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

Preliminary Report

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

Submitted: 2nd March 2018
Word count: 2113

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: 02nd March 2018

Table of Contents

Introduction

With the advent of government policies towards digitization and providing effective healthcare there are processes and policies implemented to capture the data at several levels.

Healthcare data as a landscape has several types such as:

- Prescription data
- Practices and patient size data
- Public Health profiles
- Dispensing data
- Prescription Cost Analysis data

This poses some crucial questions such as:

- How do we manage such large volumes of data?
- What insights can be generated from it?
- Can this data be used to formulate the policies effectively and predict trends in turn result in improving healthcare system?

Background Study

As part of the research, the focus was three-fold as below:

- To look at data sources of healthcare data
- To look at the existing systems which are designed around the data sources under consideration. As a system, either it can merely act as a data repository or provide as well certain processing and analytical capability.
- To look at the existing research done by other organizations or institutions in this arena.

A list of data sources and systems found during the research are as below:

1. NHS Digital – This is a digital platform providing information systems including infrastructure for health and care. It provides several datasets in relation to prescriptions. However only below two datasets are considered for this project:
    - Clinical Commissioning Group (CCG) prescribing data
    - Practice Level Prescribing in England

2. NHS Business Service Authority's Information Portal – This has "reports and data to help NHS customers track trends, inform decisions and support policy". Below datasets are considered.
    - BNF Code Information
    - Patient List Size Information
    - It also provides Detailed Prescribing Information as offline data (same as – NHS Digital Prescription data)

3. Public Health Data England or FingerTips data

It provides profiles which "are a rich source of indicators across a range of health and wellbeing themes that has been designed to support JSNA and commissioning to improve health and wellbeing, and reduce inequalities. With these profiles you can:

- Browse indicators at different geographical levels
- Benchmark against the regional or England average
- Export data to use locally" Retrieved from https://fingertips.phe.org.uk/ [8]

This is an important dataset, to link the prescription with the outcomes and effectiveness of health care policies.

4. Office for National Statistics (ONS Digital)
   It is the "largest independent producer of official statistics and the recognised national statistical institute of the UK". It provides several datasets including:
   - Mortality reports
   - Suicides
   - User requested data sets
   - Lookup Datasets – codes and names -https://digital.nhs.uk/organisation-data-service/data-downloads/national-statistics

5. OpenPrescribing –
   This is a system provided in collaboration between several partners namely The Health Foundation, University of Oxford, NIHR and others. It also encompasses other data sources. It also provides Open API - https://openprescribing.net/api/.
   However, during the discussion with stakeholders, it was found that the clinical questions under consideration are not still answered since the data is visualized in isolation from practise and CCG perspective but does not give information with respect to age/disease in the same view.

6. http://www.prescribing.info/
   This is another system built using Microsoft BI. This primarily focuses on prescription data and provides a drill-up and drill down views which is an important reference for the project.

In terms of research, below technical articles were referred:

- Raghupathi, W. and Raghupathi, V et al research "Big data analytics in healthcare: promise and potential" mentions that Big data "has the potential to transform the way healthcare providers use sophisticated technologies to gain insight from their clinical and other data repositories and make informed decisions." The research also suggests architecture for a big data analytics and how it can be applied to healthcare data.

- Liu, S., Cui, W., Wu, Y. and Liu, M et al "A survey on information visualization: recent advances and challenges" – This technical paper details the different ways of research data and how it can be visualized using Data Driven document D3 JavaScript library.

- van der Corput, P., Arends, J. and van Wijk, J.J. et al research "Visualization of Medicine Prescription Behavior" – This paper did research to find out the relationship between physician, patient and medicine and developed prototype with visualizations which enabled physicians to adapt their prescription behaviours.

- Laramee, B. et al research - “Time-oriented Cartographic Treemaps for Visualization of Public Healthcare Data” – This paper has focused on the NHS prescription data, showcased how Treemaps can be used to demonstrate prevalence of drugs with drugs at CCG level. This is an important reference in terms of visualization of the data under consideration at practice level and will help to discover insights for clinical questions.

Project Motivation

Based on the research, although there are systems which process the healthcare data in isolation, they lack in correlating the data between related data sets. Also, there are certain limitations on how the data is visualized i.e. either the results are shown in the form of tables or in the form of large reports making them less interactive and intuitive in use.

The focus of this project is primarily on the prescription data and related data sets to discover insights for clinical questions about drug or disease prevalence and relate them to other demographic parameters.

Objectives

The primary objective will be to experiment on the data sets available and develop a proof of concept or a pilot of a framework or a model to analyse and visualize them in more intuitive ways using modern technologies.

In terms of data sets the focus will be on:

- Prescription data (from NHS Digital) along with patient list size (from NHS BSA)
- Fingertips data from PHE will be considered at the second level for disease prevalence.

Some of the additional objectives under consideration are:

- Understand cost variation between practices
- Analyse prevalence of certain drugs
- Discover correlation between volume of patients with the volume of prescription
- What proportion of practice population has diabetes, hypertension and heart problems and how does this affect prescription practices?

Challenges

- For the data sets considered, the primary challenge is the volume of data. For example: prescription data for a month is about ~ 1.25 GB. If we consider it for a year, it will be ~ 15 GB. As part of the pilot, the data will be considered for Leicester City and CCGs part of it.
- The data sets might involve too many variables, hence its crucial to choose appropriate parameters to analyse and generate valuable insights.
- Since the project will involve a correlation between 2 or more data sets, the data available might not be in the expected format. It might have aggerated data. Hence as part of data exploration phase, need statistical approximation will be identified and sent for review.

- A core challenge will be to device visualization techniques on top of data parameters in an effective and intuitive way so that it generates value from the end-user perspective.
- The project will involve the application of data analytics and science concepts, being from a conventional software development background it will involve a learning curve to understand and implement the techniques. For the same, I am taking up Python for Data Science course UCSanDiego on edx.org. In terms of visualization, d3js is a primary contender and requires a steep learning curve.

Requirements

Aims

The primary aims of the project are as follows:

- Develop a pilot framework or methodology for analysing the prescription data and its visualization
- Data measurement which informs rather than mere comparison (for ex: between practices) with additional details such as patient/age distribution
- Use innovative ways to present the data at multi-dimensional or parameter level

To achieve the aims, the set of requirements needs to be analysed and executed in an iterative way keeping it flexible. Hence, the priority and details of the requirements might change as we progress in the project.

Essential

Priority I:

- Define data sets and the parameters for consideration
- To prepare localized data sets for pilot
- Explore the data sets with reference to clinical questions under consideration
- Explore and generate correlations of data sets
- A set of clinical questions to be considered in the order of priority are as below:
    - Practices - Patient Size and Prescription cost variation
    - Practices level insights with respect prevalence of drugs and diseases
    - CCG level view - practices & prevalence of drugs

Priority II:

- Develop a methodology and generate a model to normalize, visualize and correlate the data
- Visualizations for correlations.
- Export the correlated data in visually rich formats
    - Combining Patient Size, Prescription Volume and Patient Distribution to bring insights

Recommended

- Cost per head calculation and its relation to patient size and prescription volume

- Practices - Prescription and Distribution among drug groups for ex: Diabetes Hb1Ac with Statin usage on a certain population on QoF prevalence similar for cholesterol

Optional

- Develop a (used in epidemiology) technique for statistical calculation of prevalence - i.e. to develop a prevalence estimate which is age adjusted
- Build a technique to predict the demand for volume for a certain drug
- 3D Visualization of data from multiple datasets

Technical Specifications

In terms of a technical solution, the project will cover distinct aspects listed below:

- Analytics
  - This will involve data gathering, cleaning and exploration aspects.
  - This will also involve data transformation using the ETL methodologies.
  - The model developed will act as a base for integration with the presentation layer.
  - Python is considered as a technology for this aspect.
- Visualization
  - This is the presentation layer and will be used by the end-user
  - It will be a web application and will involve interaction with a data model developed via web services
  - For visualization Data Driven Document d3js – a javascript library will be used.
  - The technology to be used for web application has many options and decision has some flexibility. At this moment, Python Django or Spring Framework is considered as the web application framework.
- Business Layer
  - The business layer will be required to manage the communication with data sources, data model and presentation layer.
  - This will involve communicating with web services and expose the data model as web services.
  - Like the presentation layer, from a technology perspective this also many options. Either Java or python will be considered.
- Computational and Processing power
  - Since the solution will involve processing of large data sets. Although we will be localizing the data for the pilot, there might be need of high computational and processing power. If needed Microsoft Azure available through University Student subscription will be used.

To conclude, the technology will be:

- Python
- Java
- Django

- D3JS
- HTML, JavaScript
- Azure

In addition, during the development different frameworks will be selected as necessary.

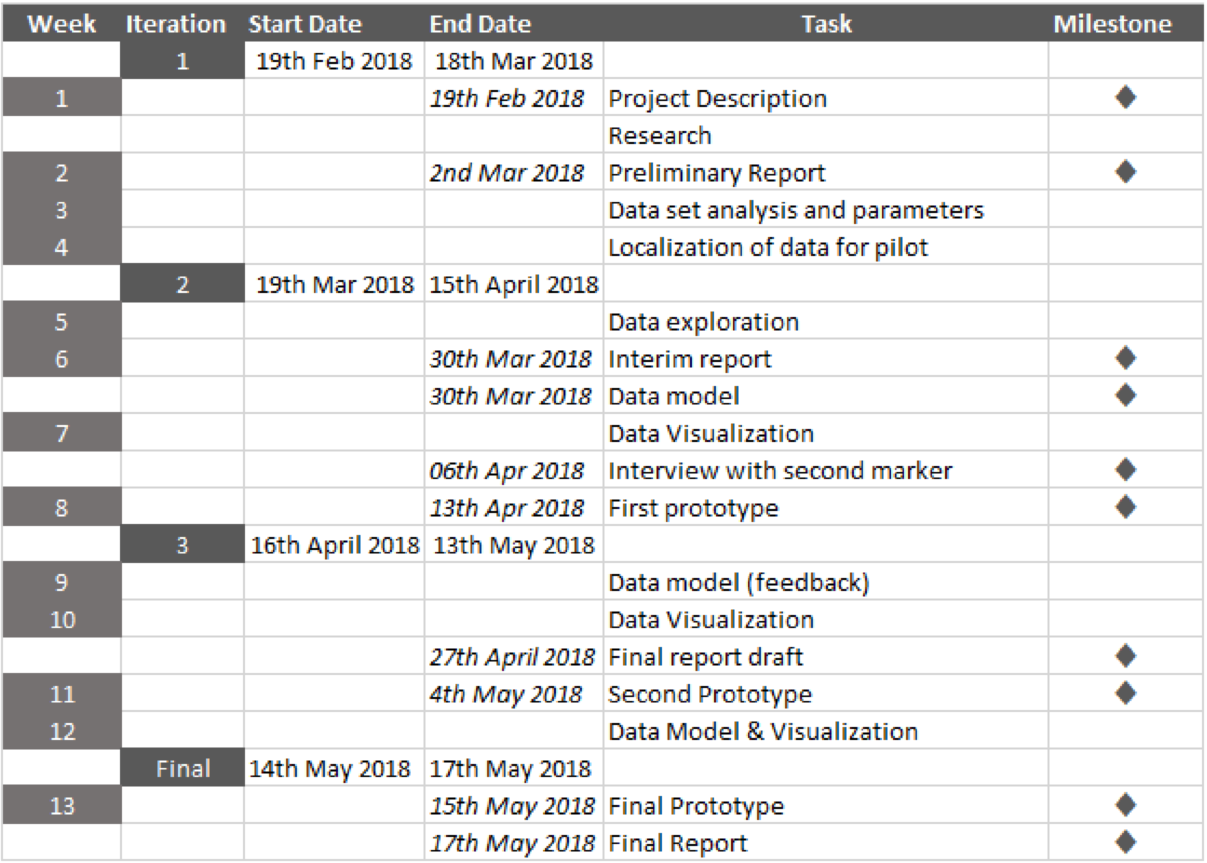
Work Plan

Considering the experimental nature of the project, it will follow an iterative model with each iteration having following phases:

- Requirements prioritization in terms of clinical questions on focus
- Data analysis and experimentation
- Prototype – an app with visualization and parameter selection
- Feedback

This project will be planned in an iterative way with each iteration spanning 4 weeks. So overall there will be 3 iterations spanning 3 months.

The high-level plan considers both project and academic commitments with milestones marked as ◆ (diamond) symbol. Priorities for data exploration, model and visualization will be decided and detailed on-going basis and hence kept at an abstract level. The details are as below:



Week	Iteration	Start Date	End Date	Task	Milestone
	1	19th Feb 2018	18th Mar 2018		
1			*19th Feb 2018*	Project Description	◆
				Research	
2			*2nd Mar 2018*	Preliminary Report	◆
3				Data set analysis and parameters	
4				Localization of data for pilot	
	2	19th Mar 2018	15th April 2018		
5				Data exploration	
6			*30th Mar 2018*	Interim report	◆
			30th Mar 2018	Data model	◆
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	◆

Figure 1 Work plan

Risks

Considering the experimental nature of the project and is primarily related to data sets being analysed following project and technical risks are foreseen.

- Dataset unavailability – it is an assumption that data sets will be retrieved with an API interface provided by the data source provided. In case of data unavailability, the data must be downloaded offline and considered for processing. This will have an impact on overall project plan and execution.
- Data structure consistency - It is assumed that the internal structure of the data will remain consistent throughout the project. In case of variances relevant assumption will be made with a review from stakeholders and data will be managed appropriately.
- Dataset correlation – One of the primary assumptions is that data sets do have the data in common to be correlated. In case it is found that with the data available correlations cannot be made, then it will be a blocker to create the appropriate data model.
- Data Visualization – The visualization of the data is limited by the technology considered. In case certain visualization is required and it's not available, either that requirement should be postponed, or additional effort should be spent to research another visualization framework.
- Computational Infrastructure availability – The technology considered (Azure) for deployment of the data model becomes unavailable, alternate options need to be considered.

References

1. Laramee, B. (2017) *Time-Oriented Cartographic Treemaps for the Visualization of Public Healthcare Data.* (Accessed: Feb 21, 2018).
2. Liu, S., Cui, W., Wu, Y. and Liu, M. (2014) 'A survey on information visualization: recent advances and challenges', *The Visual Computer,* 30(12), pp. 1373-1393.
3. Raghupathi, W. and Raghupathi, V. (2014) 'Big data analytics in healthcare: promise and potential', *Health information science and systems,* 2(1), pp. 3.
4. van der Corput, P., Arends, J. and van Wijk, J.J. (2014) 'Visualization of Medicine Prescription Behavior', *Computer Graphics Forum,* 33(3), pp. 161-170.
5. OpenPrescribing - https://openprescribing.net/
6. NHS Digital - https://www.digital.nhs.uk/prescribing
7. NHS BSA - https://www.nhsbsa.nhs.uk/information-services
8. Public Health England - https://fingertips.phe.org.uk/
9. Office of National Statistics - https://www.ons.gov.uk/