Data Sheet for Analysis of A&E Admissions

Motivation

The purpose of the dataset is to see if there are any factors which can predict A&E (Accident & Emergency) Attendances in the UK. There are several studies on risk factors for admission (<https://bmjopen.bmj.com/content/bmjopen/7/6/e011547.full.pdf>) but the process of this project is to see if there are any correlation with labelled datasets on patients attending A&E as attending A&E subsequently will lead to admission. In the world we are in at the moment the is big push on risk stratification and predictive analysis. In practical terms if we can predict determinants of attendance to A&E we might be able to help intercept them prior to attendance.

The dataset was created

* by South West London BI service who securely emailed attendances to A&E over the last 3 years of patients from my surgery only
* A data analysis tool I use in my surgery to risk stratify and link population groups to deprivation and population health metrics.

Composition

The dataset comprises labelled data comprising of the following inputs

* Gender
* Age
* Is BAME (Black Asian Minority Ethnicity)
* Actual Ethnicity (eg Black British)
* Conditions as columns eg Coronary Heart Disease (CHD), Hypertension (BP)
* Number of Long Term Conditions
* Deprivation Index based on National Census Data
* Health Index based on National Census Data
* Risk Score from in house algorithm
* Is Housbound (Bool)
* Latest Observational Values
  + BMI
  + eGFR (Renal function)
  + Cholesterol
  + Hba1c (for Diabetes)

And the following outputs

* No of A&E Attendances
* A&E Attendance in the last 3 years (Bool)

There is no missing data but from evaluating the project it might be that there are unknown or unplotted markers of A&E Attendance which will be more specific. I believe that Mental Health is an important factor of A&E Attendance and also Access to Primary Care and Confidence in the Primary Care Physician all of which can be harder to quantify but something to explore in future projects

The initial dataset is confidential and comprised of Patient Identifiable data but has been stripped of all PID information such as primary keys, NHS Numbers, Patient Addresses and other demographics and date of births. I have approached my DPO to ensure this is anonymised enough for public use. The primary key has been replaced with an autonumber which cannot be traced back to the original patient (ie it is not pseudo anonymised)

Collection Processes

The data was acquired via a search from the clinical system which filtered through the in-house tool within the boundaries of the surgery and also an extract of A&E activity from the SWL BI (South West London) Team who emailed the activity data (outputs) securely to my nhs.net email

The Sample Base Dataset is all patients who are in my practice (4495). The A&E data is about 4-6 weeks out of date so there maybe a few patients who I do not have information from who have registerted with my practice in this time, but it will be in the single digits counts ie (<0.1%)

The A&E Data are any of my patients who have attended A&E in the last 3 years. The Base Patient List are patients currently on my list with their associated input data fields. Because I have access to the full population I did not choose any sampling strategy.

Preprocessing/Cleaning/Labelling

The raw data from SWL BI was activity of each instance of attendance, date of attendance with other metrics which I did not use. I would have liked to have data around subsequent hospital admissions too but was unable to obtain this.

From this data I used pandas to rearrange the data so I had a table with 1 row per patients via their primary key and frequency of attendance. The primary key was the same as with my practice dataset so I could map this across to get the dataset of primary care inputs with A&E output data. This way I know for all my population has there been any attendance to A&E for my population in the last 3 years and what was the frequency of attendance in this time.

Uses

As well as using this dataset to establish a link between A&E Attendances and primary care factors it could also be used to see if deprivation index (output) is linked to BAME or Long Term Conditions (inputs) or more abnormal metrics (eg high Hba1cs). There are several uses for this dataset which could be explored.

The dataset could inform local healthcare providers about the demand for A&E services and help allocate resources more effectively, such as targeted community health programmes or outreach services.

A wish of future iterations would be overlaying mental health. While a small amount of mental health data is included, the dataset could be combined with external mental health datasets to explore the relationship between mental health conditions and A&E attendance.

As the dataset only represents my patients it might not be an appropriate reflection of other areas due to differences in demographics. Also, there is no data on other ethnic groups. The reason for this is that BAME is a known risk factor for poor health outcomes. It is also important to note that there might be other factors determining A&E attendance eg primary care access so future research might need to be done in this area.

Strategies for mitigations would include to conduct fairness audits to ensure that analyses do not disproportionately affect specific groups and clearly communicate the limitations of the dataset, particularly its lack of generalisability. We would also need to apply robust anonymisation techniques and ensure compliance with data protection laws.

Where the dataset should not be used for would be in **direct patient identification**. The dataset should not be used to attempt to re-identify patients, as this would violate data protection laws and ethical guidelines. Also **stigmatisation of ethnic groups** is an important feature where we should not use the data. The dataset should not be used to make sweeping generalisations about specific ethnic groups (e.g., assuming all BAME patients are high-risk for A&E attendance). This could lead to harmful stereotyping and discrimination. Finally the dataset should not be used for **overgeneralisation**. The dataset should not be used to make broad conclusions about A&E attendance patterns across the UK, as it is limited to a single practice and may not reflect national trends.

Distribution

The dataset has not been distributed and there are no plans to distribute without ensuring GDPR is maintained even though the data is totally anonymised. There is no copyright or IP Licence infringement

Maintenance

The dataset is currently maintained by myself for the purpose of this project