Using Data Science methods to explore and understand individuals' medical events within the UKBiobank

Samuel E Butterfield

May 2020

G401: Computer Science with Industrial Placement

Supervisor: Dr Paolo Missier

Word Count: 14,762

Abstract

This paper explores the use of data science methods to understand medical events within the UKBiobank study while tackling the difficulties of handling big data from the medical field.
With an understanding of the data, I used exploratory data analysis to solve problems the data presented, and engineer a feature to classify an individual's overall health surrounding a point in time.

Declaration

"I declare that this dissertation represents my own work except where otherwise stated"

Acknowledgements

I would like to thank Dr Paolo Missier for his frequent support and guidance throughout this project. I would also like to thank Phillip Darke and the rest of the project team for their assistance and input.

Table of Contents

Table of Figures

Table of Tables

Introduction

Motivation for Exploring Medical Data

In the UK, healthcare is a fundamental part of our society with the NHS providing most of the health services free of charge. However, due to the increased costs of treating the elderly & the ageing population, increased hospital admission rates and other financial constraints, the NHS's services are becoming overstretched [1] [2]. Part of this problem is the direct cost of treating patients with preventable diseases. To mitigate this, there has been an increased focus on proactive rather than reactive healthcare. A proactive healthcare model is predicting a patient's health and making necessary adjustments to their lifestyle, treatment and/or premedicating to prevent the onset of preventable diseases and illnesses; this is beneficial for the patient and is often cheaper.

One possible approach to enabling a proactive healthcare model is to apply data science methods, such as data analytics or machine learning, on medical data in order to predict a patient's likelihood of developing medical conditions [3].

The UKBiobank and the Data

All the medical data is provided by the UKBioBank, which is a charity organisation that collates medical information and supplies it to researchers to better the treatment and diagnoses of a wide range of diseases. Between 2006 and 2010 the UKBioBank recruited over 500,000 volunteer participants between the ages of 40-69 into a longitudinal study. Upon entry to the study a range of data was collected, and since then there have been additional datasets gathered that contain other medical-related information for the original participants [4].

The UKBiobank Project at Newcastle University

There is a larger project going on at Newcastle University surrounding the data provided by the UKBiobank. The focus of this project is the onset and development of Type 2 Diabetes (T2D) and Cardiovascular Diseases (CVD). CVD equates to 1 in 4 premature deaths in the UK [5] with the direct cost of treating CVD estimated at £7.4 Billion per annum [6]. Furthermore, T2D is regarded as a contributing factor in developing CVD but independently has a treatment cost of £8.8 Billion per annum [7], both conditions can mostly be prevented, having significant benefits to patients and savings to the £134 Billion yearly budget [8]. I will be working in parallel with this project but not on it.

Aim

To understand individuals' medical events within the UKBiobank study by deriving insights and knowledge from medical datasets. With this understanding, I will focus on a specific problem to solve using iterative exploratory data analysis.

Objectives

1. Learn the relevant tools and techniques used by data scientists to explore and produce insight from data.

   Due to the amount of data needed to be analysed, I will use specialist tools to permit me to do this. Technologies like Python, R and PowerBi are industry standards and incredibly powerful tools used to perform statistical analysis and visualisation of data. This objective needs to be achieved first, as I will need to know the tools I will be using to achieve the following objectives.

2. Understand the features within the datasets and what information they provide.

   The datasets are complicated and understanding each feature will help me derive the most insight from the data. To do this, I will perform feature analysis for each feature within the dataset. I should do this once I have completed objective 1.

3. Perform data quality analysis to ensure that the data is clean and ready to be used.

   Before using the datasets, I need to ensure that there are no mistakes or missing data, especially as the UKBiobank collated some of the data from multiple sources. I should do this once all prior objectives have been complete.

4. Once I understand this data and the information it provides, focus on one problem to solve using iterative Exploratory Data Analysis (EDA) techniques.

   I have a vast amount of data at my disposal, so once I understand the information these datasets provide, I can focus on a specific problem to solve. I will use EDA cycles to explore the data further and gain insights to bridge a gap to the solution of the problem I have identified. I should perform this only once all prior objectives have been complete.

5. Reflect ethically on the data used and the problem I aimed to solve.

   We in 2020 are still living in the wake of the Cambridge Analytica Scandal, which proved that data is an extremely powerful commodity. I want to examine this paper from an ethical perspective to see if it has had any unintended ethical consequences.

Background and Preparation

This chapter covers the background of data science as a discipline, what tools & technologies I chose to use, and the challenges faced when exploring medical data.

What is Data Science

The origins of data science can be traced back to the paper "From Data Mining to Knowledge Discovery in Databases" [9]. This popularised the idea that "There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data." [9]. In 2001, William S Cleveland combined data mining with computer science to create data science [10] [11].

Data science is now a broad subject area with the Journal of Data Science defining it as "almost everything that has something to do with data: Collecting, analysing, modelling..... yet the most important part is its applications --- all sorts of applications." [12]. This definition is broad as a data scientists' goal is to make as much impact with data as possible, and they do this through deriving insights from data or producing data products [11]. Figure 1 shows a pyramid that summarises different roles a data scientist potentially performs with a focus on implementing machine learning models.

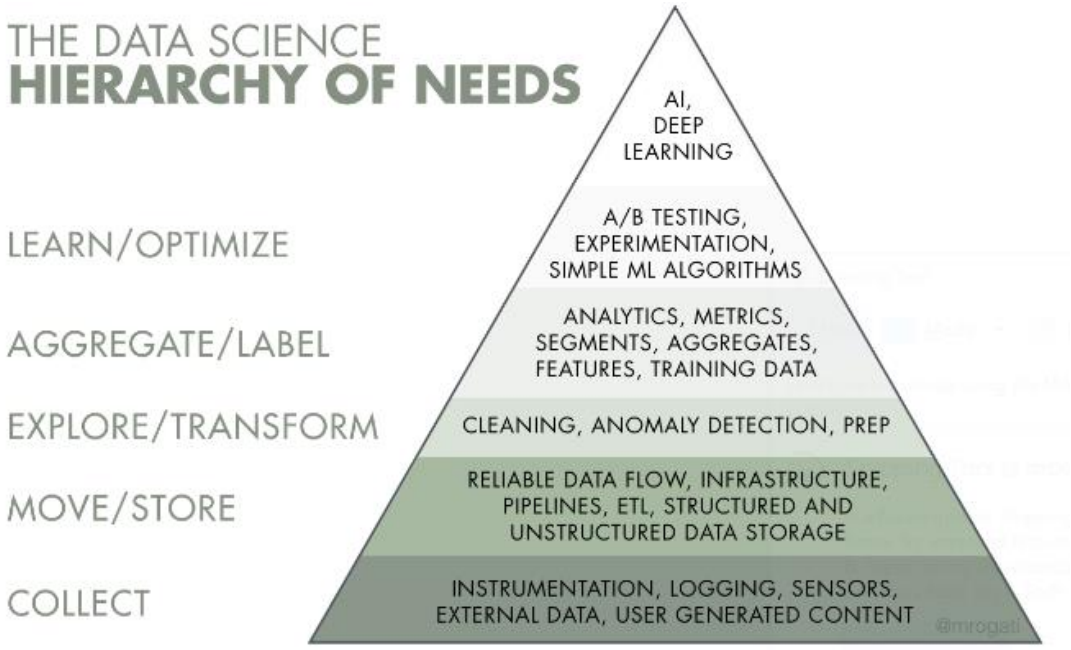


Figure 1: Data science hierarchy of needs [13]

What Tools to Use

There is a large and ever-expanding toolbox available to data scientists, therefore identifying the right tools to use is essential. This paper aims to gain an understanding and explore the medical history of individuals within the UKBiobank study, so I chose tools that fulfilled this purpose.

Programming Language – Python

There are two major languages in data science, R and Python. Both languages are similar in functionality. However, I chose Python solely on my personal desire to learn the language as it is increasing in popularity [14]. The specific libraries I used are: Pandas for data storage, NumPy for multi-dimensional arrays and Matplotlib for producing visuals.

Platform – Newcastle Universities Rocket High-Performance Cluster (HPC)
The datasets are large and require computing services with a lot of memory and power to manipulate them. The HPC has a Linux CLI interface that allows me to SSH in where I can schedule my Python scripts.

Dashboarding – PowerBi
While on placement, I was involved with different Business Intelligence projects where PowerBi was the primary tool used, it offers quick visualisation, quick manipulation of data and quick drilldowns.

Spreadsheets and CSV – Microsoft Excel
Some data from the UKBiobank was provided in XLSX format, so Excel was the obvious choice. Excel is often perceived as a basic tool, but there are some tasks that are simple and can be completed faster using tools in Excel rather than Python.

Exploring Electronic Health Records

Electronic Health Records (EHR) are digitised versions of paper medical records, EHR's are the main source of my medical data in this project. Deriving information from EHR data does not come without problems, an article "Can AI Fix Medical Records?" gives an overview of the challenges the United States faced after transitioning to EHR's. The article details how doctors had a rough transition to the managing patients EHR's and inputting patient information due to clunky UI's, in rare cases patients' deaths have been traced back to issues relating to EHR systems. As well as frontline issues, the article also highlights problems faced when using EHR data to build data-driven healthcare models. Academics from the University of Chicago and Stanford University have teamed up with big tech firms like Google to overcome these challenges, part of the solution they came to be standardising the different EHR formats. [15]

The UKBiobank first released EHR data to researchers in June of 2019 [16], because of how recent this has been, there are very few publications that have utilised this data. However, before distributing this data, the UKBiobank standardised the format. This was a significant hurdle faced in looking at EHR data in the United States, with this already done it will allow researchers to focus on other potential problems of EHR data, such as noise.

Understanding the Data

The first step I took was to understand the data I had at my disposal; this involved feature analysis to tell me what information the dataset provides and research into the Read medical vocabulary.

Overview

The UKBiobank has supplied a wide range of datasets that provide a multitude of different medical, genomic, biological and demographical information. Due to the size and complexities, I decided to focus on three datasets:

Baseline – upon entry to the UKBiobank study, the 502,645 participants undertook a baseline assessment where participants answered a questionnaire, had a face-to-face interview and had various measurements taken [17].

Activity Tracker Reading – over 104,000 out of the 502,645 participants wore high-resolution accelerometers for a week that tracked their activity levels. These accelerometers were developed in Newcastle Universities Open Lab [18]. This dataset stores when an individual wore the accelerometer and not the outcome of how active participants were.

GP Clinical – the UKBiobank has collated participants primary healthcare data from GP surgeries. It holds data that captures a broad range of medical and non-medical information. This dataset holds the vast majority of an individual's healthcare information [19].

The datasets are flat files, stored in CSV or TSV formats. Each dataset has a common feature 'eid' that is a unique identifier for an individual in the UKBiobank study. This feature allows me to link records for the same individual across different datasets. Each record in the datasets has a timestamp, Figure 2 is a visualisation of individuals record over time.

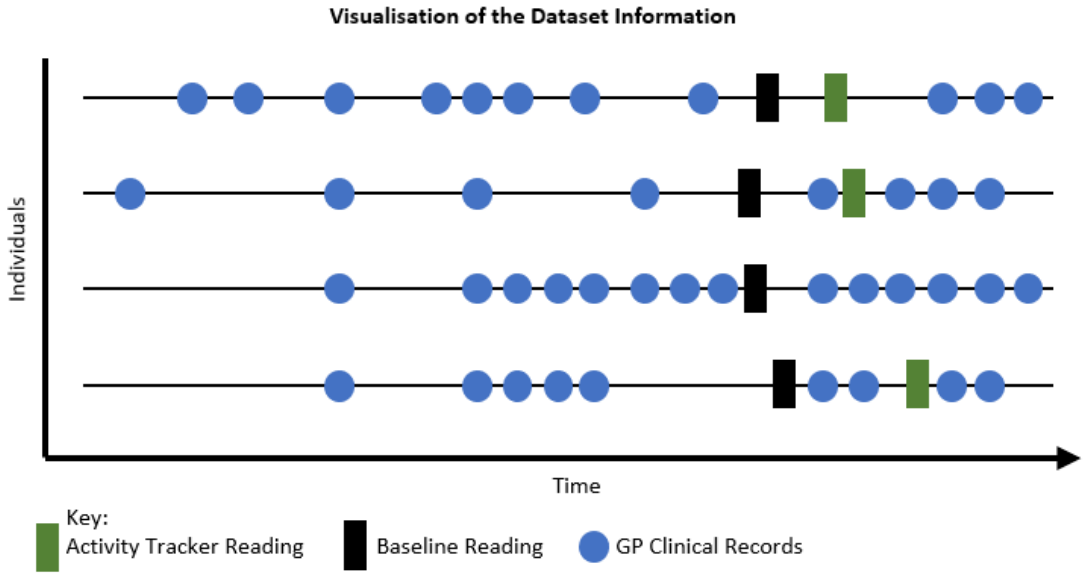


Figure 2: Visual representation of the information in the datasets

Baseline

All 502,645 individuals in the study have a single observation in this dataset. Individuals have a maximum of 541 features; this varies because over the four years of recruitment, more tests were added, and not all of the 22 test centres had the same testing capabilities. Newcastle University has only acquired features relevant to the project space; there are more features within the UKBiobank's databases.

Due to the vast number of features and the variation of features per individual, selecting a specific subset of these features that are relevant is necessary. The only feature I used was 'HEALTHY_CVD_T2D1_T2D2' which is a health classification that classifies all individuals as: Inconclusive, healthy, have CVD, have T2D or have CVD & T2D. This feature relates to the larger UKBiobank project at Newcastle; I will use it to compare other datasets health classification demographics that are subsets of the whole population.

Statistical summary of baseline

Table 1 is a breakdown of the full population by the Baseline Health Classification. This acts as a good reference when looking at other datasets or sub-populations to see if they are disproportionately biased according to this classification.

Baseline Health Classification	Number of Individuals	Percent of Individuals
Inconclusive	208,606	41.5%
Healthy	127,502	25.36%
CVD	146,394	29.12%
T2D	5,178	1.03%
Both CVD & T2D	14,965	2.98%

Table 1: Baseline health classification

Activity Tracker Readings

The activity tracking data is extremely complex, and another part of the larger project at Newcastle is deciphering the activity outcomes. The data was collected between the start of 2014 and the end of 2015 and has three features: eid, start date and end date of when the individuals wore the high-resolution accelerometers. There were over 104,000 individuals that wore the trackers, and I have 45,846 individual's data. Table 2 is a breakdown of the individuals by the health classification. I believe the Activity Tracker Reading dataset provided has already had the individuals classified as Inconclusive removed. However, based on Table 1's proportions, that would bring the total individuals to 89,006, which is a shortfall of 14,994.

Baseline Health Classification	Number of Individuals	Percent of Individuals
Inconclusive	0	0
Healthy	22,891	49.93
CVD	20,751	45.26
T2D	625	1.36
Both CVD & T2D	1,579	3.44

Table 2: Activity Tracker Readings health classification

GP Clinical

As of doing this project, the UKBiobank has only released around 45% of the participants GP Clinical records [20]; this explains why the unique number of individuals expressed in Table 3 is 230,105. The individuals within the GP Clinical dataset is a subset of the individuals in Baseline; comparing Table 1 with Table 4, GP Clinical is representative of the full population of individuals in terms of health classification.

Statistical summary of GP Clinical

Number of events	123,669,371
Unique Individuals	230,105
Average events per person	537.45
Mode events per person	1
Min events per person	1
Max events per person	10,396
Standard Deviation of events per person	474.72
non-unique events	9,875,164

Table 3: Event Summary for GP Clinical

Baseline Health Classification	Number of Individuals	Percent of Individuals
Inconclusive	95,646	41.57%
Healthy	57,804	25.12%
CVD	67,567	29.36%
T2D	2,339	1.02%
Both CVD & T2D	6,749	2.93%

Table 4: GP Clinical health classification

Levels of Healthcare In the UK

To understand what type of medical information primary healthcare presents, I researched the levels of healthcare within the UK. The NHS provides a vast range of services to millions of people around the country. This care is broken down into Primary, Secondary and Tertiary. Primary care is the first port of call for people in need of health services and is often provided by GP's, health centres, Pharmacists or Opticians. Secondary care is usually based in hospitals, and often a patient will get referred to Secondary care by a primary care professional if they cannot resolve the medical issue. Tertiary care is provided in specialist centres, these are for highly specialised treatments such as transplants or consultant care [21].

Feature analysis of GP Clinical

Each observation in the dataset is a medical event, and this medical event has eight features where a maximum of seven could be populated:
Mandatory – eid, data_provider, event_dt, read_2 OR read_3.
Optional (Read code dependant) – value1, value2 and value3.

eid	data_provider	event_dt	read_2	read_3	value1	value2	value3
XXXX432	2	02/03/1974	246..	NaN	units	val	val
XXXX321	3	17/05/1994	NaN	2469.	val	NaN	NaN
XXXX876	2	08/07/1980	B1…	NaN	NaN	NaN	NaN
XXXX876	2	19/11/1980	B1a2.	NaN	NaN	NaN	NaN

Figure 3: Example head of GP Clinical dataset

eid – everyone in the UKBiobank study is anonymised and represented as a 7-character numerical string.

data_provider – this is a categorical feature that can be values of: 1, 2, 3 or 4. There is no information on what these values represent but it looks to be linked to the 'read_2' and 'read_3' feature. 'read_2' can only be classified as: 1, 2 or 4, whereas 'read_3' can only be classified as: 3. I believe this feature to be linked to the source of the data rather than adding useful information to the event.

event_dt – Is the day, month and year of an event's occurrence, this is in the format of DD/MM/YYYY.

Read codes – are an encoded way to describe what the event is within the data. These codes can be decoded into a textual description using a mapping document to better make sense of them. There are two versions, Read 2 and Read 3, they both have a length of 5 characters long. An event can either use Read 2 or Read 3. Figure 3 shows an example head of the GP Clinical dataset, where the top two observations are for a Blood Pressure reading. The codes differ based on the version of Read code. There is a bias of Read 3 codes where 70.75% of events using this version of code.

Value – There are three value features: 'value1', 'value2' and 'value3'. Whether they are populated depends on the events Read code. In Figure 3, the bottom two observations are diagnostic codes typically not having values associated, and the top two observations being test codes that typically have values associated. The columns populate from 'value1' to 'value3', meaning you would never have a populated 'value2'/'value3' without the preceding value field. Read 3 codes tend to only have the first value feature populated, Read 2 tend to be more inconsistent when populating the value features even among identical codes, this is represented in the top two observations in Figure 3.

Mapping Document

The UKBiobank provided an Excel spreadsheet that maps Read codes to textual descriptions, Read code versions and other clinical encoding libraries.

Top 20 Read 2 & Read 3 Codes

Read codes are the main source of information for each event within GP Clinical, as the Read code describes what the event was about. Figure 4 and Figure 5 are visuals of the top 20 Read codes for Read 2 & Read 3. Looking at the top 20 codes provides some insight into what the whole dataset holds, I am taking this approach because there are too many events with too many different Read codes to observe with no summarisation.

Occurrences – Top 20 Read 3 have higher occurrences than Read 2, this is expected due to 70.75% of events in GP Clinical are in version Read 3.

Nature of events – Both versions of the codes top 20 are populated with test codes. I believe test and recorded biomarkers like these are performed frequently by GP's to understand a patient's health and to track health over time, meaning I could potentially use similar methods to view a patient's health over time. I expect codes relating to diagnosis of conditions to have fewer occurrences; this is because a patient is diagnosed once each time they have a condition, and I believe that conditions will not reoccur as often as tests such as a Blood Pressure reading. Most of the tests observed are generalised, and I expect a lot of individuals to have occurrences of these general tests, whereas diagnostic codes are specific and can only occur if an individual has a specific condition.

Nature of codes – There is a noticeable correlation in what the code is and what it represents. In Read 2 2469. is O/E – Systolic BP Reading and the next increment 246A. represents O/E – Diastolic BP reading. As I am not a domain expert within the medical field a lot of these medical terms are difficult to understand, so if there is an underlying structure behind these codes, understanding that could prove to be extremely beneficial to understanding medical events within GP Clinical.

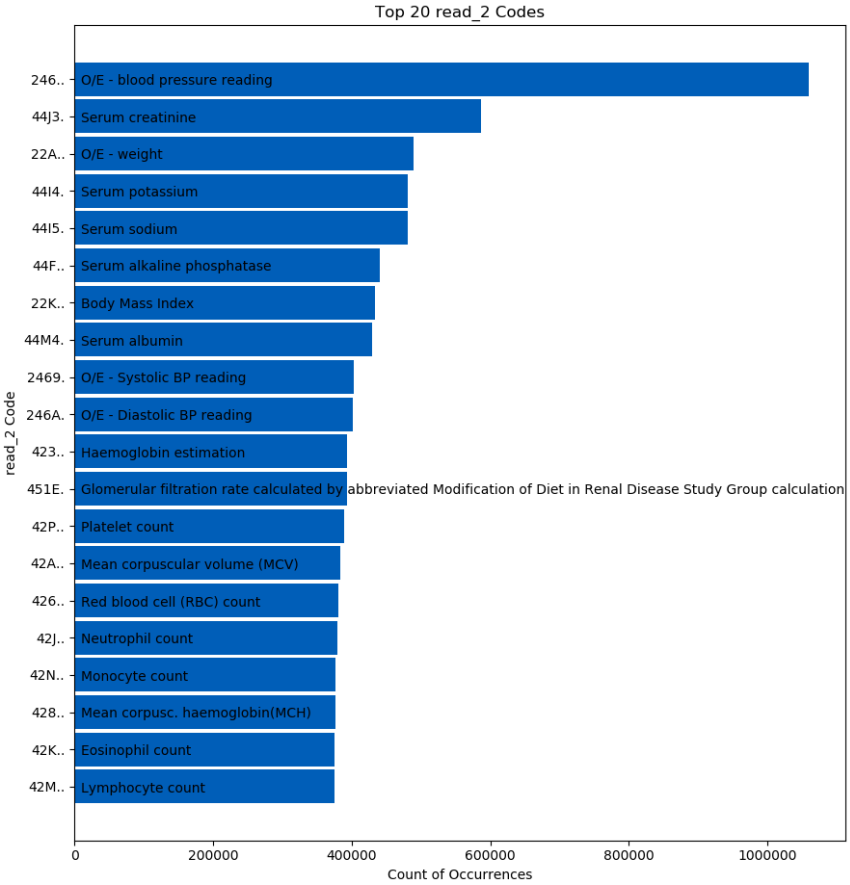


Figure 4: Top 20 read_2 codes for the GP Clinical dataset

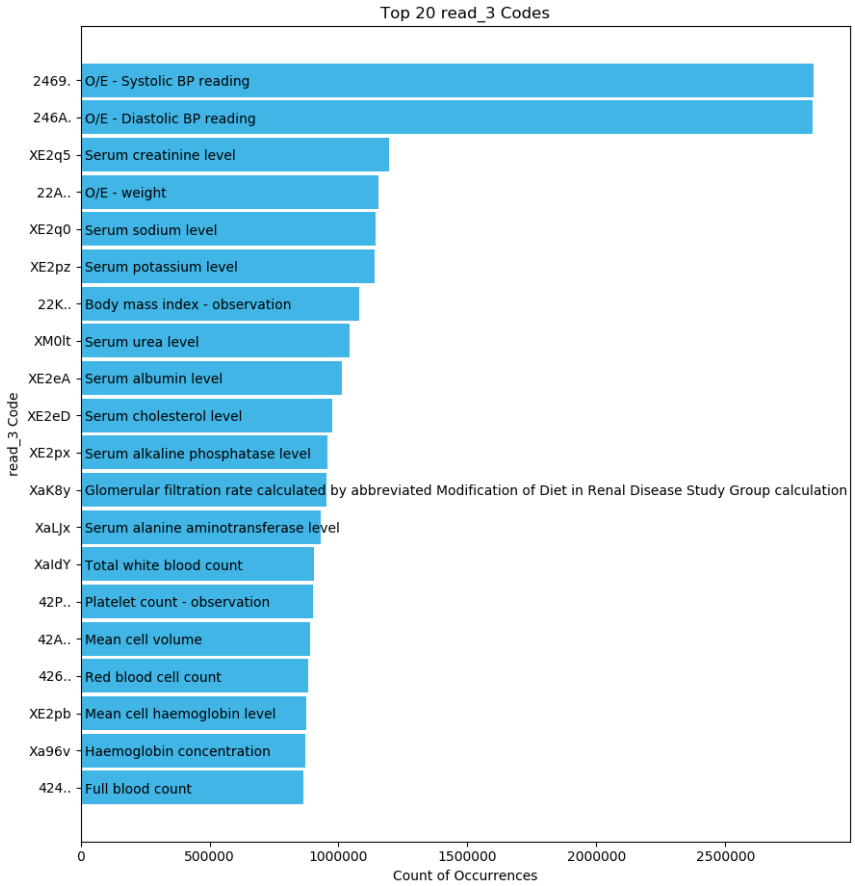


Figure 5: Top 20 read_3 codes for the GP Clinical dataset

Time Distribution of events in GP Clinical

Figure 6 plots the number of events each year. The main body of events spans from 1937 to 2017, this is explained by the maximum age of individuals during recruitment was 69 and 2017 was when this dataset was collated. The number of events increase overtime; I believe this is because as people age they visit their GP's more often [22] and before computers they used paper health records, which will have a higher chance of being lost. There is a group of events in the future and group of events around 1900, these are explained in documentation provided by the UKBiobank to be placeholders [23].

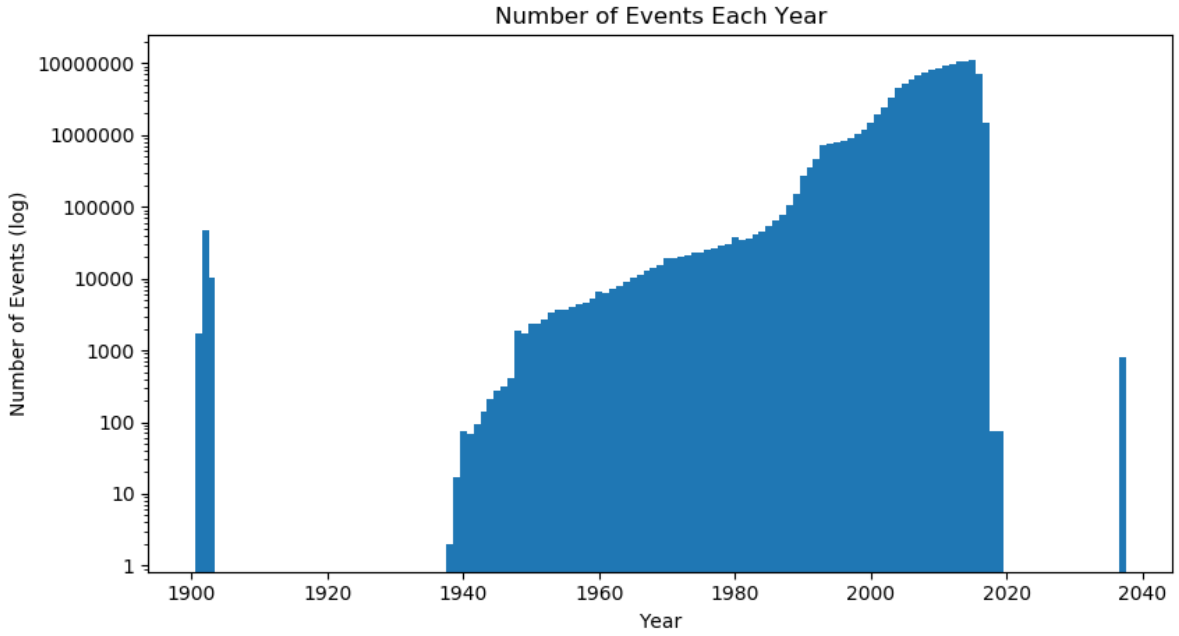


Figure 6: Number of Events Each Year

Distribution of how many events individuals have in GP Clinical

Figure 7 visualises how many individuals have X number of events in GP Clinical. The distribution backs up the metrics in Table 3 where the average number of events for individuals to have is 537.45, mode being 1 and the standard deviation is 474.72. There are outliers where 3272 individuals have over 2000 events.

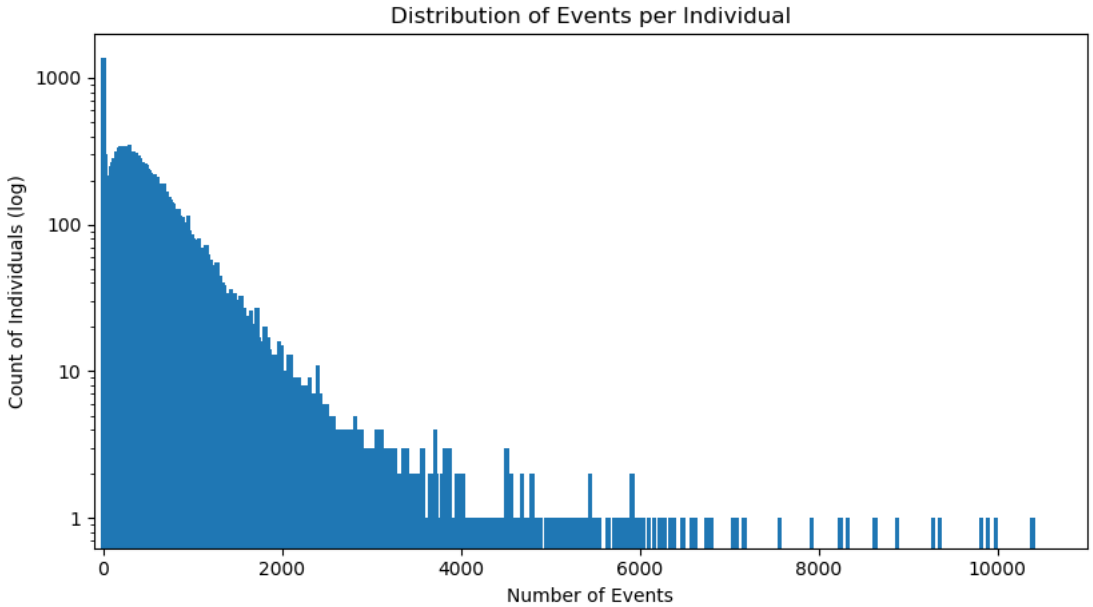


Figure 7: Distribution of how many events individuals have in GP Clinical

Understanding Read Codes

I have identified that the key source of information is found in the Read code of an event. This section outlines my research into understanding Read codes.

Origin

Read codes were created in the early 1980s by Dr James Read, a GP who was working with Abies Informatics Ltd [24]. These codes were designed to classify and capture primary care data from GP's and store this information within a computer system. The Read Version 1 had a hierarchical structure using a 4-byte code, which theoretically had 1,679,616 different codes available. The codes were deemed a success and in 1988 were recommended as the standard for encoding medical information in GP's, later in 1990 the rights were bought by the UK department of health [25].

As the benefits of encapsulating patient information became apparent, the scope and detail of what information to capture expanded, which made Read Version 1 obsolete. In 1990 Read Version 2 was released which was an expanded 5-byte encoding with a similar hierarchical structure to Read Version 1. Read Version 3 was released in 1994 which was a 5-byte code but used a polyhierarchical structure [25]. Read Version 2 and Read Version 3 have been the standard encoding from their introduction till now, they are currently being phased out and replaced by a new encoding system SNOMED CT [26].

Read Version 2 (Read 2)

Read 2 is a 5-byte code that utilises a hierarchical tree structure to represent information about a patient or drugs prescribed. The code is filled in from left to right with the starting byte being the chapter, with the proceeding bytes describing something in greater detail; not all bytes can be filled in and the non-fill character is ' . '. This version was an expansion from Version 1 and theoretically can support 656,356,768 different codes [25].

Chapter Codes

The first byte in the Read 2 code is known as the chapter code and it describes the information that is stored with the subsequent bytes.

History, Examination, Procedures and Administration Chapters:

Chapter 0 – Occupation
Chapter 1 – History/Symptoms
Chapter 2 – Examination/Signs
Chapter 3 – Diagnostic Procedures
Chapter 4 – Laboratory Procedures
Chapter 5 – Radiology/Medical Physics
Chapter 6 – Preventative Procedures
Chapter 7 – Operations, procedures, sites
Chapter 8 – Other therapeutic procedures
Chapter 9 – Administration

Diagnostic Chapters:

Chapter A – Infectious and Parasitic Diseases
Chapter B – Neoplasms (Cancers)
Chapter C – Endocrine, nutritional, metabolic and immunity disorders (Includes Diabetes)
Chapter D – Diseases of blood and blood-forming organs
Chapter E – Mental Disorders
Chapter F – Nervous system and sense organ diseases
Chapter G – Circulatory system diseases

Chapter H – Respiratory system diseases
Chapter J – Digestive system diseases
Chapter K – Genitourinary system diseases
Chapter L – Complications of pregnancy, childbirth and the puerperium
Chapter M – Skin and subcutaneous tissue diseases
Chapter N – Musculoskeletal and connective tissue diseases
Chapter P – Congenital anomalies
Chapter Q – Perinatal conditions (Fetal and neonatal conditions)
Chapter R – [D]Symptoms, signs and ill-defined conditions
Chapter S – Injury and poisoning
Chapter T – Causes of injury and poisoning
Chapter U – [X]External causes of morbidity and mortality
Chapter Z – Unspecified

Chapter Code Information sourced from [27], [28] and the mapping document.

Hierarchy

The hierarchy of Read 2 codes are represented within the codes themselves making them easy to read, any subsequent child byte would be under the parent in the hierarchical structure.

Example:
1-byte depth: 'C....' Endocrine, nutritional, metabolic and immunity disorders
2-byte depth: 'C2…' Nutritional deficiencies
3-byte depth: 'C24..' Vitamin A deficiency
4-byte depth: 'C247.' Vitamin A deficiency with other ocular manifestation
5-byte depth: 'C2470' Vitamin A deficiency with xerophthalmia

Hierarchy information sourced from [27], [28] and the mapping document.

Read Version 3 (Read 3)

Read 3 uses the same 5-byte format, but the code itself does not represent the hierarchy. Instead, it uses a separate table listing all binary parent-child relationships. The way the relationships work in Read 3 allows for a polyhierarchy with unlimited depth, but still constrained to a 5-byte code [29].

Finding information on Read 3 was a lot more difficult compared to Read 2. I do not have access to the table that lists the parent-child relationships, which means I cannot create or analyse information based off the hierarchy of Read 3 codes.

Description Prefixes

Some code descriptions have prefixes that add additional information to the code description, they come in the format of '[X]' which in this case means Mental and Behavioural [27].

Synonyms

Codes can be duplicated many times, but the descriptions would be a synonym of each other, this is because there are multiple ways of describing a clinical concept. Every code has one description that is preferred, and the rest are marked as synonyms. These synonym terms are useful for providing additional medical information to GP's but are not useful to a non-medical professional; because of this, I only used the Prefered Terms [27].

Comparison

The differences between the codes can be observed in Figure 4 & Figure 5 for Read 2's '44I5.' Serum Sodium and Read 3's 'Xe2q0' Serum Sodium Levels. Based on the adjusted occurrences, they appear to be the same test; however, Read 3's code makes it difficult for a non-medical professional to determine roughly what this code is by looking at the code itself. This makes the analysis of large quantities of Read 3 codes a lot more difficult compared to Read 2 codes.

The main trade-off between Read 2 & Read 3 is the level of detail. Read 2 offers less detail about specific events, whereas Read 3 is a lot more granular due to its polyhierarchy and unlimited depth. However, Read 3 is difficult to read due to its hierarchy not being represented within the code, and there is the problem of too much information, Read 3 might be more useful to a GP looking at an individual whereas most of the medical information is lost on a non-medical professional.

Other Health Coding Systems

Read Codes were designed specifically for the UK medical system and the only other country I could find that uses Read Codes at the point of care delivery was New Zealand [30]. However, other clinical vocabularies exist.

ICD – International Classification of Diseases

ICD is published and maintained by the World Health Organisation (WHO). It is a global standard for the classification of medical conditions and many of the encoding systems like Read map to it. The WHO use ICD as an international standard for statistical reporting on different diseases and health conditions around the word.

SNOMED CT

The company SNOMED International is a non-profit organisation that owns and maintains the clinical terminology SNOMED CT [31]. The NHS has started to replace Read Codes with SNOMED CT across Primary and Secondary levels of care [32]. Just like Read Codes, SNOMED CT is just a clinical terminology and not a full system that manages EHR's, but the NHS provides Systems of Choice that Primary and Secondary providers can choose from [33].

Data Quality Analysis

This chapter expands on Understanding the Data, where I investigated the quality of the data used; this is crucial to ensure the integrity of the data so any insights drawn will not be inaccurate.

Baseline

The only feature I use from this dataset is the Health Classification. There are 502,645 observations which is one per each individual in the study; this means that each individual will only have one Health Classification and no errors must exist for the purposes I am using this dataset.

Activity Tracker Reading

This dataset was created by another member of the project team at Newcastle University by running a script over the folder that stored JSON files on everyone's activity metadata. This script extracted the Start & End date from these files. To check the number of individuals was correct at 45,846, I ran a script that counted the number of metadata files in this folder, and it was 45,846.

GP Clinical

NaN Analysis

I started by looking for NaN (blank) values. I did not look in the value features as I know that NaN is expected, this is shown In Table 5.

Feature:	**eid**	**data_provider**	**event_dt**	**read_2**	**read_3**
No. NaN:	0	0	163,222	87,493,722	36,175,649

Table 5: GP Clinical NaN summary

The feature eid has 0 NaN values, this is good as all events can be linked to an individual. data_provider is linked to the individual which explains why it also has 0 NaN values. There are 163,222 events with a NaN value for event_dt, this is a large number but proportionally only equates to 0.13% of the overall dataset. The Read features have many NaN values; this is expected as only one of these features is supposed to be populated per event. When added together, there are 123,669,371 which is equivalent to the number of events in GP Clinical, so every event has one of these features populated.

Duplicate Analysis

This dataset is a collation of data from many different GP systems and because of this there is a margin for mistakes to be made; one of these mistakes is duplication of data. From Table 3 I learnt that there were 9,875,164 non-unique events meaning that duplicates exist within this dataset. I investigated this further to see if the duplicates logically make sense in the context of the event and are supposed to exist or if these are errors. Documentation provided by the UKBiobank flags duplication as a known problem when individuals move from different data suppliers [23], but this doesn't provide a lot of information on the extent of the duplication.

The GP Clinical dataset has a lot of duplicates, 7.99% of all events are non-unique. I performed a high-level visual analysis and I observed a possible trend: if there was a Read code that typically has a value associated to it, there would be two almost identical entries where one entry was valueless and another entry that had values. Without including the value columns, there are 35,552,885 non-unique events which is 28.75% of the events in GP Clinical. Given my current understanding of the dataset, I cannot rule out that value holding Read codes are the only cause of duplication, because of this I continued to look at duplicates across all features within the dataset.

Duplicates by Individual

Figure 8 plots the 209,182 individuals' who have duplicate events. The average number of times an event is duplicated is 2.19, and the average number of non-unique events per person is 47.21. There are a few extreme cases where one individual has 2097 duplicated events with an average duplication of 161.31. This shows how prolific duplication is across the GP Clinical dataset and could explain why some individuals have a high number of events in Figure 7.

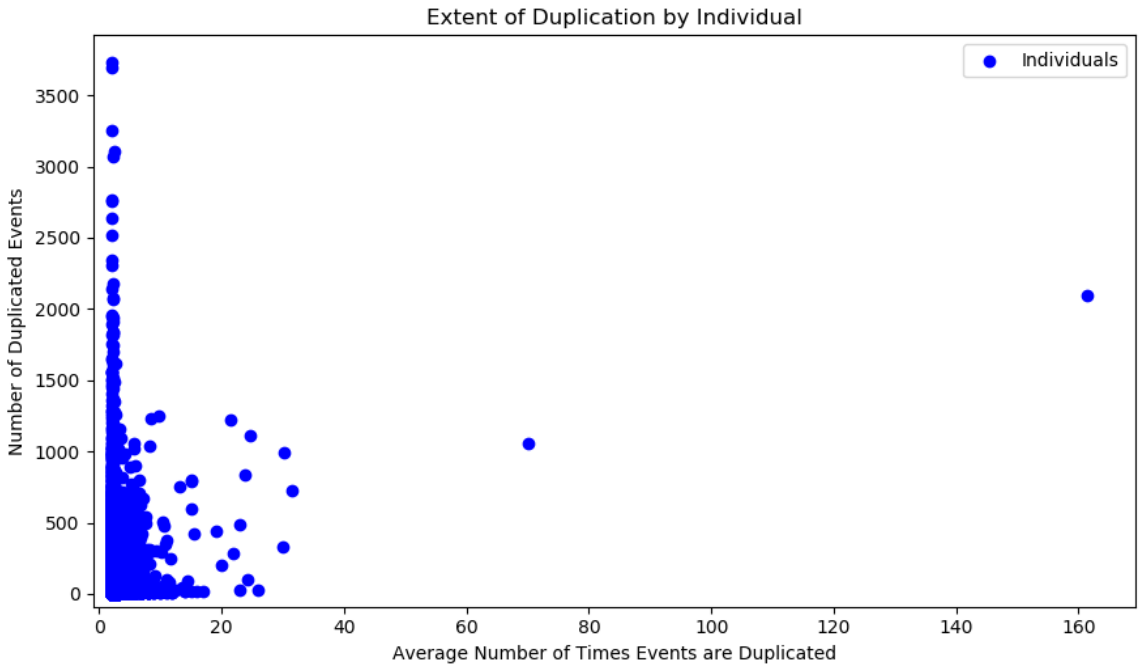


Figure 8: Extent of duplication by individual

Duplicates by Code

I looked at the top 20 most duplicated codes for both Read versions to understand if there are any codes that make logical sense to be duplicated, to see if duplication is limited to a subset of codes and if code duplication is correlated to code occurrence.

Logical duplication – I believe there are five Read 2 codes that make sense to be duplicated. These codes are the codes relating to medication reviews or test request (8B3S., 8B3x., 8B314, 8B3V. and 413..). These tests make sense because this is an event in GP Clinical that is tracked by day and it is unlikely individuals will visit a GP twice in a single day. However, they might have multiple prescribed treatments that come under review or multiple tests requested in this single GP visit. There is a similar situation for Read 3 codes where two codes (XaK6t and XaF8d) for test request medication reviews exist.

Subset of duplication – From Figure 9 there appears to be no clear subset of codes in Read 2 that stands out; when visually looking over the top 100 codes there is a steady decrease in duplications. Figure 10 shows that the top code (Y0384) is disproportionately higher than any other code; there is no mapping for this code and after investigating this further I discovered there is no textual mapping for any Read 3 code beginning with a 'Y'.

Correlation with top 20 codes – For Read 2 top 20 comparing Figure 9 to Figure 4 there are eight shared codes. The proportion of these eight codes occurrences in GP Clinical that are non-unique is 6%-10%. For Read 3 top 20 comparing Figure 10 to Figure 5 there are ten shared codes; the proportion of these ten codes in GP Clinical that are non-unique is 1.7% - 5.9%. I believe there is a loose correlation between code occurrences and the number of non-unique codes.

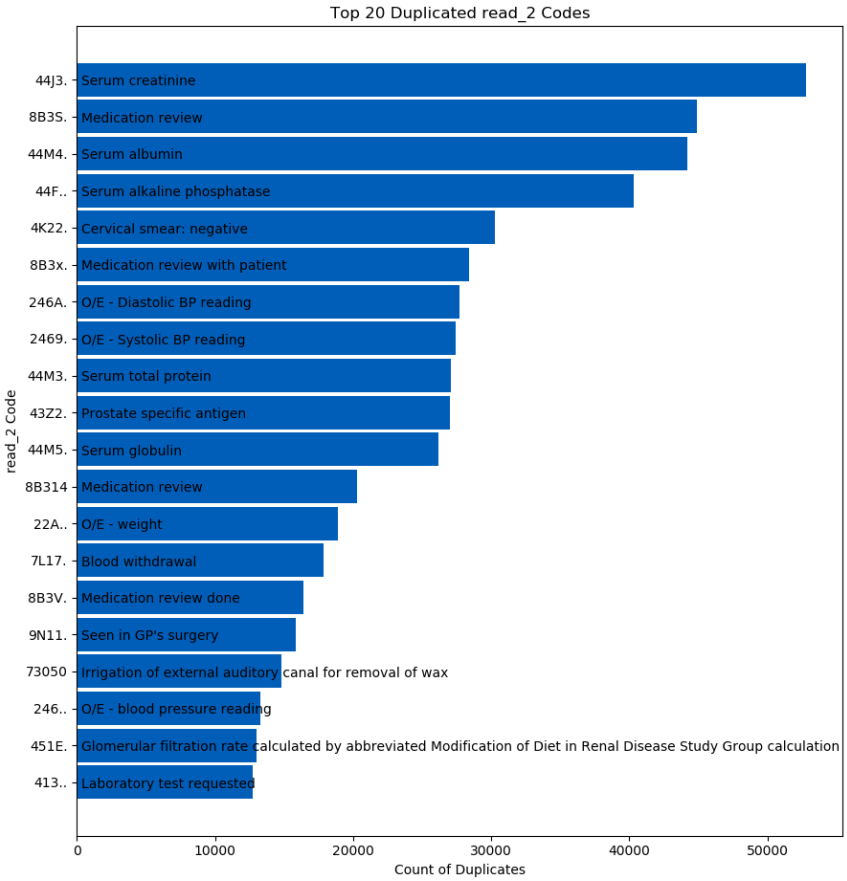


Figure 9: Top 20 Duplicated read_2 Codes for the GP Clinical dataset

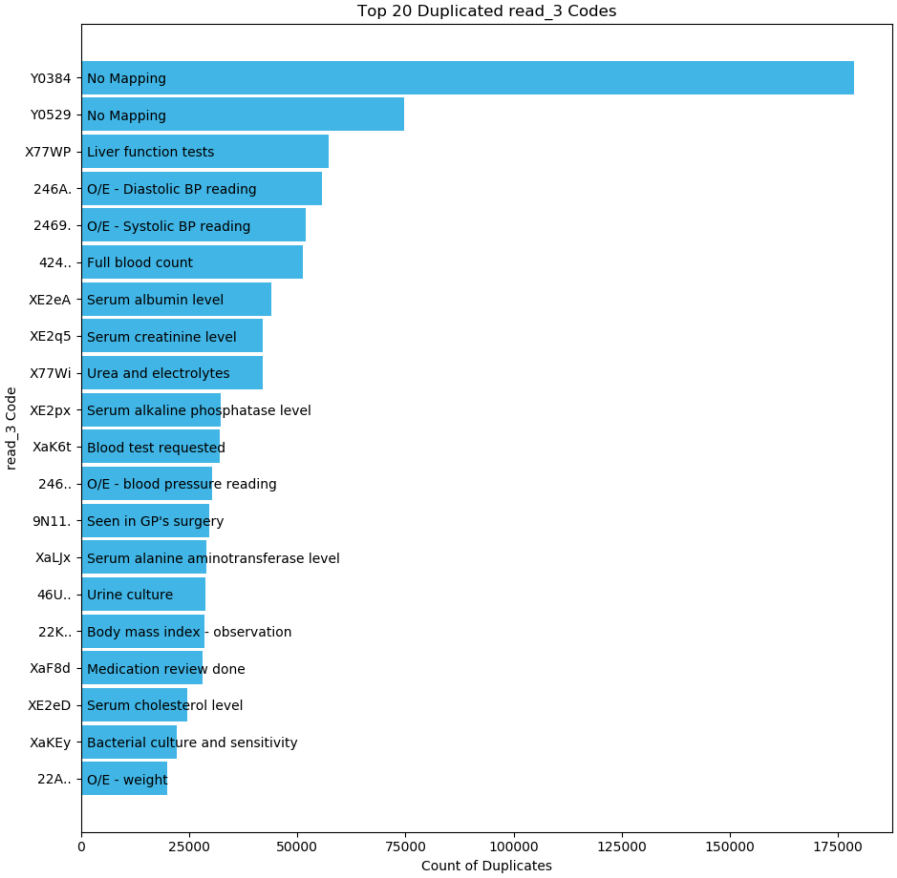


Figure 10: Top 20 Duplicate read_3 Codes for the GP Clinical dataset

Individuals with a single event in GP Clinical

The mode number of events for individuals to have is one. I would find any individual having one event surprising, so I would deem 1,372 individuals a cause for investigation. Most individuals in this subset use Read 3 codes. The most popular Read 2 code in this subset is “Patient registered with FPC” and the most popular Read 3 code is “Main spoken language English”. After doing a visual look over these codes, I conclude that no codes make sense to be entered only once and no anomalous trends appear.

Read codes with no textual mappings

There are some codes that have no textual description mappable in the mapping document, this means that the specific code is not providing specific information on the event. There are 2,859,299 events in GP Clinical that have textual description, this is 2.31% of the dataset. Out of the 2,859,299, 2,453,212 are Read 3 codes that begin with ‘Y’ meaning that this issue is mostly contained to this sub-problem of not having any mapping for the ‘Y’ codes.

GP Clinical Corrections

I have identified several issues that need to be addressed before I go on to use this dataset.

NaN event dates corrections

For events that had no date I assigned 01/01/1920, this is because no other event has this date or dates 10 years either side and will allow me to look at the events with no date in a similar way to events in the future. Figure 11 shows this.

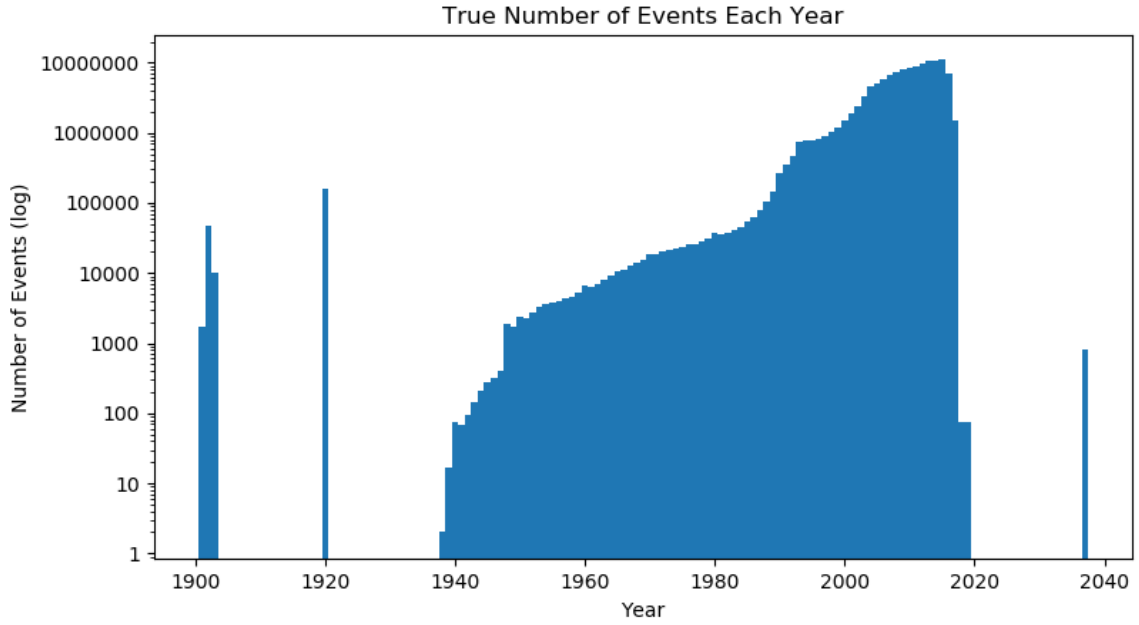


Figure 11: True Number of Events Each Year

Duplicated events corrections

I did not uncover a clear reason to why duplication was taking place on such a large scale. However, I did discover some events that have a logical purpose for being duplicated. My understanding of this is limited to the top 20 duplicated codes and the only way to fully understand this is by examining every duplicated code which is infeasible, so I will treat these as regular duplicates. To resolve the duplicate problem, I will drop all but one non-unique event, Figure 12 is a representation of GP Clinical with duplicates and Figure 13 visualises my way of removing duplicates. My method will not resolve the issue of duplication of events where one event has the value fields populated and duplicates have no values associated, this is shown with grey & blue events in Figure 12 & Figure 13.

eid	data_provider	event_dt	read_2	read_3	value1	value2	value3
XXXX432	2	02/03/1974	246..	NaN	units	val	NaN
XXXX321	3	17/05/1994	NaN	2469.	val	NaN	NaN
XXXX432	2	02/03/1974	246..	NaN	NaN	NaN	NaN
XXXX321	3	17/05/2000	NaN	2469.	val	NaN	NaN
XXXX432	2	02/03/1974	246..	NaN	NaN	NaN	NaN
XXXX321	3	17/05/1996	NaN	246A.	val	NaN	NaN
XXXX432	2	02/03/1974	246..	NaN	units	val	NaN
XXXX432	2	02/03/1974	246..	NaN	NaN	NaN	NaN

Figure 12: Representation of GP Clinical with duplicates

eid	data_provider	event_dt	read_2	read_3	value1	value2	value3
XXXX432	2	02/03/1974	246..	NaN	units	val	NaN
XXXX321	3	17/05/1994	NaN	2469.	val	NaN	NaN
XXXX432	2	02/03/1974	246..	NaN	NaN	NaN	NaN
XXXX321	3	17/05/2000	NaN	2469.	val	NaN	NaN
~~XXXX432~~	~~2~~	~~02/03/1974~~	~~246..~~	~~NaN~~	~~NaN~~	~~NaN~~	~~NaN~~
XXXX321	3	17/05/1996	NaN	246A.	val	NaN	NaN
~~XXXX432~~	~~2~~	~~02/03/1974~~	~~246..~~	~~NaN~~	~~units~~	~~val~~	~~NaN~~
~~XXXX432~~	~~2~~	~~02/03/1974~~	~~246..~~	~~NaN~~	~~NaN~~	~~NaN~~	~~NaN~~

Figure 13: Representation of GP Clinical with duplicates deleted

Individuals with a single event corrections

I kept these individuals in the dataset as they are such a small proportion of GP Clinical and they did not appear to be anomalous.

Read codes with no textual mapping's corrections

There is an issue as events with no textual description provide no information, but from my research into Read codes I learnt that Read 2 codes still give some information based on the chapter. Therefore, I decided to keep these codes within my dataset. I also decided to keep Read 3 codes that begin with 'Y', as an entire chapter was missing I assumed that the mapping document was not up to date and that the UKBioBank could update it at any time.

Summary of corrections

Table 6 shows the effects of my corrections; the only corrections that made an impact on the metrics were dropping the duplicates. The GP Clinical dataset had 9,875,164 non-unique events, which is 7.99% of the dataset. Depending on the severity of the individual event duplication there was potential to lose a large proportion of the dataset, however, removing the duplicates achieved a loss of only 4.37% which is 5,400,285 events.

	GP Clinical	GP Clinical Corrected
Number of events	123,669,371	118,269,086
Unique Individuals	230,105	230,105
Average events per person	537.45	513.99
Mode events per person	1	1
Min events per person	1	1
Max events per person	10,396	10,043
Standard Deviation of events per person	474.72	453.22
non-distinct events	9,875,164	0

Table 6: GP Clinical correction summary comparison

Building a Clear Picture Surrounding Individuals High Resolution Accelerometer Readings

Once the data was understood from an information and integrity point of view, I moved onto finding a solution to a specific problem identified. I did this with iterative EDA's which incrementally built my knowledge or part of a solution to solve the problem.

Overview

The specific problem I will be focussing on is building a Clearer Picture of an individual's medical status surrounding their High-resolution accelerometer reading (CPH) using medical events within the GP Clinical dataset.

Rationale

The larger UKBioBank project at Newcastle University is focussing on CVD & T2D, and for part of this there will be a study into the activity levels of individuals based on their health classification. Every individual in the study has a health classification, which can be: Healthy, CVD, T2D, CVD & T2D and Inconclusive. The purpose of this classification is to solely classify an individual into those groups and is not a good representation of an individual's overall health, as Healthy is a byword for not having a T2D or CVD diagnosis.

Problem

There is no process for investigating the general health of an individual or any features that classify an individual's overall health. This becomes a problem when examining the activity outcomes from the high-resolution activity trackers. The health classification does not provide any information on other ailments that would impact an individual's activity outcome, this means drawing conclusions on their health classification would be flawed, as individuals may have other ailments beyond CVD & T2D that prevent them from being active.

Proposed Solution

Perform iterative EDA's on the GP Clinical dataset to further the current understanding and find insights that will help engineer a feature that will classify an individual's overall health. This feature should be able to determine if an individual has other ailments, and therefore should be excluded when looking at the activity outcomes of the health classification populations.

EDA - Overview

EDA has varying definitions depending on what data you are exploring and the type of outcome you are looking for, therefore I will outline my reasoning choosing my methodology. The methodology of Cross Industry Standard Process for Data Mining (CRISP-DM) [34] is an effective way of performing EDA, but it is overkill for my niche problem and it is oriented with the end goal of deploying machine learning models [35]. My method will take inspiration from CRISP-DM and a cycle-oriented problem-solving methodology I used extensively while on industrial placement, Plan Do Check Act (PDCA) [36].

Plan – Outline the cycles aim.
Do – Perform the necessary steps to process the data or produce metrics.
Check – Analyse the actions performed in the Do section.
Act – Evaluate the outcome of this cycle against the aim of the cycle, what actions are to be taken.

Targeting the Data

From initially looking at the data, I learnt that both GP Clinical and the Activity Tracker Reading datasets are subsets of the baseline population. I only want to be looking at data that exists in both these datasets. Figure 14 shows a visual of the targeted population. The dataset GP Prescriptions is a dataset holding prescription information; I originally had planned to use it, but due to time constraints I was unable to.

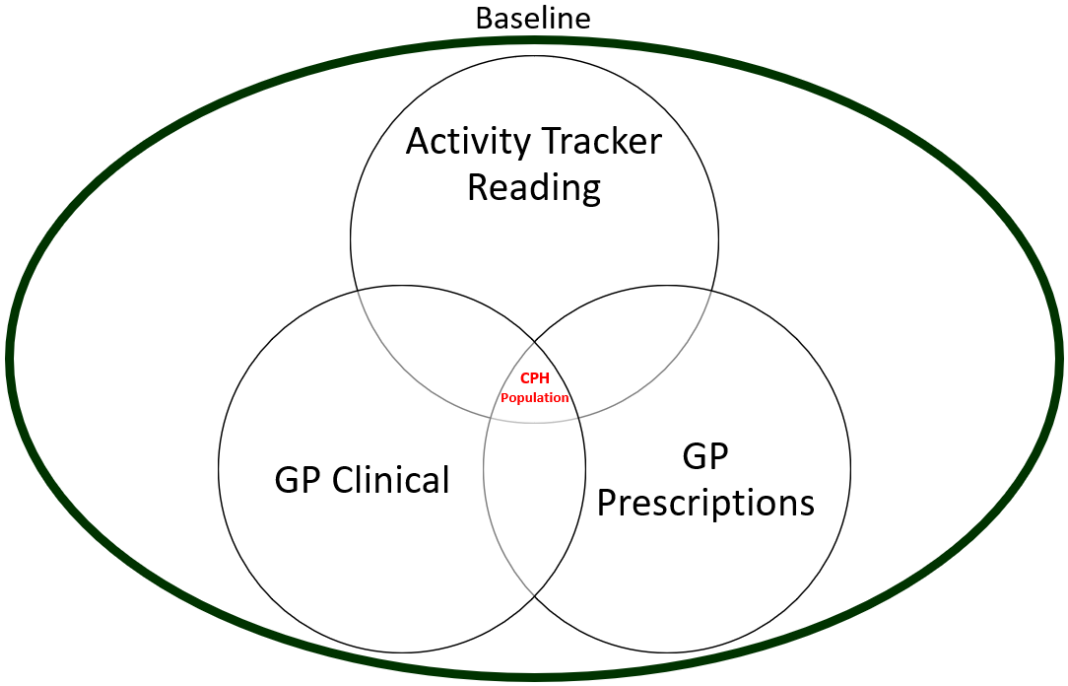


Figure 14: Visual of Target Population

EDA1 – Understanding the CPH Population

EDA1 – Plan

Aim – Understand the data loss from intersecting the three datasets by comparing metrics.

EDA1 – Do

Health Classification	GP Clinical FULL	Intersected Population	Population Loss %
Inconclusive	95,646	0	100%
Healthy	57,804	9,695	83.23%
CVD	67,567	9,190	86.4%
T2D	2,339	276	88.2%
Both CVD & T2D	6,749	691	89.76%
TOTAL:	**230,105**	**19,852**	**91.37%**

Table 7: CPH GP Clinical summary of population loss

	GP Clinical FULL	Intersected Population	Percentage Change
Number of events	118,269,086	10,133,630	-91.43%
Average events per person	513.99	510.46	-0.69%
Mode events per person	1	1	0%
Min events per person	1	1	0%
Max events per person	10,043	9,757	-2.85%
Standard Deviation of events per person	453.22	443.93	-2.05%

Table 8: CPH event summary for GP Clinical

EDA1 – Check

Table 7 shows a 91.37% reduction of the population, this is still representative of the overall GP Clinical population as the main classification have proportionally similar reductions. The UKBiobank has only released around half of the GP Clinical dataset, which is why the intersected population is roughly half of the 45,846 in the Activity Tracker dataset. The impact of also intersecting GP Prescriptions is minimal, it only excludes 833 individuals that exist in GP Clinical and the Activity Tracker datasets.

Table 8 shows the summary of events in the full GP Clinical dataset compared with the intersected population. The reduction in events are in line with the reduction of individuals and the percentage changes in the other metrics are minimal.

EDA1 – Act

I have understood the data loss and I can conclude that the intersected population is not misrepresentative of the full GP Clinical population. With this, I could proceed to investigate the intersected populations Read codes. I will refer to the intersected GP Clinical population as the CPH population.

EDA2 – Investigating Read Codes

EDA2 – Plan

Aim – Build on my knowledge of the top 20 read codes and find Read codes that can be used to determine an individual's health.
Approach – Looking at the top 20 codes by occurrence was not enough information to conclude how useful a code is in the context of this EDA's aim, so I built a script that produced a statistical summary of each Read code. I also built a PowerBi report on this summary that allowed me to quickly transform, drill down and analyse the summarised data. There are two main sets of metrics: one set summarises if the Read code is spread across the population and the other summarises how many times the code appears in individuals' records.

EDA2 - Do

I decided to divide the data up into 4 separate reports, these were:
read_2_val – Read 2 codes that hold values
read_2_noval – Read 2 codes that do not hold values
read_3_val – Read 3 codes that hold values
read_3_noval – Read 3 codes that do not hold values

Appendix 1 is a screenshot of one of the reports, which is broken down into different visuals that produce information based on the two sets of metrics observed.
1) Metric Overview – summarises key metrics tracked for each Read code, this table can be sorted by each metric independently to observe the tops codes depending on the metric.
2) Filter by Description – filters the entire report by specific descriptions.
3) Filter by Read Code – filters the entire report by specific Read Codes.
4) Total Number of Events – displays the total number of events within this section of the report.
5) Number of Different Codes – displays the number of different codes in this section of the report.
6) Percent of Individuals with Code – displays how many people within this section of the report have the code.
7) Distribution of Events by Individuals – visualises the min, max, average and standard deviation of event occurrences. This shows the characteristics of the codes and can provide information on how often a code appears within an individual's events.

EDA2 – Check

To summarise the learnings from the PowerBi reports, I have broken the Check section for each report into a high-level observation, the top 5 when sorting by percentage of individuals with the code and top 5 when sorting by total occurrences.

read_2_val

High level observations

There are 1,509,454 individual events spread over 5,652 different codes with 5,635 different descriptions, there are 3 codes that map to no descriptions and these codes appear in 5 events. From Figure 4 I learnt that most of the codes that take a value are commonly performed tests, this is still seen in this report where most the codes observed are from Read 2 chapters 2 (Examinations/Signs) or 4 (Laboratory Procedures).

Summary by percent of individuals with code

Only 35 codes exist in over 50% of individuals, Table 9 shows the top 5 codes when sorting by Percent of Individuals with Code. I was expecting a lot more codes to exist in most individuals, but from my observations I found that this was not the case.

code_description	read_2	avg_occurrences	pct_inividuals_with_code	total_occurrences
O/E - weight	22A..	6.931172	95.46948	38,267
O/E - height	229..	3.732464	94.17257	20,327
Serum creatinine	44J3.	8.637446	88.56995	44,241
Serum potassium	44I4.	7.957978	87.23846	40,148
Serum sodium	44I5.	7.991277	87.22117	40,308

Table 9: read_2_val top 5 codes by percentage of individuals with code

Summary by total occurrences

Table 10 shows the top 5 codes when sorting by total occurrences, there is a large jump down from code '246..' to code '44J3.'. There are codes that appear to be used much more frequently than others as only 40 codes have over 10,000 occurrences and 160 codes that have over 1,000 occurrences.

code_description	read_2	avg_occurrences	pct_inividuals_with_code	total_occurrences
O/E - blood pressure reading	246..	18.90086	82.67335	90,365
Serum creatinine	44J3.	8.637446	88.56995	44,241
Serum sodium	44I5.	7.991277	87.22117	40,308
Serum potassium	44I4.	7.957978	87.23846	40,148
O/E - weight	22A..	6.931172	95.46948	38,267

Table 10: read_2_val top 5 codes by total occurrences

read_2 codes with no value

High level observations

There are 1,266,999 individual events spread over 17,387 different codes with 16,836 different descriptions. There are 476 codes that map to no descriptions and these codes appear in 8,559 events. I have no reference but based on my understanding of the Read 2 codes I expected to find a much broader range of codes and a lot more codes from the diagnostic chapters, however there is a lot of noise that is polluting the data.

Summary by percent of individuals

Table 11 shows the top 5 codes when sorting by Percent of Individuals with Code, there are a lot of codes that are prevalent in up to 55.76% of the population, but they are either ambiguous or for administrative purposes. When looking over the top 100 codes there are only four that are from the diagnostic chapters.

code_description	read_2	avg_occurrences	pct_inividuals_with_code	total_occurrences
Notes summary on computer	9344	1.157916	85.76377	5,734
Never smoked tobacco	1371.	5.39238	62.27918	19,391
Telephone encounter	9N31.	5.956938	61.53446	21,165
Bowel cancer screening … test normal	686A.	1.4885	56.47731	4,854
Referral for further care	8H...	3.164021	52.37271	9,568

Table 11: read_2_noval top 5 codes by percentage of individuals with code

Summary by total occurrences

Table 12 shows the top 5 codes when sorting by total occurrences. There is very little useful information to be derived here. When looking at the top 100 codes, none are from the diagnostic chapters, and most of the codes come from the administrative chapter.

code_description	read_2	avg_occurrences	pct_inividuals_with_code	total_occurrences
Consultation	9Na..	15.56153	24.48909	22,004
Telephone encounter	9N31.	5.956938	61.53446	21,165
Patient reviewed	6A...	7.169798	48.85694	20,226
Never smoked tobacco	1371.	5.39238	62.27918	19,391
Laboratory test requested	413..	7.27903	40.71701	17,113

Table 12: read_2_noval top 5 codes by total occurrences

read_3 codes with value

High level observations

There are 4,220,775 individual events spread over 1,813 different codes with 1,690 different descriptions, there are 124 codes that map to no descriptions and these codes appear in 14,423 events. There is an increase in the number of codes, and this is to be expected as most individuals within the study use Read 3 codes.

Summary by percent of individuals

Table 13 shows the top 5 codes when sorting by Percent of Individuals with Code, there is a much higher coverage for individuals when comparing to the read_2_val population. Blood Pressure had 82.67% coverage in Read 2 but the equivalent code in Read 3 covers 98.75% of the population. Only 45 codes exist in over 50% of the population and 72 codes that exist in over 20% of the population.

code_description	read_3	avg_occurrences	pct_inividuals_with_code	total_occurrences
O/E - Diastolic BP reading	246A.	20.48303	98.75045	280,044
O/E - Systolic BP reading	2469.	20.52798	98.74323	280,638
O/E - weight	22A..	7.591099	96.72806	101,660
Body mass index - observation	22K..	7.124916	96.44637	95,139
O/E - height	229..	4.802067	96.44637	64,122

Table 13: read_3_val top 5 codes by percentage of individuals with code

Summary by total occurrences

Table 14 shows the top 5 codes when sorting by total occurrences, there is a much higher number of occurrences due to more individuals using Read 3 rather than Read 2. When compared to Table 10, we see similar codes existing, the main difference being that blood pressure is split into two different codes pushing weight to the 6th position. There is also a similar steep drop-off in occurrences from blood pressure to the next code.

code_description	read_3	avg_occurrences	pct_inividuals_with_code	total_occurrences
O/E - Systolic BP reading	2469.	20.52798	98.74323	280,638
O/E - Diastolic BP reading	246A.	20.48303	98.75045	280,044
Serum creatinine level	XE2q5	8.59061	90.15529	107,228
Serum sodium level	XE2q0	8.435257	89.97472	105,078
Serum potassium level	XE2pz	8.42973	89.93861	104,967

Table 14: read_3_val top 5 codes by total occurrences

read_3 codes with no value

High level observations

There are 3,127,843 individual events spread over 36,582 different codes with 34,567 different descriptions. And there are 2,011 codes that map to no descriptions and these codes appear in 174,513 events. It is a similar situation to the read_2_noval, but it is more difficult to see the type of code due to read_3 not showing the code hierarchy within the code itself.

Summary by percent of individuals

Table 15 shows the top 5 codes when sorting by Percent of Individuals with Code, similarly to Table 11 there are a number of codes with high prevalence in the population but they are not useful.

code_description	read_3	avg_occurrences	pct_inividuals_with_code	total_occurrences
Notes summary on computer	9344.	1.186646	87.10745	14680
Patient allocated named accountable general practitioner	XacWQ	1.085318	80.71398	12441
Liver function tests	X77WP	6.206093	80.6647	71097
Never smoked tobacco	XE0oh	5.476217	74.46134	57911
Urea and electrolytes	X77Wi	7.123951	73.84875	74716

Table 15: read_3_noval top 5 codes by percentage of individuals with code

Summary by total occurrences

Table 16 shows the top 5 codes when sorting by total occurrences, similarly to Table 12, it does not provide any useful information apart from there are a lot of codes that are not useful.

code_description	read_3	avg_occurrences	pct_inividuals_with_code	total_occurrences
Urea and electrolytes	X77Wi	7.123951	73.84875	74,716
Liver function tests	X77WP	6.206093	80.6647	71,097
Medication review done	XaF8d	6.997633	65.45557	65,050
Never smoked tobacco	XE0oh	5.476217	74.46134	57,911
Blood sample taken	XaEJK	5.458204	52.30953	40,549

Table 16: read_3_noval top 5 codes by total occurrences

EDA2 – Overall Summary

The EDA emphasised that there is a 4-way split within the codes: one split between Read 2 & Read 3, and a second split between codes that hold values and codes that do not. It revealed that there were a lot of codes that did not map to a description, this is mainly down to Read 3 codes beginning with a Y having no description mapping and most of the other codes that do not map begin with an X. A code that does not have a mapping to a description means I have no way of knowing what that code represents, and therefore it is worthless, and I should consider removing it.

When looking at no-value codes, there was a lot of codes that were not useful and were polluting the underlying diagnostic information I needed to look at. When looking at Read 2 no-value codes it was clear to see what type of codes were polluting the picture by looking at the chapter character. However, when looking at Read 3 codes it is difficult to understand what the code is as there is no hierarchical relationship displayed within the code itself. I noticed some no-value codes that I would expect to have values. I investigated this and found that Read 2 codes like 242.. (Pulse Rate) had both entries with and without values, I believe this is the issue I touched on in this In the GP Clinical duplicate analysis section where value codes would be duplicated in a way where there would be an event without a value and another event with a value.

When looking at value codes, some codes had good coverage across individuals, especially for individuals using Read 3, but as I am not a medical professional it is difficult for me to understand what the test is, what is it testing for and what is a 'good' result. There were very few tests that had a high coverage, high average occurrences and a low standard deviation, meaning that some tests exist that were consistently performed on the population.

EDA2 – Act

I was successful in building on my prior knowledge; however, I did not identify any codes that I could use to determine an individual's health. Even with the summary metrics there was too much data and information to analyse, there are 17,997 different Read 2 codes and 38,209 different Read 3 codes in the CPH population, most of which was not useful and simply polluting the picture.

EDA3 – Investigating Read Chapters

The Read 2 codes have a hierarchical structure that is represented within the code, and the starting character is what chapter the code belongs to. This chapter will give me a high-level understanding of the code. The major issue is Read 3 codes do not have this hierarchical structure represented within the displayed code.

EDA3 – Plan

Aim – Map Read 3 codes to Read 2, and analyse the events by chapter in the CPH population.

EDA3 – Do

Mapping read_3 codes to read_2

The mapping file that was supplied by the UKBiobank is not engineered for large scale data mapping and needed to be cleaned to make this possible. The main issue was the file had a many-to-many relationship. This duplication was caused by synonym codes, which I highlighted in Understanding Read Codes. To solve the duplication in Read 3 I dropped all synonym codes which created a one-to-many relationship and kept the same number of distinct Read 3 codes.

Producing Metrics & visuals

I repurposed code from EDA2 to produce similar metrics but across chapters rather than individual codes, and I used python to plot bar charts of the chapters to compare occurrences.

EDA3 – Check

Table 17 is a summary of each chapter and Figure 15 is a visual of the total occurrences of codes. From these, it is clear to see that the most popular codes have numeric chapter codes '0-9' with the most popular chapter being '4'.

Figure 15 also shows that there have been 134,033 events under chapter '_' that have been mapped across as '_NONE' or '_DRUG'. I expected some Read 3 codes not to be mapped, I investigated this and found that 83.19% of '_NONE' codes came from Read 3 codes that began with an 'X'. I believe that codes that map to '_NONE' to be the result of the Read 3 hierarchy having an unlimited depth, I have observed that a detailed Read 3 code often starts with an 'X' suggesting that once the code depth reaches 5 it rewrites the character from the 1-byte position. The total loss of information from these codes is only 1.32% of the total codes in the CPH Population.

The event occurrences in chapters '.', '-', 'X' and 'Y' equate 302,376, this is 2.98% of the total CPH population.

Chapter	Individuals with code	% Individuals with code	Occurrences	Avg Occurrences
Z	6,304	31.75499	20,665	3.278077
Y	12,811	64.53254	168,085	13.12037
X	1	0.005037	1	1
U	571	2.876285	714	1.250438
T	3,713	18.70341	5,256	1.415567
S	10,444	52.60931	23,517	2.251723
R	8,883	44.74612	23,692	2.667117
Q	96	0.483578	102	1.0625
P	866	4.362281	1,099	1.269053
N	14,624	73.66512	71,189	4.867957
M	12,760	64.27564	44,515	3.488636
L	2,854	14.37639	5,382	1.885774
K	10,097	50.86137	31,726	3.142121
J	9,094	45.80899	25,108	2.760941
H	11,538	58.12009	43,961	3.810106
G	11,346	57.15293	50,680	4.466772
F	12,449	62.70905	43,134	3.464857
E	5,635	28.38505	15,959	2.832121
D	1,476	7.435019	2,526	1.711382
C	6,609	33.29136	22,535	3.409744
B	7,207	36.30365	14,914	2.069377
A	9,209	46.38827	20,747	2.252905
_	13,192	66.45174	134,033	10.16017
.	169	0.8513	257	1.52071
9	19,694	99.20411	830,373	42.16376
8	19,316	97.30002	553,941	28.67783
7	17,983	90.58533	140,799	7.829561
6	19,340	97.42091	647,865	33.49871
5	15,042	75.7707	77,150	5.128972
4	19,300	97.21942	4,945,710	256.2544

3	16,674	83.99154	136,647	8.195214
2	19,538	98.4183	1,388,026	71.04238
1	19,686	99.16381	636,217	32.31825
0	4,874	24.55168	7,105	1.457735

Table 17: Chapter metric summary

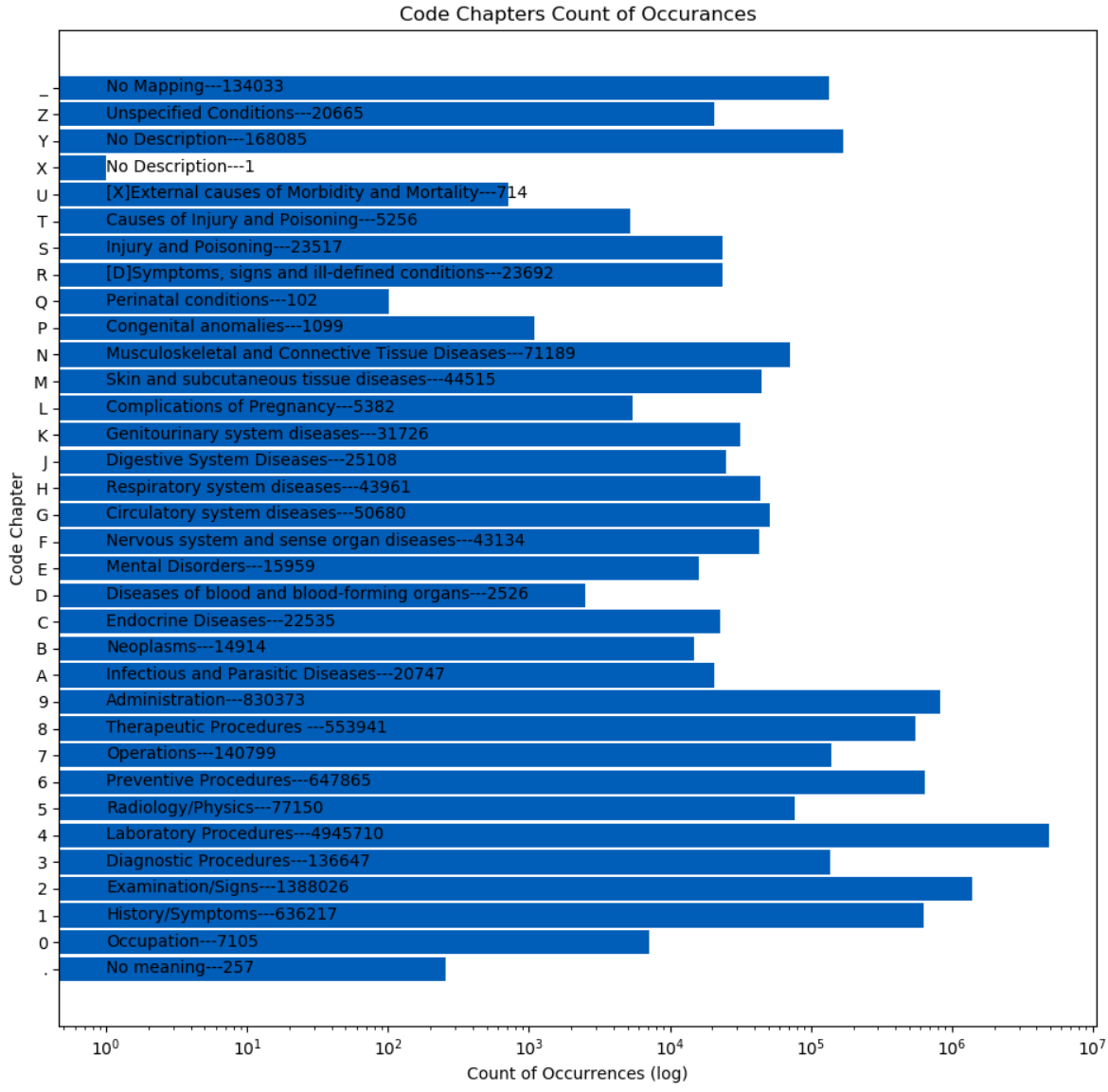


Figure 15: Chapter Occurrence Summary

EDA3 – Act

I was successful in mapping Read 3 to Read 2 with minimal loss of data, I was also successful in analysing events by their chapter. However, these chapters are too high level to determine if the event is enough to impact an individual's overall health.

EDA4 – Exploring Read Codes to determine optimal byte depth

Building on EDA3, this cycle explores the individual chapters by determining what chapters can be excluded, and for the included chapters, what byte depth to truncate the codes at.

EDA4 – Plan

Aim – Evaluate each chapter and determine if it should be included, if so at what byte depth. With this, summarise the events with the truncated codes within the CPH population.

EDA4 – Do

Chapters to be excluded

0 – Occupation

Solely describes an individual's occupation, Table 17 shows this chapter exists in 24.55% of individuals at an average occurrence of 1.46. Arguably an individual's occupation could determine their activity because someone with a desk job will be less active than a personal trainer, it can also generalise their health as some people could have jobs that impact their health, such as a miner. However, I am dropping this code due to the low percentage of individuals who have it, and the large number of people who will be retired at the time of the reading.

1 – History/Symptoms

Is a broad chapter that covers a vast number of different topics surrounding medical history, Table 17 shows this exists in 99.16% of individuals at an average occurrence of 31.32. This chapter is too granular and tells me information on the individuals past rather than their current situation.

5 – Radiology/Medical Physics

Holds codes surrounding X-rays, ultrasounds and other medical screenings, Table 17 shows this chapter exists in 75.77% of individuals at an average occurrence of 5.13. Most of the procedures in this chapter are difficult to understand as a non-medical professional and the tests I do understand, like X-rays, go into such granular detail they would only be useful at a 4/5-byte depth.

8 – Other therapeutic procedures

Is a broad chapter that covers Post Operation Monitoring to Physiotherapy. Table 17 shows this chapter exists in 97.30% of individuals at an average occurrence of 28.68. I investigated this and found that Medication Review (8B3V.) accounts for 30% of the entries in this whole chapter.

9 – Administration

Hold all general and medical administrative information, Table 17 shows this chapter exists in 99.20% of individuals at an average occurrence of 42.16.

L – Complications of pregnancy, childbirth and the puerperium

Describes conditions solely related to pregnancy, Table 17 shows this exists in 14.37% of individuals at an average occurrence of 1.89. Given the age of the individuals of this study, I will not take this code.

P – Congenital anomalies

Are conditions present from birth such as Heart Defects or Cleft Lips, Table 17 shows this exists in 4.36% of individuals at an average occurrence of 1.27. Given the age of the individuals of this study, I will not take this code.

Q – Perinatal conditions
Describes conditions immediately before or after giving birth, Table 17 shows this exists in 0.48% of individuals at an average occurrence of 1.06. Given the age of the individuals of this study, I will not take this code.

R – [D]Symptoms, signs and ill-defined conditions
Stores vague symptoms a medical practitioner is unsure of a condition and wants to submit symptoms into the system, Table 17 shows this exists in 44.75% of individuals at an average occurrence of 2.67.

T – Causes of injury
Is associated with chapter S, where this chapter holds the reason behind an entry. Table 17 shows this exists in 18.7% of individuals at an average occurrence of 1.42.

U – [X]External causes of morbidity and mortality
This chapters stores how an individual died based causes outside of the medical field, such as Car Crash, Table 17 shows this exists in 2.88% of individuals at an average occurrence of 1.25.

Z – Unspecified conditions
Is an assortment of codes that add additional information, such as Normal Pregnancy, Table 17 shows this exists in 31.75% of individuals at an average occurrence of 3.28.

Chapters to Include

3 – Diagnostic Procedures
Is a broad chapter that covers a vast range of medical procedures or tests, Table 17 shows this chapter exists in 83.99% of individuals at an average occurrence of 8.2. However, it is difficult to understand why the procedure has been performed but there is one subchapter that describes tests for disabilities, I am not concerned about the outcomes as if someone is being tested I assume they have been observed to have physical ailment:

- Disability assessment – Physical (39***)

6 – Preventative Procedures
Can be vaccines, counselling and stroke prevention, Table 17 shows this exists in 97.42% of individuals at an average occurrence of 33.5. I took these subchapters:

- Chronic disease – general (661**)
- Cardiac disease monitoring (662**)
- Respiratory disease monitoring (663**)
- Gout monitoring (669**)
- Diabetic monitoring (66A**)
- Obesity monitoring (66C**)
- Alcohol disorder monitoring (66e**)
- Cardiovascular disease monitoring (66f**)
- Arthritis monitoring (66H**)
- Chronic pain review (66n**)
- Further diabetic monitoring (66o**)
- Warfarin monitoring (66Q**)
- Lipid disorder monitoring (66X**)
- Primary Prevention ischaemic heart disease (6C***)
- Stroke prevention (6F***)

7 – Operations & Procedures

Any operation or procedures performed will be in this chapter. Table 17 shows this chapter exists in 90.58% of individuals at an average occurrence of 7.82. Operation & Procedures can be anything from clearing an ear canal or a full lung transplant. Separating the two extremes is not possible at a 2/3-byte depth as hierarchy is separated in this way, this is seen in the subchapter Lung and Mediastinum Operations (745**) where in this subtree can have codes for Lung Transplant & Nicotine Replacement Therapy using Gum. My solution to this was to take some codes at a 2-byte depth and classify the possible severity with (High {H}, Medium {M}, Low {L}) depending on the potential procedures impacting an individual's health or their ability to be active.

- Nervous system operations (70***) {H}
- Endocrine system and breast operations (71***) {H}
- Eye Operations (72***) {M}
- Ear operations (73***) {L}
- Respiratory tract operations (74***) {H}
- Mouth Operations (75***) {M}
- Upper digestive tract operations (76***) {M}
- Lower digestive tract operations (77***) {H}
- Other abdominal organ operations (78***) {H}
- Heart Operations (79***) {H}
- Artery and Vein Operations (7A***) {H}
- Urinary operations (7B***) {M}
- Male genital organ operations (7C***) {H}
- Lower female genital tract operations (7D***) {H}
- Upper female genital tract operations (7E***) {H}
- Obstetric operations/Pregnancy operations (7F***) {H}
- Skin Operations (7G***) {M}
- Soft Tissue operation (7H***) {H}
- Skull or spine, bone and joint operations (7J***) {H}
- Other bone and joint operations (7K***) {H}
- Miscellaneous operations (7L***) {H}
- Diagnostic imaging, testing and rehabilitation (7P***) {L}

A – Infection and Parasitic Diseases

Is the diagnostic chapter for infectious and parasitic diseases like Syphilis, Table 17 shows this exists in 46.38% of individuals at an average occurrence of 2.25. This is a diagnostic chapter, and I took all subchapters at a 2-byte depth.

B – Neoplasms (Cancers)

Is the chapter for different types of cancer related diagnosis, Table 17 shows this exists in 36.3% of individuals at an average occurrence of 2.07. I took this code at a 2-byte depth as I can differentiate the cancerous neoplasms from the benign neoplasms.

C – Endocrine, nutritional, metabolic and immunity disorders

Is a large and diverse chapter holding large amounts codes for diabetes and vitamin deficiencies, Table 17 shows this exists in 33.29% of individuals at an average occurrence of 3.41. This chapter is extremely technical and granular, from what I understood most of the key differentials are done at 5-byte depth, such as differentiating complications with Type1 and Type 2 diabetes. I took this chapter at a 3-byte depth and acknowledged that these subchapters are still diverse.

D – Diseases of blood and blood-forming organs
Is where conditions like Anemia are held, Table 17 shows this exists in 7.44% of individuals at an average occurrence of 1.71. I took all subchapters at a 2-bytes depth.

E – Mental disorders
Conditions such as Psychosis or Learning Disabilities, Table 17 shows this exists in 28.38% of individuals at an average occurrence of 2.83. I took all subchapters at a 2-bytes depth.

F – Nervous system and sense organ diseases
Holds information on Meningitis related conditions and sense organ diseases such as Deafness, Table 17 shows this exists in 62.71% of individuals at an average occurrence of 3.46. I took all subchapters at a 2-bytes depth.

G – Circulatory system diseases (Cardiac/Heart Diseases)
Holds codes relating to the Heart or Circulation of Blood, Table 17 shows this exists in 57.15% of individuals at an average occurrence of 4.67. I took all subchapters at a 2-bytes depth.

H – Respiratory system diseases
Holds conditions such as Pneumonia and Lung Diseases, Table 17 shows this exists in 58.12% of individuals at an average occurrence of 3.81. I took all subchapters at a 2-bytes depth.

J – Digestive system diseases
Holds conditions relating to the Mouth, Stomach and Lower Digestive Tract, Table 17 shows this exists in 45.81% of individuals at an average occurrence of 2.76. I took all subchapters at a 2-bytes depth.

K – Genitourinary system diseases
Holds conditions relating to the Genital and Urinary Glands, Table 17 shows this exists in 50.86% of individuals at an average occurrence of 3.14. I investigated and found that most codes were for UTI's and conditions relating to the female anatomy. I took all subchapters at a 2-bytes depth as you can differentiate between male, female and common codes.

M – Skin and subcutaneous tissue diseases
Describes the conditions such as Sore Skin or Ulcers, Table 17 shows this exists in 64.28% of individuals at an average occurrence of 3.49. The codes in the chapter are mostly mild so I took this chapter at a 1-byte depth.

N – Musculoskeletal and connective tissue diseases
Holds conditions such as Arthritis, Table 17 shows this exists in 73.67% of individuals at an average occurrence of 4.89. I took all subchapters at a 2-bytes depth.

S – Injury and poisoning
Holds codes for injuries such as Broken Bones and Overdoses, Table 17 shows this exists in 52.61% of individuals at an average occurrence of 2.51. This chapter has a large range of severity between the codes, from Fractured Skull to Wasp Sting. I took all subchapters at a 2-bytes depth as I can differentiate the severity of the subchapters at this level.

Chapters to consider

2 – Examinations / Signs

Holds a lot of value-codes, Table 17 shows this exists in 98.42% of individuals at an average occurrence of 71.04. There is a lot of information that can be derived in this chapter that would be useful, but I do not know what a good value or exactly what the test is, so I excluded this chapter.

4 – Laboratory Procedures

Is the where codes for any laboratory tests are held, Table 17 shows that this chapter has the most occurrences and it exists in 97.22% of individuals at an average occurrence of 256.25. This chapter has a wealth of information, but I do not know what a good value is, so I excluded this chapter.

Code Extraction

The best way I found to extract the codes was to divide the chapters into sections by what byte depth to take. For chapters that had multiple byte depths or did not use all the codes, I divided these sections independently by selected codes or different byte depths.

EDA4 – Check

Table 18 shows the impact of removing and truncating chapters. There is a small reduction in the number of individuals, this is to be expected. There is a drastic reduction in the number of events, this is partially down to the amount of information I deemed not useful, but this is mostly down to the exclusion of chapters 2 & 4 which have the highest number of occurrences. Most notably, I have reduced the number of different codes from 28,109 to 215. This is a significant reduction in the amount of information needed to process.

	CPH Population Full	**CPH Population Truncated**	**% Change**
Individuals	19,852	19,617	-1.18%
events	10,133,630	702,955	-93.06%
Unique Codes	28,109	215	-99.24%

Table 18: Analysis of truncating Read codes for GP Clinical

EDA4 - Act

I have been successful in selecting the chapters to include and at what byte depth, I have also been successful in evaluating events based on my truncated codes. However, 215 different codes is still a lot of information, and I am looking for a concise feature to determine an individual's overall health.

EDA5 – Final Iteration: Engineer the feature Health Impact Score

EDA5 – Plan

Aim – Assign a code impact score to each of the 215 different Read 2 codes, take events within a timeframe surrounding an individual's activity timestamp and then calculate the individual's health impact score based on this.

EDA5 – Do

Timeframe

I decided 6 months before the activity reading and 1 month after.

Assigning a code impact score

This needed to be numerical and reflect that minor events will occur more frequently than major events. Like my approach to chapter 7 in EDA4, I will use a tiered system to numerically classify the event based on impact to health, High {100}, Medium {10} and Low {1}. I will take the average of these scores to normalise the effects individuals having larger numbers of events from High Impacts chapters. With the higher the value, the less likely they are to be active.

Individuals with no events surrounding the activity reading

Some individuals had no events within my specified time frame. I will assign these individuals with a 0 health impact score and assume they have nothing medically wrong with them.

EDA5 – Check

Table 19 shows the impact of taking codes surrounding individual's activity tracker readings. As expected, there was a reduction of individuals and events, there was also a reduction in unique codes, meaning I could have assigned impacts scores to 171 codes instead of 215.

	CPH Population Truncated	**Events Surrounding Activity Tracker**	**% Change**
Individuals	19,617	8,833	-54.97%
events	702,955	27,737	-96.1%
Unique Codes	215	171	-20.47%

Table 19: Analysis of taking events surrounding individual's activity tracker readings

Table 20 shows the number of individuals within the score ranges. The distribution is what I expected, around 75% of individuals do not have an event with a High impact score.

Individuals Health Impact Score Range	**Number of individuals in range**
0	11,019
>0 to <=10	3,468
>10 to <=50	2,010
>50	3,355

Table 20: Health Impact Score Ranges

Using this feature to identify individuals who are unable to be active due to other ailments can be done by excluding individuals over a certain threshold. Table 21 shows the impact of excluding individuals with a health impact score over 10, which means they most likely will not have a High impact event.

Health Classification	CPH Population	Health Impact Score > 10	% Excluded
Inconclusive	0	0	N/A
Healthy	9,695	1,783	18.39%
CVD	9,190	3,158	34.36%
T2D	276	128	46.38%
Both CVD & T2D	691	296	42.84%
TOTAL:	**19,852**	**5,365**	**27.02%**

Table 21: Example of excluding individuals

EDA5 – Act

I was successful in achieving this cycles aim, and I believe I have built a clearer picture surrounding an individual's high-resolution accelerometer reading by engineering the feature health impact score.

Overall Evaluation

In this project, I set out to explore the and understand the datasets provided by the UKBiobank, and then I focussed on building a clearer picture surrounding individuals high-resolution accelerometer readings. I have achieved this with varying levels of success, and I will use this section to explore this in greater detail.

Objective 1 – Learn the relevant tools and techniques used by data scientists to explore and produce insight from data

Rocket HPC – This project was a fantastic opportunity to use a remote cluster computer to run scripts, it was a steep learning curve which involved a lot of trial & error with setting how many CPU's, how much RAM and the best node to allocate jobs too.

Python – All my scripts were written in Python, I found it to be an extremely powerful language with simple syntax. Most of my knowledge gained centres around the Pandas library.

Efficiency – Some of the data files I was manipulating were over 5GB in .TXT format, so even on an extremely powerful machine I needed to code efficiently to avoid running out of memory or to complete the job in a reasonable time.

PowerBi – I have had a lot of experience of using PowerBi before this project and I had fully intended to use it a lot more. However, I encountered an issue where PowerBi would randomly change the cases of the Read codes [37], meaning I couldn't build true reports that pivoted off Read codes. In EDA2 I did find a compromise, where everything was pivoted off the code descriptions rather than the actual code, there were some codes that mapped to the same description, but I believed this to be acceptable.

Excel – There were a lot of 'low-hanging fruits' in this project where I could utilise Excel, especially when I was working on the mapping documents. I also found Power Query to be exceptionally powerful when cleaning the mapping document and even for detecting dataset encodings.

From my research into data science, I learnt that a data scientists role is to produce insights from data and how they achieve this is irrelevant, so I haven't learnt the entire toolbox just the tools that have helped me derive insights from the data on this project.

Objective 2 – Understand the features within the datasets and what information they provide

In this paper I have included the relevant information on the datasets format and type of information in enough detail to understand the specific problem I set out to solve. This is because of the sheer amount of information and time it takes to investigate these datasets was a lot more than I expected, so I had to settle for relevance over a full comprehensive report on these datasets. I did underestimate the size of this task and I believe this is mainly down to inexperience, but I did deliver more than the minimum viable product in deriving enough information in order to solve a specific problem within the data.

Objective 3 – Perform data quality analysis do ensure that the data is cleaned and ready to be used

I performed a full data quality analyses on the Activity Tracker Reading and the features I used from Baseline, but this was because they were relatively simple and straightforward. I did not finish the quality analysis for GP Clinical, specifically on the extended problem of duplicates because I faced lengthy investigative times and there were more quality issues than I anticipated. However, I believe the issues discovered were dealt with sufficiently to ensure the integrity of any insights derived from this dataset.

Objective 4 – Building a clearer picture surrounding individuals high resolution activity readings by using iterative EDA techniques

I delivered a feature that assigned a score to individuals that can be used to understand their overall health surrounding a point in time, and from that if they are healthy enough to be active.

Method Evaluation

Truncating Read Codes – EDA2 emphasised that there was too much information to analyse, my method of truncating the codes reduced the different number of codes from 28,109 to 215, therefore making analysis of the population possible. I believe this method of converting to Read 2 and truncating the byte depth for better analysis will be transferable to other aspects of the UKBiobank project. The amount of information is massive and often too medically specific, so a lot of analysis is better done with codes truncated at 2-3 bytes. I was able to apply this method to produce EHR chapter summaries for participants in the UKBiobank that tested positive for Covid-19, see Appendix 2.

Time Surrounding Accelerometer Reading – I decided to take codes 6 months before the reading because this would capture any conditions prior that might affect an individual's activity, and 1 month after so anything that had not been diagnosed at the time but could still affect an individual's health would be accounted for. However, this is arbitrary as every unique condition would be better represented with its own timeframe, but this would require a lot of time and effort to complete.

Code Impact Score – This score is entirely pivoted off my medical knowledge, which is limited. The 215 codes were derived from EDA4, the code impact scores assigned to these codes are what drive the health impact score for individuals, and they need to be classified correctly to ensure accuracy.

Health Impact Score Calculation – To calculate the health impact scores I chose to average each events code impact score; upon reflection I believe this to be flawed. It does not fully reflect the occurrences of codes, if an individual has 1 High risk code and 100 Low risk codes, the average score would be 1.98, which is lower than an individual with 1 High impact event. Also, someone with 100 Low impact events will have the same health impact score as someone with 1 Low impact event. I believe this should either have a more complex aggregation technique or a much simpler binary classifier to indicate if an individual has a High impact code or not.

Health Impact Score Evaluation

What is healthy? – the WHO defines healthy as "Health is a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity" [38]. This definition is very difficult to capture with a single value built off GP Clinical records as: diagnostic events don't capture severity, there is no information about social well-being, mental health is unique to an individual and a simple encoding cannot relay this information, but most importantly I believe 'health' is individual to a person and cannot be classified on a population level perfectly.

Activity – Health impact score will be used to identify individuals who are too unwell to be active, however not being ill does not correlate to being active. Some individuals will have a less-active lifestyle than others, even if they have a severe ailment.

Other Data

I had to limit myself to the GP Clinical dataset due to time and scope constraints. However, there is a lot more data that can be incorporated into the health impact score to improve it.

Objective 5 – Ethical reflection

There is currently around 40 zettabytes of data that exists, and current trends from 2010 show that this value is doubling every year [39]. With this exponential growth in data and the speed at which computing methodologies & technologies are advancing to derive information, there is a real cry for the ethics of data-driven projects to be scrutinised. The Cambridge Analytica scandal is one of the biggest examples of data misuses in history, where millions of Facebook users unknowingly had their data sold. This data was then used to create personality profiles on American citizens which were then used to target ads for Trump's 2016 presidential bid and arguably helped him secure office [40].

When I reflect on engineering the feature 'health impact score', I am only reflecting on what I have created and how I am going to use it, when in reality this feature can be used for many different purposes. Uber is a data-driven tech company, it has many different uses for data in order to improve its service to drivers and riders. A program called Violation Of Terms Of Service (VTOS) identifies users who are misusing the service, when it identifies users in breach of VTOS it displays a fake UI with fake cars on their device and they will never be able to hail a taxi. This is good as drivers will not get harassed and legitimate users will have a better chance of getting a car, however this program was also being used to identify and avoid law enforcement officers in cities where Uber was non-complaint [41].

Uber and the VTOS program acts as an example of why I should also reflect on the root functionality of what this feature does and speculate its potential uses, this should help identify any unintentional ethical impacts. A potential use would be automation within the welfare system, people who are applying for disability would need to meet the threshold of this datapoint. Automation within systems such as welfare have huge ethical questions associated with it, and because of this there is potential for data driven products to be produced for such systems but under different pretences to avoid these complex ethical questions.

This is an extreme example, but I believe this line of thinking belongs in the project proposal/ethics stage to prevent projects kicked off under False Flags. Newcastle University is a trusted institution with an independent ethics approval process, and I am confident that work would not be re-produced for morally questionable reasons, but I do not have the same confidence in corporate organisations such as Uber. Although there are well-publicised data governance laws and even a code of ethics for working with health data in the UK [42], with the ever-growing amount of data collected I believe there needs to be more oversight to ensure nothing unethical is being conducted. This oversight should be independent of political, public and private organisations to ensure that unethical data usage does not take place, as I highly doubt the engineers who created VTOS were fully aware of its morally questionable intended purpose.

Conclusion

One of the main learnings from this project has been what data science is, when selecting it as a theme I very much assumed that it was a standalone field within computer science. But only after starting this project and immersing myself into the data science community did I really learn how it is a intersect between Computer Science, Maths and Domain of Work, which is Medicine in this project Figure 16.

The aim of this project was to understand individuals' medical events within the UKBiobank study by deriving insights and knowledge from medical datasets, with this understanding focus on a specific problem to solve using iterative exploratory data analysis. My main shortcomings were my lack of medical knowledge and my inexperience when handling big data, because of this, I did not produce as much as I initially intended. But overall, I believe I have been successful in fulfilling the aim.

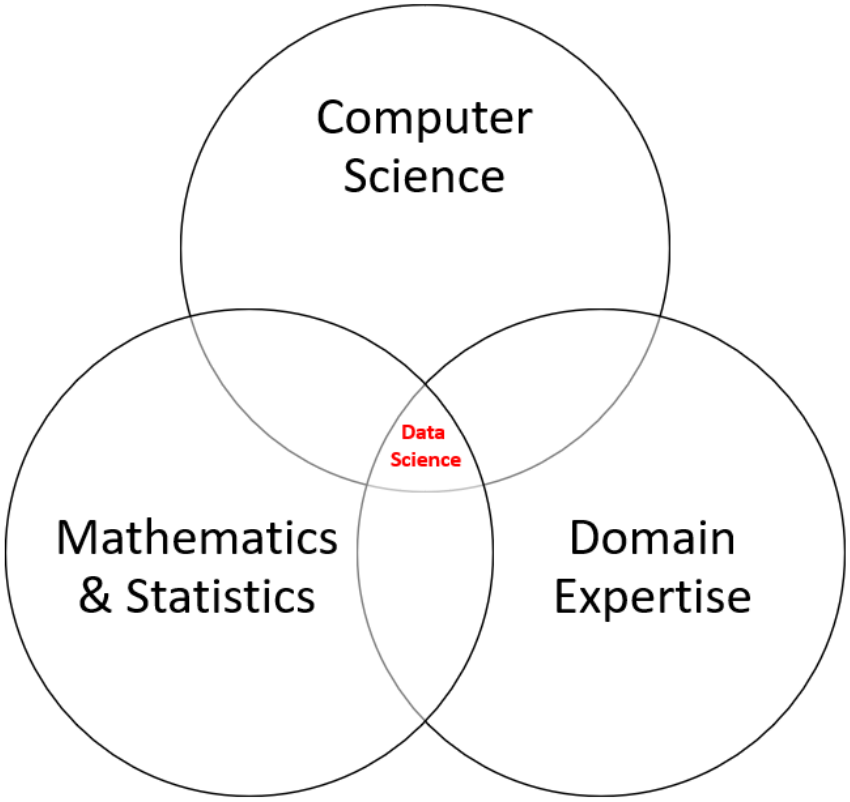


Figure 16: Data Science Discipline

What I would do differently

The experience gained from this project has given me an appreciation of the challenges of working with big data from a specialist field, with the experience I have gained I would not repeat this project in the same manner.

Approach – I had focussed a lot of time and effort into understanding all of the data, I do not think there is anything wrong with this and understanding data is extremely important. However, I would try and focus my efforts towards a specific problem much sooner as I only utilised part of my overall knowledge base of the data. To facilitate this, I would:

- Perform a much quicker analysis of the datasets just to give me a high-level textual understanding of what data they can provide.
- Use a methodology like CRISP-DM from the start of the project life cycle, doing this would assist in focussing on the problem to solve.
- Have a more structured data clean-up.

Setting realistic targets – When selecting the Data Science theme, I did some initial research and I learnt that data scientists build machine learning models to solve real-world problems. However, I got fixated on this and a lot of my early ideas involved machine learning despite there being no need for such an implementation. Now knowing that there is more to data science than building machine learning models I would:

- Look at the problem and evaluate the different possible options I have to solve it.
- I believe the solution I worked towards was the right one for the problem, but I would set clearer targets to evaluate success.

Future work

Dataset understanding and quality analysis – There are many more datasets that the UKBiobank has supplied to Newcastle University, I would analyse and understand these to see if they can improve my solution. Specifically relating to my data quality analysis on GP Clinical:

- Finishing the duplicate analysis, I would perform a full study on duplication across features excluding value fields. Doing this will conclude whether this is an issue or not.
- Top 20 codes per code chapter would be a good metric to know, with this I could then see what the most popular codes are within each code chapter which will add more context to chapter summaries.

Health impact score – I believe my overall method to be good, but there is some work that needs to be done to ensure and improve the accuracy.

- Have a medical professional review the chapter I have included and at what byte depth.
- Have a medical professional review the code impact scores I have assigned.
- Deal with events that have allocated dates in 1900, 1920 and 2017. These are anomalous and have been assigned these dates arbitrarily, but they could have occurred at any point in the individual's timeline, I would look for codes that would impact an individual throughout their lifetime, such as being in a wheelchair.
- Review the time frame of taking events, when classifying each codes impact score, I would also specify how long before the activity tracker reading to look for these events. This is because different illnesses have a different impact on the body.
- Once the activity outcomes have been calculated, compare it to the health impact score. I would look for individuals who are highly active, but my score is predicting them to be unable to be active. with this re-evaluate how effective the health impact score is.
- Incorporate information from chapters 2 & 4, these hold biomarkers such as BMI which can be used to get a clearer picture on an individual's health leading up to the accelerometer reading. Specifically, there would be benefit from learning if an individual's BMI is improving or worsening leading up their activity tracker reading.
- Incorporate more information from baseline, such as age.
- Include GP Prescriptions by lining up prescription events with clinical events. I would do this by using the type of drug prescribed and the dosage to gauge condition severity.

Bibliography

[1] N. Triggle, “10 charts that show why the NHS is in trouble,” BBC, 24 05 2018. [Online]. Available: https://www.bbc.co.uk/news/health-42572110. [Accessed 27 04 2020].

[2] C. Ham, J. Dixon and N. Edwards, “Call to strengthen NHS finances: letter to the Editor,” Kings Fund, 27 06 2017. [Online]. Available: https://www.kingsfund.org.uk/publications/articles/call-strengthen-nhs-finances. [Accessed 27 04 2020].

[3] Health Catalyst, “Predictive Analytics Solutions,” Health Catalyst, [Online]. Available: https://www.healthcatalyst.com/predictive-analytics. [Accessed 27 04 2020].

[4] UKBiobank, “About UK Biobank,” UKBiobank, [Online]. Available: http://www.ukbiobank.ac.uk/about-biobank-uk/. [Accessed 27 04 2020].

[5] NHS, “Cardiovascular disease (CVD),” NHS, [Online]. Available: https://www.england.nhs.uk/ourwork/clinical-policy/cvd/. [Accessed 27 04 2020].

[6] J. Waterall, “Health Matters: Preventing cardiovascular disease,” Public Health England, 14 02 2019. [Online]. Available: https://publichealthmatters.blog.gov.uk/2019/02/14/health-matters-preventing-cardiovascular-disease/. [Accessed 27 04 2020].

[7] NHS, “NHS Diabetes Prevention Programme (NHS DPP),” NHS, [Online]. Available: https://www.england.nhs.uk/diabetes/diabetes-prevention/. [Accessed 27 04 2020].

[8] Full Fact, “Spending on the NHS in England,” Full Fact, 09 07 2019. [Online]. Available: https://fullfact.org/health/spending-english-nhs/. [Accessed 27 04 2020].

[9] U. Fayyad, G. Piatetsky-Shapiro and . P. Smyth, “From Data Mining to,” *AI Magazine,* vol. 17, no. 3, pp. 37-54, 1996.

[10] W. S. Cleveland, “Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics,” *International Statistical Review,* vol. 69, no. 1, pp. 21-26, 2001.

[11] l. Meazzini, “What is really data about?,” Towards Data Science, 15 05 2019. [Online]. Available: https://towardsdatascience.com/what-is-really-data-about-a60a2af1cfaa. [Accessed 27 04 2020].

[12] Journal of Data Science, “About JDS,” Journal of Data Science, [Online]. Available: http://www.jds-online.com/about. [Accessed 27 04 2020].

[13] M. Rogati, “The AI Hierarchy of Needs,” Medium, 01 08 2017. [Online]. Available: https://medium.com/hackernoon/the-ai-hierarchy-of-needs-18f111fcc007 . [Accessed 27 04 2020].

[14] Guru99, “R Vs Python: What’s the Difference?,” Guru99, [Online]. Available: https://www.guru99.com/r-vs-python.html. [Accessed 27 04 2020].

[15] C. Willyard, “Can AI Fix Medical Records?,” *Nature,* vol. 576, pp. 59-62, 18 12 2019.

[16] UKBiobank, “Data-Field 42040 - GP Clinical event records,” UKBiobank, [Online]. Available: http://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=42040. [Accessed 29 04 2020].

[17] UKBiobank, “Baseline assessment 2006-2010,” UKBiobank, 25 02 2019. [Online]. Available: https://www.ukbiobank.ac.uk/background-to-the-project/. [Accessed 27 04 2020].

[18] Newcastle Univeristy, “Open source, open data, Open Lab,” Newcastle Univeristy, 06 02 2017. [Online]. Available: https://www.ncl.ac.uk/press/articles/archive/2017/02/ukbiobank/] . [Accessed 27 04 2020].

[19] UKBiobank, “General Practice,” UKBiobank, [Online]. Available: https://www.ukbiobank.ac.uk/general-practice/. [Accessed 27 04 2020].

[20] UKBiobank, “Category 3000 - Primary Care,” UKBiobank, [Online]. Available: http://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=3000. [Accessed 29 04 2020].

[21] NHS Providers, “The NHS provider sector,” NHS Providers, [Online]. Available: https://nhsproviders.org/topics/delivery-and-performance/the-nhs-provider-sector. [Accessed 27 04 2020].

[22] L. Marcinowicz, T. Pawlikowska and M. Oleszczyk, “What do older people value when they visit their general practitioner? A qualitative study,” *European Journal of Ageing,* vol. 11, no. 4, p. 361–367, 2014.

[23] UKBiobank, “Primary Care Linked Data,” 09 2019. [Online]. Available: http://biobank.ndph.ox.ac.uk/showcase/showcase/docs/primary_care_data.pdf. [Accessed 27 04 2020].

[24] T. Benson, “The history of the Read codes: the inaugural James Read Memorial Lecture 2011,” *Journal of Innovation in Health Informatics,* vol. 19, no. 3, pp. 173-182, 2011.

[25] T. Bentley, C. Price and P. Brown, “Structural and lexical features of successive versions of the Read Codes,” in *Annual Conference of the Primary Health Care Specialist Grou*, Worcester, 1996.

[26] NHS, “Retirement of Read Version 2 and Clinical Terms Version 3,” NHS, 02 08 2018. [Online]. Available: https://digital.nhs.uk/services/terminology-and-classifications/read-codes. [Accessed 27 04 2020].

[27] Scottish Information Management in Practice, “Read Terms for General,” 15 10 2014. [Online]. Available: https://www.scimp.scot.nhs.uk/wp-content/uploads/SCIMP-Guide-to-Read-Codes-1-6.pdf. [Accessed 27 04 2020].

[28] GP Training, “Using 5-byte Read codes,” GP Training, [Online]. Available: https://www.gp-training.net/it/synergy_archive/synergy/readcode.htm. [Accessed 27 04 2020].

[29] D. Robinson, P. B. E. Schulz and P. Colin, “Updating the Read Codes: User-interactive Maintenance of a Dynamic Clinical Vocabulary,” *Journal of the American Medical Informatics Association,* vol. 4, no. 6, pp. 465-472, 1997.

[30] New Zealand Ministry of Health, “Migrating from READ codes to SNOMED CT for better information in primary care,” New Zealand Ministry of Health, 20 12 2017. [Online]. Available: https://www.health.govt.nz/news-media/news-items/migrating-read-codes-snomed-ct-better-information-primary-care. [Accessed 27 04 2020].

[31] SNOMED International, SNOMED International, [Online]. Available: http://www.snomed.org/. [Accessed 27 04 2020].

[32] NHS, “SNOMED CT,” NHS, 17 01 2020. [Online]. Available: https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct. [Accessed 27 04 2020].

[33] NHS, “GP Systems of Choice,” NHS, 14 01 2020. [Online]. Available: https://digital.nhs.uk/services/gp-systems-of-choice. [Accessed 27 04 2020].

[34] W. Vorhies, “CRISP-DM – a Standard Methodology to Ensure a Good Outcome,” Data Science Central, 26 07 2016. [Online]. Available: https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome. [Accessed 27 04 2020].

[35] J. Stirrup, “What’s wrong with CRISP-DM, and is there an alternative?,” Jen Stirrup, 01 07 2017. [Online]. Available: https://jenstirrup.com/2017/07/01/whats-wrong-with-crisp-dm-and-is-there-an-alternative/. [Accessed 27 04 2020].

[36] K. Ross, “PDCA: The Scientific Method or the Artistic Process?,” Toyota, 13 09 2013. [Online]. Available: https://thetoyotaway.org/pdca-the-scientific-method-or-the-artistic-process/. [Accessed 27 04 2020].

[37] C. Webb, “Power BI And Case Sensitivity,” Chris Webbs BI Blog, 06 10 2019. [Online]. Available: https://blog.crossjoin.co.uk/2019/10/06/power-bi-and-case-sensitivity/. [Accessed 27 04 2020].

[38] World Health Organisation, “Constitution,” World Health Organisation, [Online]. Available: https://www.who.int/about/who-we-are/constitution. [Accessed 27 04 2020].

[39] C. Petrov, “Big Data Statistics 2020,” Tech Jury, 22 03 2019. [Online]. Available: https://techjury.net/stats-about/big-data-statistics/#gref. [Accessed 27 04 2020].

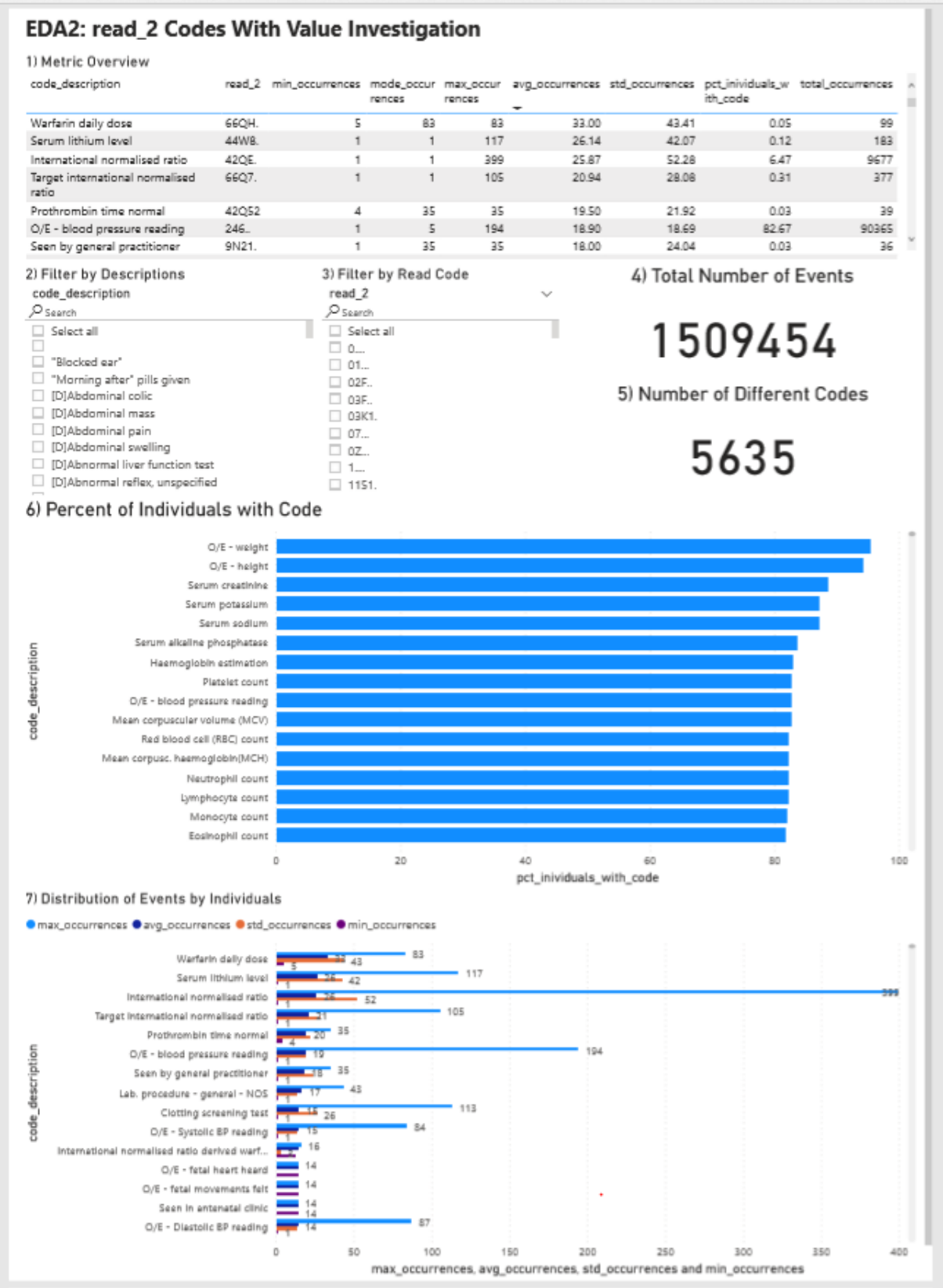
[40] I. Lapowsky, “How Cambridge Analytica Sparked the Great Privacy Awakening,” Wired, 17 03 2019. [Online]. Available: https://www.wired.com/story/cambridge-analytica-facebook-privacy-awakening/. [Accessed 27 04 2020].

[41] M. Isaac, “How Uber Deceives the Authorities Worldwide,” The New York Times, 03 03 2017. [Online]. Available: https://www.nytimes.com/2017/03/03/technology/uber-greyball-program-evade-authorities.html. [Accessed 27 04 2020].

[42] UK Government, “Code of conduct for data-driven health and care technology,” UK Government, [Online]. Available: https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology. [Accessed 02 05 2020].

Appendices

1. Exploratory Data Analysis of GP Clinical Read Codes within the CPH Population



2. Truncated code analysis applied to Covid-19 analysis