NHSX Report: Value of Commercial Product Sales Data in Healthcare Prediction

Elizabeth Dolan
N/LAB, University of Nottingham, UK

1 Executive Summary

This report summarises work carried out as part of the NHSX PhD internship project entitled "Value of Commercial Product Sales Data in Healthcare Prediction".

The primary aim of the project was to apply the novel variable importance technique, MCR (Model Class Reliance), to machine learning models which could predict registered respiratory deaths in the UK. The objective was to assess the value of commercial health data in healthcare predictions compared to other available datasets.

In order to apply MCR, a set of optimal models have to be created which can successfully make the required predictions. The project managed to achieve this outcome with the machine learning model PADRUS (Predicting the amount of deaths by respiratory disease using sales). PADRUS is a random forest regressor which makes accurate weekly predictions of respiratory deaths in 314 local authorities across England 17 days in advance. The models' features are created from the following dataset types: week number, commercial sales, weather, indices of multiple deprivation, age and population, demographics, housing, and land use.

MCR was applied to PADRUS showing the highest and lowest impact variables had on predictions across all instances of the model. Grouped MCR was also employed in order for variables to be evaluated in concert as a collection of features created from a dataset type. The MCR results implied model instances of PADRUS were using variables in different ways to achieve the same predictive results, and suggested where variables could be interchangeable or critical to predictions.

2 Introduction

Respiratory disease (ICD 10 coding: J00 - J99) was the underlying cause in 369,900 deaths in England and Wales in the years 2015 to 2019 [1]. In 2020 the COVID-19 pandemic became the leading cause of death in England and Wales [1]. COVID-19 is now reported on 167,927 UK death certificates to date [2]. With COVID-19 appearing to become a dominant and long-standing disease affecting the respiratory system [2, 3] it is now critically important to investigate how to forecast the impacts of Influenza-like Illnesses (ILI) on a population. In answer to this need an increased effort has been made to include social and behavioural data to produce integrated disease models, which may outperform those from a traditional epidemiological standpoint [4, 5, 6, 7, 8, 9]. In their review of integrated disease models, Bedson et al. report that there is a gap in using these social and behavioural datasets due to the lack of access to operational data that would be useful in real-world dynamic circumstances [4]. Predominantly these additional social and behavioural datasets are sourced through qualitative methods relying heavily on surveys and self-reports [4, 5, 6, 7, 8]. Survey data and sample size

is often limited by budget and capacity, and captures one moment in time. Self-reports, often used in apps [5, 10] or monitored through analysis of social media accounts [11,12] can also be subject to reporting bias due to memory failures and social desirability. This suggests the need to capture objectively recorded sources of dynamic behavioural data, that supplement and complement qualitative data, and can be applied to real-time incidents. Types of datasets which meet these requirements have already been applied to covid modelling, mainly in the form of mobility data with some success [13, 14]. An alternative dataset which could also meet these requirements is shopping data, especially pertinent purchases of healthcare products. No recent integrated disease models in the UK have included transactional data in the form of retail sales that we are aware of, and its use in health surveillance of ILI is not currently prevalent [15, 16, 17].

Recorded and stored by UK retailers, personal transactional data offers national longitudinal time-stamped data on customer purchases at a high geographical granularity. Transactional retail data is updated in real-time and can be used to gain insights into population health [18, 19, 20, 21]. Shopping data is considered a traditional dataset for use in health surveillance systems, especially medication (16, 22, 23), and was revealed as a potential indicator of influenza in the community by Welliver et al. in 1979 [24]. Hogan et al. showed a high correlation between respiratory and diarrheal outbreaks in children and sales of over-the-counter electrolyte products [25]. Socan, Erčulj and Lajovic showed a moderately high correlation between medicines for sore throat, antitussives, decongestants and mucolytics, and weekly influenza incidence; the sales of antitussives alone could predict increased incident rates [26].

However, Al-Tawfiq et al.'s review of health surveillance systems for emerging respiratory viruses reported mixed results from monitoring over-the-counter drugs [27]. Pivette et al. systematic review of drug sales in outbreak detection did find the potential for earlier alerts, yet noted challenges in selecting indicator drug groups and poor-quality clinical surveillance data [15]. In the UK, Davies and Finch's 2003 study using Nottingham City Hospital NHS, Boots and Reckitt Benckiser pharmacies data for the three winter periods, showed over-the-counter cough/cold remedies may give two weeks warning of peaks in admissions for respiratory illness [28]. In contrast, at a UK national scale Todd, Diggle et al. show there was no significant correlation between retail sales of symptom remedies and cases for the 2009 influenza outbreak [29]. However, lower scale geospatial areas were not considered, with England and Wales divided into only 6 regions, and correlation was looked for between purchases and flu cases rather than hospital admissions or deaths [29].

The aim of this study was to investigate whether shopping data in the UK has value in models used to forecast deaths by respiratory disease. Yet, establishing the predictive capability of sales data in itself, is not enough to show its value. To warrant the investment in the use of shopping data in integrated disease models, we must also investigate its value via comparison against other variables. After initial exploratory analysis showed the potential of sales data in predicting deaths by respiratory disease in the UK, a new experimental design to compare the use of sales data against other datasets used in the prediction of respiratory deaths was created. In line with Hofman et al.'s recommendations on computational social science we create an appropriate baseline model, perform out-of-sample testing and devote attention to combining prediction and explanation [30]. We establish this by the inclusion in the predictive models of input variables

(features) created by multiple datasets outside of sales data. We then explain how these datasets compare in importance to making the models predictions. To ensure the robustness of this variable importance analysis, we use the novel technique of Model Class Reliance (MCR) to examine the impact of variables across all instances of a model [31, 32, 33]. To achieve a valid implementation of MCR we first had to create a model which could return accurate predictions. This in turn seeds the creation of a set of best performing model instances called the "Rashomon set" [34]. Each model instance in this set utilizes the input variables in different ways to make predictions with the same accuracy. This occurs when input variables share or overlap in the predictive information they contain with regard to the output variable. This results in competing explanations of the phenomena which, when each is used, lead to the same predictive accuracy (explain the phenomena equally well). This existence of multiple differing explanations for the phenomena which are all equally well supported by the data is known as The Rashomon effect [34]. While the automatic enumeration of all competing equally valid explanations is currently intractable, MCR provides an indication of the least (MCR-) and most (MCR+) each variable is used in any of these competing explanations. This provides an indication of the absolute requirement (MCR-) and maximal utility (MCR+) of each variable rather than the variable's (somewhat arbitrary) utility in a randomly selected model from the Rashomon set.

Weekly historical data on respiratory deaths in England and sales data from a high street retailer, including sales of cough, throat and decongestant medicines, were used to train and test a machine learning model. Additional model features were made from information on week number, weather, indices of multiple deprivation (IMD), age/population level, demographics, housing and land use. We created an optimised random forest regressor model named PADRUS (Predicting the amount of deaths by respiratory disease using sales). PADRUS predicts deaths by respiratory disease in 314 Lower Tier Local Authority Areas (LTLAs). We find variable permutation importance bounds, computed by MCR, show optimal predictions by the PADRUS model could not be achieved without sales data variables.

3 Results and Discussion

Results and Discussion to be published after peer-review

4. Methods

4.1 Open-source software. We used the following open-source software in the analysis:

Matplotlib: https://matplotlib.org/
MCRForest: https://github.com/gavin-s-smith/mcrforest
Numpy: https://numpy.org/
Pandas: https://pandas.pydata.org/
Seaborn: https://seaborn.pydata.org/
Scikit-learn: https://scikit-learn.org/
SHAP: https://shap-lrjball.readthedocs.io/en/latest/index.html#

4.2 Ethics information

The health data in this study was used under the terms of usual practice for research as defined in the UK Policy Framework for Health and Social Care Research, with the study designed to investigate the health issues in a population in order to improve population health [35]. To keep health data de-identified and extractable from the National Commissioning Data Repository (NDCR), death counts below 5 within an LTLA were suppressed and reported as 5. Sales data used in the study was limited to the number of product types sold at a store. No sales transactions were linked to individual customers, and no personal data was used in this study.

4.3 Prior Exploratory Design and Data Analysis

Prior to the design and investment in the following experimental design, initial data analysis and machine learning modelling took place using open source ONS data on registered deaths from respiratory disease (as the underlying cause) in England and Wales from 7th December 2009 to 13th April 2015 [36]. The analysis of registered deaths from respiratory disease was framed as a regression task with the number of deaths predicted (the target class) a continuous output integer variable (y). The models' task was to predict national weekly respiratory deaths in advance. The models inputs used 'week number' (1 to 52) of registered deaths and the following features created from commercial sales data only: weekly sales of cough, dry cough, mucus cough, decongestant and throat healthcare products. Commercial sales data was from a UK high street retailer, and was a sample of transactions recorded through 2,702,449 loyalty cards. Due to the typical demographics of loyalty card holders approximately 87.4% of the sample were female. Data was stratified into a training (70%) and test set (30%), with the performance of the models examined against the held-out test data with time series cross validation (TSCV) applied. TSCV is an alternative to k-fold cross-validation which would cause data leakage in time series data, and creates a walk-forward validation in order to evaluate the models' ability for generalisation. Two metrics were used to score the regression task. These were RMSE (root mean squared error), and R^2 (coefficient of determination). RMSE is the square root of the average squared distance between the target (y) and the predicted score. R^2 measures the proportion of the variance in predictions (from the average of y) that can be explained by X (the model input variables). Several linear regression models were trained and evaluated in order to test for the optimal time lag in days between sales and registered deaths, for increased accuracy in predictions. Lag refers to the time lapse between the date of registered deaths being predicted and the date of reported sales.

4.4 Experimental Design

The experiment was designed to run in two phases. Phase 1 was to create a set of models to predict registered deaths from respiratory disease. These models used commercial sales data and a wide range of other variables, which have shown associations with deaths from respiratory disease (Table 1). Phase 2 was to explain the models by identifying the different impact variables inputted have on the models' predictions, including commercial sales data. Phase 2 implements the novel variable importance tool MCR for random forest regressor [32, 33], and phase one needed to be achieved in order for valid implementation of MCR.

Data

Data was collected for this study via working partnerships with NHSX, a UK commercial high street retailer, and via open-source data. Data sources and descriptions used in this study can be seen in Table 1 which lists the data used for the models' target for predictions and features (variable inputs). Commercial sales data was sales units of all in-store transactions with store location only; no customer information was linked to transactions. Data was collected or aggregated to the geographic regions of LTLAs (Local Tier Local Authorities). LTLAs were used because as geographic areas used by government data they would enable the inclusion of a wide range of demographic, environmental and socioeconomic data previously found to be significant to the risk of death from respiratory disease (Table 1). LTLAs were chosen over other spatial divisions, including census areas MSOAs (Middle Layer Super Output Areas) and LSOAs (Lower Layer Super Output Areas), in order to find a balance between limiting data suppression of deaths done to maintain data de-identification, and maintaining a high enough level of spatial granularity for data to have a significant impact on predictions [50,51]. Data availability at LTLA level for open-source data for areas in the UK outside of England was limited. Due to this limiting feature creation and the capacity to compare a wider range of data variables, analysis was restricted to 314 LTLAs in England alone. Even with this restriction other variables under consideration for inclusion such as search engine trends and temporal pollen, traffic, and air pollution counts, could not be included as there was no access to these datasets at this level of geo-spatial granularity within the time limits of the project. Data matching, the time series of available target predictions and available dynamic datasets of sales and weather, gave the dataset its timeframe of the 18th of March 2016 to 27th March 2020 for weekly deaths by respiratory disease. The decision was made not to include registered deaths from respiratory disease with a 17 day lag, as in a real-world scenario this data would not yet be available [52].

Prediction Target

The data used for the target predictions (y) was all registered deaths from respiratory disease (ICD 10 coding: J00 - J99) by LTLA on a weekly time basis. Death counts below 5 were suppressed and reported as 5. Suppression was between ⅕ and ¼.

Feature Engineering

All raw data was aggregated to 314 LTLA spatial resolutions across England for each of 56 input features. Features can be categorised as dynamic or static variables and were grouped by the following: week number, commercial sales, weather, indices of multiple deprivation, age/population level, demographics, housing and land use (Table 1). These groupings were used for the Group-MCR (Model Class Reliance). Dynamic (temporal) features were also aggregated by total weekly figures. How the 56 features were created can be seen in Table 1 which lists any data manipulation from the original source, for example averages and percentages.

Feature Engineering Sales Data

More complex data manipulation was needed for feature creation from commercial sales where the national transactional data was given by store location. Sales from 2,354 stores distributed across England had to be allocated to 314 English LTLAs. A store catchment model was developed that can attribute the proportion of sales occurring at each store to any particular MSOA (Middle Lower Super Output Area - a governmental geographic area) served. Using a naive Gaussian kernel based "influence" approach, we model the proportion of sales at a store which should be attributed to an MSOA. We established a two-dimensional Gaussian kernel over each store, centred at the store's longitude and latitude, and using a variance reflecting a theorised catchment radius (metres) corresponding to the store type (Regional Mall 15,000m, Hospital / Transport Hub 10,000m, Retail Park / Shopping Centre 8,000m, High Street / Supermarket 5,000m, Health Centre / Community 3,000m). We examined the "influence" the store has over a MSOA, by examining the value of the store's probability density function at the region's population weighted centroid. Due to the fact that Gaussian functions have an infinite range, a store will be assigned a non-zero "influence" for every MSOA. In order to simplify analysis, the "influence" attributed to a store for a given MSOA is zeroed if it is below a threshold of 3.449937e-318. The "total influence" of each store is then calculated by examining the sum of "influence" the store has over all 8000 MSOAs. The sales proportion each store attributes to a given MSOA, is the "influence" the store has over the MSOA, divided by its "total influence" over all regions. This process is run for all 3,000 stores (examining all 8,000 MSOAs). In this way, we obtained a probability that the sale from any store was by an individual living in a particular MSOA. Using lookup tables, the apportioned sales were then aggregated to the LTLA level for analysis, and reduced to LTLAs within England.

The total weekly sales units of key product types were chosen to create individual model features. These consisted of all cough medicines, dry cough medicines, decongestant medicines and throat medicines. Products were selected due to reports from previous literature associating a rise in their sales with an increase in respiratory illnesses [26, 28], and the prior exploratory analysis. Sales and weather feature inputs were lagged at least 17 days prior to deaths for effective forecasting, again in line with the findings from the prior exploratory analysis. Further features were created from the sales data by normalising product sales to create local and national sales ratios (See Figure 1)

Modelling Approach

The analysis of registered deaths from respiratory disease was framed as a regression task with the number of deaths predicted (the target class) a continuous output integer variable (y). The model's task was to predict weekly respiratory deaths 17 days in advance for each of the 314 LTLA areas in England from 18th March 2016 to 27th March 2020. The model was named PADRUS (Predicting the amount of deaths from respiratory disease using sales data). Data was stratified into a training set (45,844 data points from 18th March 2016 to 28th December 2018), approx. 70%) and test set (20,410 data points from 4th January 2019 to 27th March 2020, approx. 30%), with the performance of the models examined against the held-out test data. In addition an extra test set, to test models on the timeframe after the COVID-19 outbreak (18,840 data points from 27th March to 21st May 2020), was also created once the dataset became

local_sales_ratio = (S_L / T_L) S_L = product X sales at local level (LTLA) (weekly)

T_L = total sales (all products) at local level (weekly)

national_sales_ratio = (Sc / Tc) S_C = product sales at national level (weekly)

T_C = total sales at national level (weekly)

local_sales_multiplier = (SL / TL) / (SC / TC)

<1 *sales in locality lower than the national expectation*

=1 *sales in locality the same as the national expectation*

>1 *sales in locality higher than the national expectation*

Figure 1. Formula for creating features from normalising sales using local and national ratios

available during the experiment. Three metrics were used to score the regression task. These were RMSE (root mean squared error), MAE (mean absolute error) and R^2 (coefficient of determination). MAE is the mean of all differences (errors) between the target(y) and the predicted score. A random forest regressor model was trained and evaluated. Meta-parameters for the model were optimised using a time series cross-validation (TSCV, 4 splits) grid-search on training data to prevent over-fitting. The model using optimised meta-parameters was then re-fit to the training dataset, and evaluated against the held-out test set.

Creating a Baseline

In order to assess whether the PADRUS model created was valuable, an appropriate baseline model was created. This baseline model was created using the same methodology; a random forest regressor optimized using a time series cross-validation grid-search on training data. The baseline model performed the same task of predicting weekly respiratory deaths 17 days in advance for each of the 314 LTLA areas in England from 18th March 2016 to 27th March 2020. Data was stratified using the same splits on the training and test data (45,844 training data points, 20,410 test data points), and model predictions on the held out test data were scored using RMSE, MAE and R^2. The difference between PADRUS and the baseline model was feature inputs. The baseline inputs consisted of only two input features. The first the week number (1 to 52) for a dynamic input, and secondly the static input of the population of over 65s in each LTLA. The first feature was chosen because of the known seasonality of deaths by respiratory disease [37], yet used alone the prediction accuracy levels were very low. Therefore, the second feature was chosen due to the assumption that the size and age of population would greatly influence the number of deaths, with 90% of deaths from respiratory disease in Europe occurring in those aged 65 and over [43].

Variable Importance Tools

In order to compare the variables included in the PADRUS model, we conducted a feature importance analysis. This scored the importance of each feature (variable) in producing the model's predictions. We used a standard variable importance tool within the scikit-learn python library. This was the permutation importance function which measures feature importance by the expected increase in error after damaging that feature column [53]. On top of this inbuilt variable importance tool, a SHAP (SHapley Additive exPlanations) Analysis [54] was conducted. SHAP carries out a more complex calculation of feature importance. The Kernel Explainer SHAP tool used in this analysis applies weighted linear regression to determine the importance of each feature based on Shapley values from game theory and coefficients [54]. Both tools were applied to the training data set on an arbitrary instance of the PADRUS model. The SHAP analysis is computationally expensive and therefore was run on two random samples of data points of 100 and 1000. This enabled a comparison of results between sample sizes to ensure a large enough sample had been evaluated.

By running these variable importance tools on one arbitrary model, misleading results can be given due to the stochastic nature of the model's machine learning algorithms. This is a problem because different instances of an optimised model class can use different variables (features) and in different ways to achieve the same model predictive performance. To address the problem of standard variable importance tools only evaluating one instance of the PADRUS model, MCR (Model Class Reliance) was applied. MCR was developed by Fisher et al. to compute the feature importance bounds across all optimal models called the Rashomon set for Kernel (SVM) Regression (polynomial run-time) [31]. Smith, Mansilla and Goulding introduced a new technique that extends the computation of MCR to Random Forest classifiers and regressors [32]. MCR builds on permutation for a single model, computing the permutation feature importance bounds (MCR-, MCR+) for an input variable across all instances of PADRUS; calculating the minimum and maximum impact a variable could have on the predictions across all instances of the model (See Figure 2). This initial MCR analysis evaluated each feature individually for its importance, for example the importance of 'minimum temperature'. It did not assess how important a dataset type used to create a number of the models' features was to predictions as a whole, for example 'weather'. Ljevar et al. introduced a Grouped Feature approach to MCR for random forest [33]. Group-MCR was created in order to calculate the effects of variable groups, measuring the importance of a collection of features together on the predictions of random forest classifiers [33]. In order to evaluate the importance of groups of variables, and their impact in concert on the PADRUS model, we apply for the first time Group-MCR to a random forest regressor. Group-MCR is achieved through a modification of the random forest MCR algorithm, which reconsiders the definition of Model Reliance to be with respect to a group of variables rather than a single variable [33].

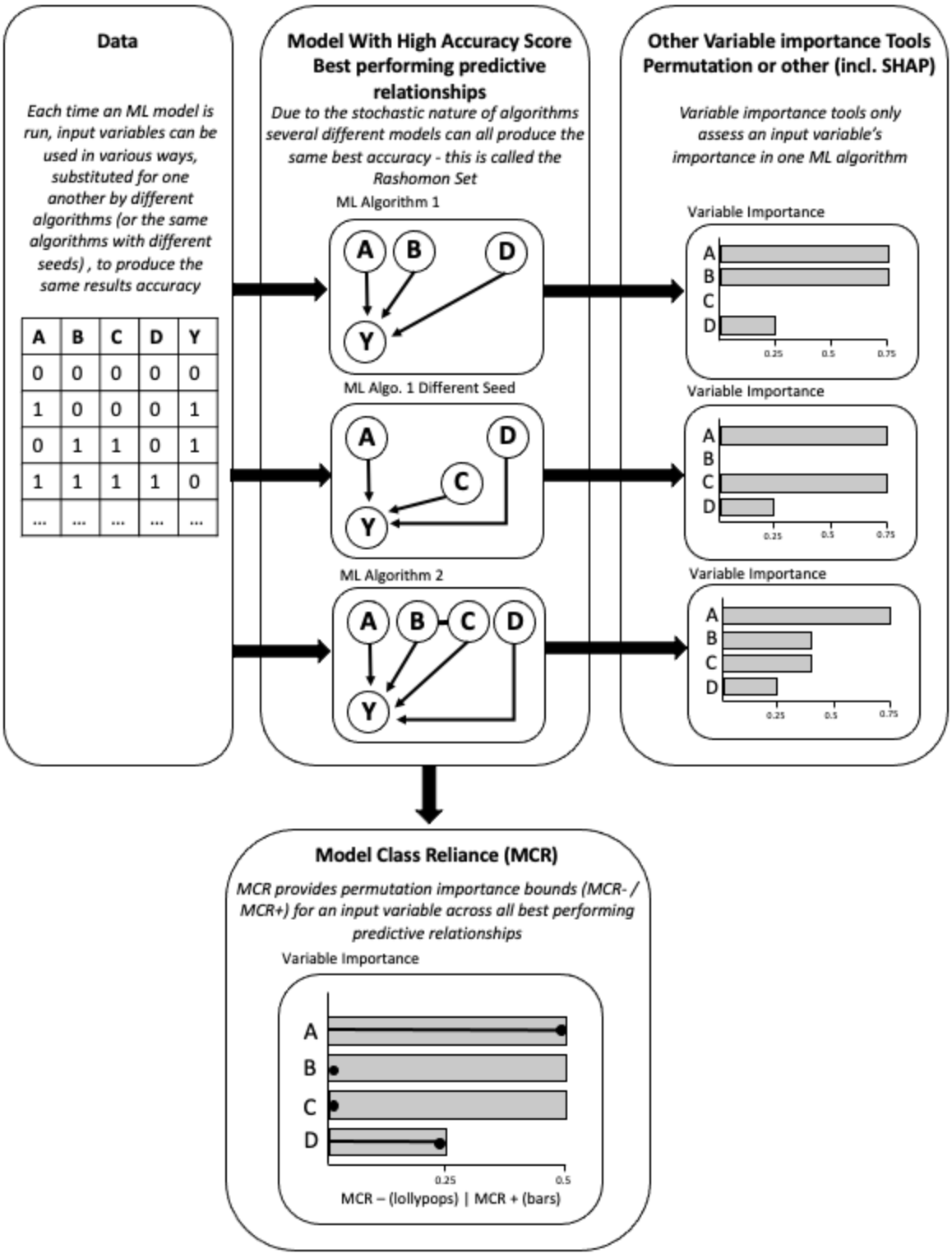

Figure 2. Diagrammatic representation of the difference between other variable importance tools and MCR

MCR can be used to establish the value of using a variable in a model being created for healthcare predictions, by suggesting which data types must be acquired for a predictive model to work effectively. In a healthcare setting it must be considered whether it is worth the investment in acquiring a type of data if another dataset works equally well, and is more affordable, or easier to access. MCR highlights which data types could be irreplaceable and which could be interchangeable. MCR results can also be used to guide the reduction of variables in order to create more transparent models for healthcare predictions. Group-MCR offers us further information by implying how groups of variables may be working together in the model's algorithms. By enabling the reduction of the number of variables, MCR can also be used to help decrease the dimensionality of the model leading to a more manageable global surface area to search for optimized meta-parameters. Creating the most accurate model possible using the chosen variables, would further evidence the value of their inclusion.

Modelling without sales data

Although MCR can explain which variables need to be included in a model to achieve the maximum accuracy rates for predictions, it cannot compute the difference in accuracy (the loss) if those variables were to be excluded from the model. In order to determine the loss in accuracy if sales data were to be left out of the model, the model PADRUNOS (Predicting the amount of deaths from respiratory disease using no sales) was created as a comparison to PADRUS. PADRUNOS was created in line with the baseline and PADRUS model methodology, as a random forest regressor optimized using a time series cross-validation grid-search on training data. MCR and Group-MCR was applied to PADRUNOS to calculate how the variables were used differently.

Testing the models in the Pandemic

During the course of the experiment, new sales data became available which facilitated testing the models PADRUS and PADRUNOS on a time period where the COVID-19 outbreak, and pandemic occurred. Registered deaths from respiratory disease and weather data were also available for a proportion of this timeframe. This enabled models to be tested on a new held-out dataset from the 3rd of April 2020 to 21st April 2021.

5 Data availability

The health data used in this study is not publicly available but can be requested via NHS England and Improvement NCDR [55]. The shopping dataset used in this study is commercially sensitive and therefore not available for access. All other datasets are open source and can be accessed via the website links given in Table 1.

6 Code availability

The analysis code developed for this paper can be found online at https://github.com/nhsx/commercial-data-healthcare-predictions

7 References

1 Office for National Statistics. Deaths from respiratory disease from 2015 to 2020 and influenza and pneumonia in 2020. https://www.ons.gov.uk/aboutus/transparencyandgovernance/freedomofinformationfoi/deathsfromrespiratorydiseasefrom2015to2020andinfluenzaandpneumoniain2020 [Accessed on 15 December 2021]

2 GOV.UK Coronavirus (COVID-19) in the UK Deaths in United Kingdom https://coronavirus.data.gov.uk/details/deaths [Accessed 15 December 2021]

3 Marini, J. J. & Gattinoni, L. Management of COVID-19 respiratory distress. Jama 323, 2329-2330 (2020).

4 Bedson, J. et al. A review and agenda for integrated disease models including social and behavioural factors. Nature human behaviour, 1-13 (2021).

5 Allen, W. E. et al. Population-scale longitudinal mapping of COVID-19 symptoms, behaviour and testing. Nature human behaviour 4, 972-982, doi:10.1038/s41562-020-00944-2 (2020).

6 Atchison, C. et al. Early perceptions and behavioural responses during the COVID-19 pandemic: a cross-sectional survey of UK adults. BMJ open 11, e043577 (2021).

7 Betsch, C. How behavioural science data helps mitigate the COVID-19 crisis. Nature human behaviour 4, 438-438 (2020).

8 Kalanidhi, K. B. et al. Development and validation of a questionnaire to assess socio-behavioural impact of COVID-19 on the general population. Diabetes & Metabolic Syndrome: Clinical Research & Reviews 15, 601-603 (2021).

9 Steinegger, B., Arola-Fernández, L., Granell, C., Gómez-Gardeñes, J. & Arenas, A. Behavioural response to heterogeneous severity of COVID-19 explains temporal variation of cases among different age groups. Philosophical Transactions of the Royal Society A 380, 20210119 (2021).

10 ZOE COVID study. https://covid.joinzoe.com/ [Accessed on 15 December 2021]

11 Bastani, P. & Bahrami, M. A. COVID-19 related misinformation on social media: a qualitative study from Iran. Journal of medical Internet research (2020).

12 Islam, M. S. et al. COVID-19–related infodemic and its impact on public health: A global social media analysis. The American journal of tropical medicine and hygiene 103, 1621 (2020).

13 Kraemer, M. U. et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. Science 368, 493-497 (2020).

14 Buckee, C. O. et al. Aggregated mobility data could help fight COVID-19. Science (2020).

15 Pivette, M., Mueller, J. E., Crépey, P. & Bar-Hen, A. Drug sales data analysis for outbreak detection of infectious diseases: a systematic literature review. BMC infectious diseases 14, 1-14 (2014).

16 Margevicius, K. J. et al. Advancing a framework to enable characterization and evaluation of data streams useful for biosurveillance. PLoS One 9, e83730 (2014).

17 Park, H.-A., Jung, H., On, J., Park, S. K. & Kang, H. Digital epidemiology: use of digital data collected for non-epidemiological purposes in epidemiological studies. Healthcare informatics research 24, 253-262 (2018).

18 Nevalainen, J., Erkkola, M., Saarijärvi, H., Näppilä, T. & Fogelholm, M. Large-scale loyalty card data in health research. Digital health 4, 2055207618816898-2055207618816898, doi:10.1177/2055207618816898 (2018).

19 Davies, A., Green, M. A. & Singleton, A. D. Using machine learning to investigate self-medication purchasing in England via high street retailer loyalty card data. PloS one 13, e0207523-e0207523, doi:10.1371/journal.pone.0207523 (2018).
20 Aiello, L. M., Schifanella, R., Quercia, D. & Del Prete, L. Large-scale and high-resolution analysis of food purchases and health outcomes. EPJ data science 8, 1-22, doi:10.1140/epjds/s13688-019-0191-y (2019).
21 Uusitalo, L., Erkkola, M., Lintonen, T., Rahkonen, O. & Nevalainen, J. Alcohol expenditure in grocery stores and their associations with tobacco and food expenditures. BMC public health 19, 787-787, doi:10.1186/s12889-019-7096-3 (2019).
22 Lutz, C. S. et al. Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. BMC Public Health 19, 1-12 (2019).
23 Lombardo, J. S., Burkom, H. & Pavlin, J. ESSENCE II and the framework for evaluating syndromic surveillance systems. Morbidity and Mortality Weekly Report, 159-165 (2004).
24 Welliver, R. C. et al. Sales of nonprescription cold remedies: a unique method of influenza surveillance. Pediatric research 13, 1015-1017 (1979).
25 Hogan, W. R. et al. Detection of pediatric respiratory and diarrheal outbreaks from sales of over-the-counter electrolyte products. Journal of the American Medical Informatics Association 10, 555-562 (2003).
26 Sočan, M., Erčulj, V. & Lajovic, J. Early detection of influenza-like illness through medication sales. Cent Eur J Public Health 20, 156-162 (2012).
27 Al-Tawfiq, J. A. et al. Surveillance for emerging respiratory viruses. The Lancet Infectious Diseases 14, 992-1000 (2014).
28 Davies, G. & Finch, R. Sales of over-the-counter remedies as an early warning system for winter bed crises. Clinical microbiology and infection 9, 858-863 (2003).
29 Todd, S., Diggle, P. J., White, P. J., Fearne, A. & Read, J. M. The spatiotemporal association of non-prescription retail sales with cases during the 2009 influenza pandemic in Great Britain. BMJ open 4, e004869 (2014).
30 Hofman, J. M. et al. Integrating explanation and prediction in computational social science. Nature 595, 181-188 (2021).
31 Fisher, A., Rudin, C. & Dominici, F. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. J. Mach. Learn. Res. 20, 1-81 (2019).
32 Smith, G., Mansilla, R. & Goulding, J. Model Class Reliance for Random Forests. Advances in Neural Information Processing Systems 33 (2020).
33 Ljevar, V., Goulding, J., Smith, G. & Spence, A. Using Model Class Reliance to Measure Group Effects on Non-Adherence to Asthma Medication. [Paper to be presented at IEEE Big Data 2021]
34 Breiman, L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science 16, 199-231 (2001).
35 DefiningResearchTable_Oct2017 http://www.hra-decisiontools.org.uk/research/docs/DefiningResearchTable_Oct2017-1.pdf [Accessed on 15 December 2021]
36 Office for National Statistics. Deaths registered weekly in England and Wales, provisional. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/datasets/weeklyprovisionalfiguresondeathsregisteredinenglandandwales [Accessed on 15 December 2021]

37 Moriyama, M., Hugentobler, W. J. & Iwasaki, A. Seasonality of respiratory viral infections. Annual review of virology 7, 83-101 (2020).
38 Ministry of Housing, Communities and Local Government. The English Indices of Deprivation
2019 (IoD2019)
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/835115/IoD2019_Statistical_Release.pdf [Accessed on 15 December 2021]
39 Patel, J. et al. Poverty, inequality and COVID-19: the forgotten vulnerable. Public health 183, 110 (2020).
40 Bennett, J. E. et al. Contributions of diseases and injuries to widening life expectancy inequalities in England from 2001 to 2016: a population-based analysis of vital registration data. The Lancet Public Health 3, e586-e597 (2018).
41 Webb, E., Blane, D. & de Vries, R. Housing and respiratory health at older ages. J Epidemiol Community Health 67, 280-285 (2013).
42 Pannullo, F. et al. Quantifying the impact of current and future concentrations of air pollutants on respiratory disease risk in England. Environmental Health 16, 1-14 (2017).
43 Union, O. E. (OECD Publishing Paris/European Union, Brussels, 2018).
44 Pevalin, D. J., Taylor, M. P. & Todd, J. The dynamics of unhealthy housing in the UK: A panel data analysis. Housing studies 23, 679-695 (2008).
45 Ellaway, A. & Macintyre, S. Does housing tenure predict health in the UK because it exposes people to different levels of housing related hazards in the home or its surroundings? Health & place 4, 141-150 (1998).
46 Gartner, A., Farewell, D., Roach, P. & Dunstan, F. Rural/urban mortality differences in England and Wales and the effect of deprivation adjustment. Social Science & Medicine 72, 1685-1694 (2011).
47 Prats-Uribe, A., Paredes, R. & Prieto-Alhambra, D. Ethnicity, comorbidity, socioeconomic status, and their associations with COVID-19 infection in England: a cohort analysis of UK Biobank data. medRxiv (2020).
48 Hajat, S., Bird, W. & Haines, A. Cold weather and GP consultations for respiratory conditions by elderly people in␣ 16 locations in the UK. European journal of epidemiology 19, 959-968 (2004).
49 Nichols, G. L. et al. Coronavirus seasonality, respiratory infections and weather. BMC infectious diseases 21, 1-15 (2021).
50 Office for National Statistics. 6. Super Output Area (SOA)
https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography#super-output-area-soa [Accessed on 20 January 2022]
51 Local Government Information Unit. Local government facts and figures: England
https://lgiu.org/local-government-facts-and-figures-england/ [Accessed on 20 January 2022]
52 Office for National Statistics. User guide to mortality statistics.
https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/methodologies/userguidetomortalitystatisticsjuly2017 [Accessed on 15 December 2021]
53 Breiman, L. Random forests. Machine learning 45, 5-32 (2001).
54 Lundberg, S. M. & Lee, S.-I. in Proceedings of the 31st international conference on neural information processing systems. 4768-4777.
55 NHS Health Data & Analytics - NCDR https://www.ardengemcsu.nhs.uk/services/business-intelligence/ncdr/ [Accessed on 15 December 2021]

Table 1. Model data used for the target for predictions and feature creation (variable inputs). (All inputs are aggregated to 314 LTLAs across England, and weekly rates if the variable is dynamic with a 17 day lag unless stated otherwise.)

Data Description	Data Source	Research relating variable to respiratory disease	Target (y)	Temporality	Group for MCR
ONS (Office for National Statistics) deaths registered weekly in England from diseases of the respiratory system (ICD-10 Coding J00–J99)	NHSX/Digital receives data through request from ONS - https://digital.nhs.uk/services/primary-care-mortality-database.	N/A	Weekly deaths from respiratory disease in a LTLA	dynamic	N/A
-	-	-	**Feature (X)**	-	-
Week number	N/A	Moriyama, M., Hugentobler, W. J., & Iwasaki, A. [37]	Number of week from 1 to 52	dynamic	week
Weekly sales data from a UK high street retailer with stores distributed across the UK	University partnership with UK Retailer	Davies & Finch [28]	Total weekly sales	dynamic	sales
			Weekly sales of decongestant	dynamic	sales
			Weekly sales of throat meds	dynamic	sales
			Weekly sales of dry cough meds	dynamic	sales
			Weekly sales of all cough meds	dynamic	sales
			Total weekly sales 24 day lag	dynamic	sales
			Weekly sales of decongestant 24 day lag	dynamic	sales
			Weekly sales of throat meds 24 day lag	dynamic	sales
			Weekly sales of dry cough meds 24 day lag	dynamic	sales
			Weekly sales of all cough meds 24 day lag	dynamic	sales

			Local Ratio of Weekly sales of decongestant	dynamic	sales
			Local Ratio of Weekly sales of throat meds	dynamic	sales
			Local Ratio of Weekly sales of dry cough meds	dynamic	sales
			Local Ratio of Weekly sales of all cough meds	dynamic	sales
			Multiplier of Weekly sales of decongestant	dynamic	sales
			Multiplier of Weekly sales of throat meds	dynamic	sales
			Multiplier of Weekly sales of dry cough meds	dynamic	sales
			Multiplier of Weekly sales of all cough meds	dynamic	sales
English indices of deprivation 2019. Index of Multiple deprivation ranks Lower-layer Super Output Areas in England from 1 (most deprived area) to 32,844 (least deprived area). Combines data from seven domains: Income Deprivation (22.5%), Employment Deprivation (22.5%), Education, Skills and Training Deprivation (13.5%), Health Deprivation and Disability (13.5%), Crime (9.3%), Barriers to Housing and Services (9.3%), Living Environment Deprivation (9.3%). Summaries are available at LTLA level. Living Environment measures the quality of both indoor and outdoor local environments. The 'indoors' living environment measures the quality of housing; while the 'outdoors' living environment contains measures of air quality and road traffic accidents. [38]	https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019	Patel et al [39], Bennett et al. [40], Webb, Blane, and de Vries [41], Pannullo et al. [42]	LTLA IMD score for Living Environment Deprivation Domain	static	IMD
			LTLA IMD score for Crime Domain	static	IMD
			LTLA IMD score for Barriers to Housing and Services Domain	static	IMD
			Population weighted average of the combined IMD ranks for the LSOAs in the LTLA	static	IMD
			Population weighted avg of the combined IMD overall score for the LSOAs in the area of interest	static	IMD
			Extent of deprivation within a local authority	static	IMD
			Concentration of deprivation within a local authority	static	IMD
2019 mid year population estimates based on the ONS census available at LTLA level and above.	https://www.nomisweb.co.uk/datasets/pestsyoala	OECD/European Union. "Mortality from respiratory	Population aged 16 to 24 in an area	static	age
			Population aged 25 to 49 in an area	static	age

		diseases." Health at a glance: Europe 2018: state of health in the EU cycle [43]	Population aged 50 to 64 in an area	static	age
			Population aged over 64 in an area	static	age
			Population density for the LTLA (people/square km)	static	demo
			Percent of LTLA that are male	static	demo
			Percent of LTLA that are female	static	demo
Housing age data from the Valuation Office Agency 2020	https://data.cdrc.ac.uk/dataset/dwelling-ages-and-prices/resource/dwelling-age-group-counts-lsoa	Pevalin, Taylor, and Todd [44], Ellaway and Macintyre [45].	Percent of houses in LTLA built prior to 1919	static	housing
			Percent of houses in LTLA built prior to 1940	static	housing
			Percent of houses in LTLA built prior to 1973	static	housing
			Percent of houses in LTLA built prior to 1983	static	housing
Land use in England from 2018 live tables from the Department for Levelling Up, Housing and Communities and Ministry of Housing, Communities & Local Government	https://www.gov.uk/government/statistical-data-sets/live-tables-on-land-use	Gartner et al. [46]	Percent of land use associated with community buildings	static	land_use
			Percent of industrial land use in LTLA	static	land_use
			Percent of residential land use in LTLA	static	land_use
			Percent of land used by transport and utilities in LTLA	static	land_use
			Percent of agricultural land use in LTLA	static	land_use
			Percent of natural land use in LTLA	static	land_use

			Percent of land used by outdoor recreation (e.g., sports fields and parks) in LTLA	static	land_use
Property type/ethnicity/household composition ONS 2011 census datasets	https://www.nomisweb.co.uk/sources/census_2011	Prats-Uribe, Paredes and Prieto-Alhambra [47]	Percent of people in LTLA of non white ethnicity	static	demo
			Percentage of lone parent families in LTLA	static	demo
			Percent of "other children" in families in LTLA	static	demo
			Percent of detached houses in LTLA	static	housing
			Percent of semi-detached houses in LTLA	static	housing
			Percent of terraced houses in LTLA	static	housing
			Percent of flats in LTLA	static	housing
Weather. ERA5 data from European Centre for Medium-Range Weather Forecast	https://copernicus.eu/	Moriyama, Hugentobler & Iwasaki [37], Hajat, Bird & Haines [48], Nichols et al. [49]	Weekly average rainfall in LTLA	dynamic	weather
			Weekly total rainfall in LTLA	dynamic	weather
			Weekly minimum temperature in LTLA	dynamic	weather
			Weekly average temperature in LTLA	dynamic	weather
			Weekly maximum temperature in LTLA	dynamic	weather