UNIVERSITY OF
LEICESTER

Department of Informatics University of Leicester

CO7201 Individual Project

Mine, Analyse and Visualise Healthcare Data

Shashidar Ette
se146@student.le.ac.uk
169050065
MSc Advanced Software Engineering

Interim Report

Project Supervisor: Prof Reiko Heckel
Second Marker: Dr Rayna Dimitrova

Submitted: 30th March 2018
Word count: 1073

DECLARATION

All sentences or passages quoted in this report, or computer code of any form whatsoever used and/or submitted at any stages, which are taken from other people's work have been specifically acknowledged by clear citation of the source, specifying author, work, date and page(s). Any part of my own written work, or software coding, which is substantially based upon other people's work, is duly accompanied by a clear citation of the source, specifying author, work, date and page(s). I understand that failure to do this amounts to plagiarism and will be considered grounds for failure in this module and the degree examination as a whole.

Name: Shashidar Ette

Date: 30th March 2018

Table of Contents

Overview

This document details out the present status of the project. It will discuss the approaches considered and individual tasks status. In addition, it will list the tools and technologies selected for the project. The document concludes with the next steps to be executed for the success of the project.

Project Status

Based on the work plan detailed in the preliminary report. The project has 3 iterations (4-weekly). The updated work plan with status is as below:

Week	Iteration	Start Date	End Date	Task	Milestone	Status
	1	19th Feb 2018	18th Mar 2018			
1			*19th Feb 2018*	Project Description	◆	Completed
				Research		Completed
2			*2nd Mar 2018*	Preliminary Report	◆	Completed
3				Data set analysis and parameters		Completed
4				Localization of data for pilot		Completed
	2	19th Mar 2018	15th April 2018			
5				Data exploration		Completed
6			*30th Mar 2018*	Interim report	◆	In Progress
			30th Mar 2018	Data model	◆	In Progress (delayed by 3 days)
7				Data Visualization		
			06th Apr 2018	Interview with second marker	◆	
8			*13th Apr 2018*	First prototype	◆	
	3	16th April 2018	13th May 2018			
9				Data model (feedback)		
10				Data Visualization		
			27th April 2018	Final report draft	◆	
11			*4th May 2018*	Second Prototype	◆	
12				Data Model & Visualization		
	Final	14th May 2018	17th May 2018			
13			*15th May 2018*	Final Prototype	◆	
			17th May 2018	Final Report	◆	

The details of the completed tasks along with tools & techniques selected will be detailed in next section. Currently the second iteration is in progress. In terms of tasks Data Model experimentation is in progress but it is lagging by about 3 days. The delay should be covered by the end of the iteration i.e. 13th April 2018.

Important updates

Based on the preliminary report and datasets selected. The details of the tasks accomplished are as follows:

- Dataset analysis and parameters, this task involved acquisition of data. The datasets were
  - NHS Digital
    - Practice level prescription data for 2017
  - NHS BSA
    - Patient List Size for 2017
    - GP Count
  - Public Health England PHE) 's
    - Practice indicator profiles both at England and Practice level
    - Diabetic indicator profiles both at England and Practice level
    - Cardiovascular diseases (CVD) indicator profiles both at England and Practice level
    - IMD (Index of measure of deprivation) profiles both at England and Practice level

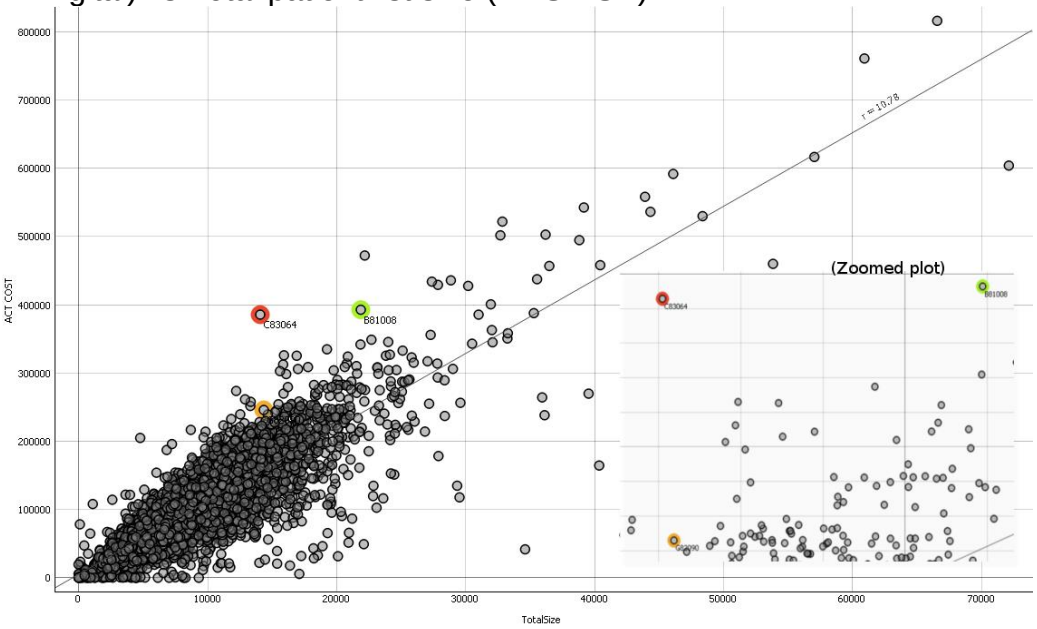
Although PHE data is available through fingertips open API Most of the data acquired are in the format of CSV files. The CSV files make it easier in terms further analysis or exploration.

- Data localization
  The NHS prescription-level data for each month has ~1.25 GB of data and ~10 million data. To simplify the data exploration. CCG area for Q59 (Lancashire and Leicester) is considered. Every CCG has set of practices, thus data such as patient list size, GP count and PHE indicators were considered with respect to those practices.

- In parallel to the data localization activity, I have also attended a course online Python for Data Science (it's in progress) it is provided by the team at the University of California, San Diego. Based on the course labs and recommendations, I have selected “Jupyter notebooks” as a tool of selection for data exploration and analysis. It provides an intuitive way of capture the analysis through notebook interface. It has the support of powerful libraries such as pandas, numpy and matplotlib visualization.

- Data Exploration
  The initial focus of exploration was around cost and patient list size.
  The exploration was broadened to consider PHE indicators such as diabetic and CVD indicators specifically a number of patients with disease in each of the practice.
  One of insight is as below:

  - As mentioned above GP patient list size along with total prescription cost (ACT cost) in Dec 2017 were considered as common factors for comparison between practices. Below is the scatter plot of total prescription costs for Dec 2017 (NHS Digital) vs Total patient list size (NHS BSA).

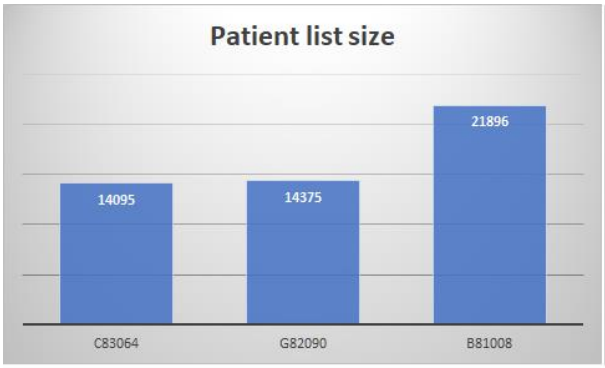


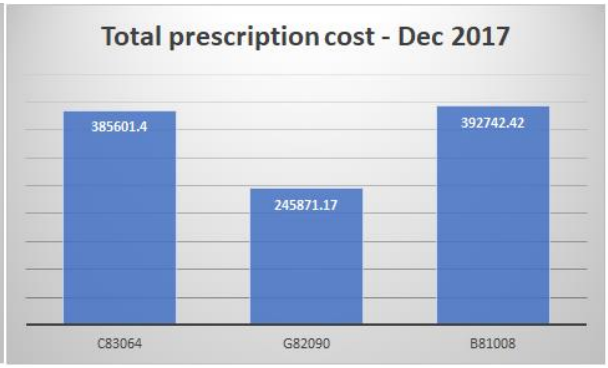
the above plot, there 3 three sample GPs were considered for further analysis. There are chosen based on the factors of:

- similar patient list size but varying costs

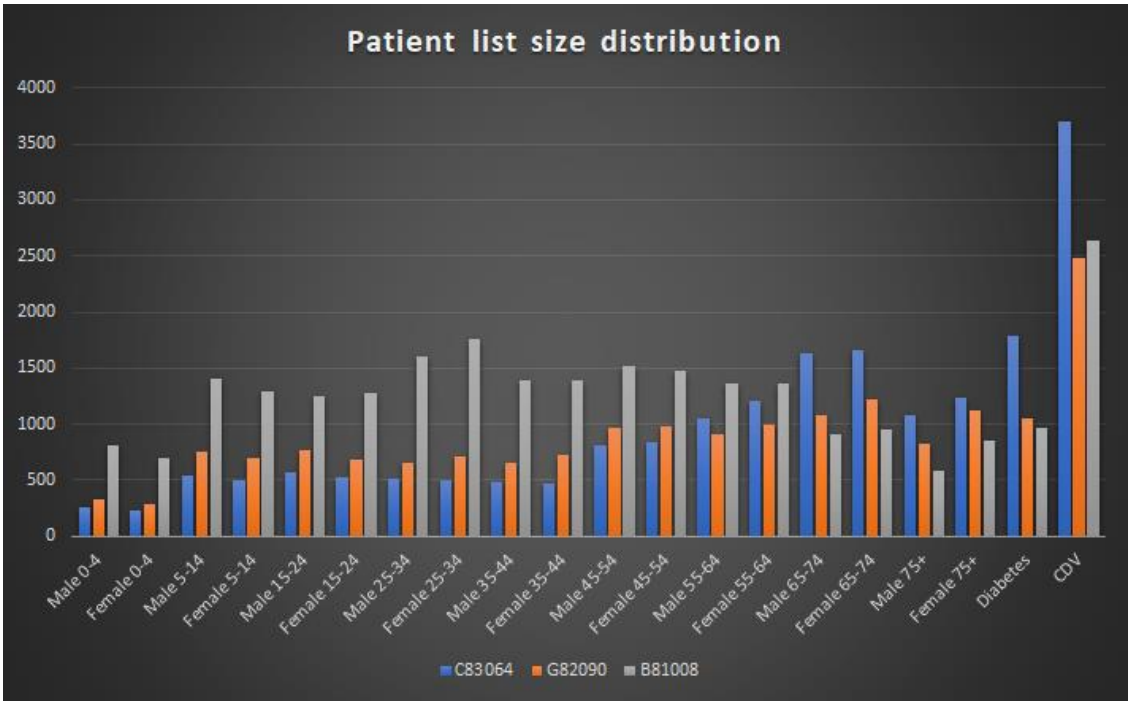
- varying patient list size but related costs

The selected GPs are: C83064, G82090, B81008. C83064 has a patient list size similar to G82090 but prescription costs similar to B81008





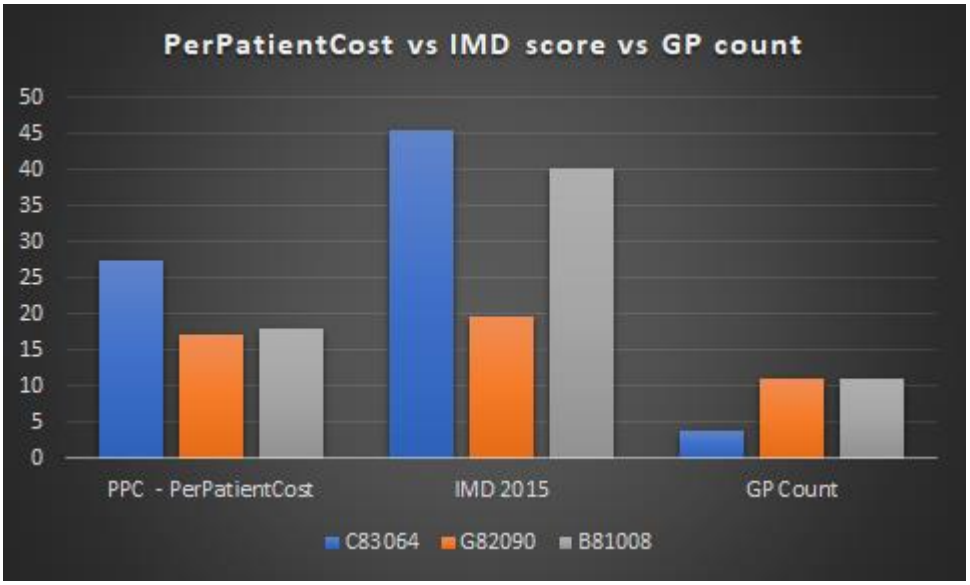
To understand the reasons for the variance, the distribution of patient list size, additional factors such as disease prevalence (PHE: Diabetes, CVD profiles) were considered:



From the above plot, C83064 differs from other two GPs:

1. It has a higher patient list size in range 55 to 75+
2. Diabetes and Cardio Vascular Disease significantly higher than others.

To explore additional parameters, Index of multiple deprivations (PHE: IMD 2015) and GP Count (NHS BSA) were considered.



From the above plot, with 19% IMD - G82090 is less deprived than C83064 which has a higher deprivation index. However, B81008 also highly deprived area but in terms of GP count, C83064 has only 4 GPs as compared to 11 in others.

To conclude the higher prescription cost at C83064 could be justified based on the patient list and presence of higher diabetic and CVD patient. The next step would be to dwell on prescription costs by chapter and at the prescription level in terms of generic and expensive drugs used at the practice level.

- Since we do not have target classification of the data and what patterns or output might be discovered from the data. I will be using the concepts of unsupervised learning and apply machine learning algorithms using Scikit-learn python library within Jypyter notebooks. The base ML algorithms to be used will be cluster analysis and association analysis.

- Based on the findings the idea is to form clusters based on patient list size, disease profile indicators and compare them with respect prescription cost overall or at specific chapter or medicine level. The details and differences of prescription and medicines are present in BNF (British National Formulary). The domain knowledge and its guidance will be needed.

- Data model and the experimentation is the current task. This is delated by about 3 days and I would target to recover the same before the end of the iteration i.e. 13th April.

- The data model will require test cycles with different samples of data for validation and build confidence on it.

- In parallel to the data model, I am also analysing a 3d visualization library based on python to be used in the prototype.

- Although a week in advance, I also had the first meeting with the second marker. I provided an overview and discussed the status of the project. I also shared the next steps involved in the project.

- The findings and outcomes from data model testing will be discussed with Supervisor and stakeholders. The feedback suggested will be analysed and incorporated accordingly.

Next steps

Based on the work plan the next steps will be:

- Data model testing and validation with reference to independent and dependent variables.
- Data visualization prototype with datasets incorporating the data model.
- Demo and deploy the prototype in stakeholder's environment to capture feedback.
- Incorporate the feedback into the Data Visualization application.
- Final project report
- The report will capture the learning and insights. In addition, it should also document the future developments possible.

References

1. NHS Digital - https://www.digital.nhs.uk/prescribing
2. NHS BSA - https://www.nhsbsa.nhs.uk/information-services
3. Public Health England - https://fingertips.phe.org.uk/