University of Strathclyde Business School

DEPARTMENT OF MANAGEMENT SCIENCE

Stratified Medicine Scotland

Creation of Synthetic data and Visualisation for Exemplar Projects

Simon Fong

201771505

MSc Data Analytics

Academic Supervisor: Dr Sarah J.E. Barry

Industrial Supervisor: Craig McCallum

University of Strathclyde

21/09/2018

Declaration

Except where explicitly stated, all the work in this dissertation – including any appendices – is my own and was carried out by me during my MSc course. It has not been submitted for assessment in any other context.

Signed:
Print Name: Simon Fong
Date: 21/09/2018

Summary

Stratified Medicine Scotland (SMS) aims to advance precision medicine in Scotland by encouraging collaboration between industries, researchers and clinicians. Data is very important in precision medicine and SMS has a digital platform for secure data storage, visualisation and analysis. R-Shiny apps can be created in the platform for data visualisation, but it was found not many users were creating R-Shiny apps. A demo dataset and app were required to showcase the platform and the benefits of R-Shiny apps to existing and potential clients, this was done for the SteatoSITE project. In addition, an app for the Future MS project was required to visualise around two years worth of data for the very first time.

In order to build an effective app, data visualisations techniques were investigated for different type of data, it was found boxplot and histogram were effective to visualise continuous data and bar charts for categorical data. There was no clinical decision support system (CDSS) capabilities on the SMS platform and it was a feature SMS would like to explore. Machine learning techniques within CDSS were explored and they performed well with a good accuracy rate from past research. A decision tree was implemented in the SteatoSITE app to predict the likelihood of a patient having non-alcoholic steatohepatitis (NASH).

Research was conducted in the relationship between different variables for the SteatoSITE synthetic dataset to create a realistic dataset. The SteatoSITE app was well received especially the decision tree as it was something the chief investigator, Prof. Fallowfield was very interested in implementing with real data in the future. The app had much more features and visualisation compared to the other demo apps on the demo workspace, this would give the SMS team a better showcase app to existing and potential clients.

The FutureMS section of this report includes the data cleaning process to help readers understand the steps taken in preparing the data so that changes to the app can be made in the future by other members. The FutureMS app allowed the team to finally visualise the data they have been collecting for around two years. The patient timeline visualisation was found to be very useful especially for patient visits as clinicians could see from a glance, all the important events for the patient.

Acknowledgements

I would not have been able to complete this dissertation without the help of others, I would like to take this opportunity to thank them. Firstly, I would like to thank Stratified Medicine Scotland for giving me this opportunity and my industrial supervisor Craig McCallum who answered all my questions and supported me throughout the placement. I would also like to thank Dr Rachel Dakin for helping me understand the projects and setting up feedback meetings with key project members which ultimately improved the synthetic dataset and apps. Lastly, special thank you to my academic supervisor Dr Sarah Barry for all the help and guidance for this dissertation.

Ethics

This dissertation involves two current projects at SMS, the SteatoSITE and FutureMS, both managed by project manager Dr Rachel Dakin. Permission was granted to write and discuss these two projects in this report.

There are no ethical issues concerning the SteatoSITE dataset as the synthetic data was created without any real patient data. Discussions related with this project with project members did not involve specific patients.

The FutureMS dataset contained real patient data but the data used for this project are anonymised and without the identifier key, no personal information of the patient data can be found. The project was completed on the SMS platform and there were security measures in place such that no data was taken from the platform onto the local machine. Permission was granted to use the data for visualisation purpose from Dr Dakin and results from the app could be part of this report as well. One of the requirements was that this dissertation would not be published online.

Table of Contents

Project Context

Client Report

List of Figures and Tables

Project Setting

1 Client Background

Stratified Medicine Scotland Innovation Centre was established in 2013 by the Scottish Funding Council and their focus is progressing precision medicine in Scotland by bringing together industries, researchers and clinicians to collaborate on precision medicine projects. Dr Diane Harbison is the Chief Executive Officer of Stratified Medicine Scotland (SMS), and the organisation consists of different teams. The laboratory team consists of two scientists and a lab manager, whom provide laboratory services. There is also the project management team which oversees the client projects and business development team which seeks out business opportunities. Finally, there is the head of IT, Craig McCallum who oversees all the IT developments and services within SMS.

SMS has several partners including industry partnership with Aridhia Informatics and Thermo Fisher Scientific. They also have partnerships with University of Glasgow, Edinburgh, Aberdeen and Dundee and with NHS Grampian, Tayside, Lothian and Greater Glasgow and Clyde.

A key element of most precision medicine projects is integration of large data sets. Some of which are existing such as information from patient health records and some are generated as part of research studies such as whole genome sequencing or MRI scans. By securely combining these types of information a new resource is generated for use in research and development and ultimately to benefit patients.

The Stratified Medicine Scotland Innovation Centre Platform (SMS Platform) is a digital health platform and is a white label product of AnalytiXagility by Aridhia. The SMS platform has a range of features such as collaborative workspaces, de-identification service, hosting and data management to give their clients the capabilities for precision medicine. SMS can create a lot of workspaces and clients are usually given at least one workspace. Each workspace can be tailored to the client such as maximum storage space and different types of services.

Depending on the client's project requirement, they can choose which services to include on their workspace. The main attribute of the platform is the security of the data storage. There is an audit trail on the platform and users of the workplace can be given certain levels of permission, allowing for a collaborative workspace where users can be invited in and given the correct level of permission for example, a contributor can upload data onto the workspace but cannot see the data stored on the workspace. If a user of the workspace wants to download any data, the request would need to be approved by a workspace admin first.

The SMS Platform has in-built R environment for analysis and users can create interactive R-Shiny apps on the platform for visualisation and analysis purposes. A R-Shiny app uses the Shiny package in R to create interactive web apps. A basic shiny app has two components, the UI and Server. The UI is responsible for the layout and appearance of the app. The server component contains the code required to build the app (Shiny.RStudio 2017). The UI.R and Server.R files are the minimum two files required for a R-Shiny app but as the app becomes more complex, more files can be added for example, a modules R file can be used to create functions in the Shiny environment (Cheng 2017). A shiny app can be used to interactively visualise data and users can change input parameters and the plots dynamically update.

SMS would like to promote the use of data analysis and visualisation in the platform as in the past, clients were mainly using the platform for data storage. At inception of this project, there was no one in SMS that could perform analysis and visualisation of the data for clients and for some projects, the client might already have their own bioinformaticians to analyse the data. Bioinformaticians usually use virtual machines to access the data and then use specialised programs to analyse the data. In the past, if a client wanted an R-Shiny app created then SMS would ask the data scientists from Aridhia to create the app for them. SMS would like to have their own data scientist and are in the process of hiring an IT developer that can perform data analysis and write R-Shiny mini-apps on the platform to enhance client services.

2 Project Background and Motivation

Clients rarely used the platform for visualisation purposes with the use of R-Shiny mini-apps. The main aim of this project was to build apps on the platform that can showcase the advantages of writing R-Shiny apps for the platform to visualise data.

The SMS platform has a demo workspace in which there are a number of demo apps made for showcasing the platform to users. The apps on the platform were created by Aridhia a few years ago and the apps have limited features and visuals. On the demo workspace there was also a limited number of dataset available to write mini-apps on. With this project, a larger and more realistic synthetic dataset would be created and stored on the demo workspace to show potential and existing clients.

In this placement, apps are created for the SteatoSITE and FutureMS projects. The SteatoSITE included creation of synthetic data and an app onto the demo workspace. There was no real data available for the SteatoSITE project hence the need to create a synthetic dataset. Research was conducted to investigate the relationship between the variables to create a realistic dataset for visualisation.

The work on FutureMS involved creation of an app to visualise around two years worth of data collected as part of the project. This project involved anonymous patient data and permission was granted from the project team to access the workspace and view the data.

3 Project plan

In the first week of the project, a project plan for the 12-week placement was created and discussed with industrial supervisor. Chart 1 shows the final project plan for the 12-week period after discussing with academic supervisor as well. The academic supervisor meetings and feedback sessions were added into the project plan.

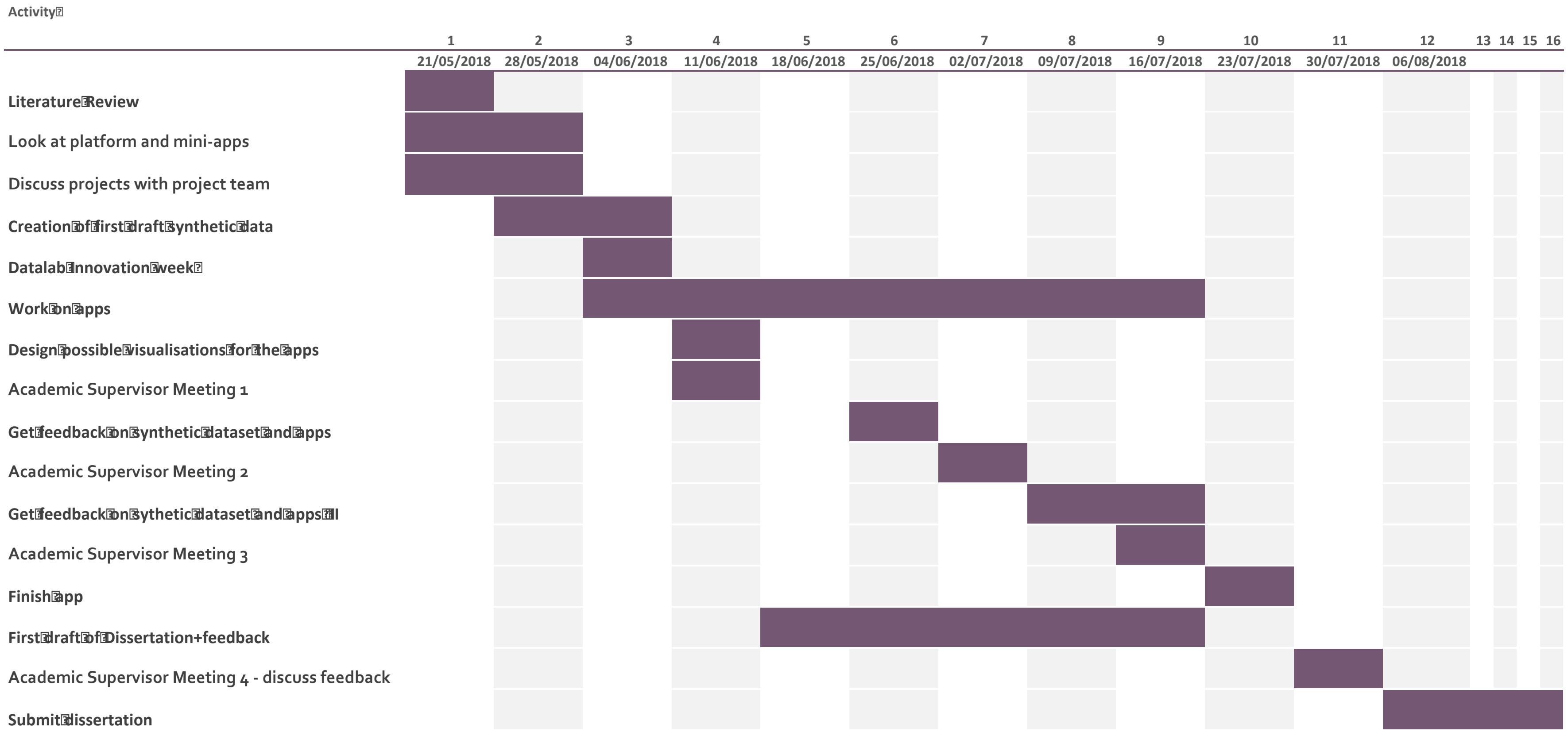


Chart 1 : Project Plan for Placement

University of Strathclyde Business School

DEPARTMENT OF MANAGEMENT SCIENCE

Stratified Medicine Scotland®

Creation of Synthetic data and Visualisation for Exemplar Projects

Simon Fong
201771505
MSc Data Analytics
Academic Supervisor: Dr Sarah J.E. Barry
Industrial Supervisor: Craig McCallum
University of Strathclyde

21/09/2018

Table of Contents

List of Figures and Tables

1 Executive Summary

This report includes details about the steps taken into the creation of two R-Shiny apps on the SMS platform for two projects, SteatoSITE and FutureMS. In addition to the apps, a synthetic dataset was created for SteatoSITE project and data cleaning was carried out for the FutureMS project. The intention was that the SteatoSITE synthetic dataset and app would be added onto the demo workspace on SMS Platform which would allow SMS to show the app to existing and potential clients. The FutureMS app would allow the FutureMS team to comprehensively visualise the data they have collected for the first time.

Effective data visualisations techniques were investigated for different type of data, it was found boxplot and histogram are good to visualise continuous data and bar charts for categorical data. Similar digital healthcare platforms were looked at to find features not on the SMS platform. There was no clinical decision support system (CDSS) capabilities on the SMS platform and it was a feature SMS would like to explore. Machine learning techniques within CDSS were explored and they performed quite well with a good accuracy rate.

Research was conducted into relationships between variables to create a realistic dataset for the SteatoSITE project, based on a clinical specification. The SteatoSITE app was well received especially the decision tree of the app as it was something the chief investigator, Prof. Fallowfield was very interested in implementing with real data in the future. The app had much more features and visualisation compared to the other demo apps on the demo workspace. This would give the SMS team a better showcase app for clients during presentations. The link to the SteatoSITE app can be found here: https://simonfong.shinyapps.io/steatosite_app/

The FutureMS app allowed the team to finally visualise most aspect of the data they have been collecting for around two years. Certain insights were found from the app including the lower than expected vitamin D in the blood test result and the irregularities in the height measurements. The patient timeline visualisation was found to be very useful especially for patient visits as clinicians could see from a glance, all the important events for the patient.

2 Acknowledgements

I would not have been able to complete this dissertation without the help of others, I would like to take this opportunity to thank them. Firstly, I would like to thank Stratified Medicine Scotland for giving me this opportunity and my industrial supervisor Craig McCallum who answered all my questions and supported me throughout the placement. I would also like to thank Dr Rachel Dakin for helping me understand the projects and setting up feedback meetings with key project members which ultimately improved the synthetic dataset and apps. Lastly, special thank you to my academic supervisor Dr Sarah Barry for all the help and guidance for this dissertation.

3 Introduction

Data visualisation is important for exploratory data analysis and presenting results to users. The use of interactive visualisation allows decision makers to gain further insights in the data. Within the Stratified Medicine Scotland (SMS) Platform, R-Shiny mini-apps could be created for project members to visualise the data. However, most clients were not using the platform for visualisation and analysis purposes. The objective of this project was to create a synthetic dataset and mini-apps to showcase the platform capabilities. The SMS demo workspace, which is used to demonstrate the platform to potential clients, has a limited number of apps and SMS would like to have a more complete app on the demo workspace to show clients. SMS has two projects where an app for visualisation was required, SteatoSITE and FutureMS.

The SteatoSITE project will involve creating a data common for Non-Alcoholic Fatty Liver Disease and have not started the data collection yet. The study is headed by chief investigator, Prof. Jonathan Fallowfield who provided feedback on the synthetic dataset and app along with fellow investigator Dr Timothy Kendall. The study investigates causes behind the increasing risk of non-alcoholic fatty liver diseases (NAFLD) in Scotland. The main aim of this study is to develop new treatments and tests for non-alcoholic steatohepatitis (NASH) by sharing data and resources to users such as clinicians and researchers ('Why We Need a NASH Data Commons' 2018). The SteatoSITE app would allow the project members to see a range of visuals and give them an idea of how their apps might look like in the future once they obtain the data. Due to the early stages of the project, there was no real data to work with and so a synthetic dataset was created. The app and synthetic dataset could also be used in the demo workspace to show clients. Several functions were written in R to generate data for the synthetic dataset and research was conducted to find the relationships between the variables. The app would include a series of plots to visualise the data and include interactive plots to allow the user to analyse the data. A decision tree was created using the data to explore clinical decision support system (CDSS) in the app. Multiple feedback sessions were conducted to improve the app and synthetic dataset.

The FutureMS project is a Scotland-wide study involving collecting clinical data from patients recently diagnosed with relapsing-onset multiple sclerosis (MS) in Glasgow, Edinburgh, Aberdeen Dundee and Inverness. Scotland has one of the highest prevalence of MS in the world and the main aim of the FutureMS project is to use all this data to create a tool that can predict how severe the patient's MS will become so that the patient can be put on the correct treatment plan. The type of data collected includes a number of clinical physical tests, blood test results, and questionnaires.

There were no dedicated visualisation tool for the FutureMS project on the workspace and so project members could not see any visualisation of the data they've collected. The FutureMS has been collecting data for nearly two years and the data has to be cleaned before it can be visualised. The app would allow users to see at a glance the data they've collected and identify any irregularities in the data collection process for the very first time.

4 Literature Review

Technology has played an important role in healthcare in recent years, it has allowed gene sequencing to be computed in less time and cost. With advancements in technology, it has become cheaper to store and easier to collect data. Digital storage has allowed clinical data such as blood test results, patients information and gene sequencing to be stored securely and easily accessible. With access to so much data, there has been a correlation with precision medicine where it aims to use data to provide better treatment plans, early preventions and better diagnosis (Hopp, Li, and Wang 2018). There are a number of digital healthcare platforms which aims to facilitate the advancement of precision medicine.

4.1 Digital Healthcare Platforms

There are similar digital platforms to the SMS platform, such as Orion Health ('Solutions for Precision Medicine' 2018) that offers a number of products in healthcare. One of the products similar to SMS platform was the Amadeus platform which allows the storage of a range of data such as clinical and from medical devices. One of the biggest differences between these two platforms is that the Amadeus platform has built-in machine learning tools and APIs which allows other third-party developers to use these data to build their own innovative apps for precision medicine.

Cohesion Medical offers similar services to the SMS platform such as User data entry system which allows users to add and view data dependent on the level of access (CohesionMedical). The Stratified Medicine Initiatives services offers their clients a number of in-built models such as patient, disease, incidents and treatment models, which uses predictive algorithms to predict outcomes (CohesionMedical). For example, the treatment model can be used to analyse effectiveness of treatments for patients.

The Eagle Genomics platform is another platform used in healthcare similar to SMS but with a focus on genomic data (Eaglegenomics). The smart data management platform allows users to gain insight from large amount of data in an intuitive way ('The Automated Data Scientist' n.d.). Eagle

Genomics also has a number of different modules on the platform including a dedicated genome browser called e(nsembl) which is a tool for searching genomic data and visualisation ('e[nsembl]' n.d.).

4.2 Visualisation of Clinical Data

Visualisation of data is usually the first step in understanding the data before analysis can begin. The use of graphics is good for data exploration and allows users to quickly interpret the results. Chun-houh Chen (2008) suggested that there are two types of graphics, presentation and exploratory graphics, and proposed that presentation graphics are usually static, while exploratory graphics are interactive graphics used for looking for results in the data.

For plotting single variables, the most common plots for continuous variables were the boxplot and histogram. For categorical variables, bar charts were one of the most common visuals and stacked bar charts can be used to compare data categories (Sahay 2017). Scatterplot can be used to visualise the relationship between two quantitative variables, the correlation coefficient can also be calculated to give an indication of the relationship of variables (Sahay 2017).

Individual patient’s data could be visualised to allow clinicians to understand the patient’s history and identify any trends or events which result in the patient to react differently to the rest of the sample. One example of the visualisation of patient history was LifeLines (Plaisant et al. 1998), which gives an overview of the medical history of the patient. The events could also be colour coded to determine the severity of the event. This kind of information could be useful for clinicians to identify if a patient on a certain medication could reduce severity of events.

Genomic data has been fundamental in the progression of precision medicine and with the cost of gene sequencing going down, sequencing data are becoming easier to access. For example, tumour genome structure can be used to determine a course of treatment (Jensen et al. 2017). Prasad and Ahson (2006) investigated the different visualisation techniques for genomic data with advantages and disadvantages for each technique. The heat map and dendrograph visualisations seemed to be popular visualisations as they allowed users to easily see the information and identify subgroups of genes and patients. One of the drawbacks of both these methods was that for very large data set, more computational power maybe required (Prasad and Ahson 2006). The use of principal component analysis (PCA) and scatterplot could also be useful cluster the patients into groups based on the gene expressions (Prasad and Ahson 2006).

4.3 Clinical Decision Support Systems in Healthcare

Clinical Decision Support Systems (CDSS) are computer systems that are designed to support clinical decisions for health professionals (Berner and La Lande 2016). CDSS could help doctors and clinicians to diagnose health conditions. With access to an increasing volume of patient data such as electronic records, more personalised treatments can be applied to patient. For example, early screenings for certain conditions can be personalised if a patient is part of a risk group, possibly a certain gene mutation or family history of the health condition. CDSS can be used to help decision makers with diagnosis, what treatment course to take and what dosage of treatment (Musen, Middleton, and Greenes 2014). There are many categories of CDSS and one of them is known as 'non-knowledge based systems' (Berner and La Lande 2016) which uses machine learning, a type of artificial intelligence and the main principle is for computer systems to learn from data (Faggella 2017).

The use of decision trees as a high level tool has been investigated with their use in healthcare and medicine by Podgorelec et al. (2002). Decision tree is popular classification method and one algorithm for generating decision trees is the Classification and Regression Tree (CART) algorithm which splits the data into subsets based on particular variables (Geron 2017). Shaikhina et al. (2017) used decision trees and random forests to predict the risk factors which would result in a patient rejecting a kidney transplant from a sample of 80 subjects where it achieved an accuracy rate of over 85% for both models and identified key predictors. Iyer, Jeyalatha, and Sumbaly (2015) used Naïve Bayes and decision trees for diabetes diagnosis with promising result, correctly identified more than 75% of the instances for both Naïve Bayes and decision tree models. These are two examples of precision medicine, identifying patients more likely to reject a kidney transplant and better diabetes diagnosis.

Digital healthcare platform are useful tools for users to store, access and analyse data. There are some platform geared towards a specific area such as genomic data. Visualisation of data are important for users to see the results of the data and to gain insights into the data. Therefore, choosing the correct visualisation techniques are important to effectively visualise the data. Clinical decision support system can be used to help decision makers by using machine learning techniques.

5 SteatoSITE

5.1 Introduction

The SteatoSite project headed by chief investigator Prof. Jonathan Fallowfield aims to investigate causes behind the increasing risk of non-alcoholic fatty liver diseases (NAFLD) in Scotland. In recent years, there has been an increase in research investigating NAFLD, as it has become apparent that NAFLD is becoming a big problem and the prevalence of NAFLD has been increasing yearly (Younossi et al. 2017). NAFLD has been linked to a higher risk of health conditions such as diabetes and hypertension (Lonardo et al. 2018)

Once a patient has been diagnosed with NAFLD, it is a chronic liver disease and NAFLD is the first stage of the progression. Hepatocellular Carcinoma is the last stage of the disease and is the most common cause of death for people with cirrhosis (Forner, Llovet, and Bruix 2012). Since cirrhosis can't be cured, if the patient does not succeed in improving their condition to stop the progression, then the patient may require a liver transplant. It has been estimated that the quarter of the world's population has NAFLD and by 2020, NAFLD would be most common reason for liver transplant in America (Jarvis 2016).

The SteatoSITE project has partners including Universities of Glasgow and Edinburgh, NHS Scotland and Eagle Genomics. The SteatoSITE study team plans to collect records from NHS patients which include blood test results and perform genome sequencing on liver samples from UK Biobank. The patient's data will be stored on the SMS platform. Eagle Genomics will be responsible for the preparation and analysis of the gene sequencing data on their own platform. The main aim of the project is to provide a data resource for researchers to better study NAFLD and ultimately produce better treatments and improve diagnosis tools for this disease.

For this project, a synthetic dataset was created to try and replicate the type of data they might obtain in the future and build an app to show the capabilities of apps on the SMS Platform. This would give the SteatoSITE project team an impression of the app that could be created with the real data in the future. The app along with the synthetic data would also be used as a demo app for the demo workspace to showcase the SMS platform with existing and potential clients.

5.2 Methodology

The methodology for this project included interviews with various team members to understand the type of visualisations that are important, looking at existing apps on the demo workspace, and researching relationships between different variables for the synthetic dataset.

5.2.1 Platform

There are a number of different apps on the SMS platform demo workspace including genome viewer. A screenshot of the genome viewer app is shown in Figure 1

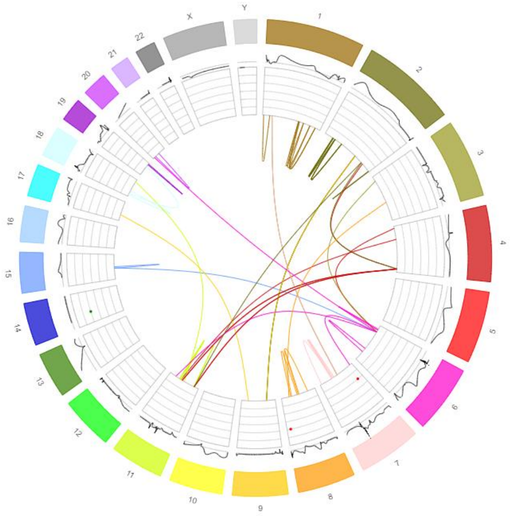


Figure 1 : Screenshot taken from Genome Viewer from SMS Demo workspace (with permission)

The visualisation of the complete genome of a patient looks very good visually but from speaking to members from SMS, it was found that analysis of the complete genome of an individual patient might not be very useful as most studies involve comparison of a whole cohort. A better visualisation for the gene expression would be to use a genomic heatmap to better visualise the sample genomic data rather than focusing on an individual genome data.

5.2.2 Interviews

There were several instances where informal interviews were conducted with data scientists from Aridhia, Harry Peaker and Fraser Black. Topics discussed included how to use the platform and the type of data and visualisation the users would like to see in the platform. There was also discussion on best practices when using the R Shiny app, such as modularisation of the codes, which is similar to writing functions in R but in R-Shiny environment.

At the start of the project, an initial meeting with the industry supervisor, Craig McCallum, and project manager Dr Rachel Dakin was conducted to learn more about the project and what kind of visualisation they would like to see. Surprisingly, simple information such as the number of participants in the study was useful for the clinicians and even simple histograms could give the user a quick overview of the demographics of the patients. There were no specific displays suggested but the plan was to create a first draft and try to create as many different ways of visualising the data as possible to generate ideas.

5.2.3 Synthetic Data Creation

Research was conducted on the different synthetic data creation on R, but most of the information available was concerned with creation of synthetic data from a real dataset. For example, the Synthpop package (Nowok, Raab, and Dibben 2016) in R allows users to create a synthetic versions of original data by changing sensitive information by using techniques such as classification and regression to create synthetic data. However, for this project, there was no real data to work with and so this idea was discarded.

Distribution functions were investigated next such as Poisson and exponential to generate random data. This seemed like a good idea at first but a big flaw of this was that for most of the data columns, it must be within a certain range of values. In order to achieve this, the parameters of those distribution functions would have to be altered for each data column which would be time consuming. Since each data columns have a set range of values and distribution spread not known, distribution functions were discarded as well.

A series of functions was written in R to create data, there were two functions used to generate continuous and categorical data which can be found in Appendix 8.1 and Appendix 8.2. The inputs for the generation of continuous data are the number of rows to generate, a vector containing probabilities, minimum value, maximum value and number of intervals to split the range of values. For example, by inputting a vector of ‘c(0.25,0.5,0.25)’ with minimum value 1, maximum value of 9, and 4 (1 plus the amount of probabilities), there is a 25% chance of the output data to be in the range 1-3 and-7-9 and a 50% chance of it being in 4-6, having a uniform distribution. The categorical generation function works in a similar way but instead of inputting the minimum, maximum value and the number of splits, only a vector of categorical values and probabilities are required. For example, to generate an even split of gender for the dataset, the user would enter in the vector c(‘M’,’F’) and the probability vector of c(0.5,0.5).

At the start of the project, a minimal complete dataset requirement report was given that contained the different data they want to obtain for this project. The report contained a series of data columns and colour coded in terms of mandatory, consultative and informational. A pathology report was also given which detailed the scoring for the pathology scores. Both of these reports were used to create the first draft of the dataset. It was decided at the start to have a hands-on approach, where the first draft of the dataset was created without in-depth research into the relationships between the variables. This was done to get as much done on the app as possible before the first feedback meeting.

5.2.4 Feedback Meetings

The first feedback meeting was at the Edinburgh University Hospital on 27th of June with Prof. Fallowfield and Dr Kendall from university of Edinburgh, along with Dr Dakin and Craig McCallum from SMS. This meeting provided very good feedback on the synthetic dataset which is included in the following section. From the feedback meeting, a number of correlations between variables for the synthetic dataset was suggested. Appendix 8.4 shows the correlation discussed from the feedback meeting. Appendix 8.4 provided very useful correlation to implement into the synthetic dataset, for example BMI had many correlations between the different variables. More research was conducted into exploring the relationship between the variables after the first feedback meeting.

New plots were added in involving the gene expressions after talking to Dr Dakin where it was discussed there would be some genome sequencing data available in the future. Dr Dakin thought it might be interesting to add in some visualisation for genomic data. Since it would be very difficult to create random data for the sequencing data, it was suggested to look for a gene expression dataset from Gene Expression Omnibus (GEO).

There was another feedback session from Prof. Fallowfield and Dr Kendall via email correspondence for the second draft of the synthetic dataset and app. The second draft of dataset was considered substantially improved and there were further suggestions for the dataset and app. Prof. Fallowfield suggested adding in Estimated Glomerular Filtration Rate (eGFR) and Creatinine as there may be a link between chronic kidney disease with non-alcoholic steatohepatitis (NASH), so they would most likely collect this data as part of the data commons. Another suggestion from both was to add in liver related and cardiovascular events into the dataset. These events might include hospital visits where depending on the circumstances, it might be noted down as liver related visit. A patient could have more than one event. It was decided not to include in these events as the list of events would be too long and difficult to assign correct events depending on the patient's health.

5.3 Synthetic Data Creation

As stated in the methodology section, a series of functions were used to create the synthetic data. Subsetting the population were used to give different distribution spreads depending on a particular variable. For example, people with a high BMI would have a greater probability of being diabetic compared to low BMI. Appendix 8.3 shows a table with details on all the columns in the final synthetic dataset. The probabilities used to generate the data were estimates based on available data from literature and logical assumptions.

5.3.1 StudyID, Age and Gender

The first step was to create a study id for the patients and since there was no certain requirement of the format of the StudyID, it was in the form AA00001 to AA00999. The StudyID would be used to as a primary key to connect between the main dataset and the comorbidities medication table. The age ranged between 18 to 80. There was no certain requirement for the age distribution, but most of the patients are aged between 30 to 70. The synthetic dataset had an even split of male and female sample.

5.3.2 Alcohol Intake

The minimal complete dataset specification stated that patients with an alcohol intake of more than 21 units per week for men and over 14 units for women should be excluded. For the first draft of the synthetic dataset, the assumption that patients had a very low alcohol intake was made resulting in a left skewness for the histogram of the alcohol unit. But from the first feedback meeting, it was found that in theory this would be the type the data they hope to get but in reality, there would be an opposite skewness towards a higher alcohol intake. The assumption that males tends to drink more alcohol units than females was made.

5.3.3 Triglycerides

The triglycerides had a range between 100 to 700 (MayoClinicStaff 2015). Since the SteatoSITE project focus on patients with NAFLD, the patients would most likely have a higher BMI. Most of the triglycerides was in the range of 150 to 350 as anything above 350 was considered extreme.

5.3.4 BMI

From the feedback meeting, it was ascertained that there was a correlation between BMI and triglycerides. The next step was to use this to determine the BMI, the dataset was first split into low and high triglycerides levels, where low level has a range between 100 to 200 and over 200 are considered high (WebMD 2016).

Splitting the dataset up into two meant that different probabilities and range can be assigned for the two different subsets. For the low subset, the BMI range was between 20 to 35 with left skewness. For the high subset, it was between 25 to 50 and most of the entries (around 60%) had a BMI between 25 to 35. Upon creating this data, it had a mean value of 33.81 which was in the range of expected average values of 30-35 given by Prof. Fallowfield and Dr Kendall.

5.3.5 Diabetes and NASH

Narayan et al. (2007) investigated the effect of BMI on the lifetime risk for diabetes and obtained risk probabilities for different demographics as shown in Appendix 8.5. Williams et al. (2013) found patients with type 2 diabetes had a prevalence of NASH at 63–87%. Patients with diabetes were allocated a 75% probability of having NASH, since this was near the average of 63 to 87.

There were patients with quite high BMI without Diabetes. The subset of patients with no diabetes was further subset into high BMI (over 25) and low BMI (under 25). Those with high BMI and no diabetes had the same probability of having NASH as those with diabetes. Whilst low BMI with no diabetes only had a 30% chance of having NASH.

5.3.6 AST and ALT

Following Appendix 8.4, the next step was to generate data for aspartate aminotransferase (AST) and alanine aminotransferase (ALT). From Sorbi, Boynton, and Lindor (1999), “Patients with NASH had a mean AST to ALT ratio of 0.9 (range 0.3-2.8, median 0.7)”. Based on this study, the full dataset was split into those with NASH and without NASH. The AST/ALT ratio could also help to determine the likelihood of the stage of liver disease of the patient. For those with no NASH, the AST levels should be lower than those with NASH (Sanyal et al. 2015). It was difficult to generate values for both AST and ALT such that the corresponding AST/ALT ratio was over a certain value for those with NASH.

A solution of this problem was to first generate AST and the AST/ALT ratio, and then using both these values, the ALT value was calculated. Therefore, the patients with NASH could be assigned a higher AST and AST/ALT ratio than no NASH. For the no NASH diagnosis, the AST/ALT ratio was between 0.5 to 0.8 and for patients with NASH, the ratio was between 0.6 to 1.5.

5.3.7 NASH-CRN (Kleiner) fibrosis stage

Based off the first feedback session, it was found that the AST/ALT ratio was a good predictor of fibrosis, where it was estimated that a ratio of over 1.0 would most likely result in some form of fibrosis. From Sorbi, Boynton, and Lindor (1999), "Subset analysis of patients with NASH revealed mean AST to ALT ratios of 0.7, 0.9, and 1.4 for subjects with no fibrosis, mild fibrosis, or cirrhosis, respectively."

The NASH-CRN (Kleiner) fibrosis stage score was generated next with the use of the ratio. The full dataset was split into the different ratios and then the categorical function was used. Table 1 shows the probability of the fibrosis stage assigned based on the ratio.

Ratio Range	Kleiner Fibrosis Stage	Probability
Under 0.8	0	100%
0.8-1	0, 1a	40%, 60%
1-1.2	1a, 1b, 1c	33.3%, 33.3%, 33.3%
1.2-1.4	1c, 2, 3	20%, 40%, 40%
1.4-1.5	3,4	20%, 80%

Table 1 : Scores and Probabilities assigned on given ratio

5.3.8 Steatosis Brunt Score

From the pathology specification, it seems that this score was derived from some imaging metric from a liver biopsy scan. It was decided to use the NASH diagnosis to determine the Steatosis Brunt score. For a patient with no NASH, they would be assigned a Steatosis Brunt Score of either 0 or 1, 60% probability of having a score of 0 and 40% of score 1. For patients with NASH, they had a probability of 10% of score 1, 50% of score 2 and 40% of score 3.

5.3.9 Ballooning Score

Brunt (2016) stated that for simple steatosis, there are no ballooning, but ballooning must be present for steatohepatitis. The Hepatocyte ballooning score has a score between 0 to 2, and a score of 0 correspond to no ballooning. Therefore, patients with no diagnosis of NASH has a ballooning score of 0. For those with NASH, they were assigned the probabilities shown in Table 2 based on logical assumptions.

NASH Kleiner Fibrosis Score	Hepatocyte Ballooning Score
1a	Score 0 (30% probability), Score 1 (70%)
1b	Score 0 (60%), Score 1 (40%)
1c	Score 1 (100%)
2	Score 1 (80%), Score 2 (20%)
3	Score 1 (60%), Score 2 (40%)
4	Score 1 (20%), Score 2 (80%)

Table 2 : Assignment of Hepatocyte Ballooning Score based on Kleiner Fibrosis Score

5.3.10 Lobular Inflammation Score and NAFLD Activity Score (NAS)

Brunt (2016) stated there might be some inflammation for simple steatosis but this score was derived from clinical data. Therefore, assignment of these scores were based off logical assumptions similar to Hepatocyte ballooning score. The inflammation score has a score between 0 to 3.

The NAFLD Activity Score (NAS) is the sum of the Kleiner steatosis score, hepatocyte ballooning and lobular inflammation scores. NAS has a range of values between 0 to 8.

5.3.11 Chronic Kidney Disease (CKD)

From the feedback meeting with Prof. Fallowfield and Dr Kendall, it was noted that NALFD was correlated with Chronic Kidney Disease. Marcuccilli and Chonchol (2016) stated that there was strong evidence linking the NAFLD to development and progression of chronic kidney diseases. Yasui (2011) found that patient with NASH had a higher prevalence of CKD compared to patients not diagnosed with NASH. The data set was split into NASH and no NASH patients and NASH patients were given a 21% probability of having CKD while no NASH patients only had a 6% probability, probabilities taken from Yasui (2011).

5.3.12 Hypertension

This was one of the comorbidities that both Prof. Fallowfield and Dr Kendall wanted to add into the dataset in addition to diabetes and chronic kidney disease. There were several studies that suggested that NAFLD increases the risk of hypertension and vice versa (Lonardo et al. 2018). Ryoo et al. (2014) found that the incident of hypertension increases with the progression stage of NAFLD with an incident rate of 14.4% for normal stage of NAFLD, 21.8% for mild and 30.1% for moderate to severe NAFLD, these probabilities were used to determine the probability of hypertension. The NAS was used to categorise the NAFLD stage, where a NAS score of 0-2 for normal, 3-5 for mild and 6-8 for moderate to severe NAFLD.

5.3.13 NAFLD and Comorbidities Diagnosis Date

The NAFLD diagnosis date was between 1 January 2009 to 1 January 2018 because the minimal data set specification stated that most patients will be post 2009. While the diagnosis date of patients with the comorbidities was all generated randomly in an 18-year period between 1 January 2000 to 1 January 2018, this was because the patient could be diagnosed with those health conditions before the diagnosis of NAFLD.

5.3.14 Cholesterol

The LDL (Low Density Lipoprotein) is known as the bad cholesterol, where the lower the value, the better the result. HDL (High Density Lipoprotein) is known as the good cholesterol and a higher value is better. The Friedewald formula (HeartUK 2015) was used to calculate the total cholesterol given by Equation 5.3.14-1.

$$\mathrm{LDL} = \mathrm{Total\ Cholesterol} + \mathrm{HDL} + \frac{\mathrm{Triglycerides}}{2.19}$$

Equation 5.3.14-1 : Friedewald Formula

The first step was to generate data for the HDL and LDL values and then calculate the total cholesterol value. Since BMI seemed to be a strong indicator of a variety of diseases, this was used to split the dataset. Rather than only splitting this into two group, it was decided to split it up into three groups with the BMI ranges of 20-30, 31-40, 41-50. Shamai et al. (2011) found that HDL was inversely correlated with BMI but could not find any correlation between LDL and BMI. While Shirasawa et al. (2013) found that LDL had a positive correlation with BMI in a study involving Japanese schoolchildren, assumed LDL is correlated with BMI for this synthetic dataset.

Schwartz (2017) found that age was another big factor in cholesterol levels, with older people are at a higher risk of high cholesterol levels. Gender was also another big factor with women at menopause stage having a much greater chance of high cholesterol. Therefore, the datasets were split into male and female subjects, then further split by age, it was 45 for male and 55 for female.

For HDL, the subgroup of male under 45 with low BMI would have higher HDL readings than the other male subgroups. It was the opposite for LDL, where male over 45 with high BMI would have higher LDL reading that other male subgroup. Similarly, the same for female for HDL and LDL.

5.3.15 Platelet Counts

Yoneda et al. (2011) found the platelet counts of patients from various stages of fibrosis as shown in Figure 2 and this was used to generate platelet counts for patients in the different stages of fibrosis with the Kleiner Fibrosis stage data column.

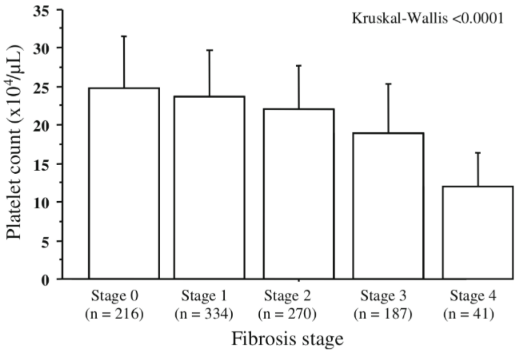


Figure 2 : Boxplots of Platelet Counts at various stage of Fibrosis (taken from Yoneda et al. (2011))

5.3.16 Glucose Count (Fasted)

Diabetes was a big factor for the glucose count, the population was split into diabetic and non-diabetic populations. The fasted glucose count for normal patients was below 100mg/dl, prediabetics has a range between 100-125mg/dl and diabetics has a value over 126mg/dl ('Blood Sugar Level Ranges' n.d.). It was decided to give non-diabetics a range between 50-135mg/dl and 125-250mg/dl for diabetic patients. There seemed to be no reason for which the glucose count should be higher than the normal ranges for non-diabetics unless patients did not fast before the blood test.

5.3.17 Collagen Proportionate Area

From Masugi et al. (2018), the following scatter plot and boxplot shown in Figure 3 was obtained and used to generate collagen area for the patients. From the boxplots, the estimates of important values were taken such as the maximum and minimum values to replicate this kind of data. The values used to generate the collagen data for patients can be found in Appendix 8.6.

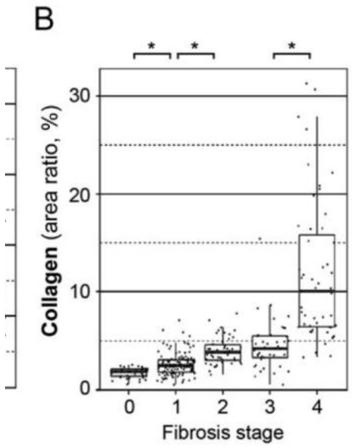


Figure 3 : Boxplot of Collagen for different Fibrosis Stage (taken from (Masugi et al. 2018))

5.3.18 Creatinine

Creatinine is the waste product produce by muscles activity ('Understanding Your Lab Values' 2017) The kidney would usually remove this waste product but if there was damage to the kidney then it would lead to an increase in the creatinine levels.

From ('Creatinine Levels Chart'), there was a difference of creatinine levels between male and females, with males with slightly higher level. Age also affected the creatinine level. The first step was to split the dataset into patients with and without CKD and then further split the tables into gender and age of 55 years old. The normal values for adult male was 0.6 to 1.2 mg/dl and 0.5 to 1.1 mg/dl for female. Since older people may have lower creatinine levels, the following values were assigned as shown in Table 3.

Demographic	**Creatinine range of values**
Male with no CKD under 55	0.75 - 1.2 mg/dl
Male with no CKD over 55	0.45 - 1 mg/dl
Female with no CKD under 55	0.6 - 1.1 mg/dl
Female with no CKD over 55	0.35 – 0.9 mg/dl

Table 3 : Assignment of Creatinine range of values for subpopulations with no Chronic Kidney Disease

Table 3 shows the range of values assigned for those without CKD. For those with CKD, most of the creatinine values were assigned to be in the range 1-6 mg/dl.

5.3.19 eGFR

The estimated glomerular filtration rate can be calculated from the equation from (Levey et al. 2009). The eGFR was calculated using the creatinine value generated before.

5.3.20 Medication

For some patient, they may have more than one comorbidity and they may also take more than one medication. In the first draft of the synthetic dataset, the assumption that patients only took one medication was made but from the first feedback session, it was quite common for patients to take more than one medication for their comorbidities.

There were three comorbidities in this dataset, diabetes, chronic kidney diseases and hypertension. Rather than putting the medication information into the main dataset, it was split into another dataset because not all patients would take the same number of medications and so there would be a lot of null values on the dataset. The studyID of the patient was used as a key to connect both these datasets. It was very difficult to include all the medications and getting all the combination of medications a patient could take.

5.3.20.1 Diabetes

The average number of type 2 diabetes medication was found to be 4.1 ± 1.9 from a study by Grant et al. (2003). There are over 50 different diabetes medications which was a very long list. It was decided to list the common classes of the medication rather than listing the individual drugs.

The most common diabetes classes list was taken from ('Diabetes Drugs' n.d.) and the following probabilities were based on logical assumptions. Metformin is the most common diabetes medication, and so those with a BMI over 35 had an 80% chance of being on metformin. Those under BMI of 35 had a 65% chance of being on metformin. Then depending on the severity of diabetes, other treatment might be prescribed for the patient.

('Type 2 diabetes' n.d.) gives a common course of treatment for diabetic patients. For those not on metformin, there will be some who might not be on any medication if they are able to keep their disease in check with diet and exercise. For those on Metformin, patients might be also on Sulphonylureas with a 40% probability. Sulphonylureas could also be used as an alternative to metformin treatment, for those not on metformin, they had a 15% chance on being on Sulphonylureas on.

Thiazolidinedione was another common medication for diabetes, the combination of treatments is Thiazolidinedione, Thiazolidinedione + Metformin or Thiazolidinedione + Sulphonylureas.

For DPP 4 inhibitors, the treatment course could be DPP 4, DPP 4 + Metformin, DPP4 + Sulphonylureas, DPP 4 + Metformin + Thiazolidinedione. Finally, for insulin, it had the only treatment course of Metformin and Insulin for this dataset.

5.3.20.2 Chronic Kidney disease

There are no medication to cure chronic kidney diseases, only treatments to help with the associated symptoms (NHS 2016a). There were five main medications to help with symptoms such as high blood pressure, high cholesterol, water retention, anaemia, bone problems (MayoClinicStaff 2018). Those with hypertension and chronic kidney disease would be on high blood pressure medications. Those with high cholesterol (over 240 mg/dL) would be on cholesterol medication. For the rest of the medications, water retention, anaemia and bone problem, there was a 25% probability of the patient being assigned that medication.

5.3.20.3 Hypertension

It seemed that for people under 55, they would either be treated with ACE inhibitors or angioteinsin-2 receptor blockers (NHS (2016b). While for 55 or over, they would usually be given calcium channel blockers.

The dataset was first split into those with and without hypertension and then further split into two groups, those older than 55 and under 55. For those under 55, there was a 50% chance of them being put on ACE inhibitors and those not on ACE inhibitors are on Angiotensin-2 receptor blockers. For over 55, assumed all were on Calcium channel blockers (NHS (2016b).

Since beta blockers are now less commonly used as they are considered less effective than other treatments, beta blockers was discarded as a hypertension medication for this dataset (NHS 2016b). For diuretics, since this medication is used to release more water from the body, if a patient is taking chronic kidney disease medication for water retention then they might also be taking diuretics medication, assumed this was true.

5.4 App

The SteatoSITE app that I developed contained six main tabs: homepage, visualisation, analysis, gene expression, view dataset and completeness dataset. Homepage has a short summary of the dataset such as a five-value summary of the variables in the dataset which includes the minimum, maximum, median, first quantile, third quantile, mean and standard deviation. In the following sections, each tab was looked at and the plots discussed. A link to the app can be found here: https://simonfong.shinyapps.io/steatosite_app/

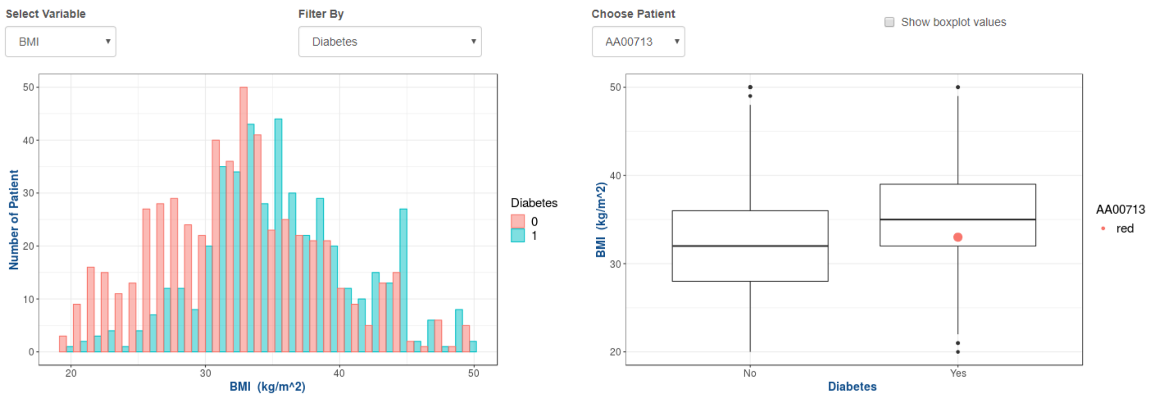
5.4.1 Visualisation Tab

This tab was sub divided into five sub tabs, demographics, blood test result, comorbidities, NAFLD scores and timeline.

5.4.1.1 Demographics

The visualisation of the demographic data was important as it allowed the user to check the demographics of the dataset to gain insights such as whether the dataset is a fair representation of the population or if they achieved a certain demographic they were targeting. The dataset only contained four demographic variables, which were age, BMI, gender and alcohol intake. Most of these were numeric data and so the best representation of this data was a histogram and boxplot to show the statistical information at a quick glance.

The app also allows the user to filter the data by a number of criteria such as gender, diabetes, chronic kidney diseases and the different NALFD related scores. This gives the user the ability the apply some filter and check whether there is any correlation with the sub-population, for example, if the user filters by diabetes, then they find that there are a greater number of patients with diabetes as the BMI value increases. A patient select was included to check the patient’s value on the boxplot, this gives the user an idea whether the patient was within the ‘normal’ ranges of the population or if it was an outlier. Figure 4 shows the demographic visualisation page with BMI selected and filtered by diabetes.


Select Variable: BMI

Filter By: Diabetes

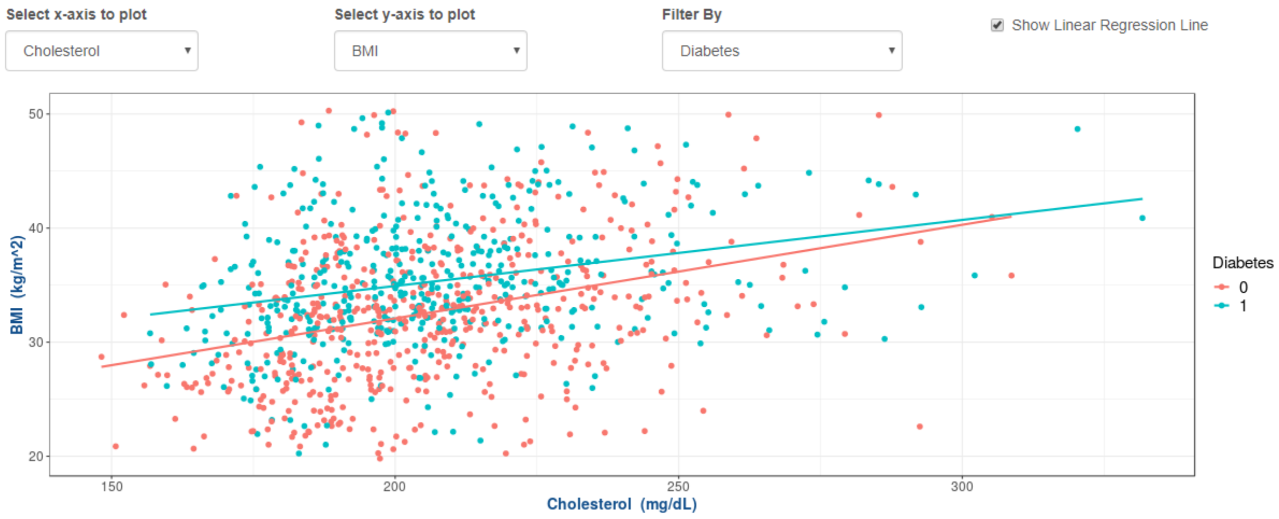
Choose Patient: AA00713

□ Show boxplot values

Figure 4 : Screenshot of Demographics visualisation tab

5.4.1.2 Blood Test Results

This tab contained three plots, scatterplot, histogram and boxplot. The histogram and boxplot were similar to the demographics tab. In the dataset, there were a number of blood test measurements created such as the cholesterol levels, AST and Glucose levels, which were all continuous variables. The user could pick variables for the x-axis and y-axis for the scatterplot. The choices for the x and y axis are the blood test results columns but also included age, BMI and alcohol intake. Again, for the scatter plot, user could filter by some variable to see if there is a difference between groups. Another function was a simple linear regression line which was a toggle option for the user to let them decide if they want to see the regression line. If user applied a filter, then there would be multiple lines depending on the number of categorical values of the filter by variable. For example, Figure 5 had two linear regression lines for patients with and without diabetes.



Select x-axis to plot: Cholesterol ▾

Select y-axis to plot: BMI ▾

Filter By: Diabetes ▾

☑ Show Linear Regression Line

Figure 5 : Scatterplot example from blood test result visualisation tab

5.4.1.3 Comorbidities

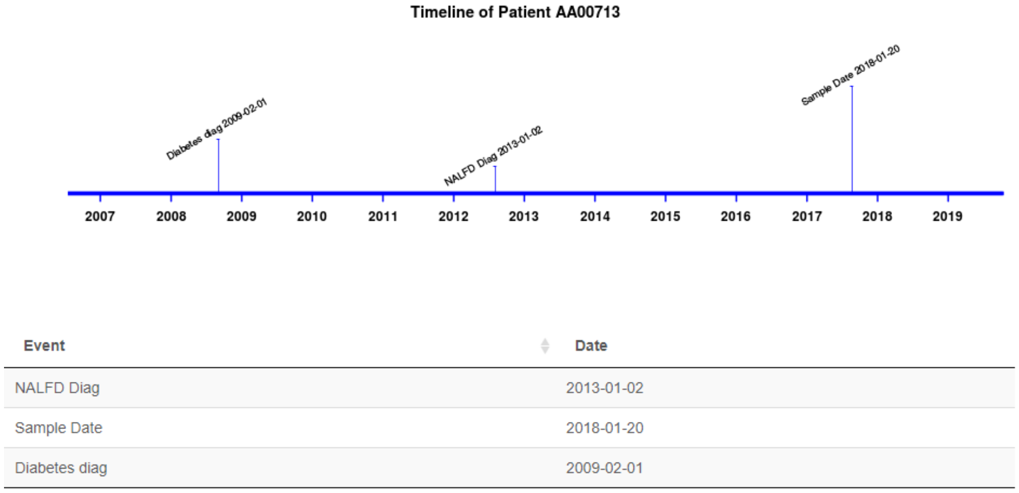
Another visualisation which was important was the comorbidities information from the dataset due the potential relationship between these comorbidities with NAFLD. This tab contained a two bar plots, one shows how many patients have the comorbidities and the other shows the medication taken.

5.4.1.4 NAFLD Scores

This tab shows a simple bar plot that counts the number of patients of the various NAFLD related scores such as the NAFLD activity score and the ballooning score.

5.4.1.5 Timeline

This tab allows the user to select a patient and then it would plot the timeline of that patient. The timeline includes date of NAFLD diagnosis, sample taken date, and the diagnosis date of any comorbidities. This could help the clinician understand the events in the patient's life and determine if the events could contribute to the progression of the NASH disease. Figure 6 shows the timeline of a patient with a table detailing the events. In the real dataset, there will be liver related and cardiovascular related events and these would be on the timeline as well but as mentioned previously, these events were not added into the synthetic dataset.



Event	Date
NALFD Diag	2013-01-02
Sample Date	2018-01-20
Diabetes diag	2009-02-01

Figure 6 : Timeline plot example

5.4.2 Analysis

The analysis tab contains tools which can help users with analysis of the data and gain insights. The sub tabs are Correlations, Decision Tree and Subset Summary tab.

5.4.2.1 Correlations

The correlation for continuous and categorical data could be found in this tab. For continuous variables such as demographic information and blood test results, the Pearson's R coefficient was used to find the correlation between the continuous variables. The user could also find the individual Pearson value by selecting the variables from the drop-down menu, as the value might not appear in the heatmap. Then a correlation plot was plotted to visualise the results, the user can choose either heatmap or a number plot, shown in Figure 7.

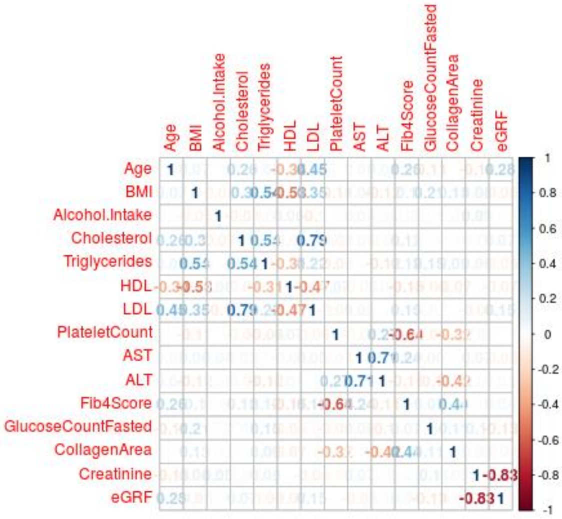


Figure 7 : Screenshot of the Correlation heat map

For categorical data, the Pearson's R test was not applicable and upon further research found that the Chi square test of independence (Nayak and Hazra 2011) was found to be suitable for this problem. Rather than showing the user a big table of p-values, the user can pick the two variables they want to investigate to see if the variables are dependent or not. Once values are selected, it would give the Chi Square P-value as output and then if the P-value was under 0.05, it would state that the selected variables are dependent on each other.

Lastly, there was an option for users to find out if there was a difference in the subgroups for the continuous variables. The nonparametric Kruskal-Willis test (McDonald 2014) was used to find the p-value, if the p-value was under 0.05 then the subgroups are not equal and vice versa.

5.4.2.2 Decision Tree

Decision trees have been used in the past for decision support systems which was something SMS was interested in as well. The decision tree was picked as a classification method as the result can be visualised easily.

The variables used for the decision tree were Cholesterol, triglycerides, HDL, LDL, Platelet Count, AST, ALT, Fib4Score, Glucose Count, BMI, Creatinine, and eGRF. The decision tree classified whether a patient was likely to be diagnosed with NASH or not based on those variables. Since decision trees are all randomised if the seed was not set on R, the user could set the seed to generate different decision trees. The user had to choose what variables to use for the decision tree. An example of the decision tree was shown in Figure 8. The decision tree tab also had a guide on how to read the decision tree.

The dataset was first split into training and testing set and then a decision tree was trained on the training data. From the example in Figure 8, the seed was set to 79 and the testing accuracy was 85.2%. This decision tree could be useful for a clinician reading blood test results, following the decision tree from top to bottom, the clinician could assess the likelihood of the patient having NASH. This type of test should give some indication whether further tests are required or if the patient is at risk.

The decision tree has an accuracy score output which is computed using the test data. The accuracy score was equal to the sum of the diagonals divided by the sum of the confusion matrix. The diagonals of the confusion matrix are the true positives and true negatives, and these are the correctly predicted values.

The decision tree was well received in the first feedback meeting and it was something they think might be very useful to have. Since the decision tree was computed using synthetic data and it gave a surprisingly high accuracy, the accuracy might not be as high with real data.

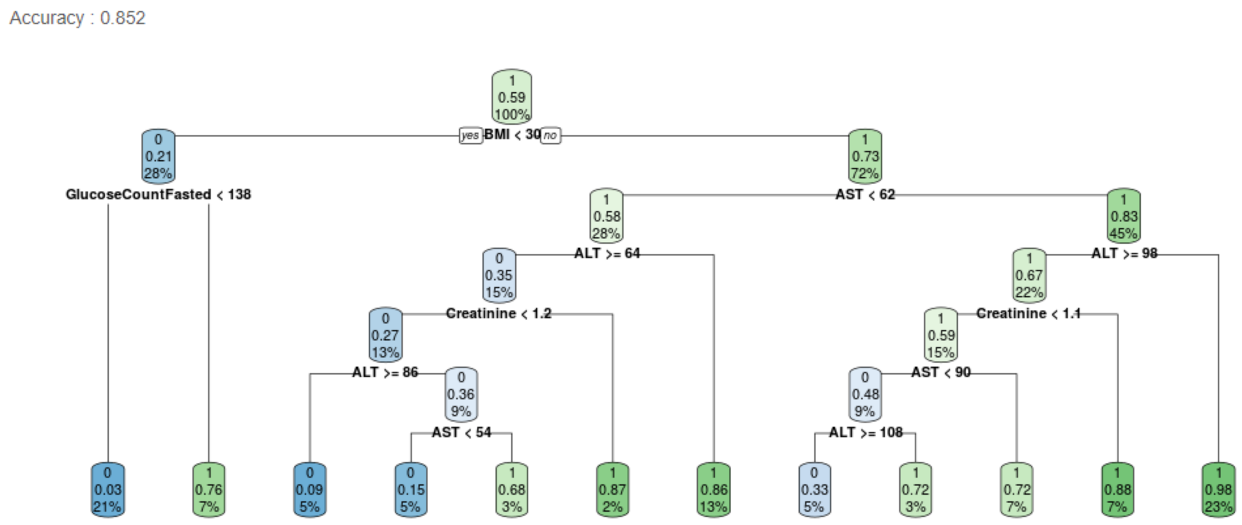


Figure 8 : Decision Tree example for NASH diagnosis

5.4.2.3 Subset Summary

This tab contains a summary table similar to the summary table on the homepage, it has the same columns such as median and mean values. However, user could pick the subgroup and then the categorical value. For example, the summary statistics can be found for diabetes, gender or the various NAFLD related scores. This would allow the comparison of different groups, it can be used with the Kruskal-Willis test from the correlation tab.

5.4.3 Gene Expression

The SteatoSITE project will have genomic sequencing data in the future from the Biobank and so it would be good to have some visualisation on the app with genomic data. However, creating synthetic genomic data was deemed too difficult and so another suggestion was to use a publicly available dataset online from the GEO (Gene Expression Omnibus) which is a database of previous genome sequence data from past research projects. Selecting the right data was difficult as the files were in IDAT format and there were difficulties trying to load these files into R and interpret the results. The datasets from GEO could be processed or unprocessed raw data dependent on uploader. Therefore, even after loading the IDAT files, the data might still need to be normalised. It was decided to use a processed dataset. The dataset chosen to use for this project was (GSE61260) which contained liver sample from 134 subjects with BMI between 14.8-70.2 and between 10-85 years old. This data was chosen as it was liver related sample and it contained a CSV file of normalised gene expression values. The CSV file was around 13mb with around 20,000 entries of gene expressions, rather than using the whole dataset, a subset was taken with selected gene expression for visualisation purposes. A list of gene was chosen from (Wood, Miller, and Dillon 2015), which identified genes associated with non-alcoholic fatty liver disease from various studies. The full list of gene expressions chosen can be found in Appendix 8.7.

5.4.3.1 Heatmap

This tab also includes a list of selected gene expressions for the user and allowed the user to select the number of samples to use. Figure 9 shows the gene expression heat map with all the samples, the heat map was created using the 'pheatmap' function in R (Kolde 2018).

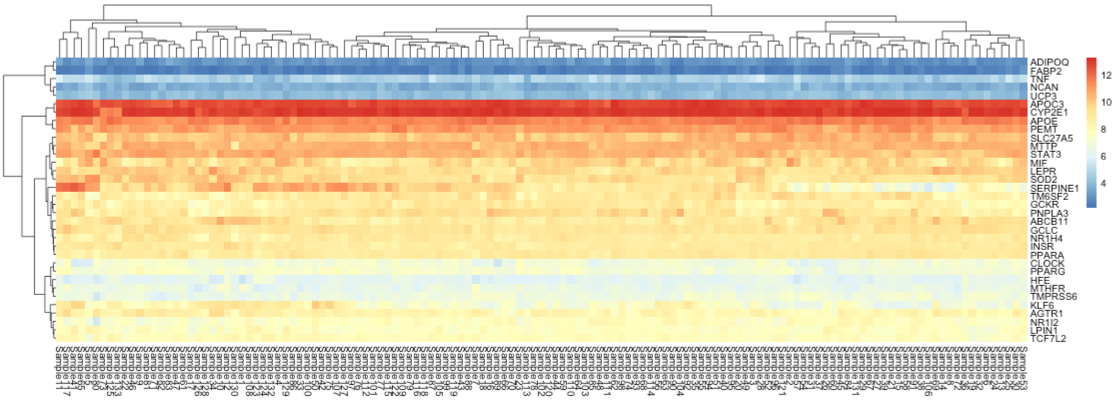


Figure 9 : Gene Expression Heatmap

The columns are the different samples and rows are the gene expression. The heatmap could show the user if there are certain patterns within the dataset can be used to group the patients. From speaking to Dr Marc Jones, lab manager at SMS, heatmaps are useful when the sample contains patient taking a certain medication compared to those not taking medication as it can be shown clearly on the heatmap. From Figure 9, the SERPINE1 gene seems to be different as there a wider range of values compared to the other genes where values are within the same range.

There are dendrogram on the top and left-hand side of the heat map, which can group together similar gene expressions and patients. Since this dataset was not connected to the synthetic dataset, the samples could not be filtered by other variables such as diabetes and so on.

5.4.3.2 PCA and K-means Clustering

PCA could be used for visualisation for high dimensional data. For this visual, K-means clustering was used to clusters the points together into a number of clusters which the user can choose. To plot the points, the distance matrix between the points was computed and then classical multidimensional scaling was applied. Figure 10 shows the cluster plot with four clusters and these were distinct clusters. This meant that there are four distinct groups of patients.

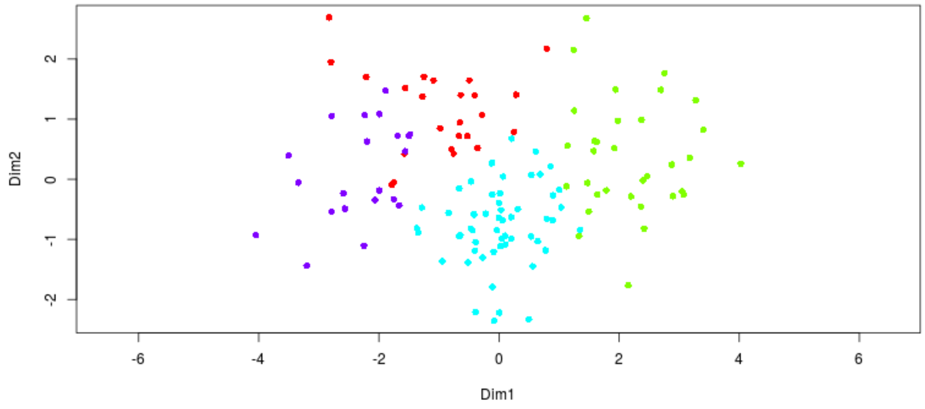


Figure 10 : PCA and K-means Cluster plot

5.4.4 View Dataset and Completeness Dataset

The view dataset tab allows the user to view the data set, there are a couple of options in which the user can choose to either view the full dataset or view a subset of the dataset. Subsets include filter by gender, comorbidities, and the different NALFD scores.

The completeness dataset tab was added into the app from interviews at the start of the project to understand what features are important to users. It was found that being able to select columns to view and focus on patients with some certain criteria was good for clinicians. For example, if diabetes, chronic kidney disease and hypertension was selected then only patients with all those health conditions are shown. The user would also select the columns to display rather than viewing the full table.

5.5 Conclusion and Future Work

A synthetic dataset was created for this SteatoSITE project which required extensive research into the relationships between the data columns of the dataset to ensure realistic data. Then a R-Shiny app was created to visualise the synthetic data and to showcase the potential of mini-apps in the platform for existing and potential SMS clients. The synthetic dataset and app were improved through a series of feedback sessions.

The app included a number of plots including interactive plots to present the findings to the users in an effective way. Users could see at a glance information about the dataset and filter by some criteria to compare different groups. A decision tree implemented in the app used the blood test results to predict whether a patient has NASH or not. Since this did not use real data, it was hard to find out how accurate the decision tree would be with real data.

At the time of this project, no data has been obtained for the SteatoSITE project and hence the need for the creation of synthetic data. The synthetic dataset was very clean which is most likely not the case for real data. It would be interesting how well the code translates to the real data, whether only a few minor changes are required. A comparison of the plots of the real and synthetic data would be interesting to see how close the synthetic data was to the real data.

In the pipeline of the SteatoSITE project was sequencing data, and if these data had been available at the time of this project then more machine learning techniques would have been explored towards precision medicine. If multiple blood test results were available and the progression of the patient through the various stages of NAFLD was noted, then it might be possible to predict the rate of progression of NAFLD with regression. This would be useful as an early prevention tool to help give better treatments and advice to patients.

6 FutureMS

6.1 Introduction

Multiple Sclerosis (MS) is highly prevalent in Scotland. The FutureMS project focuses on people recently diagnosed with relapsing onset multiple sclerosis in Scotland ('About FutureMS' n.d.). The FutureMS study plans to collect data from Glasgow, Edinburgh, Dundee, Aberdeen and Inverness. They have been collecting data for almost two years and are still in the process of data collection. So far, they have more than 300 participant entries for the baseline visits.

The project includes a multitude of clinical data collection including questionnaires, brain MRI imaging scans and genomic sequencing data (transcriptomics and genotyping). The main aim of the FutureMS project is to use all this data to create a tool that can predict how severe the patients MS will become so that the patient can be put on the correct treatment plan for their personal disease.

The project involves two sets of clinical data collection whereby each patient comes in for a baseline visit to complete a series of tests, including MRI brain scans and questionnaires. Then after one year, the patients attend a follow-up visit, during which the same tests and questionnaires are completed. On both visits, a series of scores is calculated such as the impact of MS on their lifestyle and mood. Brain MRI imaging scans are also taken on the baseline and follow-up visits. The MRI scans are being analysed by a team from University of Edinburgh to look at the lesions in the brain and how they change over the one-year study period. The group are building a series of algorithms to computational quantify brain and lesion volumes, these end points can be combined with the other datasets to determine the MS disease course in patients. The genetic sequencing has yet to be started.

The main objective of this app was the visualisation of the data as they have been collecting the data for almost two year, but they have very little idea what the data, and therefore cohort, looks like and there is no dedicated app working directly from the FutureMS database for visualisation.

6.2 Methodology

The methodology of this project included conducting interviews with the FutureMS team and taking ideas from the SteatoSITE app. The methodology follows a similar methodology with the SteatoSITE project without the synthetic dataset. A first draft of the app was created and then the app was shown to various team members for feedback to improve the app and add any requests.

The first feedback meeting with clinical project manager, Mrs Shuna Colville and clinical nurse, Dr Liz Elliot provided essential feedback including clarifying any questions such as the best metric to determine if a patient MS condition has worsened or improved. It was found that the Expanded Disability Status Scale (EDSS), an internationalised standardised system for measuring disability due to MS, was most likely the best metric to determine the clinical impact of MS. There were over 200 tables in the workspace, it was discussed what other tables might be useful to use with the minimally complete dataset, three more tables were identified as applicable which were medication, blood test results and the relapse table. Dr Elliot also suggested that a timeline plot of patients could be useful as clinicians could see from a glance whether a medication was having any effects on the number of relapses or the time in between relapses. These were taken into account and added into the app.

The second feedback on the app was from Dr Dakin on behalf of Dr Peter Connick following their discussion. Overall good feedback was given on the app, Dr Connick noted the timeline plot would be of huge benefit in a clinical setting. There was a request to add the different clinical scores on to the blood test result scatter plot to see if vitamin D had any effect on the clinical scores between visits as has previously, though inconsistently, been reported.

6.3 Data Preparation

The FutureMS data was stored in the FutureMS workspace in the SMS-IC innovation platform and permission was granted to use the data. From speaking to Dr Dakin, it was found the 'minimally_complete_dataset' was best to use, which contained data from the various datasets into one 'master' dataset and was updated monthly. The minimally complete dataset contained a number of different questionnaire scores and clinical measurements including EDSS.

The minimally complete dataset had 468 rows of patient entries and 164 columns. Out of the 468 entries, there were 343 baseline entries and 125 follow-up entries. The last column of the minimally complete dataset contained a column called 'n_missing' which was the number of missing column entries for a patient. There were a few entries with around 150 missing values, upon further discussion with the project manager, it was found that these entries might be due to patients missing appointments or participants not completing the questionnaire.

The minimally complete dataset was explored, and each column closely examined. Columns which offered no useful information were identified and removed from the dataset. Seven columns were identified, and the reasons are given in Appendix 8.8.

There was also a problem where a patient might have more than one entry for the baseline visit, this might be due to the patient having missing information on the first visit and then additional information was added in at a later date, hence the second entry. It was found that only two instances where a patient had more than one baseline visit. For both entries, the visit type was stated as baseline visit, but the appointment type was 'planned review'. It seemed these were mistakes where the user input the visit type for the planned review as 'new' instead of 'return', these entries were corrected accordingly.

Within the data set there were more than one measurement taken for some variables and a mean of the multiple measurements calculated. Only the mean measurement was used for those variables. For example, for the timed walk there was two measurements and a mean of those two values, only the mean value column was used for visualisation.

Patients with over 100 missing values were removed from the table and this reduced the number of entries down to 442 with 323 baseline and 119 follow-up entries. The 'colSums' function was used with the 'is.na' function on the remaining 442 entries, there were 27 missing entries for EDSS column. Since EDSS was an important metric to determine if a patient's MS worsened, it was decided to remove these entries. This reduced the total entries down to 415 with 303 baseline entries and 112 follow-up entries. It was decided with Dr Dakin to only focus on patients with baseline and follow-up visits to visualise the difference in the scores between visits due to time constraints. There were 110 patients with baseline and follow-up visits and these were extracted into a new table. This indicated there were 2 follow-up entries that no longer had a baseline entry, due to null value for EDSS. The project team are working to add these missing values which will be recorded in source documentation.

6.3.1 Blood test results

The blood test data was only taken at the baseline visit and so there are a lot more entries (n=232) compared to patients with both baseline and follow-up visits. The entries of patients with baseline and follow-up visits were extracted from the full blood test dataset and there were 106 patients with baseline and follow-up visits with blood results available. There were 4 patients missing blood test results data, this is usually due to issues in the clinic obtaining blood or in the laboratories processing the tests.

The column sums function was used again with the ‘is.na’ function to look at the number of null values for all the columns of the dataset, it was found that there were 40 null values for the ‘TCO2’ column and 30 for Mean cell HB concentrate column. This is expected as these variables are only measured in certain health boards. There were other variables with fewer than 5 missing values, as these were just for visualisation purposes, and the ggplot function would ignore null values when plotted, these were kept in the dataset. From discussions it seemed that vitamin D is of the greatest interest from the blood test results due to previous studies investigating potential links to MS.

6.3.2 Medication Table

The medication data table contained entries for up to 16 medications per each patient, although most patients do not take all 16-different medications. The table gives information on the medication, date, dosage, units and frequency of medication. The date entries were confusing as its not known whether it was the date from which the patient started taking the medication or the date of which this information was added into the table on the platform. It was found that the dates of the medication table correspond to the date of entry into the medication table rather than the date the patient started on the medication or treatment.

From the feedback meeting, Mrs Colville and Dr Elliott only wanted data on disease modifying therapies (DMTs) which are prescribed to directly treat MS. A list was given by Dr Elliott that contained the current DMTs, this was used to extract entries from the medication table.

There were 279 rows for the medications patient pivot data table, since only the patient with baseline and follow-up visits were considered for this app, those entries were extracted. The next step was to extract only the entries related to DMTS. The list of MS treatment and medication given by Dr Elliott can be found in Appendix 8.9.

To extract the information, one must go through all 16 medication names columns and then extract out the MS related medication. The column number of the medication name was first identified for all 16 columns. Only the medication name columns were first extracted, and it was found that for columns 12 to 16, it was all null values, therefore these were omitted from the extraction. Since the date, dosage, frequency and units might be of interest, these data columns were added into the new dataset. At this stage the new dataset had 110 rows and 56 columns (patient ID and 11*5 (name, date, dosage, frequency, unit)).

The row sums function with is.na was applied to count the number of null values for the entries, it was found that a high number of entries with 54 null values, which meant these patients did not take any MS related treatment or medication. These entries were removed which reduced the number of rows from 110 to 56 entries. The number of null values was further looked at for these entries and it was found that the highest number of null values of an entry was 52 and minimum was 47. This suggested most patients were on 1 to 2 MS related medications. Therefore, further extraction was required to remove these null values.

A new table was created again, for each entry, all 11 of the medication name columns was checked. If it contained a non-null value, then it would add that medication details in the new dataset. If it is a null value, then it would skip to the next medication name column. The final medication table contained 56 rows and 11 columns.

6.3.3 Relapse Table

The relapse table include information about patient's relapse such as if they were hospitalised, date of relapse and whether steroids was prescribed. These were considered important information for the project team as a timeline of the patient's relapse could be obtained from this information and determine if certain treatment had any effect on the time between relapses. The relapse table contained the relapse history of the patient not only relapses between baseline and follow-up visit. The entries for patients with baseline and follow-up visits was extracted and a new column was added into the data frame to count the number of relapses for each patient.

6.3.4 Dates

The dates from the minimally complete, medication and relapse tables were extracted for the timeline plot. For the medication table, the date was in the format '2018-03-19T12:01:29+00:00', but only the date was required, so these entries were reformatted to remove the time of day. A simple solution was to the use of the substring function in R to only select the first 10 characters of the date string.

The date column for the relapse table was quite unclean since not all entries included the full date, some entries only had the year and month, and some only had year. This could be due to the patient only giving a rough estimate of the date of relapse. There were 7 incomplete entries and ’01’ was entered in for month and day for completeness.

6.4 App and Result

The FutureMS app has a homepage which gives details about the dataset such as the number of entries in the dataset, the last date the dataset was updated. The link to the FutureMS app was not given due to security reasons and only members of the FutureMS workspace can view the app in line with regulatory approvals. Permission was granted to take screenshots of app for this report with sensitive information such as patient identifier and dates retracted. There are seven main tabs which were Visualisation, Blood test, Medication, Relapse, Patient Viewer, Summary and View Dataset.

6.4.1 Visualisation

The visualisation tab of the app allowed the user to visualise the wide range of information on the minimally complete dataset table. The user could see the overall study sample scores and clinical results but also compare a patient’s information against the population, in a similar way to the SteatoSITE app. This would give the user information such as if the selected patient was an outlier compared to the population. For most of these visuals, the data was split into baseline and follow-up visits to see the difference in variables between visits, for example Figure 11.

The visualisation would also allow users to identify any irregularities in the dataset. For example, there seems to be an inconsistency for the height measurement in some records. The height readings from baseline and follow-up patients are shown in Figure 11.

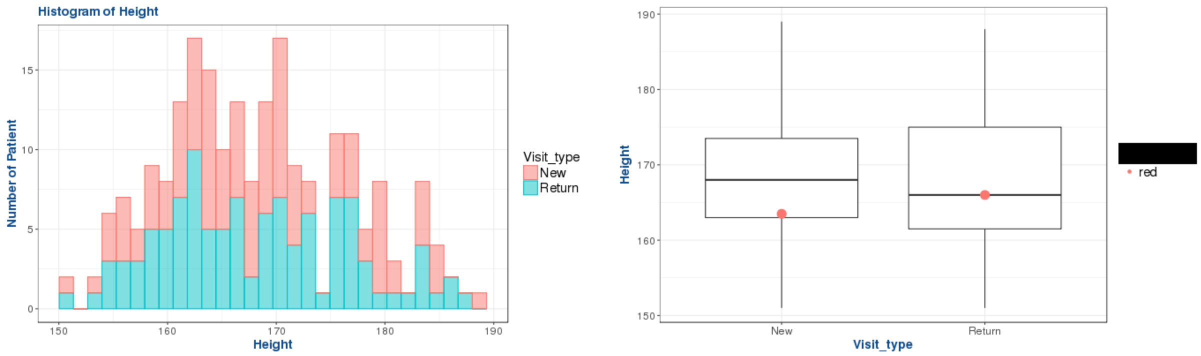


Figure 11 : Histogram and Boxplot for Height

With the histogram, one would expect the baseline and follow-up height measurements to be equal but for some of the patients, there was a big change in height value between the visits as shown on the boxplot for one of the patients. Therefore, there seems to be some irregularities in the measurement of height.

The EDSS was used to determine if the relapsing MS worsened for the patient between baseline and follow-up visit. An increase in the EDSS score indicated that the impact of the MS on the patient was greater than at baseline visit. While a decrease in EDSS indicates MS had less of an impact on the patient compared to baseline visit. Around 40 of the patients had no change in the EDSS score, there was no change in the impact of MS between visits indicative of stable disease. Appendix 8.10 shows the histogram for the change in EDSS score which was calculated by taking the follow-up score minus the baseline score.

The dataset contained a series of questionnaire questions concerning the impact of MS on the patients including lifestyle impact and anxiety depression questions. These scores usually ranged between 0 to 5 or yes/no answer. A lower score on the 0-5 questions is better as it means there is less impact from MS. For those categories with scores, the app user had the option to view the population mean or median scores. Figure 12 shows the means scores of the baseline and follow-up visit of the population for the anxiety depression questionnaire.

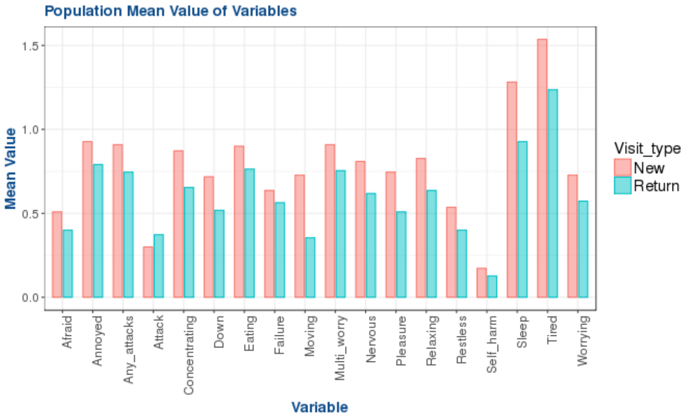


Figure 12 : Population mean scores for Anxiety Depression questionnaire

For all the different criteria, there was a decrease in the score, meaning that they felt better than the baseline visit. One of the possible reasons for was that they might have been diagnosed quite recently before the baseline visit and were feeling down from the diagnosis. Then for the follow-up visit, they had more time to come to terms with the diagnosis and felt better.

There was an increase in the alcohol units of the patient for the lifestyle questionnaire. It might be interesting to investigate whether for those patients with increase in alcohol unit was due to their MS condition and if there was a large increase then possible help could be offered to those patients to help them better manage their MS condition. Recording alcohol intake is well known to be inaccurate as patients don't understand what a unit is or want to admit to healthcare professionals how much they drink. However, in this setting where the same person is asked the same question twice we could presume a similar level of inaccuracy or mis-reporting from both answers. Figure 13 shows the histogram of the alcohol unit intake and the population mean between visits. There was an increase in the number of patients with more than 10 units of alcohol compared to the baseline visit.

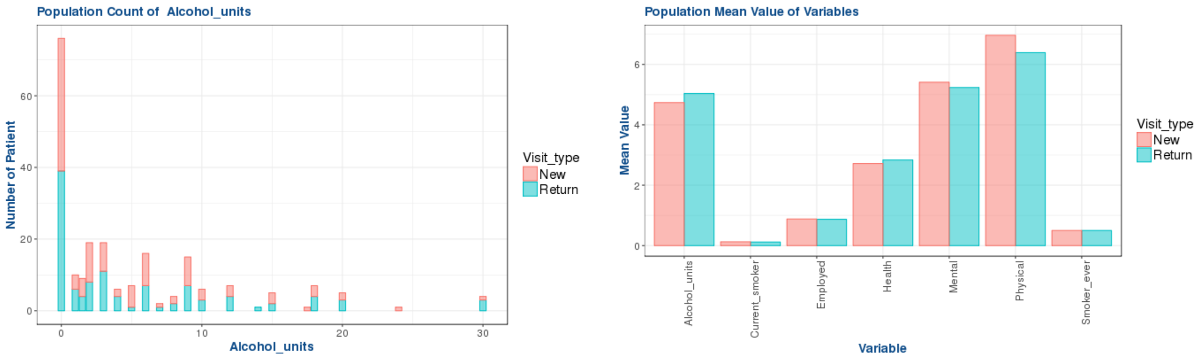


Figure 13 : Alcohol Unit Bar plot and mean values of Lifestyle variables

For some of these questionnaire tabs, Individuals patient scores were plotted as well as shown in Figure 14. The left plot shows all the scores of the selected patient and the right plot shows a clearer picture of the change in value between follow-up and baseline value. The user could quickly see at a glance whether there was an increase or decrease in value, an increase would mean the follow-up score is higher than baseline which means MS has impacted more for that variable. The importance of being able to look across a population, as in Figure 13, and at an individual patient's data (Figure 14) was highlighted to be useful to the project as both options are regularly required.

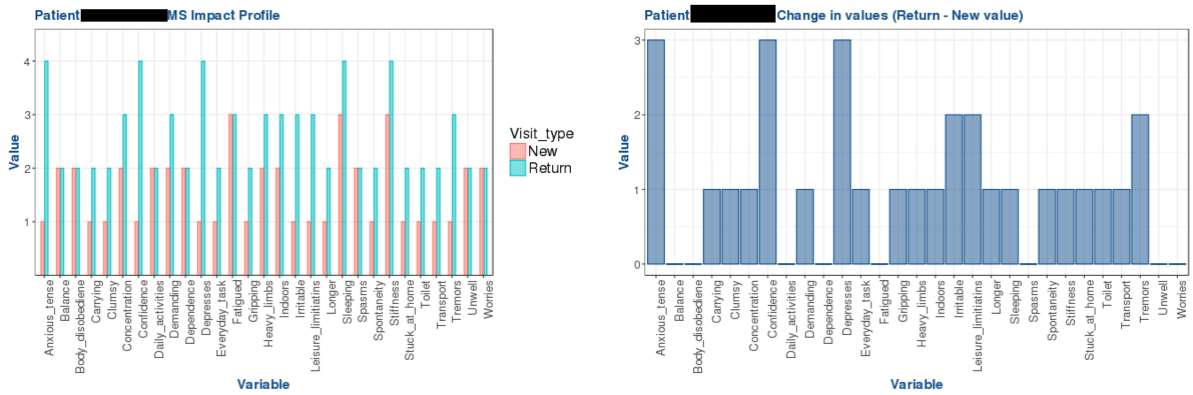


Figure 14 : Individual Patient Scores Plots for Impact questionnaire

6.4.2 Blood test Results

This tab was similar to the blood result tab from the SteatoSITE app with a scatter plot, histogram and boxplot for the various blood test results. The only difference was the filtering option where instead of filtering by gender or diseases like SteatoSITE app, user could filter by the various change in scores such as EDSS (Expanded Disability Status Scale), PASAT (Paced Auditory Serial Addition Test) and so on. These filter options had three different values, negative, positive and no change, users could use these filters to determine if those patients with a negative change in EDSS (improvement) have a different relationship compared to those with positive change. There was also a toggle option to show a simple linear regression line. For example, Figure 15 shows the Triglyceride against Vitamin D with EDSS derived score filter.

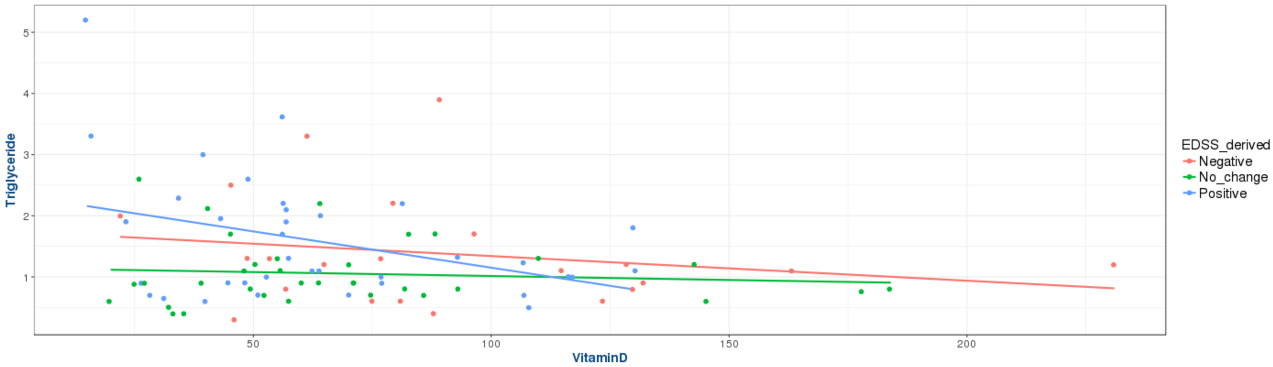
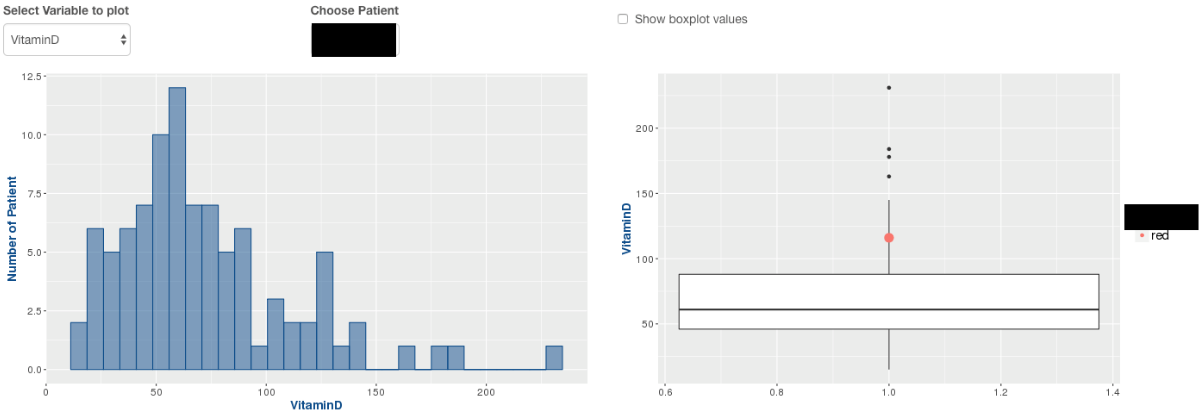


Figure 15 : Scatter plot example for Blood test result tab

A feedback on the scatter plot was that it would be good to be able select the change in scores to plot in the x or y axis as well, this was implemented into the app.

For the histogram and boxplot, the user can select the variable to focus on and show patient value on it, Figure 16 shows the vitamin D of the dataset at baseline visit.


Select Variable to plot: VitaminD

Choose Patient: [redacted]

□ Show boxplot values

[redacted]

Figure 16 : Histogram and Boxplot of Vitamin D

As noted before, vitamin D is potentially important in MS aetiology and in the second feedback meeting Dr Connick was surprised that the vitamin D across the cohort was quite low from the histogram, a result unknown before the visualisation.

6.4.3 Medication

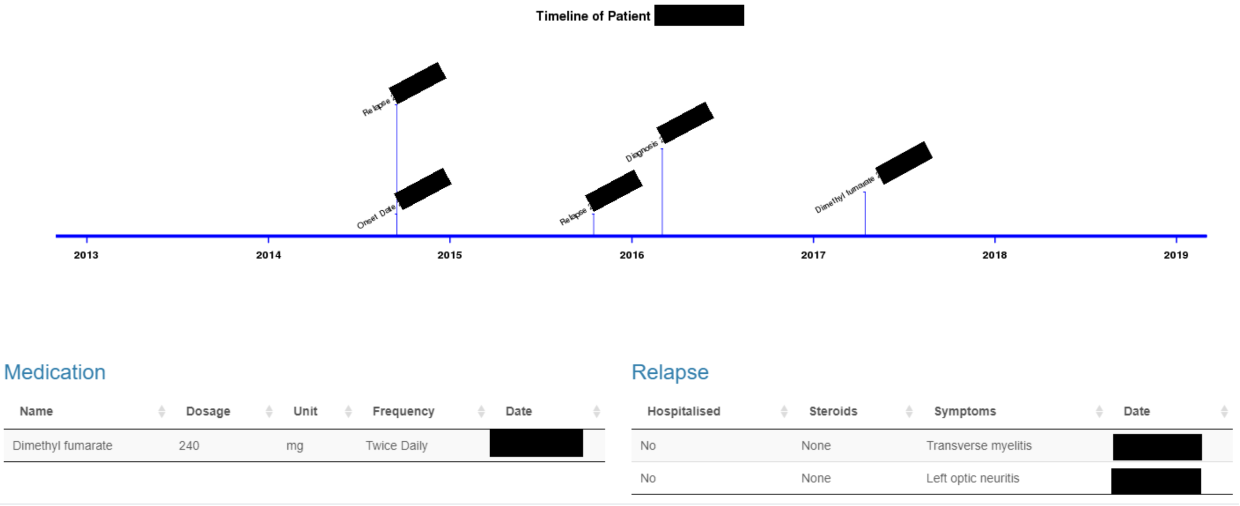
There was only one plot for medication as shown in Appendix 8.11, the most frequent medication taken by patient was Dimethyl Fumarate with 35 count. The other medications were taken by 5 or less patients. There were only 56 patients who are on MS related medications for patients with baseline and follow-up visits. Therefore, around half of the total 110 patients was not on any MS related medication. Analysis involving these medications might not be representative due to the limited data.

6.4.4 Relapse

There were 108 out of 110 patients with relapse information. Appendix 8.12 shows the count plot of the number of relapses throughout the patient's history. The majority of patients has two and three relapses which makes sense since this study focuses on those recently diagnosed with relapsing MS and for most, they might not be referred to a neurologist until after the first relapse and diagnosed with MS.

6.4.5 Patient Viewer

The patient viewer tab plots the timeline of patients where the user can pick which patient to view. It was found that for most patients, they had several relapses before the diagnosis date. An example of the timeline plot for a patient is shown on Figure 17, there are also details about the medication taken and relapse information such as whether patient was hospitalised, or steroids given.



Medication

Name	Dosage	Unit	Frequency	Date
Dimethyl fumarate	240	mg	Twice Daily	[redacted]

Relapse

Hospitalised	Steroids	Symptoms	Date
No	None	Transverse myelitis	[redacted]
No	None	Left optic neuritis	[redacted]

Figure 17 : Timeline plot example

This timeline tab was well received by Dr Connick from the second feedback meeting. He liked the amount of information given for each patient and said it would be useful to have this kind of visual when a patient visits the NHS clinic, as well as for research, as he often has very limited time with patients and trying to get a quick overview of their disease history is difficult. Visualising key clinical factors e.g. dates of starting a medication and/or relapses is something that is not available in the current system.

6.4.6 Subset Summaries

The summary tab of the app allows user to see the summary statistics (min, first quartile, median, mean, third quartile, maximum value, and standard deviation) of the change in patient's clinical scores between visits including EDSS and PASAT. Users could select a criterion such as medication taken, number of relapse attacks and filter by site and gender.

The mean and median EDSS scores are looked at for the different populations in the following section. From the 110-total number of patients, there were 84 female patients and 26 male patients roughly in line with the current diagnosis rate of 2.3 females:1 male ('National Report 2017' 2017) as shown in Table 4.

Gender	**Median**	**Mean**
Male	0	0.19
Female	0	0.15

Table 4 : Mean and Median EDSS value by Gender

The male patients seem to have a greater increase EDSS score compared to female patients. This could be due to more female patients or there was a bias within the male population, for example a couple of patients whose EDSS score has increased a lot.

Table 5 shows the mean value of the change in EDSS score for the subset of patient on a certain medication compared to those not on the medication,

Gender		**Dimethyl Fumarate**	**Fingolimod**	**Alemtuzumab**	**Natalizumab**	**Interferon beta-1a**	**Teriflunomide**	**Peginterferon beta-1a**
Number of patients		Female: 26 Male: 9	Female: 3 Male: 1	Female: 3 Male: 2	Female: 2 Male: 3	Female: 4 Male: 1	Female: 2 Male: 3	Female: 1 Male:0
Female	Median	Yes: 0 No: 0	Yes: 0.5 No: 0	Yes: -0.5 No: 0	Yes: 0.25 No: 0	Yes: 0 No: 0	Yes: 0.75 No: 0	Yes: 0.5 No: 0
	Mean	Yes: 0.23 No: 0.12	Yes: 0.5 No: 0.14	Yes: -0.5 No: 0.18	Yes: 0.25 No: 0.15	Yes: 0.38 No: 0.14	Yes: 0.75 No: 0.14	Yes: 0.5 No: 0.15
Male	Median	Yes: 0 No: 0	Yes: 0.5 No: 0	Yes: -0.5 No: 0	Yes: 1 No: 0.09	Yes: 0 No: 0	Yes: 1.5 No: 0	Yes: N/A No:
	Mean	Yes: 0.17 No: 0.21	Yes: 0.5 No: 0.18	Yes: -0.5 No: 0.25	Yes: 1 No: 0	Yes: 0 No: 0.2	Yes: 0.83 No: 0.11	Yes: N/A No:

Table 5 : Mean and Median EDSS value by Gender and Medication

For all patients, it seems that those on dimethyl fumarate increased EDSS score by a mean of 0.21 between the baseline and follow-up visit. Whilst those not on dimethyl fumarate had an increase of 0.14, this might suggest that dimethyl fumarate is given to those with more severe case of MS.

There was a considerable decrease in the EDSS for the very small subset of patients taking Alemtuzumab medication. From (MSTrust 2017b), it seems that patients are given this medication if there are baseline lesions on the MRI scan or whilst taking other treatment, the patient suffered another relapse recently. There were 4 patients on this medication and the dates of visits and medication entry was examined, it was found that for all patients the medication date entry was on or before the follow-up visit date. Therefore, patients most likely got on this treatment in the period between baseline and follow-up visit which may have contributed to the reduced EDSS score.

Patients on teriflunomide had a larger EDSS increase between baseline and follow-up visit compared to those not on the medication. This treatment seems to be given to patient with active cases of relapsing MS (MSTrust 2017a). Although they are given the treatment, their disability score continued to worsen, it would be hard to tell if there would be an even larger increase in EDSS if they didn't take the medication. The final data set, which will include those on no medications will help elucidate if these changes in EDSS would have been expected.

For the rest of the medication, there was an increase in the EDSS mean score between baseline and follow-up visits. For majority of these drugs there was not large number of patients on them and so the data might be misleading for analysis on the effectiveness of the drug at his point. One recommendation was to collect more data of patients on the different medications which could happen if the study continues in the future or by comparing to patients from other countries where prescribing rates are higher.

6.4.6.1 Relapse

Gender		1	2	3	4	5	6	7
Number of patients		Female: 12 Male: 5	Female: 30 Male: 9	Female: 24 Male: 5	Female: 10 Male: 4	Female: 4 Male: 1	Female: 2 Male: 0	Female: 1 Male:1
Female	Median	0.25	0	0	0	0.	1	0
	Mean	0.33	0.12	0.1	-0.05	0.25	1	0
Male	Median	0	1	0	-0.25	1	N/A	0
	Mean	0.3	0.5	-0.1	-0.25	1	N/A	0

Table 6 : Mean and Median EDSS value by Gender and Number of Relapse

There were only 2 patients with 7 relapses and 6 relapses and the majority of the sample had one to four relapses. It seems that patients with 1 and 2 relapses seem to have a bigger increase in EDSS between visits compared to those with 3 or 4 relapses. Interestingly, male patients with 4 relapses had a small decrease in EDSS score between visits which could be due to higher prescribing rates in those with more relapses.

6.5 Discussion and Future Work

The FutureMS app allowed users to visualise the large amount of data collected to date as part of the study. There was positive feedback from the members of the FutureMS team who were happy to be able to visualise the data collected, especially the timeline plot. Several irregularities were found in terms of clinical data collection such as the measurement of height which will now be followed up.

Also, from the FutureMS project description, patients recently diagnosed with relapsing-onset MS were considered. But from the patient's timeline, some patients had a big gap between the onset date and the diagnosis date with relapse in between, this could be a problem as most patients were recently diagnosed but the duration of their MS might be vastly different.

From the subset summary tab, different sub populations were compared, and it was found that male subjects had a higher mean change in EDSS score compared to female. Due to the low number of patients taking other medications, the findings might not be reliable and so more data should be collected.

This app was used as a preliminary analysis of the available data which were clinical tests, blood test and questionnaire responses. The FutureMS project is still at the early stages of data collection with more patients expected to complete the study. It would have been interesting to use the imaging and genomic data which will be generated in the future to avoid batching effects. Machine learning techniques could be used on the Brain MRI imaging scans to identify certain characteristics of the lesions of the brain and used to determine the severity of their MS or help develop a better diagnosis tool. For example, a convolutional neural network could be created that could classify the severity of MS based on the lesions of the brain. This would also help clinician make decisions on what type of treatment to give to the patients.

The genome sequencing could be used in conjunction with the medication table to determine if a patient on a certain medication responds better than other patients, and there could be certain bio markers in the transcriptomics that predict disease progression. The data collection for the FutureMS project is still on-going and as more data are added onto the tables, the app should continue to function, as demonstrated when the dataset was updated in July 2018. One potential problem might be a new MS treatment not on the list and so it would not be extracted in the data preparation stage, this can easily be solved by adding the new treatment name onto the list given in Appendix 8.9.

7 Conclusion

Within the 12-week placement, apps were created for the SteatoSITE project and FutureMS project. For the SteatoSITE project, a synthetic dataset was created with the use of the minimal complete dataset specification, pathology specification and a lot of research into the different relationship between the variables. The synthetic dataset includes most of the variables the project hopes to get in the future from NHS patient records. An app was then created to visualise the synthetic data and it was well received by the SteatoSITE team especially the decision tree. It was hoped that something similar will be achievable with the real data in the future. The synthetic dataset and app were uploaded onto the SMS platform which would allow SMS to showcase the app to potential and existing clients without showing any real confidential data.

The FutureMS project involved creating an app with real data collected over a one-year period. The main aim of the project was to build a tool which would allow the project team to visualise the data collected as previously, there was no such tool. The project involved quite a lot of data preparation before the app could even be started on. Multiple tables of data were used and sorted. The end result was an app that allowed the user to look at clinical data such as blood test results, medication, relapses and questionnaires all in one app for patients with baseline and follow-up visits. This was the first time the project team were able to look at the data and there were a few irregularities found from the app about the data collection. For example, there were irregularities between the height measurement between visits. This app will have real impact on the FutureMS project, in particular allowing early data cleaning and assisting clinicians in decision making, such as selecting appropriate medication for individual patients.

For both apps, comments were added into the files so that the apps can be easily changed by someone else for these projects. The SteatoSITE app might require more data cleaning before using the datasets with the app. The FutureMS app should require fewer changes to the app as datasets are updated each month.

8 Appendix

8.1 R Code to generate Continuous data

```r
GenContinuous = function (num_patients,prob_vector,min_value, max_value,num_steps){

    x=num_patients ###number of patients/rows of data to generate

    y=seq(min_value, max_value, length.out = num_steps)  ##create a sequence vector based on min and max value

    stepSize=y[2]-y[1] ###step size between the sequence values


range1= seq(min_value, max_value-stepSize,stepSize)  ###create a new seq vector based on min value to max-stepsize

range2= seq(min_value+stepSize, max_value,stepSize) ###create a new seq vector based on min+stepsize value to max

dat <- data.frame(min=range1, max=range2, prop=prob_vector) ##this creates a dataframe based on the intervals and attached probability of the generated value being in that interval

rows <- sample(nrow(dat), x, replace=TRUE, prob=dat$prop) ##creates a vector of n patients with generated values based on prob, determines what interval they belond in

Continuous<- round(dat$min[rows] + runif(x) * (dat$max[rows] - dat$min[rows])) ###generates the random values within that interval

return(Continuous)

}
```

8.2 R Code function to Generate Categorical Data

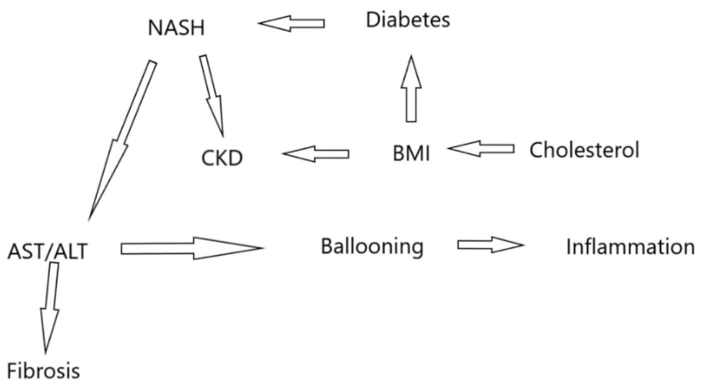
```r
GenCategorical= function (num_patients,variables,prob2){

  x=variables ##categorical variables

  y = sample(x, num_patients,replace=T,prob=prob2) ##generates the random categorical variables based off the probabilities

  return (y)

}
```

8.3 Synthetic Dataset Columns

Column	**Range of values**
StudyID	AA00000-AA00999
Gender	M/F
Age	18-80
BMI	20-50

Alcohol Intake	0 – 14 Female 0-21 Male
Diabetes	1 (yes) / 0 (no)
Diabetes Diagnosis Date	
Chronic Kidney Disease (CKD)	1 (yes) / 0 (no)
CKD diagnosis Date	
Hypertension	1 (yes) / 0 (no)
Hypertension Diagnosis Date	
Cholesterol	148.2-319
Triglycerides	100-488
HDL	21-104
LDL	71-250
Platelets	12-301
AST	30-110
ALT	23-194
Fib-4 Score	0.56-30
Glucose Count Fasted	50-250
Sample Date	
Diagnosis Date	
Liver Type	Allograft, Native
Sample Type	Explant, Needle Biopsy, Resection
NASH	1/0
Kleiner Fibrosis Stage	0,1a,1b,1c,2,3,4
Kleiner Steatosis Score	0,1,2,3
Hepatocyte Ballooning Score	0,1,2,3
Lobular Inflammation	0,1,2,3
NAS (NAFLD acitivity Score)	0,1,2,3,4,5,6,7
Collagen Area	1-20
Creatinine	0.35-6.5
eGRF	10-108

8.4 Diagram showing correlations between variables based off feedback



8.5 Effects of BMI on the risk of Diabetes for different demographics

Age group	**BMI**	**Male**	**Female**
18-45	<25	16.9%	14.5%
	25 to 30	25.5%	30.7%
	30 to 35	51.8%	48.8%
	35+	66.1%	69.3%
45-65	<25	15.9%	13.2%
	25 to 30	33.7%	27.5%
	30 to 35	47.5%	42.2%
	35+	59.4%	58.4%
65+	<25	10.2%	9%
	25 to 30	13.8%	17.3%
	30 to 35	28.3%	26.3%
	35+	33.2%	34.9%

8.6 Values used to generate Collagen data

Fibrosis Stage	**Min Collagen Value**	**Max Collagen Value**	**Median**
0	1	3	2
1 (includes 1a,1b,1c)	1.5	8	3
2	3	9	4
3	3.5	15	4
4	6	27	10

8.7 Gene Expression List for Visualisation

ABCB11, ADIPOQ, AGTR1, APOC3, APOE, CLOCK, CYP2E1, FABP2, GCKR, GCLC, HFE, INSR, KLF6, LEPR, LPIN1, MIF, MTHFR, MTTP, NCAN, NR1H4, NR1I2, PEMT, PNPLA3, PPARA, PPARG, SERPINE1, SLC27A5, SOD2, STAT3 TCF7L2, TM6SF2, TMPRSS6, TNF, UCP3

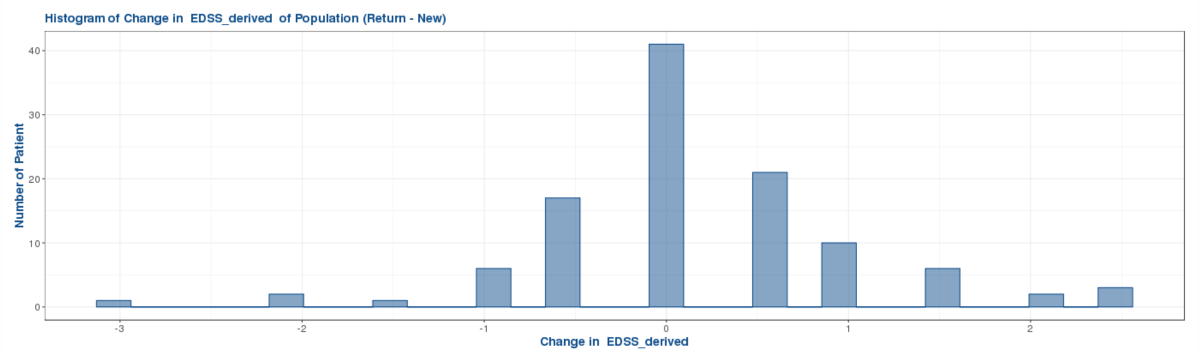
8.8 Columns removed from Minimally Complete Dataset

Column Name	**Reason for Removal**
Ethnicity	463 entries for ‘A’ and only 5 entries for ‘Z’
Date of Birth	Wrong dates, possibly on purpose for ethical issues.
Visit type clinical	Same as visit type column, repeat of another column
Principal Diagnosis	All patients are principally diagnosed with MS
Disease course	Only relapsing-remitting patients are considered for this study
Diagnosis date physician	Repeat of another column
Onset data physician	Repeat of another column

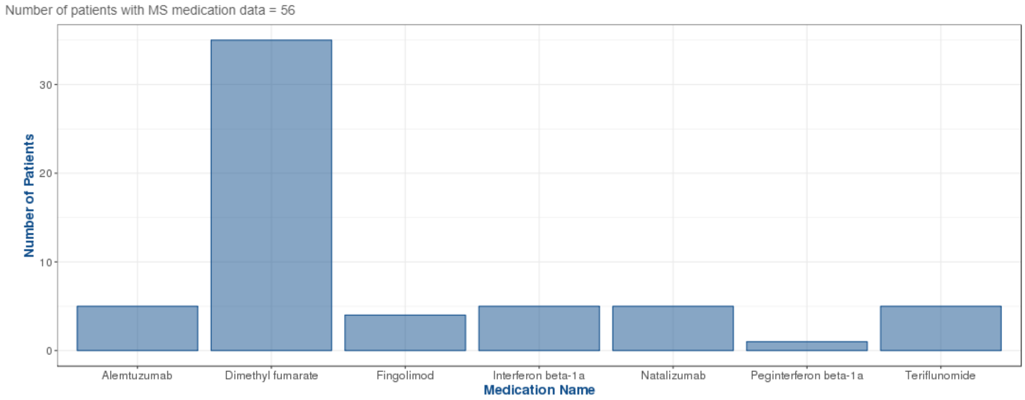
8.9 MS related Medication and Treatment

Avonex, Betaferon, Extavia, Interferon beta-1a, Interferon beta 1b, Peginterferon beta-1a, Plegridy, Rebif, Glatiramer Acetate, Copaxone, Teriflunomide, Aubagio, Dimethyl Fumarate, Tecfidera, Fingolimod, Gilenya, Daclizumab, Zinbryta, Natalizumab, Tysabri, Alemtuzumab , Lemtrada, Cladribine, Mavenclad

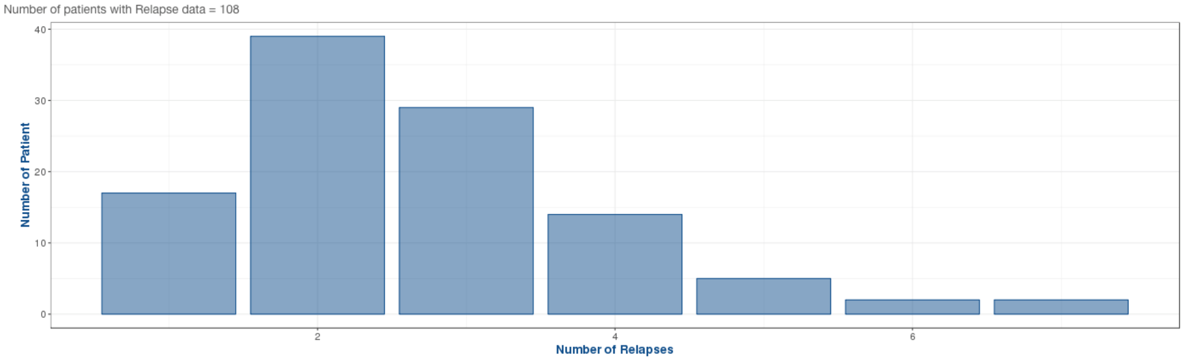
8.10 Histogram of change in EDSS score



8.11 Count plot of MS medication taken by patients



8.12 Number of relapses FutureMS



9 Bibliography

'About FutureMS'. n.d., Accessed 05-07-2018. https://future-ms.org/.

'The Automated Data Scientist'. n.d., Accessed 01-08-18. https://www.eaglegenomics.com/products/.

Berner, Eta S., and Tonya J. La Lande. 2016. 'Overview of Clinical Decision Support Systems.' in, *Clinical Decision Support Systems*.

'Blood Sugar Level Ranges'. n.d., Accessed 27-7-18. https://www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html.

Brunt, E. M. 2016. 'Nonalcoholic Fatty Liver Disease: Pros and Cons of Histologic Systems of Evaluation', *Int J Mol Sci*, 17.

Cheng, Joe. 2017. 'Modularizing Shiny app code', Accessed 12-08-18. https://shiny.rstudio.com/articles/modules.html.

Chun-houh Chen, Wolfgang Härdle, Antony Unwin. 2008. *Handbook of Data Visualization*.

CohesionMedical. 'Stratified Medicine Initiatives', Accessed 11-08-18. https://www.cohesionmedical.com/services/stratified-medicine-initiatives.aspx.

———. 'User Data Entry Systems', Accessed 11-08-18. https://www.cohesionmedical.com/services/user-data-entry-systems.aspx.

'Creatinine Levels Chart'. 2014. Accessed 19/07/2018. http://www.kidney-symptom.com/high-creatinine-level/Creatinine-Levels-Chart.html.

'Diabetes Drugs'. n.d., Accessed 08/07/2018. https://www.diabetes.co.uk/Diabetes-drugs.html#atoz.

'e[nsembl]'. n.d. https://www.eaglegenomics.com/eaglensembl/.

Eaglegenomics. 'About', Accessed 01-08-18. http://www.eaglegenomics.com/about-eaglegenomics/.

Faggella, Daniel. 2017. 'What is Machine Learning?', Accessed 11/08/18. https://www.techemergence.com/what-is-machine-learning/.

Forner, Alejandro, Josep M. Llovet, and Jordi Bruix. 2012. 'Hepatocellular carcinoma', *The Lancet*, 379: 1245-55.

Geron, Aurelien. 2017. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*.

Grant, R. W., N. G. Devita, D. E. Singer, and J. B. Meigs. 2003. 'Polypharmacy and medication adherence in patients with type 2 diabetes', *Diabetes Care*, 26: 1408-12.

GSE61260. 'Human liver gene expression data from subjects of varying ages'. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE61260.

HeartUK. 2015. 'Calculating Cholesterol', Accessed 26-6-2018. https://heartuk.org.uk/files/uploads/documents/huk_fs_mfsE_calulatingcholesterol_v2.pdf.

Hopp, Wallace J., Jun Li, and Guihua Wang. 2018. 'Big Data and the Precision Medicine Revolution', *Production and Operations Management*.

Iyer, Aiswarya, S. Jeyalatha, and Ronak Sumbaly. 2015. 'Diagnosis of Diabetes Using Classification Mining Techniques', *International Journal of Data Mining & Knowledge Management Process*, 5: 01-14.

Jarvis, Lisa M. 2016. "A silent liver disease epidemic." In, 46-52. c&en.

Jensen, M. A., V. Ferretti, R. L. Grossman, and L. M. Staudt. 2017. 'The NCI Genomic Data Commons as an engine for precision medicine', *Blood*, 130: 453-59.

Kolde, Raivo. 2018. 'pheatmap: Pretty Heatmaps'.

Levey, Andrew S., Lesley A. Stevens, Christopher H. Schmid, Yaping Zhang, Alejandro F. Castro, Harold I. Feldman, John W. Kusek, Paul Eggers, Frederick Van Lente, Tom Greene, and Josef Coresh. 2009. 'A New Equation to Estimate Glomerular Filtration Rate', *Annals of Internal Medicine*, 150: 604-12.

Lonardo, A., F. Nascimbeni, A. Mantovani, and G. Targher. 2018. 'Hypertension, diabetes, atherosclerosis and NASH: Cause or consequence?', *J Hepatol*, 68: 335-52.

Marcuccilli, M., and M. Chonchol. 2016. 'NAFLD and Chronic Kidney Disease', *Int J Mol Sci*, 17: 562.

Masugi, Y., T. Abe, H. Tsujikawa, K. Effendi, A. Hashiguchi, M. Abe, Y. Imai, K. Hino, S. Hige, M. Kawanaka, G. Yamada, M. Kage, M. Korenaga, Y. Hiasa, M. Mizokami, and M. Sakamoto. 2018. 'Quantitative assessment of liver fibrosis reveals a nonlinear association with fibrosis stage in nonalcoholic fatty liver disease', *Hepatol Commun*, 2: 58-68.

MayoClinicStaff. 2015. 'Triglycerides: Why do they matter?', Accessed 07/07/2018. https://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/in-depth/triglycerides/art-20048186.

———. 2018. 'Chronic kidney disease - Diagnosis', Accessed 08/07/2018. https://www.mayoclinic.org/diseases-conditions/chronic-kidney-disease/diagnosis-treatment/drc-20354527.

McDonald, John H. 2014. *Handbook of Biolological Statistics* (Sparky House Publishing).

MSTrust. 2017a. 'Aubagio (teriflunomide)', Accessed 03-08-18. https://www.mstrust.org.uk/a-z/aubagio-teriflunomide.

———. 2017b. 'Lemtrada (alemtuzumab)', Accessed 03-08-18. https://www.mstrust.org.uk/a-z/lemtrada-alemtuzumab.

Musen, Mark A., Blackford Middleton, and Robert A. Greenes. 2014. 'Clinical Decision-Support Systems.' in Edward H. Shortliffe and James J. Cimino (eds.), *Biomedical Informatics: Computer Applications in Health Care and Biomedicine* (Springer London: London).

Narayan, K. M., J. P. Boyle, T. J. Thompson, E. W. Gregg, and D. F. Williamson. 2007. 'Effect of BMI on lifetime risk for diabetes in the U.S', *Diabetes Care*, 30: 1562-6.

'National Report 2017'. 2017. Accessed 18-08-18. http://www.msr.scot.nhs.uk/Reports/docs/scottish-ms-register-report-2017.pdf?44.

Nayak, B. K., and A. Hazra. 2011. 'How to choose the right statistical test?', *Indian J Ophthalmol*, 59: 85-6.

NHS. 2016a. 'Chronic kidney disease - Treatment', Accessed 08/07/2018. https://www.nhs.uk/conditions/kidney-disease/treatment/.

———. 2016b. 'High blood pressure (hypertension)', Accessed 08/07/2018. https://www.nhs.uk/conditions/high-blood-pressure-hypertension/treatment/.

Nowok, Beata, Gillian M. Raab, and Chris Dibben. 2016. 'synthpop: Bespoke Creation of Synthetic Data in R', *Journal of Statistical Software*, 74.

Plaisant, C., R. Mushlin, A. Snyder, J. Li, D. Heller, and B. Shneiderman. 1998. 'LifeLines: using visualization to enhance navigation and analysis of patient records', *Proceedings of the AMIA Symposium*: 76-80.

Podgorelec, Vili, Peter Kokol, Bruno Stiglic, and Ivan Rozman. 2002. *Journal of Medical Systems*, 26: 445-63.

Prasad, Tangirala Venkateswara, and Syed Ismail Ahson. 2006. 'Visualization of microarray gene expression data', *Bioinformation*, 1: 141-45.

Ryoo, J. H., Y. J. Suh, H. C. Shin, Y. K. Cho, J. M. Choi, and S. K. Park. 2014. 'Clinical association between non-alcoholic fatty liver disease and the development of hypertension', *J Gastroenterol Hepatol*, 29: 1926-31.

Sahay, Amar. 2017. *Data visualization: Recent trends and applications using conventional and big data* (Business Expert Press: New York, New York (222 East 46th Street, New York, NY 10017)).

Sanyal, D., P. Mukherjee, M. Raychaudhuri, S. Ghosh, S. Mukherjee, and S. Chowdhury. 2015. 'Profile of liver enzymes in non-alcoholic fatty liver disease in patients with impaired glucose tolerance and newly detected untreated type 2 diabetes', *Indian J Endocrinol Metab*, 19: 597-601.

Schwartz, Jay. 2017. 'Normal Cholesterol Levels by Age', Accessed 08/07/2018. https://www.livestrong.com/article/275091-normal-cholesterol-levels-by-age/.

Shaikhina, Torgyn, Dave Lowe, Sunil Daga, David Briggs, Robert Higgins, and Natasha Khovanova. 2017. 'Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation', *Biomedical Signal Processing and Control*.

Shamai, L., E. Lurix, M. Shen, G. M. Novaro, S. Szomstein, R. Rosenthal, A. V. Hernandez, and C. R. Asher. 2011. 'Association of body mass index and lipid profiles: evaluation of a broad spectrum of body mass index patients including the morbidly obese', *Obes Surg*, 21: 42-7.

Shiny.RStudio. 2017. 'The basic parts of a Shiny app', Accessed 11/8/2018. https://shiny.rstudio.com/articles/basics.html.

Shirasawa, T., H. Ochiai, T. Ohtsu, R. Nishimura, A. Morimoto, H. Hoshino, N. Tajima, and A. Kokaze. 2013. 'LDL-cholesterol and body mass index among Japanese schoolchildren: a population-based cross-sectional study', *Lipids Health Dis*, 12: 77.

'Solutions for Precision Medicine'. 2018. Accessed 30-07-18.

Sorbi, D., J. Boynton, and K. D. Lindor. 1999. 'The ratio of aspartate aminotransferase to alanine aminotransferase: potential value in differentiating nonalcoholic steatohepatitis from alcoholic liver disease', *Am J Gastroenterol*, 94: 1018-22.

'Type 2 diabetes'. n.d. https://bnf.nice.org.uk/treatment-summary/type-2-diabetes.html.

'Understanding Your Lab Values'. 2017. Accessed 19/08/2018. https://www.kidney.org/atoz/content/understanding-your-lab-values.

WebMD. 2016. 'High Triglycerides: What You Need to Know', Accessed 1-08-18. https://www.webmd.com/cholesterol-management/high-triglycerides-what-you-need-to-know#1.

'Why We Need a NASH Data Commons'. 2018. Accessed 11-08-18. https://steatosite.com.

Williams, K. H., N. A. Shackel, M. D. Gorrell, S. V. McLennan, and S. M. Twigg. 2013. 'Diabetes and nonalcoholic Fatty liver disease: a pathogenic duo', *Endocr Rev*, 34: 84-129.

Wood, K. L., M. H. Miller, and J. F. Dillon. 2015. 'Systematic review of genetic association studies involving histologically confirmed non-alcoholic fatty liver disease', *BMJ Open Gastroenterol*, 2: e000019.

Yoneda, M., H. Fujii, Y. Sumida, H. Hyogo, Y. Itoh, M. Ono, Y. Eguchi, Y. Suzuki, N. Aoki, K. Kanemasa, K. Imajo, K. Chayama, T. Saibara, N. Kawada, K. Fujimoto, Y. Kohgo, T. Yoshikawa, T. Okanoue, and Disease Japan Study Group of Nonalcoholic Fatty Liver. 2011. 'Platelet count for predicting fibrosis in nonalcoholic fatty liver disease', *J Gastroenterol*, 46: 1300-6.

Younossi, Zobair, Quentin M. Anstee, Milena Marietti, Timothy Hardy, Linda Henry, Mohammed Eslam, Jacob George, and Elisabetta Bugianesi. 2017. 'Global burden of NAFLD and NASH: trends, predictions, risk factors and prevention', *Nature Reviews Gastroenterology &Amp; Hepatology*, 15: 11.