The Alan Turing Institute

Data Study Group Final Report: NHS Scotland

4-8 September 2017

Predicting risk of hospital admission in Scotland

Table of Contents

---

https://doi.org/10.5281/zenodo.2539563

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1

Executive Summary

Context and Remit

This report presents the output of a week-long collaboration between the Alan Turing Institute, NHS Scotland, and the National Services Scotland's Information Services Division (ISD) to investigate and update the current decision support tool for identifying patients at risk of admission - the SPARRA (Scottish Patients at Risk of Readmission and Admission) model.

Main Objectives

The main aim of the data study week was assessment and development of a state-of-art tool for admission risk ranking and prediction. More precisely, based on a representative dataset containing admission history of Scottish patients from four individual risk groups (FE = frail elderly, LTC = long-term conditions, U16 = under 16, YED = young emergency): 1) To determine suitable tasks and modelling strategies for prediction of individualized admission risk given the structure of the dataset 2) To review and quantitatively assess ranking and prediction accuracy of the existing methodology used for individual risk scores for admission to hospital, the SPARRA (Scottish Patients at Risk of Readmission and Admission) model(s). 3) To develop suitable models and modelling strategies addressing the risk prediction and ranking task(s) identified in (2)

Data overview

Data access was via a safe haven provided by eDRIS, and limited to pre-approved researchers who had undergone the necessary process for requesting access. Access to the safe haven's virtual machines was only possible from the Alan Turing Institutes local servers.

In total the data study group was provided with monthly aggregated panel data and admission time stamps for 1.8 million patients, over the largest part of a five-year period (July 2013- August 2017). To account for the limited computational resources available, analyses were conducted on random subsamples ranging in size from 1000 to 50000.

Results

I. As an important point of consideration it was flagged up that the (risk of) admissions to predict should be split between elective and emergency admissions, as the former are by definition planned and hence predictable by a GP, while the latter are not. Depending on the amount of information to be predicted and its granularity, the prediction task may be seen as probabilistic (risk) classification, ordinal target regression, count prediction, time-to-event (= "survival") modelling, or ranking versions thereof. Further distinctions in the task may be made depending on the degree of use of temporal structure in predictor/feature variables, and the presence or frequency of updating of the model as time progresses (see section 2 for details). Due to time and resource constraints, in the study only the simplest type of modelling task was considered, probabilistic risk prediction models that are static in time.

II. The SPARRA models are probabilistic classification models, one for each of the four risk groups, which are semi-static (through periodic re-fit updates). SPARRA scores were provided with the dataset for evaluation purposes. In naive experiments including both elective and emergency admission, the SPARRA score predicts risk of admissions (elective + emergency) better than an uninformed risk baseline (prediction average population risk for everyone). However, in these cases not distinguishing between elective and emergency admissions limits interpretability. In more targeted experiments, and in keeping with the SPARRA scores original purpose of predicting emergency admission risk, the results suggest the SPARRA scores are not significantly better than the uninformed baseline. (see results section for details)

III. A systematic comparison of off-shelf probabilistic risk prediction strategies shows: In all four risk groups, for both predictions of, any type of hospital admission, and prediction of emergency admissions only, logistic regression models on a larger set of pre-selected variables are more accurate than the SPARRA models and the uninformed baseline. An off-shelf random forest model is in turn either undistinguishable or significantly more accurate than the logistic regression model (see section X for details).

Main Conclusions

I. Admission risk prediction may be addressed by multiple modelling tasks, the less specific risk predictions being off-shelf and well-explored, the more granular ones touching widely open or active areas of contemporary research. (2 & 3) The SPARRA models miss the crucial distinction between elective and emergency admissions. They are sometimes slightly better than a random guess, but in all four risk groups easily outperformed by both classical and modern black-box models obtained from a standard semi-automated data processing and model selection workflow. Random forests appear to predict best, and interpretable models such as logistic regression are only slightly (but significantly) worse.

Limitations

I. The SPARRA algorithm was not provided, only its output, i.e. the risk scores calculated from the existing SPARRA models. While this is sufficient as a basis for assessment and comparison of the predictions under certain assumptions, it may yield overly optimistic estimates in comparison to a set-up where the algorithm itself is available and assessed.

II. Time and computational resource limitations only allowed for a basic experiment for the simplest modelling task, risk prediction by binary probabilistic classification. Advanced modelling tasks (such as A&E visit counts and multi-label classification) were explored and showed much promise (see sections X), but it was impossible to do so systematically.

III. Due to time and computational resource limitations, the analyses were conducted on smaller sub-samples. Hence phenomena (e.g., differences between methods) which are present in reality may appear as not statistically significant on the smaller dataset while they may be visible in the original dataset. That is, the constraints encourage type II errors (while type I errors are controlled for by the applied methodology).

Contributors

Bilal Mateen *(Moderator)*
University College London

Franz Kiraly
Alan Turing Institute

Sebastian Vollmer
Alan Turing Institute

Louis Aslett
Durham University

Raphael Sonabend
University College London

Ioanna Manolopoulou
Alan Turing Institute

Cemre Zor
University of Surrey

Nada Jankovicova
University of Warwick

Nathan Cunningham
University of Warwick

Ieva Kazlauskaite
University of Bath

Anil Rao
University College London

Chao Wang
Queen Mary University of London

Samuel Dua Oduro
NHS National Services Scotland

Introduction

The Kerr report [1], published in 2005, outlined a number of proposals for improving, and modernizing, the delivery of healthcare in Scotland.

Its key aims included:

- ensuring delivery of high quality care;

- reducing waiting times;

- providing enhanced support to rural communities, empowering clinical staff;

- reducing the health gap between rich and poor;

- providing a modern health service which is value for money through increased use of new technology.

The report emphasized the importance of viewing the NHS as a service delivered in the local community as opposed to in hospitals and identified a need to shift from a reactive system to one which anticipates the needs of patients. Risk scores can assist health care professionals in prioritizing patients with complex care needs who are likely to benefit most from anticipatory health care. Moreover, identifying high risk individuals would also allow for more effective resource allocation.

In order to identify people at risk of emergency hospital admission, the Information Services Division (ISD) developed three iterations of the Scottish Patients at Risk of Readmission and Admission (SPARRA) model, with the latest release taking place in 2012. Risk scores are predicted separately for several cohorts: U16 Under 16 year olds (U16); Younger patients who have attended the Emergency Department (YED); LTC patients with Long Term Conditions (LTC); and, Frail Elderly (FE). The original SPARRA model used between 16 (FE) and 27 (U16) features depending on the risk cohort in question. These were chosen by domain experts (public health physicians), rather than in a data-driven fashion. The SPARRA model hasn't been developed since 2012 due to other priorities and, whilst highly regarded at the time, it is felt that the model would benefit from the exploration of more innovative machine learning approaches to more accurately predict admission risk and provide actionable intelligence to clinicians.

Machine learning has proven to be effective at a wide variety of tasks traditionally associated with human intelligence. The original SPARRA model uses ordinary logistic regression on a set of preselected features to identify a specific set of risk factors which influence the risk of emergency admissions. However, recent advances in automated feature selection and classification methodologies allows us to explore new avenues in risk estimation and decision making.

Objectives

The primary objective we aimed to address in this report is:

1) To test whether machine learning methodologies can be used to improve the prediction of Admission/No-Admission within the next 12 months compared to the current SPARRA model, and compared to a dummy model in which no features are used for prediction. Here, the outcome is a binary target and the task is essentially binary classification.

The secondary objective is:

2) To test whether machine learning methodologies can be used to predict ACSC admissions, which are considered by some be preventable.

We investigated the above questions in each of the four cohorts separately, as well as in the entire sample. Furthermore, different prediction accuracy measures were calculated to reflect the different properties of the resulting models.

Data

We were provided with monthly aggregated records for 1.8 million patients over a course of five years (2013-2017), from four selected territorial Health Boards in Scotland. The data are of a panel/longitudinal type, meaning it is a combination of cross-sectional (observations are made on many patients) and time-series data (repeated observations are made at different time points).

Variables

For each month a total of 279 features are recorded for each patient, including patient-level characteristics, the characteristics of their hospital, and the derived scores from the original SPARRA model for comparison:

1) Index variables
- Unique patient identifier
- Year/month of cross-section

2) Hospital-level features
- Hospital identifiers
- Health board of hospital
- Deprivation score for health board (quintiles and deciles)

3) Demographic features
- Age at start of year
- Sex (1 = female, 2 = male)
- Date of death
- Deprivation score for patient's residence (quintiles and deciles)

4) Clinical features
- Prescriptions
  - No. of British National Formulary (BNF) sections from which medications prescribed in last year
  - No. of items dispensed for each BNF chapters and subsections in relevant month
- Cause of admission
  - Main diagnosis (defined as that diagnosis which prompted admission)
  - No. of different diagnosis groups
  - No. of alcohol/substance/fall/self-harm related admissions
  - Specialty of department admitted to
  - Indicator of certain long-term conditions: arthritis, asthma, atrial fibrillation, cancer, cerebrovascular disease, chronic liver disease, COPD, dementia, diabetes, epilepsy, heart disease, heart failure, multiple sclerosis, Parkinson's, renal failure
  - Number of outpatient attendances
- Admissions and bed days (aggregated over previous three years)
  - No. of emergency admissions/bed days

- Date of most recent emergency admission
- No. of elective admissions/bed days
- No. of daycase admissions
- Total bed days

5) SPARRA-derived features

- Risk scores by cohorts: frail and elderly (age 75+), long-term conditions (age 16-74), younger emergency department (age 16-55), and under 16

Data Quality

In the process of producing subsamples that the virtual machine would be capable of processing, several issues were encountered concerning the completeness and fidelity of the data. Those with substantial bearing on how the data was subsetted and subsampled for the modelling experiments are discussed below.

1. Completeness of records

A sample of data for time-to-event analysis for prediction of an event in 2016 was created that subsetted the individuals so that they would only be included if they had at least 1 months' worth of data in 2013, 2014, and 2015. This was predicated on the assumption that if at least 1 month existed, then in principle all 12 should be present. It was found that between 10-20% of the subsampled patients did not have a complete record through 2013-2015 i.e. assuming the patient was present in April 2013 they did not have 29 months' worth of continuous data. Typically, in these cases, there is a contiguous set of months where the records are missing, after which data collection for the individuals recommences. For example, a 3-year-old had data missing for the months Nov-Jan, but data collection restarted in Feb. As such, many of the plausible explanation posited by the group, e.g. as this was a balanced sample (50% admitted in 2016, 50% not admitted in 2016) that we may have created a sample with an inflated number of patients that die relative to the baseline population had to be discarded as they did not explain the pattern of missingness. One of the remaining theories is that these patients may have temporarily moved area or GP practice. Alternatively, a lack of recent interactions with the health service at a given point in time may be another explanation.

Our suggested approach to address this problem was to fill in the missing months by carrying forward the last observed entry, i.e. assuming that nothing changed during the unrecorded period. Code to execute this carry forward task has been saved in the safe haven. In the absence of knowledge on the cause of the missingness, we would argue this is a sensible approach. However, it does falsely concentrate any changes that occurred in that period in the first month where the ground truth is again available. We are not aware of any simple implementation for imputation (of any kind) in this situation, as any such method would require the inclusion of a threshold based on the annual aggregate variables from the next ground truth observations, so as to prevent the aforementioned ground truth aggregates being altered by the imputed values.

Furthermore, some of the features, including Main Diagnosis and Specialty, which are intuitively likely to be highly associated with admission, have very few categories and are frequently incomplete. Approximately 80% of the Specialty feature is missing and 85% of Main Diagnosis missing. We think it is important to highlight that although a number of hospital episodes may not ever be labelled with a specific diagnosis, there appears to be a field for such cases (i.e. Signs

and symptoms of illness), which should ideally be used in such cases. Given the inability of some off-the-shelf methods to handle missing data, both of these features had to be removed.

2. Accuracy of records

Our analysis identified a number of questionable records. For some unique patient IDs we found that there was more than one entry per month, suggesting the patient IDs were not truly unique, or perhaps an issue occurring during the processing of the data due to the short time period in which multiple large data extracts were carried out. These individuals appeared to have identical demographic features and IDs, but with drastically different medical histories, suggesting either a non-unique identifier or inaccurate recording of data

3. Health boards

In the admission data provided (SMR01 and SMR04), only four of the health board patients should have been represented given the description of the data provided by NHS Scotland and the ISD team. However, the data contain nine different health boards. Interestingly, the additional five health boards occurred in much smaller than expected frequencies, suggesting that most of the patients from the other five health boards have been excluded. It is not entirely clear why some patients from these other health boards are present in our data, but it is likely that patients from these 5 boards will have resided in one of the 4 selected boards at a point in time.

4. Death versus hospital admission

An additional issue identified is that death was filtered back through previous records in the SPARRA database, causing the patient to appear dead from the beginning of their records. Whilst correcting this issue was relatively simple, and code is available in the VM to execute this task, we thought it might be helpful to flag the error.

5. Patients in SMR but not SPARRA

When subsampling the data, the UPIs were used to draw samples from the larger files, so that loading the complete excel sheet could be avoided, as the task was too complex for the computational resources available. It was noted by one of the DSG participants that there are many patient UPIs which appear in the admission data (SMR01 and SMR04) but not in SPARRA. The reason for which was not readily apparent, but it should be noted that not all patients in Scotland appear in SPARRA data.

Exploratory Analysis

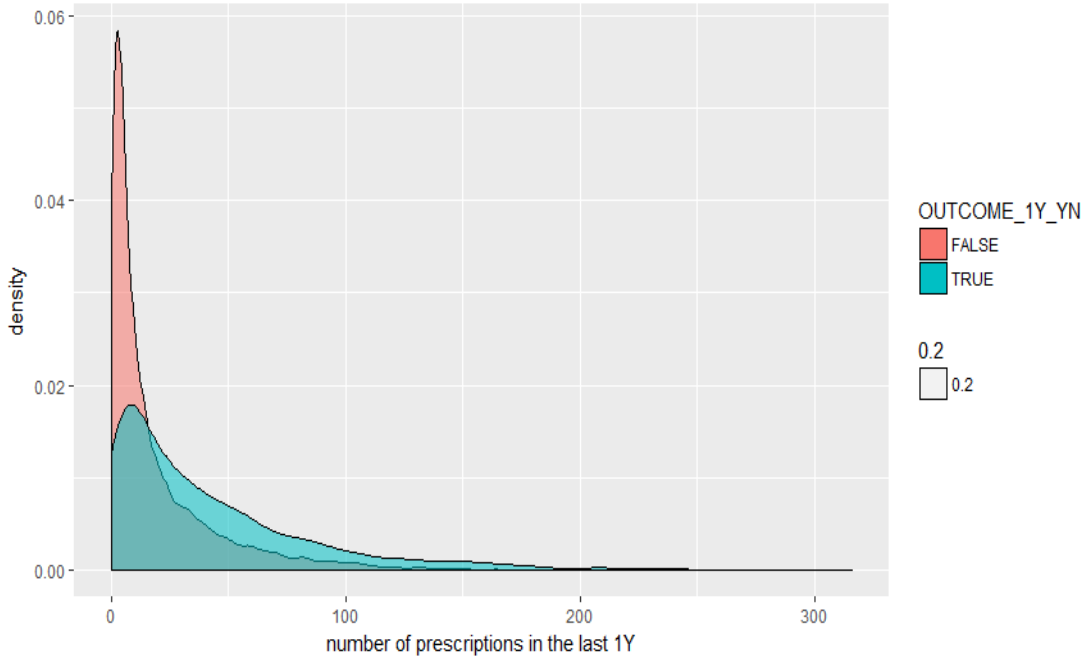
As an initial avenue of exploration we examined the impact of a number of different features on a patient's risk of future admission.

Cross-sectional Analysis

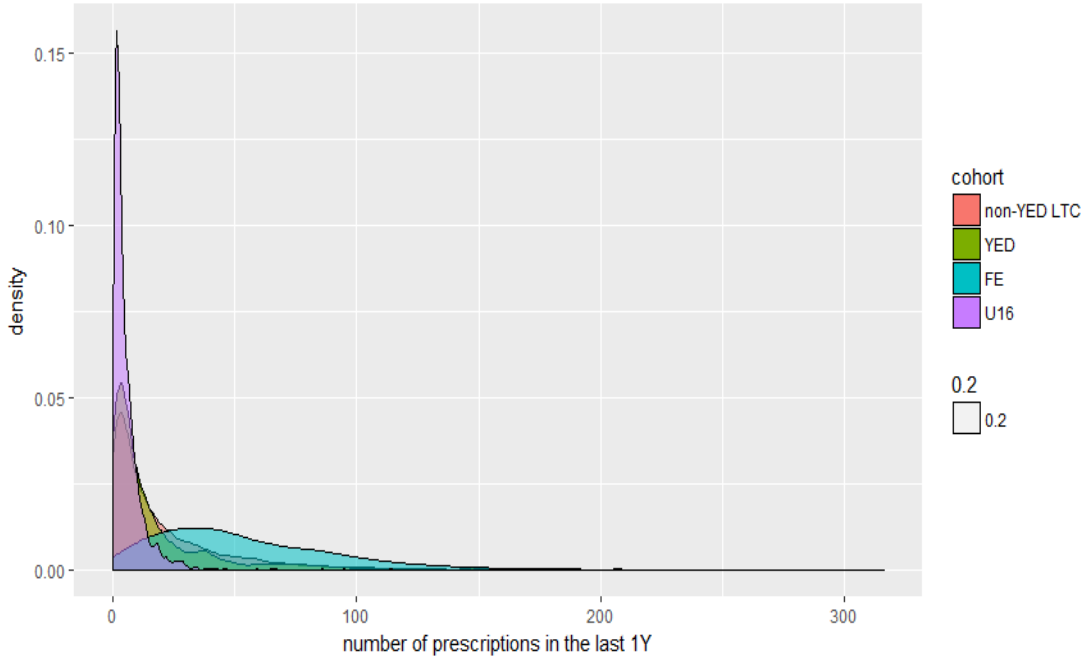
Firstly, we performed a cross-sectional analysis by examining patients at a fixed time point (April 2015). Note that for this analysis, the LTC cohort was taken to be those subjects that are members of LTC but are not members of YED (which is a subset of LTC) For that analysis, a representative subsample of 50,000 from December 2015 was used (dataset 2).

Number of dispensed prescriptions

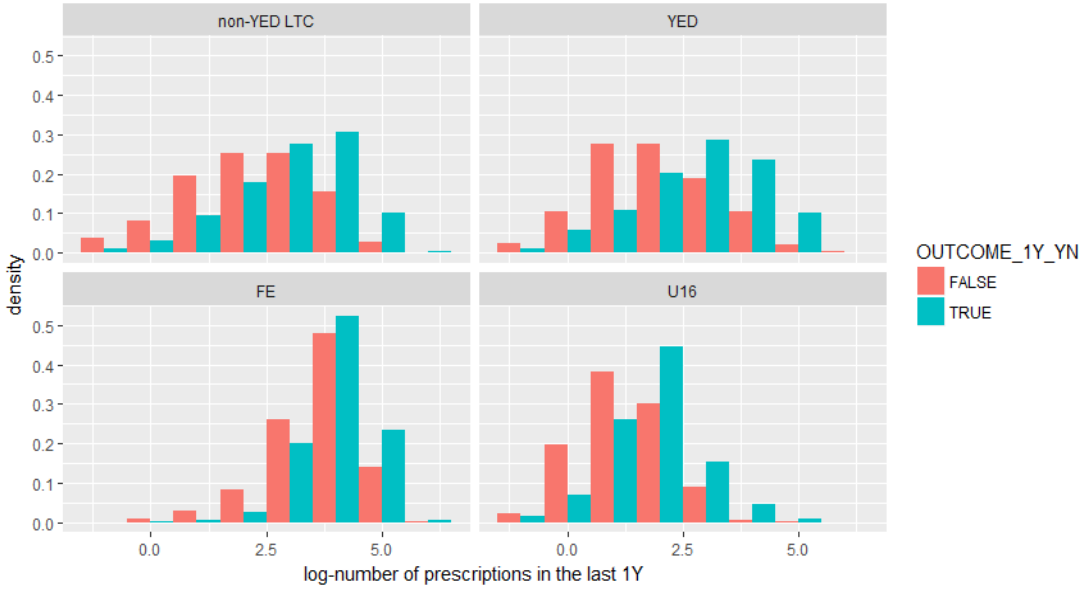
- The plot for the overall distribution of number of dispensed prescriptions in the last year (NDP) over all subjects showed that NDP was heavily skewed towards zero, and the NDP for admitted subjects is larger than for the non-admitted patients.
- However, differences could be observed in this distribution after splitting by cohort. For example, the FE cohort exhibited a less skewed distribution as this cohort tended to take a larger amount of medication overall compared to the other cohorts, while the U16 cohort was highly skewed.
- Finally, the histogram of NDP for each cohort, split by outcome, appeared to indicate a larger NDP for admitted subjects than non-admitted. However, the magnitude of this difference differs by cohort, e.g. it is smaller for the FE cohort than the others.



Density plot for number of dispensed prescriptions in the last year (NDP) for all subjects, varied by outcome of Admitted/Not-Admitted and cohort.



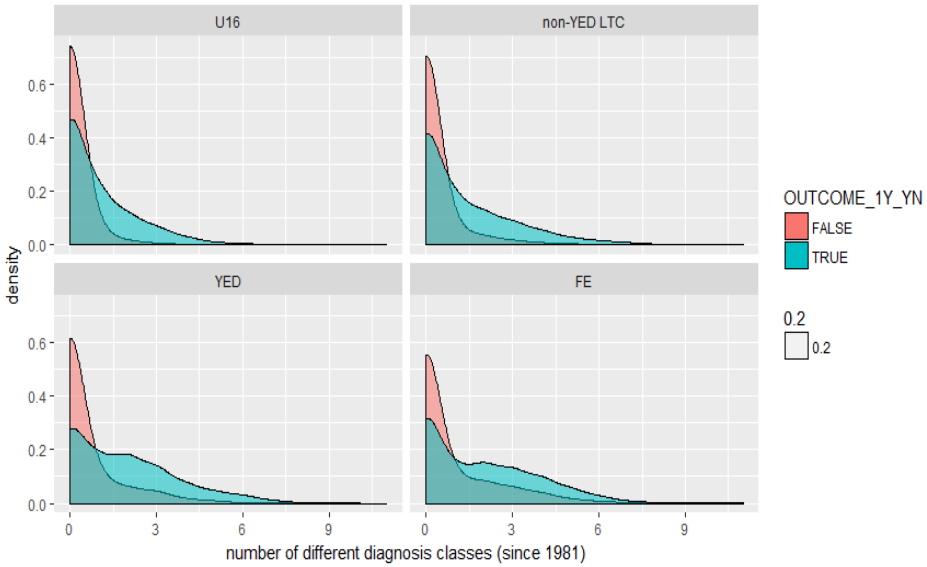
Density plot for number of dispensed prescriptions in the last year (NDP) for all subjects, varied with the outcome of Admitted/Not-Admitted.



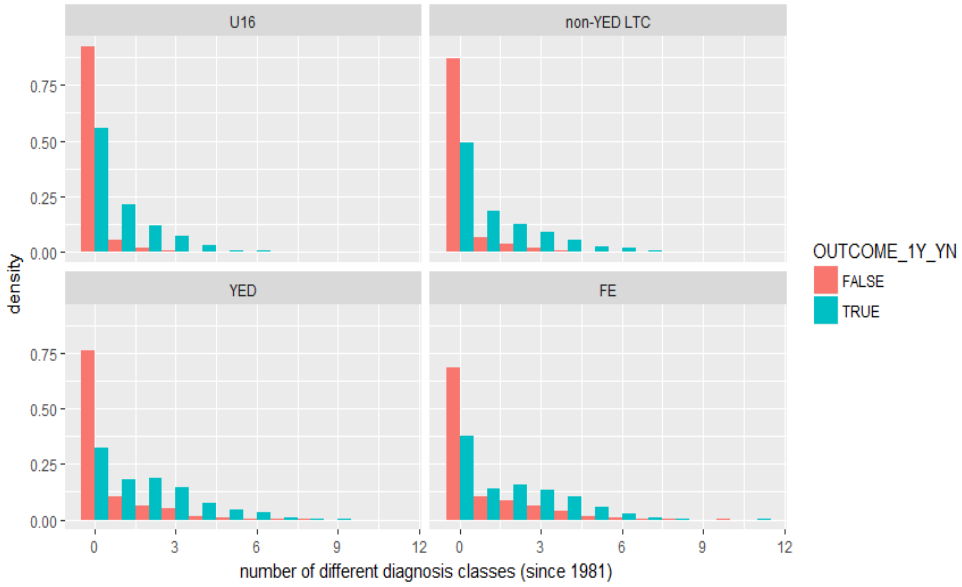
Histogram for the number of dispensed prescriptions in the last year (NDP) varied with the outcome of Admitted/Not-Admitted by cohort

Number of diagnosis'

- The density plot and histogram for number of diagnosis' suggest that the presence of multiple co-morbidities is more frequently observed in the admitted patients than the not admitted group, across all four risk groups.



Density plot of the of number of diagnosis' for all subjects, varied with the outcome of Admitted/Not-Admitted.



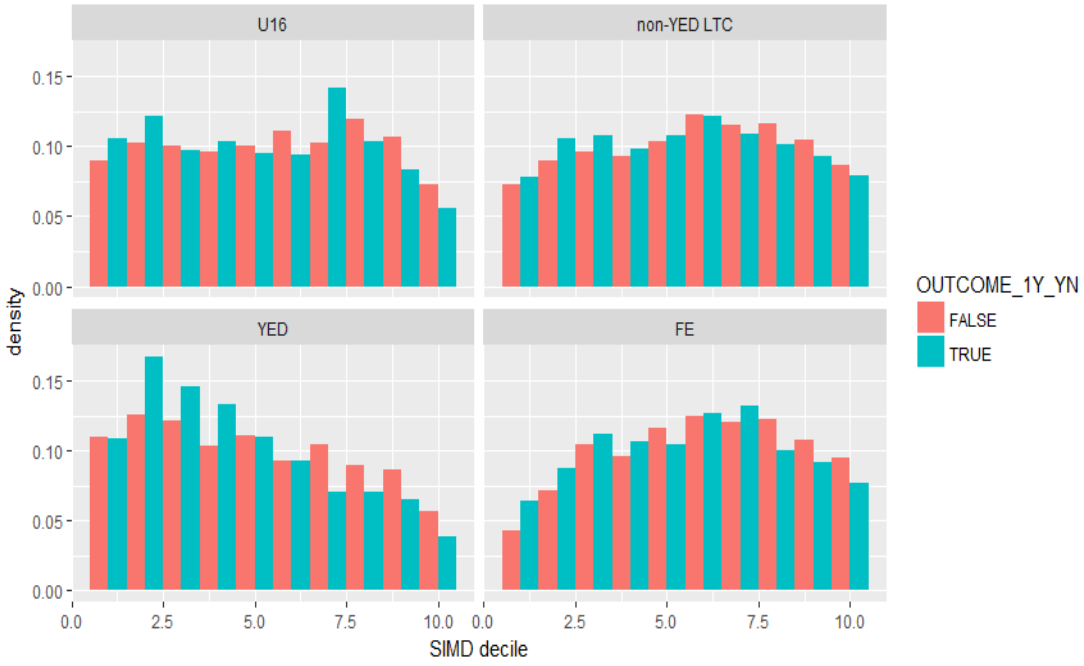
Histogram for the number of diagnosis' for all subjects, varied with the outcome of Admitted/Not-Admitted, and by risk cohort

Previous admissions

- We also considered showing the analogous plots to those above using variables describing the number of previous admissions because these would be expected to be highly related to outcome, but the plots were not informative due to the small range of discrete values for those variables.

Deprivation deciles

- The histograms for deprivation decile by risk group suggests that there is a positive correlation between deprivation and the number of inpatient admissions.



Admissions by deprivation decile for each cohort

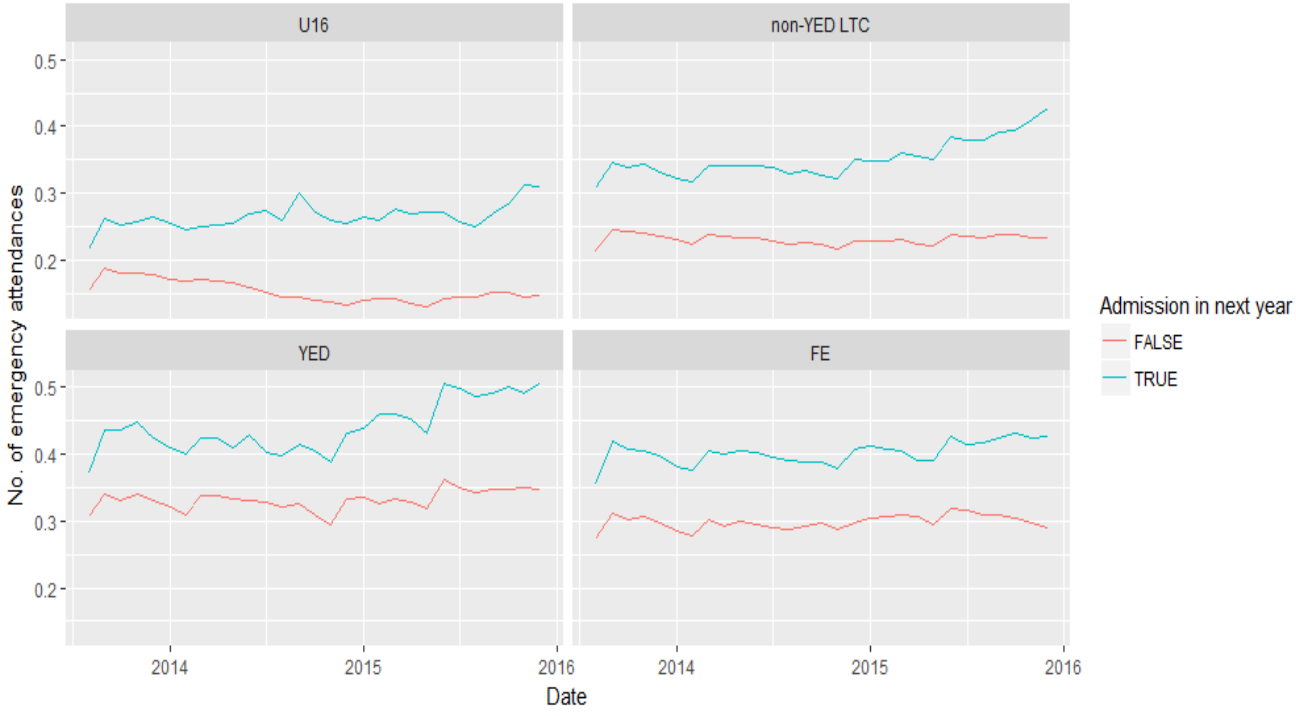
Time-Series Analysis

Time-series exploratory analysis was performed in which the outcome was taken to be Admitted to Emergency/Not Admitted to Emergency (dataset 2). In that analysis, variables were plotted over time, both by taking the average over each outcome group, and by showing the data at the subject level, for random subjects.

Some of the features in the data have a time-series element to them as observations are made on a monthly/yearly basis. We have carried out observations on patients who have data for 29 months. We have explored the trends in such data for the groups of patients who did and did not get admitted in the prediction period.

ED attendance

- Patients who were admitted on average had a larger number of ED attendances during the three-year period prior to the prediction year than those who were not admitted. Moreover, the number of ED attendance for patients who were not admitted looks relatively constant throughout the three year prior to prediction year. Finally, there appears to be a slight increasing trend in the patients who were admitted in the prediction year, especially for the LTC, YED and U16 populations.



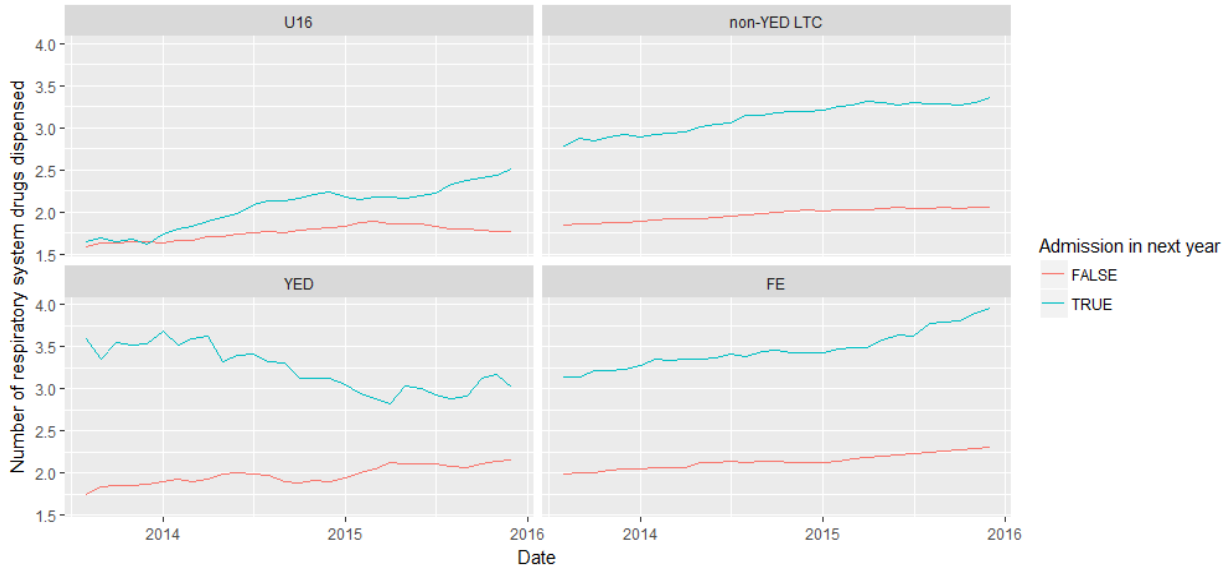
Emergency department attendance by emergency admission or not, organized by cohort

Drugs dispensed

Examination of the time-based trend in total number of drugs dispensed over the pre-prediction period suggests that although the admitted population appear to be receive a larger number of drugs, there is no clear trend over time within the four risk groups.

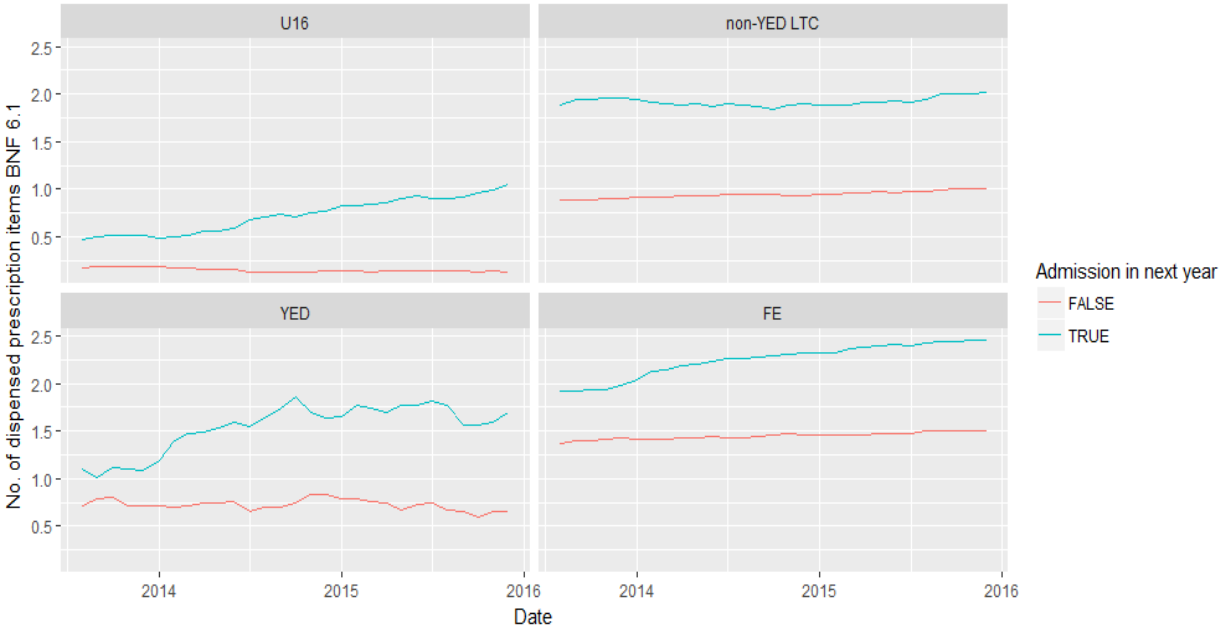
Total number of drugs dispensed varied by outcome over time

However, if we delve deeper into the specific groups of drugs dispensed, there do appear to be time-based trends. For example, with regards to respiratory system drugs, in the U16, LTC and FE populations the number of drugs dispensed seems to increase more drastically for those whom were admitted compared to those who were not. This could be a reflection of the increased burden of disease experienced by the individuals, which may explain why they were admitted and others were not. More interestingly, the decline in the number of drugs dispensed for those admitted in the YED cohort compared to the modest increase for those not admitted, could be interpreted to mean that it is poorer compliance that is driving emergency admissions in that population.



Total number of respiratory drugs dispensed varied by outcome over time

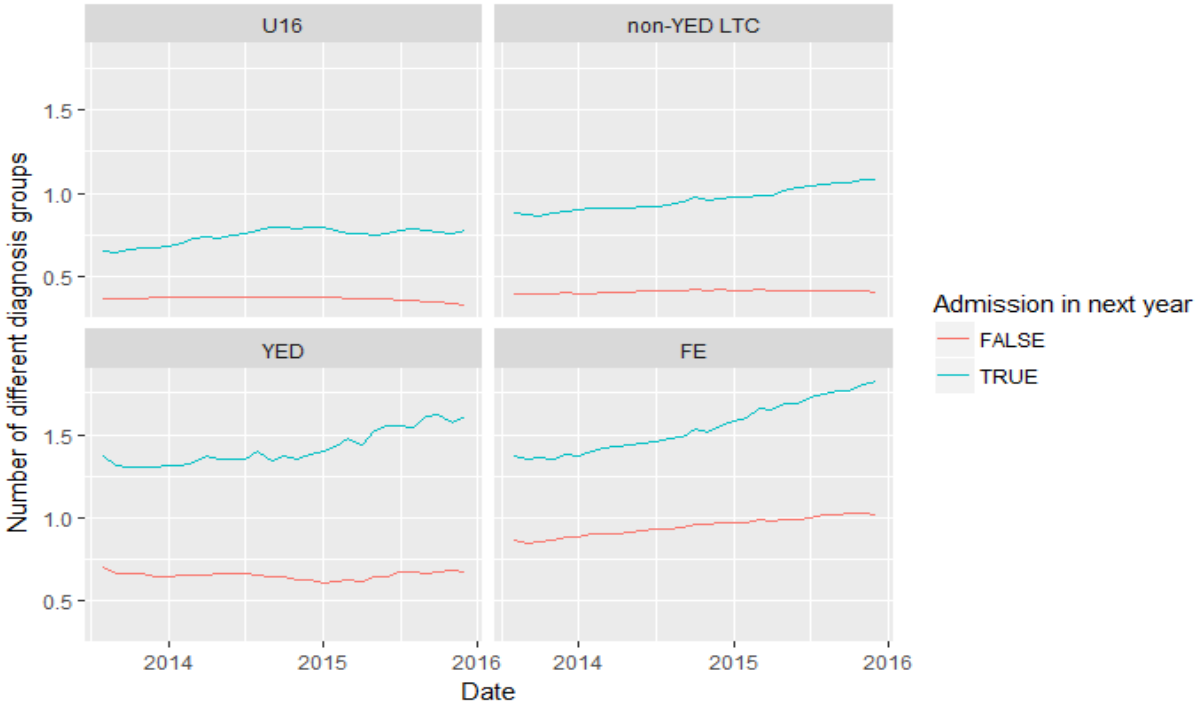
A similar pattern of admission being associated with increased disease burden, represented by increased drug prescriptions can be seen in diabetes as well, for all risk groups except the LTC population



Total number of diabetic drugs dispensed varied by outcome over time

Co-morbidities

Finally, increasing disease burden (denoted by the number of co-morbidities an individual has) appears to be a distinguishing feature between those admitted and those that aren't in the LTC and YED populations, where the burden remains constant throughout the pre-prediction period. In the FE population both the admitted and non-admitted groups appear to experience an increasing disease burden, but the magnitude of increase in the admitted population seems to be much larger. The U16 cohort is somewhat unique in that the non-admitted population appear to experience a small decrease in disease burden compared to the admitted U16s' small increase.



Total number of co-morbidities varied by outcome over time

Methodology: Supervised Classification Modelling of Hospital Admission

Our primary goal is to improve prediction of admission within a selected prediction year. This classification task can be performed on the entire data set, and on the subsets of psychiatric and non-psychiatric patients. Depending on the chosen outcome, different methodologies may be relevant. The possible outcomes include:

- Admission (emergency or elective) vs No admission

- Emergency admission vs everyone else (including elective admissions)

- ACSC admission vs Admission for other reasons.

Each of these tasks is a binary classification task.

Furthermore, we consider classifying the patients with ACSC and the patients with other conditions. One possible approach is to use two consecutive classifications, separating the patients into admitted and not admitted, and then classifying the admitted patients into ACSC and not ACSC. Alternatively, multiclass classifiers may be used to perform the two tasks at once.

Data Preparation

Establishing a Maximum Viable Feature Set

Initially the benchmark modelling experiments were run without any manual selection of variables. However, several errors were encountered by the classical modelling methods (e.g. the Logistic Regression). Namely that of a rank-deficient matrix, and an inability to converge. Upon manual examination we discovered that there are some pairs of features with different names but whose values are near-identical; we were informed that this likely resulted from the dataset being pulled from multiple sources with different variable names for the same observation. The resulting multi-collinearity was hypothesized to be the reason many of our models did not function as expected.

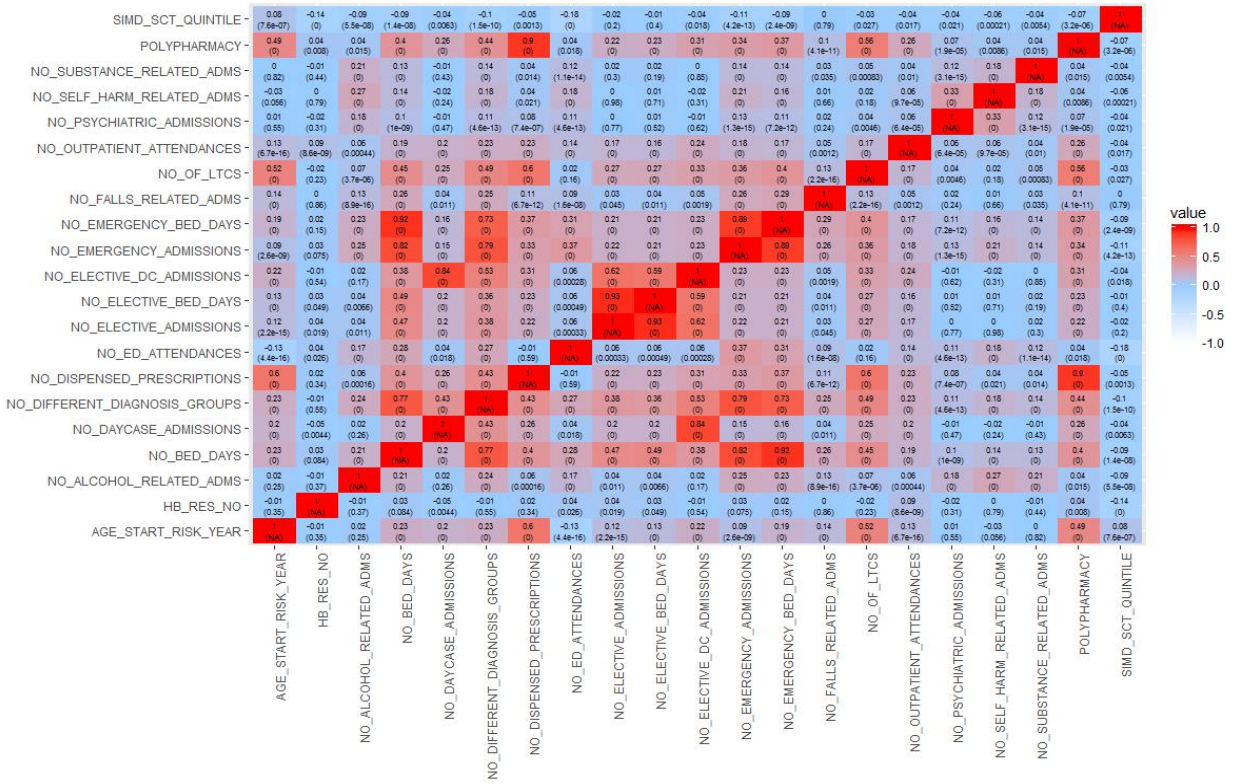
The step-wise process by which the variables that we believe to be causing the aforementioned problem were removed is as follows:

1) Remove all variables that were redundant due to coding. For example, 'Hospital Name' could be removed if 'Hospital Code' was kept.
2) Remove SCT_Decile and HB_Decile which proved highly collinear and instead keep the less granular quintiles (which are also used by the SPARRA model v.3).
3) Case-wise delete individuals with observations of NAs in the Quintile scores as these tended to have missing a substantial number of observations in other features as well
4) Remove Specialty and Main Diagnosis
5) Remove all the BNF 'Chapter' summary features as the domain experts concluded that aggregation by organ system did not accurately reflect the likely risk associated with each chapters' sub-group, and instead added noise to variables that were likely to be useful. Moreover, all of the chapter sub-groups were initially retained so no information was theoretically lost.

6) Remove all variables which upon subsampling results in all individuals selecting one factor level
7) Remove all 'count' variables, e.g. "Number of Dispensed items…" where the total number of individuals in the factor levels above "0" was not at least 15. This was to ensure that no fold of the 10-fold cross validation set-up did not, at minimum, include 2 factor levels for the variable.
8) Filtered for people who are known to be alive in 2016 then removed the Date of Death and Patient_Died columns

After automating the procedure, we were left with the features described in Appendix tables 1 through 10.

The purpose of the automation was to include the maximum possible number of variables such that all our models would work, as such we were still left with some collinear variables. In particular, we found that Polypharmacy and Number of Dispensed Predictions had a Spearman correlation of 0.9017 and Number of Emergency Bed Days and Number of Total Bed Days had a correlation of 0.9197, both of these were to be expected as they are linear combinations of each other. The correlations are visualized in the following heat map, the redder the square the closer the correlation to 1.00, the numbers in brackets are the p-values from the correlations.



Heat map of correlations for a subset of variables retained after removing duplicates, and those with insufficient observations at each factor level

Given the significant and substantial correlation that remains between some of the variables, it is clear that the multi-collinearity was not the only issue with the data preventing the models from functioning normally, as the retained variables did not cause any errors. It remains undetermined what the true cause, or causes, of the error were. Additionally, the removal of many variables was a consequence of the subsampling down to 1000 patients. With much larger samples, ensuring that sufficient numbers of individuals are present to prevent any cross validation fold not including at least one individual of the alternative factor level is likely to be much less restrictive.

Pre-processing

Numeric features were centered (mean value subtracted) and scaled (divided by the standard deviation). This is a standard statistical technique which means features can be interpreted in a cohesive manner as feature values can be interpreted in a shared way (as the number of standard deviations above the mean the individual is).

Classification Models

The following statistical models were used for the classification modelling:

- **Dummy** a null model which aims to predict admission without accounting for patient information. Used as a baseline comparison to demonstrate model efficacy.
- **Logistic regression (LogReg)** a standard classification model, similar to standard linear regression, which estimates the probability of an outcome (admission) accounting for selected patient covariates. The original SPARRA model was fit with a logistic model.
- **Linear discriminant analysis (LDA)** a classification approach which attempts to find a linear combination of the patient's features which best demarcates between those who become admitted and those who do not
- Naive Bayes
- **Neural Networks (NN)** Neural networks with one hidden layer and 3 nodes
- **Random forests (RF)** a classification method which builds a large number of decision trees (if this, then that) and outputs the outcome predicted by the majority of these decision trees
- **Linear Support Vector Machine (SVM.lin)** a classification model which determines the linear boundary that separates the two classes while maximizing the distance between itself and the examples
- **Gaussian Support Vector Machine (SVM.rbf)** a classification model which determines a nonlinear boundary that separates the two classes while maximizing the distance between itself and the examples, after projecting the examples into a non-linear space

All of the above models will output a probability of being Admitted, apart from the SVM which will give a binary admitted/Non-admitted output (though there are ways of calibrating the SVM decision function to give a probability – e.g. Platt's Scaling).

Internal Model Validation

We evaluated the models using 10-fold cross-validation meaning the dataset is divided into 10 non-overlapping folds, and each fold is used to evaluate model performance after training on the remaining data. The model goodness-of-fit statistics from each fold were then averaged to obtain the final score.

We used nested cross-validation with 3 folds to estimate model tuning hyper parameters for the modeling strategies that require this additional step for accurate estimation of their predictive accuracy (i.e. support vector machines and random forests).

We consider a number of different model metrics to evaluate the performance of our various models. It is important to consider a variety of metrics as, e.g., a perfect true positive rate (TPR) can be achieved by simply predicting all patients will be admitted in the next year. The following model assessment statistics were considered for the binary classification tasks: mean misclassification error (MMCE), true positive rate (TPR, also known as sensitivity), true negative rate (TNR), Positive Predictive Value (PPV), F1 score, area under the receiver operating characteristic curve (AUC), Brier, Logloss - True positive: High risk patients correctly identified as high risk - False positive: Low risk patients incorrectly identified as high risk - True negative: Low risk patients correctly identified as low risk - False negative: High risk patients incorrectly identified as low risk

Table 1: Deterministic and probabilistic measures of predictive performance

MMCE	**Mean misclassification error**
TPR	True positive rate/sensitivity/recall: the probability that a patient will be predicted as having an admission if they do, indeed, have an admission
TNR	True negative rate/specificity: the probability that a patient will be predicted as having no admission if they do not
PPV	Positive predictive value/precision: the probability that a patient will have an admission given that they are predicted to have one
F1	The harmonic mean of PPV and TPR
AUC	Area under the receiver operating characteristic curve, which is a measure of the trade-off between the true positive rate and the false positive rate for various probability thresholds
Brier	A measure of the difference between the predicted outcome and the true outcome
Logloss	An accuracy measure which penalizes misclassifications

Results: Supervised Classification Modelling of Hospital Admission

The following shorthand is used in the results tables:

- BL: Dummy/Baseline
- LDA: Linear Discriminant Analysis
- LR: Logistic Regression
- RSVM: Radial Basis Function Support Vector Machine
- RF: Random Forest
- NB: Naive Bayes
- NN: Neural Net
- *: denotes that that the features have been centered and standardized to unit variance before fitting the model
- ^: means that hyper parameter tuning has been implemented for the model.

Predicting Admission (Regardless of type - Elective or Emergency)

Table 2: A summary of the estimates of predictive accuracy from the predictive benchmark experiment for admission of any kind (elective + emergency) of LTC individuals in the coming year, based on subset 3.

Method	MMCE	TPR	TNR	PPV	F1	AUC	Brier	Logloss
BL	0.47 (0.05)	1 (0.14)	0 (0.00)	0.53 (0.05)	0.69	0.50	0.25 (0.00)	0.69 (0.01)
LDA*	0.24 (0.05)	0.63 (0.11)	0.90 (0.15)	0.88 (0.18)	0.73	0.82	0.17 (0.03)	0.56 (0.10)
LR*	0.28 (0.05)	0.67 (0.12)	0.78 (0.14)	0.79 (0.15)	0.72	0.74	0.24 (0.04)	5.01 (1.05)
RSVM^	0.33 (0.05)	0.89 (0.13)	0.42 (0.10)	0.63 (0.08)	0.74	0.77	0.22 (0.02)	0.63 (0.04)
RF^	0.22 (0.04)	0.66 (0.12)	0.91 (0.15)	0.89 (0.17)	0.76	0.85	0.15 (0.02)	0.44 (0.05)
NB	0.42 (0.05)	0.24 (0.07)	0.97 (0.16)	0.90 (0.35)	0.38	0.78	0.42 (0.05)	14.10 (1.78)
NN*	0.26 (0.05)	0.69 (0.12)	0.80 (0.14)	0.80 (0.15)	0.73	0.78	0.21 (0.04)	1.41 (0.38)

The results in Table 2 suggest that all of the classical and machine learning based methods are better than the baseline (featureless) classifier when compared using the deterministic and probabilistic measures of performance. Whilst a retrospective set of results for the SPARRA model has not been included due to time constraints, based on the original values quoted for the SPARRA model (i.e. TPR - 10%, PPV - 55%), it would appear that using more variables improves predictive performance (see LogReg results in table 12). And the addition of machine learning methods to the increased feature set further improves performance. However, caution must be taken with the direct comparison of these results to the SPARRA model, as the former is based on a balanced sample (50% admission, 50% non-admission) whereas the latter was built using an unbalanced sample.

Upon further discussion the domain experts in the group noted that prediction of elective admission was not necessarily useful as these admission events are by definition planned, and usually relate to conditions that have long been known to the individuals' general practioner. As such it was argued that predicting which individuals will have a planned event provides little additional information to general practioners other than highlighting those whom may require intervention but are largely stable. What is arguably more useful is the identification of individuals who are likely to experience a precipitous decline in health/wellbeing requiring emergency admission, and thus, the following modelling tasks focus only on this sub-group.

Predicting Non-Psychiatric Emergency Admissions only (Balanced Samples)

The results in tables 3-6 suggest that the SPARRA model's performance relative to the (featureless) baseline/dummy differs based on the risk group in question, and the measure of predictive performance used. Across all four groups, when using the probabilistic measures of performance (Brier and Logloss), the SPARRA is consistently worse than the baseline. However, the original SPARRA model was developed using the entire population, which we know to constitute a highly unbalanced (admission vs. non-admission) sample. As such, the significantly lower retrospectively calculated probabilistic results can be explained away by the SPARRA model not being calibrated for the balanced sample which we have used. When the SPARRA model is compared to the baseline using the deterministic MMCE, for which calibration should not be as much of a problem, it appears that it is significantly better in the U16 population (Table 5), significantly worse in the FE population (Table 3), and statistically similar in the LTC and YED populations (Tables 4 & 5).

It is worth noting that any difference between the SPARRA model and any model other than the Logistic Regression can be explained not only by model type, but also the difference in the number of features used. As such, a comparison between the logistic regression models we trained and the SPARRA model would provide some insight into the impact of using as many features as is possible instead of domain expert selected variables for each of the risk groups. Given that comparing the models based on the probabilistic measures of performance could reap potentially misleading results, the following statements relate only to the MMCE results. The retention of additional features appears to significantly improve predictions in all of the risk groups (Tables 3,4 & 6), except the U16 population (Table 5), where the difference in performance is not significant. Based on these results, all other methods in the LTC, FE and YED population if better than the logistic regression can be assumed to be so due only to the altered modelling strategy.

Across all four risk groups, it is not possible to discriminate between the classical and machine learning models we trained based on their MMCE as the differences are not statistically significant. Similarly, most of the comparisons based on the probabilistic measures are also insignificant. And for those that reach significance, the absolute difference is incredibly small, and thus it could be very easily argued that the ability to interpret white box methods (e.g. the logistic regression) is more useful than this very small improvement. The model that consistently appears to be best, although as we said previously these differences are not always statistically significant, is the random forest.

Table 3: A summary of the estimates of predictive accuracy from the predictive benchmark experiment for admission of FE individuals in the coming year, based on subset 5.

Method	MMCE	TPR	TNR	PPV	F1	AUC	Brier	Logloss
BL	0.53 (0.05)	0.50 (0.08)	0.50 (0.08)	0.43	0.32	0.50	0.25 (0.00)	0.69 (0.00)
SPARRA	0.70 (0.01)	.	.	.	.	.	0.46 (0.00)	1.29 (0.00)
LDA*	0.42 (0.05)	0.52 (0.10)	0.64 (0.11)	0.60 (0.12)	0.55	0.60	0.26 (0.02)	0.79 (0.09)
LR*	0.41 (0.05)	0.55 (0.11)	0.63 (0.11)	0.60 (0.12)	0.57	0.60	0.27 (0.03)	0.91 (0.16)
RSVM^	0.46 (0.05)	0.81 (0.13)	0.29 (0.07)	0.51 (0.09)	0.60	0.58	0.25 (0.01)	0.69 (0.02)
RF^	0.38 (0.05)	0.61 (0.11)	0.63 (0.11)	0.62 (0.11)	0.61	0.66	0.23 (0.02)	0.66 (0.03)
NB	0.42 (0.05)	0.35 (0.08)	0.80 (0.13)	0.64 (0.18)	0.45	0.60	0.42 (0.05)	8.78 (1.21)
NN*	0.43 (0.05)	0.51 (0.10)	0.63 (0.11)	0.58 (0.12)	0.54	0.59	0.32 (0.04)	1.05 (0.13)

Table 4: A summary of the estimates of predictive accuracy from the predictive benchmark experiment for admission of LTC individuals in the coming year, based on subset 4.

Method	MMCE	TPR	TNR	PPV	F1	AUC	Brier	Logloss
BL	0.52 (0.05)	0.60 (0.09)	0.40 (0.06)	0.49	0.39	0.50	0.25 (0.00)	0.69 (0.00)
SPARRA	0.47 (0.02)	.	.	.	.	.	0.37 (0.01)	1.17 (0.00)
LDA*	0.38 (0.05)	0.53 (0.10)	0.72 (0.12)	0.66 (0.14)	0.58	0.67	0.24 (0.02)	0.71 (0.08)
LG*	0.37 (0.05)	0.54 (0.10)	0.71 (0.12)	0.65 (0.13)	0.59	0.67	0.25 (0.03)	0.95 (0.22)
RSVM^	0.41 (0.05)	0.82 (0.13)	0.37 (0.09)	0.56 (0.08)	0.67	0.63	0.24 (0.01)	0.68 (0.02)
RF^	0.32 (0.05)	0.66 (0.12)	0.70 (0.12)	0.69 (0.12)	0.67	0.73	0.21 (0.02)	0.62 (0.05)
NB	0.39 (0.05)	0.32 (0.08)	0.89 (0.13)	0.75 (0.23)	0.45	0.67	0.39 (0.05)	12.62 (1.63)
NN*	0.40 (0.05)	0.54 (0.10)	0.67 (0.12)	0.63 (0.13)	0.57	0.61	0.29 (0.03)	0.96 (0.12)

Table 5: A summary of the estimates of predictive accuracy from the predictive benchmark experiment for admission of U16 individuals in the coming year, based on subset 6.

Method	MMCE	TPR	TNR	PPV	F1	AUC	Brier	Logloss
BL	0.53 (0.05)	0.46 (0.07)	0.56 (0.09)	0.44	0.31	0.50	0.25 (0.00)	0.69 (0.00)
SPARRA	0.39 (0.01)	.	.	.	.	.	0.28 (0.00)	0.86 (0.00)
LDA*	0.38 (0.05)	0.57 (0.11)	0.67 (0.12)	0.63 (0.12)	0.60	0.66	0.24 (0.02)	0.67 (0.05)
LG*	0.38 (0.05)	0.56 (0.11)	0.68 (0.12)	0.64 (0.13)	0.59	0.66	0.24 (0.02)	0.70 (0.06)
RSVM^	0.42 (0.05)	0.75 (0.12)	0.42 (0.09)	0.56 (0.08)	0.64	0.62	0.24 (0.01)	0.67 (0.02)
RF^	0.38 (0.05)	0.60 (0.11)	0.66 (0.12)	(0.64 (0.12)	0.61	0.69	0.23 (0.02)	0.65 (0.05)
NB	0.42 (0.05)	0.27 (0.07)	0.89 (0.13)	0.71 (0.24)	0.39	0.66	0.41 (0.05)	10.71 (1.43)
NN*	0.41 (0.05)	0.53 (0.10)	0.66 (0.12)	0.61 (0.13)	0.56	0.62	0.28 (0.03)	0.87 (0.10)

Table 6: A summary of the estimates of predictive accuracy from the predictive benchmark experiment for admission of YED individuals in the coming year, based on subset 7.

Method	MMCE	TPR	TNR	PPV	F1	AUC	Brier	Logloss
BL	0.55 (0.05)	0.4 (0.07)	0.60 (0.09)	0.09	0.24	0.50	0.25 (0.00)	0.70 (0.00)
SPARRA	0.53 (0.01)	.	.	.	.	.	0.36 (0.00)	1.08 (0.00)
LDA*	0.36 (0.05)	0.54 (0.11)	0.72 (0.12)	0.12	0.61	0.71	0.23 (0.03)	0.70 (0.09)
LogReg*	0.35 (0.05)	0.59 (0.11)	0.71 (0.12)	0.12	0.62	0.70	0.24 (0.03)	1.00 (0.26)
NB	0.41 (0.05)	0.28 (0.08)	0.90 (0.14)	0.14	0.40	0.70	0.41 (0.05)	13.63 (1.67)
RSVM^	0.35 (0.05)	0.82 (0.13)	0.48 (0.10)	0.10	0.70	0.69	0.23 (0.01)	0.65 (0.03)
RF^	0.32 (0.05)	0.66 (0.12)	0.71 (0.12)	0.12	0.67	0.76	0.20 (0.02)	0.60 (0.05)
NN*	0.37 (0.05)	0.57 (0.11)	0.71 (0.12)	0.12	0.61	0.64	0.29 (0.04)	1.09 (0.15)

It is possible that our subsampling down to 1000 individuals has resulted in many of the more nuanced relationships being diluted to the point of being effectively absent. Moreover, the balancing of the samples could also explain the relatively strong performance of the classical modelling methods, as they usually perform worse with heavily biased samples. As such, we are hesitant to suggest that machine learning based methods are not useful in this setting. The only way to be able to make this claim would be to run these benchmark experiments on the whole dataset, instead of the balanced subsamples we have used.

Finally, it should be noted that no statements about the clinical efficacy of the SPARRA model can be made from these results as there was no clinical baseline with which to compare performance.

Table 7: A summary of the estimates of predictive accuracy from the predictive benchmark experiment for emergency admissions in the coming year, where the risk groups are variables instead of subsetting features (Dataset 8).

Method	MMCE	TPR	TNR	PPV	F1	AUC	Brier	Logloss
BL	0.51 (0.02)	0.45 (0.03)	0.55 (0.04)	0.34	.	0.31	0.25 (0.00)	0.69 (0.00)
SPARRA	.	.	.	.	.	.	.	.
LDA*	0.37 (0.02)	0.54 (0.05)	0.72 (0.06)	0.66 (0.07)	0.59	0.67	0.23 (0.01)	0.65 (0.02)
LogReg*	0.36 (0.02)	0.54 (0.05)	0.73 (0.06)	0.67 (0.07)	0.60	0.68	0.23 (0.01)	0.65 (0.02)
NB	0.41 (0.02)	0.33 (0.04)	0.85 (0.07)	0.69 (0.10)	0.45	0.64	0.41 (0.02)	12.34 (0.79)
RSVM^	0.40 (0.02)	0.82 (0.06)	0.37 (0.04)	0.56 (0.03)	0.67	0.62	0.24 (0.00)	0.67 (0.01)
RF^	0.35 (0.02)	0.62 (0.06)	0.67 (0.06)	0.65 (0.06)	0.64	0.72	0.22 (0.02)	0.62 (0.02)
NN*	0.39 (0.02)	0.50 (0.05)	0.71 (0.06)	0.64 (0.07)	0.56	0.63	0.26 (0.01)	0.78 (0.04)

The results presented in Table 7 suggest that when the four risk groups are combined and used as variables instead of subsetting features, the ability of models to predict the outcome of emergency admission is still similar to that of the models based on each of the risk groups individually. It was thought that by combining the groups additional subtle relationships that drive emergency admission may be identified that would improve the overall predictive

capability, however this was not the case. It should be noted that this hypothesis cannot be completely rejected based on the results, as new relationships could still be identified when the sample sizes are scaled up further, currently the largest sample considered is 4000 out of a potential 1.8 million.

Predicting Non-Psychiatric Emergency Admissions only (Pure [representative] Samples)

Table 8: A summary of the estimates of predictive accuracy from the predictive benchmark experiment for emergency admissions in the coming year, for a pure/representative sample of the U16 cohort (Dataset 9).

Method	MMCE	TPR	TNR	PPV	F1	AUC	Brier	Logloss
BL	0.09 (0.03)	0 (0.00)	1.00 (0.04)	0.20	0.00	0.50	0.08 (0.02)	0.31 (0.07)
SPARRA	.	.	.	.	.	.	.	.
LDA*	0.11 (0.03)	0.02 (0.02)	0.98 (0.05)	0.17 (0.20)	0.04	0.56	0.09 (0.02)	0.36 (0.09)
LogReg*	0.10 (0.03)	0.02 (0.02)	0.99 (0.05)	0.25	0.04	0.56	0.09 (0.02)	0.41 (0.13)
NB	0.32 (0.04)	0.36 (0.20)	0.71 (0.05)	0.13 (0.06)	0.18	0.58	0.30 (0.04)	3.50 (0.74)
RSVM^	0.09 (0.03)	0 (0.00)	1.00 (0.04)	0.20	0.00	0.50	0.08 (0.02)	0.31 (0.07)
RF^	0.09 (0.03)	0 (0.00)	1.00 (0.04)	0.20	0.00	0.50	0.08 (0.02)	0.31 (0.07)
NN*	0.12 (0.03)	0.06 (0.04)	0.96 (0.05)	0.09	0.07	0.54	0.10 (0.02)	1.02 (0.41)

When a representative sample of 1000 individuals from the U16 group was utilized instead of a balanced sample (50/50 admission or not), the results (Table 8) suggest that there was insufficient information for the models that previously performed best (RSVM and RF) to learn anything useful. As such, these models learnt to act similarly to the Baseline/Dummy model. On closer examination of the data, we found that there were only 89 positive outcomes out of the 1000 individuals randomly sampled. Not only was this insufficient for the models to learn from, but from the lack of associated error estimates for certain measures it can be inferred that when this data was run through the 10-fold cross validation setup some of the prediction/test folds did not have a single positive outcome. To accurately assess the performance of classical statistical methods and machine learning methods in the representative sample setting the use of much larger sample sizes is going to be necessary to ensure sufficient data for model training is available.

Multilabel Prediction of ACSC Admissions

Part of the challenges posed to the Data Study Group was to determine whether it was possible to predict if an individual's admission would be due to an ACSC condition or not. The reason being that many consider ACSC conditions to be preventable, and therefore they represent a sub-group of admissions that should be preventable.

This task was approached in a similar manner to those above i.e. a supervised classification approach was adopted. However instead of treating the task as binary classification, a multilabel approach was utilized. In the multilabel setting the model first learns an initial binary prediction, and then subsequently learns the dependent binary predictions; in this setting that means that first the model learns to predict whether an individual will be admitted or not, and if they are an admission then it will learn to predict if they are an ACSC prediction or not. The measures of predictive performance calculated were: F1, PPV, TPR, and accuracy (= 1 - MMCE). Additionally, a multilabel specific measure was also included called Hamloss (the symmetric difference between predicted and true labels divided by the total number of labels).

In the balanced sub-sample of 50,000 (subset 1) individuals there were less than 2000 ACSC flags, and of those, only 1116 were for the first admission an individual had in the year 2016 (the prediction year), i.e. there were several hundred ACSC flags for second admission and onwards that we had to ignore for the purpose of this analysis as their inclusion would prevent easy interpretation of the results. Given the complexity of the model being trained, and the sample size required to run a sensible experiment only a logistic regression model was utilized as that is the model type used by SPARRA.

Table 9: A summary of the estimates of predictive accuracy from the predictive benchmark experiment for multilabel classification admission i.e. predicting if the individual was an emergency admission, and if so, whether they were had an ACSC flag.

Method	ACC	F1	Hamloss	PPV	TPR
LogReg*	0.690	0.692	0.162	0.739	0.594

Whilst the logistic regression results appear to be reasonably good, without any baselines to compare it too, we cannot draw any conclusions about the efficacy of the model. On closer inspection of the results (at the individual level), it was found that the model only predicted ACSC positive for two people. And when the ground truth for those two individuals was compared to the out-of-sample predictions, we found that the model was not only wrong about the ACSC flag, but neither individual had actually been an emergency admission in the prediction year. The results suggest that the model is attempting to learn, but unsuccessfully. We believe this is most likely due to the relatively small number of ACSC flags. Given that the sample is balanced and thus the number of admissions and therefore ACSC flags is proportionally much larger than one would expect for a sample of 50,000, it is likely that a much larger sample is necessary to accurately assess the multilabel approach in this setting.

Attendance versus Admission

An additional analysis of interest was to investigate patient profiles which are associated with visiting the emergency department more often (but not necessarily being admitted). To this end, we focused on the number of ED attendances in December 2015 (of the Pure/unbalanced subsample (subset 1)) and compared it against the complete profiles of patients in December 2014. We fitted a Generalized Linear Model for the number of ED attendances in 2015, using a Poisson distribution with a log-link, so that the logarithm of the mean number of ED attendances is a linear combination of all the December 2014 covariates. Numerical covariates were normalized to mean 0 and variance 1. Coefficients were penalized using Elastic Net.

The variables which were returned as associated with ED attendances were:

Variable	**Coefficient**
(intercept)	-1.59
Psychiatric admission	0.310
Ed attendance only	0.223
Prescribed item only	-0.219
No ed attendances	0.172
Age start risk year	-0.141
Simd scotland quintile	-0.088
No different diagnosis groups	0.079
No of ltcs	0.038
No emergency admissions	0.030
Analgesics (0407)	0.029
Hypnotics and anxiolytics (0401)	0.028
Anti-infective skin preparations (1310)	0.025
Vitamins (0906)	0.022
Evidence of parkinsons disease	0.015
Antiplatelet drugs (0209)	0.0063
Specialty	0.00073
Laxatives (0106)	0.00014

The results of the penalized GLM have identified a number of variables, some surprising and others not, which are associated with A&E attendance. It is outside of the scope of this report to conjecture about the reasons for these results. The purpose of this analysis was to highlight some of the reasons that drive A&E attendance. The code to run this experiment on the whole dataset is available in the safe haven. The results of that experiment should hopefully provide a robust set of high risk features for general practitioners to incorporate into their own personal assumptions.

Conclusions

The primary objective of the data study group was to determine whether it was possible to produce models capable of superior predictive performance than the SPARRA model. The preliminary results demonstrate that it is possible to produce both classical and machine learning based models which are capable of identifying individuals in the LTC population at low risk of admission (including both elective and emergency) with very high levels of accuracy (91% - Random Forest, 90% - Linear Discriminant Analysis), whilst also identifying a large proportion of those who are going to be admitted (66% - Random Forest, 63% - Linear Discriminant Analysis). When we attempt to predict just those at risk of emergency admission (Table 3-6), whilst it is much harder (illustrated by the overall worse performance of all models), the results suggest that in all of the risk groups except the U16s it is possible to generate models with superior predictive capabilities than the retrospectively calculated results for the SPARRA model.

In summary, we have found that leveraging as many features as possible allows us to improve upon the predictive performance of the original SPARRA model, but the ability of machine learning methods to further improve performance above that of the classical methods remains an open question. Moreover, whilst the results from the initial predictive modelling experiments are very promising, it is important to remember that they very are derived from balanced samples of 1000 individuals, meaning that their generalizability may be limited. As such, more work is required to ensure that the models are robust in the real world setting.

Other Modelling Strategies

1. Increasing the Granularity of the Outcome Predictions

We would have also liked to employ modelling approaches that can predict not just whether a subject is admitted, but also the time until the subject is admitted. Note that the prediction of Time to Admission was not considered in the original SPARRA model. Code to carry out the following modelling approaches was written, and is available in the safe haven, but was not run on the data samples:

- **Survival Analysis**: We would ideally use this to predict the time to admission to hospital and/or time to death given the patients characteristics. In this case, the outcome is a continuous variable, and the task is essentially a time-to-event regression analysis. The current implementation uses a semi-parametric survival model (Cox) to predict the median survival time. Other parametric survival models can also be implemented e.g. Weibull, Gompertz, lognormal, and generalized gamma to give median and mean/expected survival time.

- **Ordinal Regression Models** Here, we treat the time until a patient is admitted as a discrete ordinal variable by binning the time into discrete categories (e.g., 0-6 months, 6-12 months, and >12 months) and then the model is trained to learn these categories. It differs from multi-class classification approaches in that it accounts for the ordering of the categories.

Note that after training, both of these models would be able to perform the primary objective of predicting whether a subject is admitted within the following year. Moreover, it may have also been possible to use these models to investigate other interesting questions e.g. determining the effect of season on admissions.

2. Other Modelling Tasks

In addition to the objectives addressed in this report, which all exist in a binary classification framework, we wanted to explore predicting Not Admitted vs ACSC-Admissions Vs Other Admissions which is a multi-class classification problem. Some of the models used in this report can naturally be extended to the multi-class case.

Other models that could have been considered for performing the binary classification are:

- **Regularized Logistic Regression** (ridge, LASSO, Elastic-Net regularized)

- **Gaussian Process classification** a non-parametric method that takes into account a smoothness/periodic/etc. prior distribution. Unlike the other approaches will also incorporate uncertainty into the probability i.e. probabilities will be softened (towards 0.5) for patients whose input features are far away from those of the other subjects.

- **Generalized Additive Models logistic regression (GAM)** a logistic regression with additional spline terms to take into account non-linear effect.

Outstanding Scientific Questions

1. Feature Importance

One of the main aims set out in the preliminary discussions was to determine which features were most useful, and to identify those that were unnecessary. One way to approach this problem is to wrap the statistical models with feature selection algorithms to identify the variables of greatest predictive importance. However, we also felt it was important to seek the advice of independent domain experts to establish a clinically relevant baselines. As such, we sought the input of a general practitioner (GP) and a consultant in emergency medicine on which variables are most likely to be associated with an increased risk of future emergency admission. The results of those discussions are included below.

Eventually we agreed that the most suitable way to approach this question was to produce a set of models to compare: 1. The unrestricted use of variables; 2. The groups selected by the two medical doctors we consulted during the data study group; 3. The SPARRA public health physician selected set; 4. A series of feature selection algorithm results (e.g. Sequential floating forward/backward selection, genetic algorithms, etc.);

Unfortunately, due to time and computational resource-related limitations we were unable to run this set of experiments. Although some of the experiments conducted do allow us to draw conclusions about the likely results (i.e. the domain-expert selected variables do not produce the optimal results always), we believe it would still be useful to run this experiment to determine the optimal selection of features for the final model; a question that we were not able to address during the data study group.

The GP-identified high risk features included:

- Drugs prescribed for certain high-risk conditions: Anti-arrhythmic drugs, anticoagulants, antiplatelet, stable angina, oxygen, mucolytic, hypnotics and anxiolytics, drugs used in psychoses, obesity, antiepileptic, drugs used in parkinsonism, substance dependence, dementia, diabetes, corticosteroids, bone metabolism, cytotoxic drugs, drugs affecting the immune response, intravenous nutrition, oral nutrition, rheumatic diseases & gout, pressure plates

- Evidence of certain long-term conditions such as dementia, diabetes, heart disease - Indicators of general health

- Number of drug items dispensed in the last year, number of emergency admissions in last 3 years, no. of different diagnosis groups, no. of alcohol/substance/falls/self-harm related admissions, all LTCs (except arthritis), and looking at total outpatient attendances instead of each category, and finally, the no. of psychiatric attendances in last 3 years

Moderate risk indicators:

- Drugs prescribed for certain moderate-risk conditions.
    - Bronchodilators, corticosteroids (respiratory), cromoglycate rel. leukotriene Antagonists, respiratory stimulants and pulmonary surfectants, Mucolytic, Antidepressant drugs (low-medium), drugs used in Nausea and Vertigo, Analgesics, Antibacterial drugs, Antiviral drugs, Thyroid and antithyroid drugs, Hypothalamic & Pituitary Hormones & Antioest, Anaemias + Other Blood Disorders, Fluids and Electrolytes, Drugs used in Neuromuscular Disorders, General Anesthesia, Local Anesthesia, Ileostomy Bags, Ileostomy Sets, Stoma Caps/Dressings, Urostomy Bags, Urostomy Sets
- Indicators of general health
    - Number elective admissions in last 3 years (low-medium), number daycase admissions in last 3 years (low-medium), diagnosis of asthma, and diagnosis of epilepsy.

Clustering

At the moment the patients are split into 4 cohorts, Frail Elderly, LTC, Younger ED, and Under 16. Performing clustering on the entire dataset might give us some insight into other possible ways of splitting the patients into risk groups. One assumption is that factors such as age or the deprivation index are more predictive than any of the clinical features. Consequently, we may consider modelling the predictions within the newly discovered groups. Moreover, certain groups are likely to overlap, which would justify using the data for the entire population instead of using the cohorts. Clustering may also expose outliers in the dataset.

Possible clustering methods include k-means clustering, mixture models, Gaussian Process Latent Variable Model, and other unsupervised learning approaches.

Data Subsets

A number of balanced subsets were drawn from the data according to admission type (psychiatry or not) and cohort.

Dataset	N	Admission type	Cohort
1*	50,000	SMR1	all
2	50,000	mixed	all
3	1,000	mixed	LTC
4	1,000	SMR1	LTC
5	1,000	SMR1	FE
6	1,000	SMR1	U16
7	1,000	SMR1	YED
8	4,000	SMR1	all
9*	1,000	SMR1	U16

Sample 1 and 9 are not balanced. Instead it was a purely random subsample of the data, with no subsetting. SMR4 = psychiatric admissions SMR1 = non-psychiatric admissions mixed = SMR4+SMR1

References

[1] HTTP://WWW.GOV.SCOT/RESOURCE/DOC/924/0012113.PDF

Appendix

Admission related variables in the original SPARRA compared to those utilized to create the new ATI models, for each risk cohort. S is the shorthand used for the original SPARRA model, whereas T refers to the results of the data study group

Feature	S:FE	S:LTC	S:YED	S:U16	T:ALL	T:FE	T:LTC	T:YED	T:U16
Number of emergency admissions in last 3 years	X	X	X	X	X	X	X	X	X
Number of elective admission in last 3 years				X	X	X	X	X	X
Number of daycase admissions in last 3 years					X	X	X	X	X
Total elective and daycase admission in last 3 years	X	X	X		X	X	X	X	X
Nubmer of emergency bed days in last 3 years	X	X	X		X	X	X	X	X
Number of elective bed days in last 3 years		X			X	X	X	X	X
Total number of bed days in last 3 years					X	X	X	X	X
Number of alcohol-related admissions in last 3 years	X				X		X	X	
Number of emergency admissions with drug or alcohol diagnoses		X	X						
Number of different diagnosis groups in last 3 years					X	X	X	X	X
Number of substance-related admissions in last 3 years					X			X	
Number of falls-related admissions in last 3 years					X	X	X	X	
Number of self-harm related admissions in last 3 years					X			X	

Prescriptions, BNF Chapter Prescriptions, and Drug Prescription related variables in the original SPARRA compared those utilized to create the new ATI models, for each risk cohort. S is the shorthand used for the original SPARRA model, whereas T refers to the results of the data study group

Feature	S:FE	S:LTC	S:YED	S:U16	T:ALL	T:FE	T:LTC	T:YED	T:U16
Number of BNF sections for which patient has prescriptions in last 12 months	X	X	X	X	X	X	X	X	X
Total Respiratory System Chapter prescriptions	X								
Total Central Nervous System Chapter prescriptions	X		X						
Total Infections Chapter prescriptions	X	X							
Total Endocrine System Chapter prescriptions	X								
Total Incontinence Devices Chapter prescriptions	X								
Total Gastro-Intestinal System Chapter prescriptions				X					
(1) Corticosteroids Drugs prescriptions		X							
(2) Corticosteroids Drugs prescriptions			X						
Stoma devices Drugs prescriptions		X	X						
Dyspep & Gastro-Oesophageal Reflux Disease					X	X	X	X	X
Antispasmodic. & Other Drgs Alt.Gut Motility			X		X	X	X	X	X
Antisecretory Drugs + Mucosal Protectants			X		X	X	X	X	X
Acute Diarrhoea					X	X	X	X	
Chronic Bowel Disorders					X		X		
Laxatives					X	X	X	X	X
Local Prepn for Anal & Rectal Disorders					X	X	X	X	
Drugs Affecting Intestinal Secretions			X		X	X	X	X	X
Positive Inotropic Drugs					X	X			
Diuretics					X	X	X	X	

Beta-Adrenoceptor Blocking Drugs					X	X	X	X	
Hypertension and Heart Failure				X	X	X	X	X	
Nit,Calc Block & Other Antianginal Drugs					X	X	X	X	
Anticoagulants And Protamine			X		X	X	X	X	
Antiplatelet Drugs					X	X	X	X	
Antifibrinolytic Drugs & Haemostatics			X		X				
Lipid-Regulating Drugs					X		X	X	
Local Sclerosants						X			
Bronchodilators				X	X	X	X	X	X
Corticosteroids (Respiratory)					X	X	X	X	X
Cromoglycate,Rel,Leukotriene Antagonists				X	X	X	X	X	X
Antihist, Hyposensit & Allergic Emergen					X	X	X	X	X
Mucolytic			X		X	X	X		
Cough Preparations					X	X	X	X	X
Systemic Nasal Decongestants					X			X	
Hypnotics And Anxiolytics					X	X	X	X	
Drugs Used In Psychoses & Rel.Disorders					X	X	X	X	
Antidepressant Drugs					X	X	X	X	
CNS Stimulants and drugs used for ADHD					X				
Obesity					X				
Drugs Used In Nausea And Vertigo					X	X	X	X	
Analgesics					X	X	X	X	X
Antiepileptic Drugs					X	X	X	X	
Drugs Used In Parkinsonism/Related Disorders					X	X			
Drugs Used In Substance Dependence		X		X	X		X	X	

Dementia		X			X	X			
Antibacterial Drugs				X	X	X	X	X	X
Antifungal Drugs					X	X	X	X	X
Antiviral Drugs					X		X	X	
Antiprotozoal Drugs					X	X	X		
Anthelmintics					X				X
Drugs Used In Diabetes			X		X	X	X	X	
Thyroid And Antithyroid Drugs					X	X	X	X	
Corticosteroids (Endocrine)				X	X	X	X	X	X
Sex Hormones					X	X	X	X	
Drugs Affecting Bone Metabolism					X	X	X		
Treatment Of Vaginal & Vulval Conditions					X	X	X	X	
Contraceptives					X		X	X	
Drugs For Genito-Urinary Disorders					X	X	X	X	
Drugs Affecting The Immune Response					X				
Sex Hormones & Antag In Malig Disease					X	X	X		
Anaemias + Other Blood Disorders					X	X	X	X	X
Fluids And Electrolytes		X	X		X	X		X	X
Oral Nutrition		X			X	X		X	X
Minerals					X				
Vitamins					X	X	X	X	X
Minerals&Vitamins		X	X						
Drugs Used In Rheumatic Diseases & Gout				X	X	X	X	X	X
Drugs Used In Neuromuscular Disorders				X	X		X	X	
Soft-Tissue Disorders & Topical Pain Rel					X	X	X	X	
Anti-Infective Eye Preparations					X	X	X	X	X

Corticosteroids & Other Anti-Inflamm.Preps.					X	X	X	X	X
Mydriatics & Cycloplegics				X					
Treatment Of Glaucoma					X	X			
Miscellaneous Ophthalmic Preparations					X	X	X	X	
Drugs Acting On The Ear					X	X	X	X	X
Drugs Acting On The Nose					X	X	X	X	X
Drugs Acting On The Oropharynx					X	X	X	X	X
Emollient & Barrier Preparations					X	X	X	X	X
Top Local Anaesthetics & Antipruritics					X				X
Topical Corticosteroids					X	X	X	X	X
Preparations For Eczema And Psoriasis					X		X	X	
Acne and Rosacea					X		X	X	X
Preparations For Warts And Calluses					X				X
Sunscreens And Camouflagers					X				
Shampoo&Other Preps For Scalp&Hair Cond					X	X	X	X	X
Anti-Infective Skin Preparations					X	X	X	X	X
Skin Cleansers,Antiseptics & Desloughing						X			
Vaccines And Antisera					X				
Local Anaesthesia					X				
Arm Sling/Bandages		X			X				
Wound Management & Other Dressings					X	X	X	X	
Catheters		X			X	X			
Elastic Hosiery					X	X			
Other BNF					X	X	X		
Number of drug items dispensed in the last year					X	X	X	X	X

Psychiatric Admission related variables in the original SPARRA compared those utilized to create the new ATI models, for each risk cohort. S is the shorthand used for the original SPARRA model, whereas T refers to the results of the data study group

Feature	S:FE	S:LTC	S:YED	S:U16	T:ALL	T:FE	T:LTC	T:YED	T:U16
Number of admissions to psychiatric hospital		X	X		X			X	

Emergency Department related variables in the original SPARRA compared those utilized to create the new ATI models, for each risk cohort. S is the shorthand used for the original SPARRA model, whereas T refers to the results of the data study group

Feature	S:FE	S:LTC	S:YED	S:U16	T:ALL	T:FE	T:LTC	T:YED	T:U16
Number of ED attendances	X	X	X	X	X	X	X	X	X

Outpatients related variables in the original SPARRA compared those utilized to create the new ATI models, for each risk cohort. S is the shorthand used for the original SPARRA model, whereas T refers to the results of the data study group

Feature	S:FE	S:LTC	S:YED	S:U16	T:ALL	T:FE	T:LTC	T:YED	T:U16
Number of new outpatient appointments/outpatient attendances in last year	X	X	X						
Number of new outpatient appointments (mental health specialty)			X						
Number of Outpatient Attendances in Last Year				X	X	X	X	X	X
General Medicine (A1)					X	X			
Cardiology (A2)					X		X	X	
Dermatology (A7)					X	X	X	X	X
Endocrinology & Diabetes (A8)					X				
Gastroenterology (A9					X	X	X	X	
Geriatric Medicine (AB)					X	X			
Paediatrics (AF)					X				X
Paediatric Surgery (CA)					X				X
Neurology (AH)					X		X	X	
Respiratory Medicine (AQ)					X				
Rheumatology (AR)					X				
General Surgery (C1)					X	X	X	X	
General Surgery (excl Vascular, Maxillofacial) (C11)					X	X	X	X	
Vascular Surgery (C12)					X				
Oral and Maxillofacial Surgery (C13)					X			X	
Pain Management (C31)					X				
Ear, Nose & Throat (ENT) (C5)					X	X	X	X	X
Ophthalmology (C7)					X	X	X	X	X
Trauma and Orthopaedic Surgery (C8)					X	X	X	X	X
Plastic Surgery (C9)					X				
Urology (CB)					X	X	X	X	
Gynaecology (F2)					X		X	X	
General Psychiatry (Mental Illness) (G1)					X			X	
Psychiatry of Old Age (G4)					X	X			

Demographic variables in the original SPARRA compared those utilized to create the new ATI models, for each risk cohort. S is the shorthand used for the original SPARRA model, whereas T refers to the results of the data study group

Feature	S:FE	S:LTC	S:YED	S:U16	T:ALL	T:FE	T:LTC	T:YED	T:U16
Age	X			X	X	X	X	X	X
Gender				X	X	X	X	X	X
SIMD (quantile)	X	X			X	X	X	X	X
Health Board Residence Number					X	X	X	X	X
Care Home Residency Flag					X	X			

Variables indicating risk factor group in the original SPARRA compared those utilized to create the new ATI models, for each risk cohort. S is the shorthand used for the original SPARRA model, whereas T refers to the results of the data study group

Feature	S:FE	S:LTC	S:YED	S:U16	T:ALL	T:FE	T:LTC	T:YED	T:U16
Long Term Condition Cohort member indication					X				
Frail Elderly cohort member indication					X				
Younger ED cohort member indication					X		X		
Under 16 cohort member indication					X				

Variables indicating certain conditions in the original SPARRA compared those utilized to create the new ATI models, for each risk cohort. S is the shorthand used for the original SPARRA model, whereas T refers to the results of the data study group

Feature	S:FE	S:LTC	S:YED	S:U16	T:ALL	T:FE	T:LTC	T:YED	T:U16
Evidence of Parkinson's Disease based upon their long term condition flag or an item dispensed from BNF Section 4.9 during the pre-prediction period.	X	X			X	X			
Evidence of Multiple Sclerosis based upon their long term condition flag or an item dispensed from BNF Section 10.2 during the pre-prediction period.		X			X		X	X	
Evidence of epilepsy based upon their long term condition flag or an item dispensed from BNF Section 4.8 during the pre-prediction period		X		X	X	X	X	X	X
Evidence of Alzheimer's/dementia based upon their long term condition flag or an item dispensed from BNF Section 4.11 during the pre-prediction period			X		X	X			
Evidence of Congenital problem based upon their condition flag				X	X	X	X	X	X
Evidence of Blood Disorder based upon their condition flag				X	X	X	X	X	
Evidence of Endocrine/Metabolic conditions based upon their condition flag				X	X	X	X	X	X
Evidence of Other digestive conditions based upon their condition flag				X	X	X	X	X	X

Diagnosis indicators in the original SPARRA compared those utilized to create the new ATI models, for each risk cohort. S is the shorthand used for the original SPARRA model, whereas T refers to the results of the data study group

Feature	S:FE	S:LTC	S:YED	S:U16	T:ALL	T:FE	T:LTC	T:YED	T:U16
Arthritis					X	X	X	X	
Asthma				X	X	X	X	X	X
Atrial Fibrillation					X	X	X	X	
Cancer				X	X	X	X	X	
Cerebrovascular Disease					X	X	X	X	
Chronic Liver Disease					X	X	X	X	
COPD					X	X	X	X	
Dementia					X	X			
Diabetes				X	X	X	X	X	
Epilepsy				X	X		X	X	
Heart Disease					X	X	X	X	
Heart Failure					X	X	X		
Renal Failure					X	X			
Number of Long Term Conditions					X	X	X		X

Other variables in the original SPARRA compared those utilized to create the new ATI models, for each risk cohort. S is the shorthand used for the original SPARRA model, whereas T refers to the results of the data study group

Feature	S:FE	S:LTC	S:YED	S:U16	T:ALL	T:FE	T:LTC	T:YED	T:U16
Prescribed item dispensed					X	X	X	X	X
Hospital admission					X	X	X	X	X
Emergency department attendance					X	X	X		X
Outpatient attendance					X	X	X	X	X
Psychiatric inpatient admission					X			X	
ONLY prescribed item dispensed					X	X	X		X
ONLY had a hospital admission					X		X		X
ONLY had an emergency department attendance					X			X	X
ONLY had an outpatient attendance					X				

turing.ac.uk
@turinginst