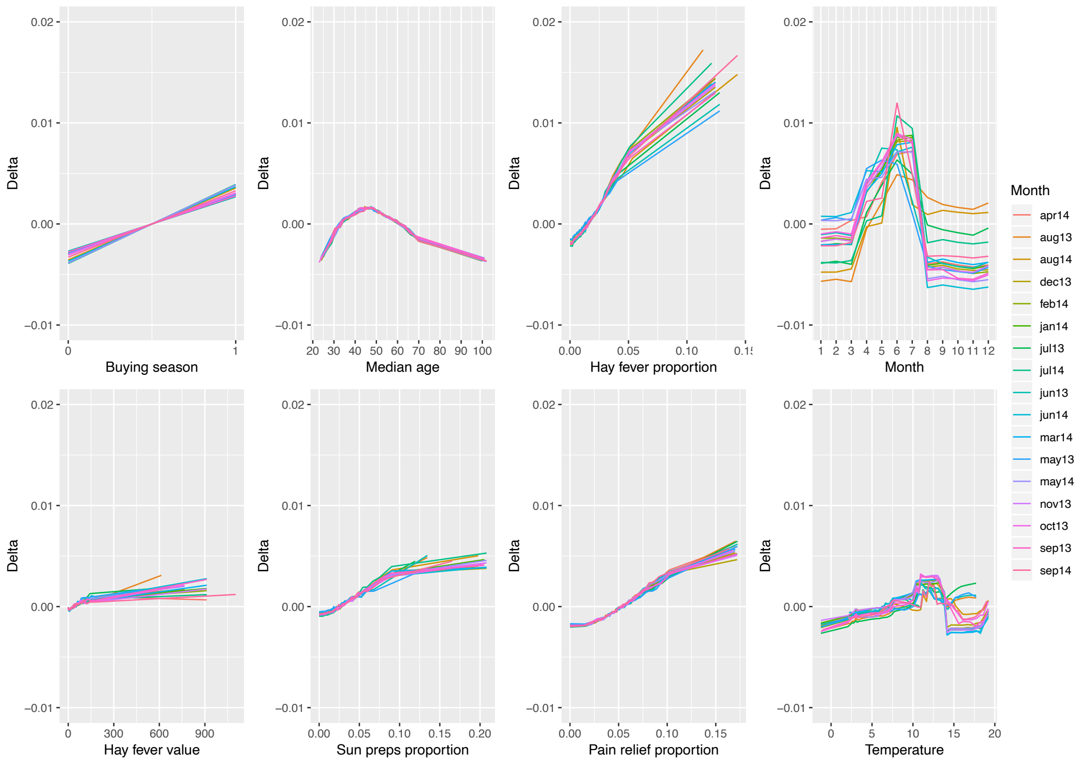

Figure 5. Accumulated local effects plot for 8 most important features (hay fever)

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 4-5) 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 1). The large increase in December (and highest Delta (figure 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 from year to 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 affect 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 1 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).

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).

Acknowledgements

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. Ethical approval was granted for this research by the University of Liverpool's Research Ethics Committee (ref: 752). The data for this research has been provided by the Consumer Data Research Centre, an ESRC Data Investment, under project ID CDRC 004, ES/L011840/1; ES/L011891/1. Contains data provided by the ESRC Consumer data Research Centre; Contains National Statistics data © Crown copyright and database right 2015–2017; Contains Ordnance Survey data © Crown copyright database right 2015–2017; Contains Royal Mail data © Royal Mail copyright and database right 2015– 2017; Contains CDRC data 2016; Contains LDC data 2016–2017; Contains NHS data 2017; Contains DEFRA data © Crown copyright 2017; Contains OSM data 2017; Contains public sector information licensed under the Open Government Licence v3.0; Acknowledgement: Climate hydrology and ecology research support system meteorology dataset for Great Britain (1961-2015) [CHESS-met] v1.2 data licensed from NERC – Centre for Ecology & Hydrology; Database Right/Copyright NERC – Centre for Ecology & Hydrology. All rights reserved; Contains material based on Met Éireann data © Met Éireann, Met Office and OS data © Crown copyright and database right 2015 and University of East Anglia Climatic Research Unit CRU. This research paper forms part of Alec Davies thesis. Further information regarding the data, methods and software packages used can be found at: https://github.com/sgadavi3/nfd_self-medication.

References

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.

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

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.

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.

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.

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

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.

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.

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.

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.

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.

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

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.

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.

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.

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.

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

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.

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

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

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.

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).

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).

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

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.

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.

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.

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.

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.

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.

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

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

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.

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

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.