City, University of London

MSc in Data Science

Project Report

2023

Using Visual Analytic workflow design to analyse variations and their implications in the Use of SNOMED by GPs Across England

Manisha Mendonca

Supervised by: Dr Mai Elshehaly

02/10/2023

Declaration

By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the assessment instructions and any other relevant programme and module documentation. In submitting this work, I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

Signed: Manisha Mendonca

Abstract

Healthcare professionals across England use Electronic Healthcare records (EHRs) to record the "incident" when a patient visits the Doctor; healthcare records often have discrepancies due to the multiple stages the data goes through. These incidents are usually recorded using medical codes such as SNOMEDCT or CTV3. This project aims to investigate the discrepancies within the medical coding at least one stage of the process- the Admin Stage.

Through the use of an analytical workflow, this report aims to show the issues in the primary care dataset Provided by the Connected Bradford Team.

The team at Connected Bradford can use the analytical workflow and dashboard to understand the extent of issues currently happening in the primary care dataset and use this information to streamline the issues mentioned in the report.

Key Words: Electronic Healthcare Records, Analytical Workflow, SNOMEDCT, CTV3 Codes

Table of Contents

Chapter 1: Introduction and Objectives

1.1 Background

Healthcare professionals rely on electronic healthcare records (EHRs) to document a patient's visit to the doctor. Based on the type of visit, the documentation is broken down into further distinct subcategories based on specific symptoms and diagnoses given to the patient on the day of the visit at every Healthcare practice in England. This categorisation further influences the types of diagnosis that the patient will receive in hospitals if further treatment is required; this data is also used in more comprehensive research that can be used to develop new strategies and trends.

Before a diagnosis is finalised in any GP surgery, the data undergoes a systematic process where the data/admin team adds data to the system. To enter data, GP surgeries use various medical codes, such as SNOMED and CTV3. However, the codes used can differ due to varying practices, leading to inconsistencies within the datasets when large datasets are merged for research purposes. This may sometimes lead to a wrong code being entered or sometimes "generic" codes with a broader meaning being selected.

In most settings, to establish uniformity and minimise human errors in data management, the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) is utilised. This structured clinical vocabulary is integrated into electronic health records and adopted across England's entire National Health Service (NHS) system.
The objective of the project is to create a visual analytics workflow to illustrate how medical codes, missing data, and data inconsistencies affect the healthcare data, leading to misdiagnosis being recorded and reducing the quality of the data set as a whole, that can be used for further research to develop advanced strategies and discover new trends within healthcare. With the use of visual graphics, this report aims to show the current inconsistencies within the data set and how these can be improved to improve the accuracy of the data within healthcare settings.
The report will use the primary care dataset provided by Connected Bradford to assess data integrity and variations of code usage across regions. To gain a complete insight into the data set, the data analytics team at Connected Bradford will be interviewed to gain insight into the quality issues and obstacles currently faced by the team. The dataset provides access to an extensive data collection across 61 healthcare practices. The successful accomplishment of the project's goal will require using multiple data-defining sources, such as the SNOMED codes descriptions, the Athena vocabulary repository, and the Open code list in conjunction with the connected Bradford dataset to understand semantic relationships and analyse them and discover anomalies and discrepancies within the data set.
The "Bradford Institute for Health Research: data sharing agreement" has been completed to gain access to the data set and design visualisations to answer the research question. The data will be accessed via Google Cloud and not shared with third parties. Data will not be stored locally or downloaded from the secure platform during any project stage. The data set is already shared in the anonymous form, and the project will not involve any re-identification of the individuals. Once the dataset is obtained, the likely results will be to generate insights that the analysts at the Bradford data centre can view as an analytical workflow report and a visual dashboard on Google Looker Studio, allowing them to check the data and interact with the graphs.

1.2 Research Question

Based on the level of discrepancies and missing data that is currently present in the NHS dataset, the quality of the dataset that could otherwise be used for high-quality research to develop trends and strategies, has been lowered; the following research question was therefore formulated:

Using visual analytic workflow design to analyse the variations and their implications in the use of structured clinical vocabulary and other discrepancies in the NHS's electronic health record.

The use of the primary care dataset, which gathers information on almost 61 GP practises around the area and the use of the multiple medical coding dictionaries such as the open code list website and SNOMED data browser from NHS will be used in conjunction for the identification of discrepancies within the data set.

1.3 Aims and Objective

The aim of the project is to identify the discrepancies that occur within the healthcare setting in general; however, in this case, the Bradford-connected Primary care dataset will be used to identify the issues and use the information to design an analytical workflow to show how the errors within the healthcare settings can be minimised by the use of structured language models that can be introduced across all healthcare settings.

Objective	Result
Identify the current issues with the healthcare dataset that are currently available. i.e., The NHS Digital dataset, Connected Bradford primary care Dataset.	Graphs and Tables to show the discrepancies that are currently present in the dataset
Get insight into the current issues within the dataset by talking to at least one person who currently works with the dataset.	Interview the data Analyst at the Connected Bradford team to gain better insight into the dataset.
Identify the Different types of healthcare codes, such as SNOMED, and their description issues and compatibility with other "codes."	Research the SNOMED and other data vocabulary repositories to identify issues such as missing codes or issues with compatibility when merged.
Connect the dataset with Google Looker Studio to gain insights and produce graphs based on the issues within the dataset.	Use Google Looker Studio and Big Query, SQL to understand and identify the dataset to produce visual graphs.
Show the effect of missing data, and null values	Use Big Query to show the impact of null values and missing data within the dataset.
Design the workflow model to show how the discrepancies can be minimised	Design the Workflow model
Create documentation as to how the designed workflow can help to minimise the discrepancies.	Write the documentation

Table 1: Project Objectives

1.4 Work Products

The deliverables that are intended to be delivered at the end of the project:

- A visual dashboard that is designed using Google Looker Studio will highlight the discrepancies that occur within the dataset, which can be used to understand and help review the issues that are currently occurring within the dataset. – The dashboard will be designed using Google Looker Studio and SQL statements, and Big Query will be used to show the inconsistencies within the dataset.

- Documentation – The documentation will provide a detail of how the data can be stored in a better way from the first point of contact at the GP's office to minimise data loss. The use of analytics will also demonstrate how making these changes suggested in the workflow will help to improve data quality for further research.

1.5 Project Beneficiaries

- Connected Bradford Team: The Connected Bradford team will be able to understand the issues they face within the dataset. With analytic workflow, the team will be able to understand the implications of the way GP's code and how it leads to discrepancies.

- NHS Team – The NHS team may use the developed work to understand the key challenges currently faced in the healthcare dataset. The report and the research regarding the discrepancies should help them design a better system to input data to retain its quality, which can be carried forward to further research.

- Researcher – The research project can be added to the portfolio to show the understanding of using visual analytic design to show the issues within large data sets. This will also help improve the user's understanding of using Big Query and SQL to retrieve the correct data type to understand its impact.

- Other Research Teams – By improving the quality of data, the accuracy of data will be enhanced significantly, which will help fuel further research and include all the data, and currently missing values and other discrepancies which are excluded from significant research studies.

1.6 Methods Outline and Project Plan

A data-sharing agreement has been completed to gain access to the connected Bradford dataset and analyse the data issues within. The Rational Unified Process (RUP) is a software development process used to design the implementation and complete the coursework (Hammad, 2020). The RUP process has multiple stages and processes that allow us to plan and design the coursework effectively. The advantages of using such a workflow are that it allows good documentation of the work and is designed to account for any risks that may arise during the project; the methodology allows testing and checking each stage of the project to ensure no issues are delaying the project with significant changes.

The workflow documentation and design depend on the dataset's access and analysis. The time taken to access the dataset may delay the project and the types of discrepancies found within the dataset. Therefore, in some cases, rapid analysis of the dataset to recognise the issues, investigate them in detail with the analysts with the Connected Bradford team and then start designing the process and the workflow may be required.

The main stage of the project flow uses the RUP approach as it allows maximum flexibility with the project within the timeframe. The four main stages are designed to include the bulk of the project and experimenting with the dataset to gain insights about the data. The report is a continuous progress that allows regular documentation of the process and the necessary implemented changes.

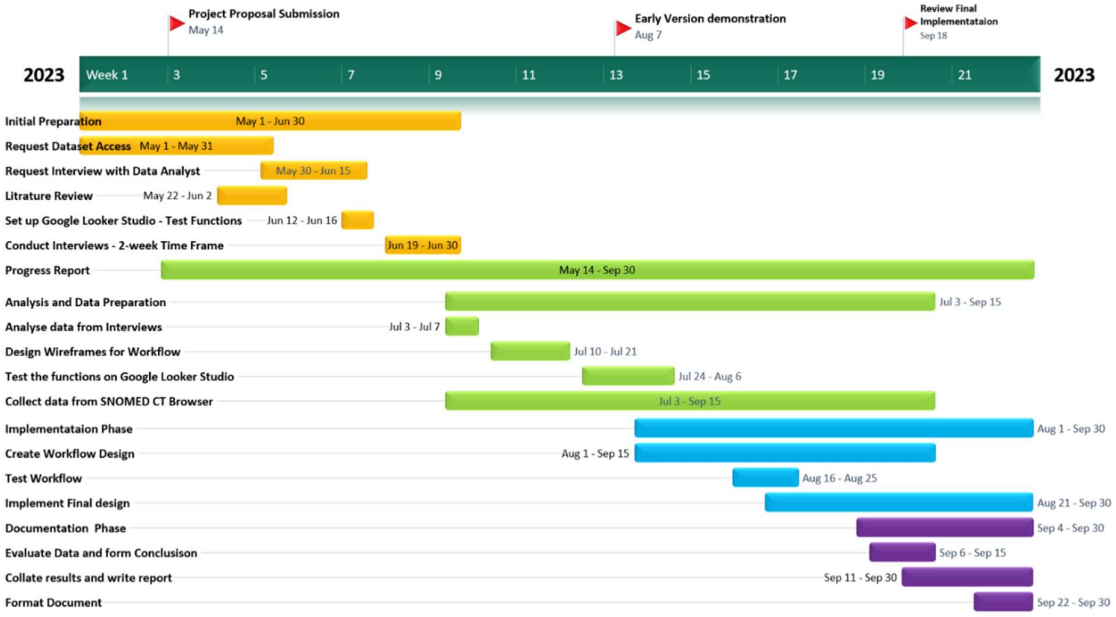


Figure 1: Project Plan

1.7 Project Report Outline

The report is organised as follows:

- Section 2 – Context describes the background of the project's research and the issues currently faced within the NHS and other healthcare datasets that limit the data used in high-quality research. The background research also describes the issues surrounding missing values and discrepancies, such as mismatches within the codes used in the datasets, such as SNOMED codes and how the use of different codes impacts the dataset when merged to create large datasets for research purposes. The context section will also discuss using Big Query and Google Look Studio to design the workflow and highlight the issues within the data set.

- Section 3 – Methods section will describe the techniques and the resources used to deploy the project during its different stages. The methods are applied throughout the project to effectively design and implement the work. The method section will also describe the types of programs considered for the design process and how the results were achieved. Information regarding the interview process and how this information was used in conjunction with research to implement the workflow will also be discussed.

- Section 4 – Results section will describe the final workflow, an implementation and a breakdown of individual components within the workflow. This section will also describe each section in detail and how it can be used to minimise the discrepancies within data sets, alongside highlighting the current issues that are faced within the Connected Bradford dataset and how the new suggestions and pointers could help to standardise the information that is currently being entered into the system at healthcare practises. The final report will have a visual analytic workflow, and a visual dashboard that the connected Bradford team can use to see data demographics.

- Section 5 – The discussion section will discuss the results with the objectives set at the beginning of this project. The report aims to discuss the issue surrounding the connected Bradford primary care data set. They are currently facing discrepancies due to how data has been entered into systems in the healthcare practises and how this further impacts data that goes for research, developing new strategies, and finding out trends within healthcare datasets that can be used to discover new software and fuel advanced research.

- Section 6 – Evaluation, Reflections, and Conclusion section will discuss the final workflow and reflect on improvements that could be done in the future and how they could serve a meaningful purpose within the connected Bradford data team to develop new strategies; it will also suggest possible implications on future related researches in the area of visual analytics and its uses in finding anomalies and trends within routinely collected medical data sets.

Chapter 2 - Critical Context

2.1 Introduction

The research element of this project spans the Data Analysis and Data Visualisation categories of the Data Science course. The analysis starts with gathering data on the connected-Bradford dataset regarding the issues currently faced within the routinely collected data for the 60 GP surgeries. Further analysis will also include questioning the Connected-Bradford team based on their experience of handling the data and the discrepancies found within the data set that need to be investigated in detail. The RUP project methodology and guidelines will be followed to collect the information and allow enough room to make changes and allocate new information where necessary throughout the project. The project aims to deliver a visual analytic workflow that can be used by the Connected Bradford team and the NHS in other healthcare systems to understand the issues faced within the data in their current systems.

Missing data, Admin error or misinterpretation of the medical codes due to different types of codes used is one of the most common issues faced in electronic healthcare records (EHRs). Visualisation helps to understand patterns in data on a large scale, especially in healthcare data, where it is challenging to gather data since it comes from many different GP surgeries. Visualisation dashboards are used widely in healthcare; however, these are mainly used to show the current results and the currently available data. Discrepancies, unless delved into deeply, often go unnoticed. When the data is sent for further research, such values are often discarded as anomalies rather than discovering the reason for those missing values or discrepancies within the data.

Visualisations such as heat maps and pie charts can help find unexpected gaps within their data and find patterns in operation codes, such as SNOMED codes, that do not match with other codes used by other practices when data sets are being merged (Ruddle, Adnan and Hall, 2022).

Design studies are used for problem-driven research where the main aim of the research is to solve a real-world problem. M.Sedlmair et al. describe problem research as *"A design study is a project in which visualisation researchers analyse a specific real-world problem faced by domain experts, design a visualisation system that supports solving this problem, validate the design, and reflect about lessons learned in order to refine visualisation design guidelines."* In the design Study methodology, the study reflects on the 9-stage framework essential to design and convey the message the visualisation is trying to depict (Sedlmair, Meyer and Munzner, 2012).

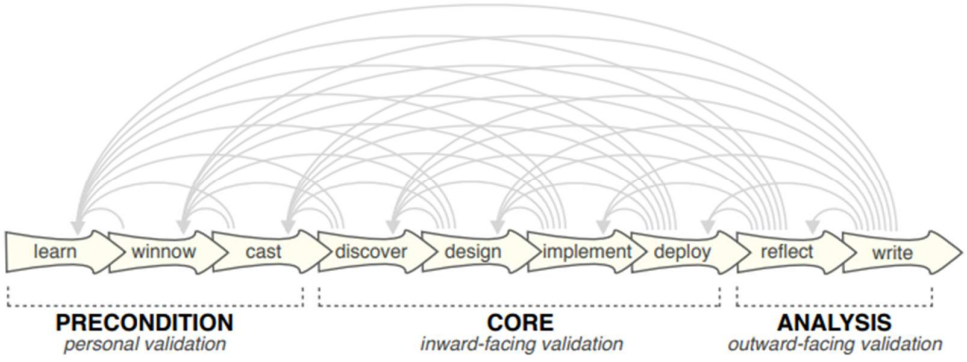


Figure 2: Nine-stage design study methodology framework – Design Study Methodology paper (Sedlmair, Meyer and Munzner, 2012).

Data visualisation is about graphic displays to show anomalies and transformations within the data set that would otherwise be very difficult to interpret and gain statistics about. Data visualisation is also helpful in exploring data structures and cleaning the data, which is essential for data mining and analysis to check the quality of the data before being passed on to analysts and further investigations. Although data visualisation is about displaying an image regarding the data, it is a complex procedure where the background of the data is important, such as the context of the data, why it was collected and the way it was collected to be able to present the data in the necessary format to give its highest value. Data visualisation has now become a commonly used strategy to analyse the strengths and weaknesses of complex data structures. It can be found within scientific publications and many other credible works to show the power of graphics in describing a bigger picture and detecting trends that were otherwise difficult to analyse (Unwin, 2020).

A research paper on "The principles of effective data visualisation" by S.R. Midway discusses the essential concepts of visualisation. Technology advancements have helped create complex visual presentations used more regularly in scientific research literature. Although visualisations are effective in conveying scientific information, they need to be designed using the correct principles and technical skills to be more effective in conveying the meaning behind the data. Misfitting visualisations can cause misleading information to be conveyed, causing unintended consequences such as confusing the readers and setting them back in the understanding of the material. The report talks about multiple principles that can be used (Midway, 2020).

To design a diagram effectively. Firstly, it is important to concentrate on the bigger picture that is trying to be displayed using the research or the report, draw wireframes of the diagrams that can be produced, followed by other features such as using the right software to convey the message this may include in learning new software to develop new techniques that are needed to convey the message of this data set or enhancing the knowledge of a software programme that is currently being used. Furthermore, using practical geometry to show the data, geometric figures such as bar plots may seem simple but are highly effective in showing complex issues within certain data sets. Each dataset can be visualised with more than one geometry; however, choosing the correct type of geometry to convey the message and concentrating on intricate features such as checking the axis and ensuring null values or missing data have been accounted for is essential (Midway, 2020).

2.2 Literature Review

Research Conducted by Ruddle et al. shows that missing data is the most common data quality issue that the NHS faces in the electronic health records (EHRs). The research study was conducted on sixteen million records and 86 fields to find the patterns in missing values. The study shows how visualisation can be used to investigate missing data patterns. To identify the meaning behind the missing data, it is essential to understand the structure of the data and then decide the type of approach needed to understand the missing data to form visualisations. Researchers also need to understand the data patterns for cohort selection and bias that may impact the data section. Although highlighting and improving the data quality is the report's primary intention, it must be ensured that any bias while selecting the variable is excluded. Presently, researchers exclude null values in the report or use advanced algorithms to compensate for the missing values if they are in low instances. In the case of the NHS digital records, a feedback report is sent in response if the missing values are greater than 30% for the hospital. However, it is now essential to investigate how those missing values in multiple fields affect the data integrity. "NHS Digital wants to define new data cleaning and cross-referencing policies". Missing values can be used to find patterns in the data. Choosing computational models has benefits and drawbacks, as essential trends can be missed. In this case, the researchers analysed 15733889 records, and the dataset was processed to flag any "Unknown" values in each field. A visual interactive dashboard was used to display the missing values, where the user can see the importance, the missing values can be added and the trends that can be discovered. Simple graphs such as bar charts, heatmaps and histograms were used to display the missing data. The study aimed to show how missing data impacts the secondary uses of the EHRs. It uses data mining and visualisation to show the value of finding trends in missing data. (Ruddle, Adnan and Hall, 2022).

Furthermore, research conducted by R.Khare et al. discusses the "data quality in large pediatric data research network". It shows a range of quality issues that were monitored over the range of 18 months. The study found 2182 data quality issues identified across multiple data cycles. Analyses of the recent data cycle included 850 issues, including more than 30 outliers and missing data points in complex domains. The study focused on data quality based on real-life accounts of studying and interpreting data quality in pediatrics CDRN so that this information can be carried forward for better analysis of data sets in the future. The study focused on the pediatric learning path system city RN, PEDSnet. The dataset contained data from over 5 million children with at least one clinical encounter and at least one code of diagnosis during or after 2009. A large-scale data set was used to understand the critical challenges faced by undocumented data in the columns, how this affects the quality of the data set, and how it can be interpreted in the future for research purposes (Khare et al., 2017).

A recent study by S. Bacon and B. Goldace discusses the "barriers to working with National Health Service England's open data". This article talks about the issues faced by researchers on a day-to-day basis regarding healthcare data that is openly available. However, only a little can be done with it due to inconsistencies with the structures of the available data sets and the type of information based on the codes used by GPs and other hospital trusts as they are too challenging to merge and get a large data set that can be used for research purposes. Researchers have shown inconsistencies within the data set, such as introducing new fields without definitions or referring to fields that do not exist in the data set. These outliers limit the potential of the data set and what could be done with it. Furthermore, there were issues with British national formulary names (BNF code); these are often used in the prescription of drugs. The study mentions that the coding system is based on BNF's old classification system, which is no longer maintained, causing the NHS to alter its version to a new version, "Pseudo BNF classification". Changes to the BNF coding took place over the years where a significant reclassification process happened every January with some drugs in the classification of the old BNF system and some using an entirely new BNF classification code, the lack of explanation/precise classification to establish the new and the old codes were not done correctly, leading to inconsistencies in the data set and making it

almost not usable for research purposes. Furthermore, the research mentions the issues within the procedures that GPs use to register the data into the system. Data quality issues such as not standardising the name of different kinds of institutions, such as drop-in centres or care homes, and apparent errors in coding and classification system have also been noticed; some data sets also contain fictional values that contains practises that have prescribing at improbable levels that far exceeded the total number of patients, it is hard to identify if it was a data entry error. This makes it hard to know if the data is being maintained reliably by NHS and even if it could be used for future research (Bacon and Goldacre, 2019).

The Imperial College report on maximising the impact of NHS data mentions the issues surrounding the NHS. The main issues in the data lay within the pre-processing of the data; since the teams of data admins mostly record the information, the data is not captured in real-time, leading to misunderstandings and missing data. NHS services say approximately 1,000,000 patients are seen every 36 hours, and almost all of the data is stored in the form of electronic records at every GP surgery; this significant amount of data is often not captured in real-time, leading to significant efforts to post-process the data and curate it to match the standards of larger data sets which makes it particularly difficult as there is currently not a very strong standardised approach used by many GP surgeries across England. Furthermore, technology and infrastructure need to be improved to move the data to the cloud, reducing the data quality. Although primary care practice management has been using software to manage data since the early 90s, most of the secondary care providing places are paper-based and only started using technology to manage the EHRs [electronic health records]. Due to this, there is a problem with how data has been managed for a long time. In the current setting, all GP data has been digitalised, and there is a route to convergence of standardised data for all GP systems. On the other hand, outside of primary care, a recent survey showed that approximately 23% of patient records in acute hospitals are entirely paper. This leads to decreased quality and reliability to use the data for further research. It also limits the amount of data that is available for research. Furthermore, almost 10% of hospitals were using multiple EHRs within the same hospital, causing duplicity and miscommunication within healthcare data. Improving the systems and digitalising the NHS records is the only way to improve access to information that can be further researched to Build new strategies on infrastructure. Although, in recent years, there has been a positive effort to improve data quality within the NHS, there still needs to be a marked difference in data quality and infrastructure across providers, making it difficult and costly to combine multiple datasets (Ghafur et al., n.d.).

A study conducted by I.Petersen et al. speaks about the vital implication of handling missing data within the UK's primary care electronic health records. EHRs are now commonly used for research as they capture information on common health indicators such as blood pressure levels and people's smoking status and alcohol consumption. Although these factors are not recorded regularly, they are still used as a significant clinical database for health research; therefore, missing, or mishandling data severely impacts the quality of the data from the NHS. The study used the Health Improvement Network (THIN) UK primary care database. It used demographics and variables such as age, gender, and other chronic illnesses such as diabetes and stroke to be fitted into linear and logistic regression models to examine the association of weight measurements and probability of having weight recorded with individuals' demographic characteristics and chronic diseases. The data suggested that women between 18 and 65 are more likely to be of the same age to have health indicators recorded. In contrast, only 60 to 80% of individuals have their height, weight and blood pressure, another critical indicator recorded during the first year of registration; as the years of registration increase, these proportions fell by 10 to 40%. In contrast, only individuals with chronic diseases for a long time were more likely to have health indicators recorded. The study suggests that missing data for common health indicators will affect statistical analysis for research studies as, in most cases, researchers exclude the available data rather

than manipulating it to make use of it. Since it is challenging to make assumptions based on missing data, this leads to NHS data being utilised and recognised as a rich source of healthcare data for Advanced Research studies (Petersen et al., 2019).

The Standardized Nomenclature of Medicine, Clinical Terms (SNOMED CT) is slowly becoming the most used clinical coding system in the NHS. The report suggests that up to 30 countries use the SNOMED coding System.

The report mainly discussed the logical equivalence between SNOMED CT and ICD 10 - PCS Surgical procedures. They compared the logical definitions of SNOMED CT concepts to the ICD-10-PCS axial components to identify overlap and gaps. The most significant discrepancy was in the surgical approach, specified in all ICD-10-PCS codes but only in 8.7% of SNOMED CT surgical procedures. Lack of matching definitions is the most significant issue leading to discrepancies, and the report points out that only 25% of the definitions of ICD-10-PCS could be fully matched.

Although the study does not discuss the mismatch in definitions of CTV3 to SMONED CT, this report is an excellent research article to show the ongoing issues surrounding the lack of definitions and descriptions that lead to discrepancies and wrong codes being selected. (Fung et al., 2018)

Chapter 3 – Methods

3.1 Introduction

The Project stages mentioned in the Design study methodology paper (Sedlmair, Meyer and Munzner, 2012) were followed closely to gather the necessary data required to collect sufficient knowledge about the data and how to conduct research to demonstrate the issues presented in this report. The project aimed to follow most of the stages from the above. However, some stages were merged to better suit the description of the study.

The project's initial phase was to gain access to the primary care dataset from the Connected Bradford team. Initial contact was made with the Connected Bradford team on the 17th of May by speaking to the Research support administrator and filling out the "Expression of interest" (EOI) document that outlines the scope of the MSc project and the how the data would be used to answer the proposed research question. Once the EOI was accepted, a Data sharing agreement was filled to gain access to the dataset, where the connected Bradford team evaluated the stated requirements and provided access to the dataset on the 3rd of July via Google Cloud console with the NHS login details and a Collaboratory workspace platform to conduct the research. Clear guidelines were stated in the data sharing agreement that prevent the data from being removed from the original platform and how the data can be used. The initial contact to gain access to the dataset was made well in advance to prevent any delays in access to the dataset and its subsequent analysis.

3.2 Dataset

The Connected Bradford dataset contains the Primary care dataset, collated data from multiple healthcare practices that record the patient's visits and additional details regarding their diagnosis. The visit details are not included in the dataset as they are confidential. The primary care dataset is divided into thirty-two different tables. Additionally, the dataset comes with additional resources such as "Vocab" and "Lookup" Tables, which help to understand subsequent analysis.

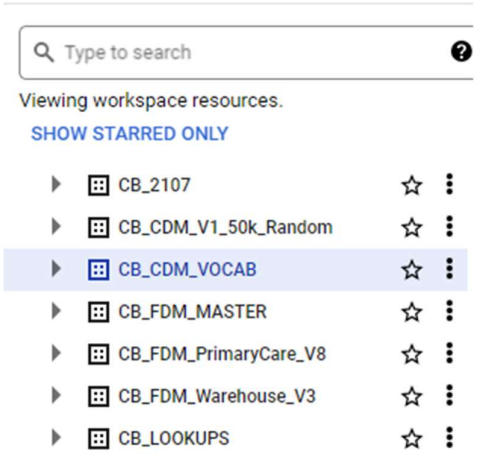


Figure 3: An image of the dataset on the secure platform.

3.3 Software Tools and Programs

Google Cloud

The dataset is accessed via Google Cloud – The Connected Bradford team already set up the workspace based on the data-sharing agreement to ensure analysts have their workspace to perform experiments, run tests, and save the analysis (Google Cloud, 2023) Some of the advantages of using the Google Cloud platform are:

- Storing large datasets for data loss prevention
- Advanced Security for maintenance and security features to store the data.
- Flexibility in Accessing the data at any time.
- Secure platform to store confidential data.
- Great for managing, updating, and maintaining large-scale datasets.
- Detailed documentation is available for each step/process.
- Easy to navigate and use the features.
- The cost-effective way to store and explore the data.
- Regular updates and new features are introduced, which can be explored to scale up the projects.
- Access to Big Query

Big Query

Big Query is an enterprise data management system that helps to explore data for data analytics and research purposes. The flexibility to assess and the ability to store data within big Query makes it a powerful tool used for data analysis. (Google Cloud, 2023)
Some of the benefits of using Big Query are:

- Fully Managed data warehouse
- Big Query command line tool – Allows to run familiar programming languages like Python SQL for data analysis.
- Run queries on data on external tables.
- Big Query Studio – Version Control available for Notebooks and saved queries
- Big Query Administration – Allows centralised management of the data and compute resources.
- Big Query Resources – Extensive documentation on the process and programme
- Ability to enable integration with other API platforms if needed in the future.

Google Looker Studio

Google Looker Studio is a free tool from Google that allows one to visualise the data. (support.google.com) It allows the connection of the data from the Big Query/Google Cloud platform, making this the ideal platform for visualisation since the data from Connected Bradford cannot be taken out/ downloaded from the secure platform. Creating an interactive dashboard was one of the deliverables of the project. Initial design concepts and wireframes were drawn to design the dashboard.

- The tool has an easy-to-use web interface.
- Creates interactive, meaningful visualisations and interactive dashboards.
- Easy to connect various data sources, including Big Query and MySQL

Jupyter Notebook (Jupyter, 2019) (www.nobledesktop.com)

Jupyter Notebook is a web-based application that allows to configure workflows for data analysis. It is a free software that allows interactive computing across all programming languages.

- Easy access platform to edit and test code/Queries.
- Can collaborate with a large range of tools and interfaces.
- Can be used for Data Visualization

Language-independent Architecture – The IDE can be used to code in any language.

3.4 Data Analysis

Once the Data Sharing agreement was approved and access to the dataset was granted via Google Cloud, the initial phase of the project stage started with learning and understanding the dataset: the features of the data, the variables, and its explanations.
The dataset was provided with a thorough Data dictionary and GitHub links that help to understand the meaning of the tables and other resources. A preliminary dataset analysis was conducted to understand the data and how the available information can be used to answer the question posed in the project.

- Run queries on the dataset to find general patterns and trends.
- Design question to ask the data research team at Connected Bradford.

The "Winnow" stage of the design study methodology paper talks about the importance of having access to the database well in advance to prevent "pitfalls" when recruiting participants and researchers to enhance the study. This stage emphasises selecting the right people to recruit, as it is a time-consuming process, and the process must ensure that valuable insights are produced. Finding the domain expert is necessary to ensure the task is being addressed and the existing approach has problems that can be addressed in the report. After the initial analysis, multiple team members from the connected Bradford research team were considered for the interview process. However, only two participants were selected for the interview based on their knowledge and expertise of the dataset.
The selection of the right people takes to the 3rd stage of the design study methodology – "Cast- Identify collaborator Roles." To proceed with the project, it was crucial to understand the backend process of how the data was collected and stored in connected Bradford, along with how the GPs select the code. Are GPs the ones to select the code in the first place? What is the role of a medical coder? To gain a thorough understanding of the backend workflow, choosing the right people to interview is essential. The two interviewees were a senior research fellow and a senior Database Manager, who have a crucial role at Connected Bradford and have hands-on experience with the data.
The interviews were scheduled for one hour each on the 8th and 10th of August. Before the interview, the participant information forms, and Consent documents were sent to interviewees to be completed. The interview was conducted and recorded using teams. Once the transcription is completed, the interview recordings will be discarded as per the data-sharing agreement.

Design Study methodology – stage 4 – Discover: Problem Characterisation & Abstraction

It is essential to understand the target domain, problems, and the requirement of the domain expert to discover how the visualisation can enable insight and discovery. Concentrating on the successful aspects of the projects rather than just concentrating on the negative aspects is essential. This stage is directly linked to talking to and observing the domain experts.

3.5 Interview Analysis

An Interview guide approach was applied, with an outline of the topics and questions that needed to be covered. The advantages of this approach are that it allows to cover all the topics, but the interviewer is free to vary the wording and the order of the questions. This helped to keep the tone of the interview conversational. One of the drawbacks of this interview style is the difficulty in analysing the interview answers, as the respondents answer variations of the same questions or concepts (Sewell, n.d.).

The interview itself started with an introduction to the research topic, an explanation of what was already discovered about the dataset based on the preliminary analysis that was conducted when access was granted to the dataset, the aim of the research, and how this interview could help to enhance the research and help to develop a visualisation that could help to better support the current practice.

The interview question was carefully curated to gain the best information on how the data is managed at Connected Bradford, understand the current issues within the practices, and what some of the implications of the current practice could be. It is essential to know the issues in general to get a better idea of the data and how the information can be used to draw meaningful insights about the data.

Questions of the following nature were asked of the interviewees based on their expertise:

1 Will you please explain your role in connected Bradford?
2 What kind of data do you deal with daily, and what are the key challenges currently faced in the healthcare data you receive/ Input?
3 How do the inconsistencies in the data collection standard decrease data reliability?
4 In your Expertise, what are the critical decisions that could benefit from routinely collected high-quality data? How would the subsequent analysis of the data impact in terms of deriving new strategies?
5 what are the key strengths that are expected from routinely collected data to support decision-making? who are your key stakeholders (i.e., people who reach out to you with requests for data insights to support their decisions)?
6 There is a lack of definitions/ or clarity to certain concepts an issue in the healthcare system, leading to a decrease in data integrity?
7 What types of Data coding Systems exist, and which ones have you come across in your line of Work? I.e., SNOMED Codes.
8 Is the way these "codes" are used impacting the quality of data in the healthcare system?
9 How do you think the difference in the "Coding" based on the GP's discretion impacts the ability of the data to be used and utilised? – Specifically, in terms of using the data for broader research and then using that research to form a generalised opinion across the entire healthcare system.
10 How does the accuracy of data affect future use of the data in terms of its analysis and research?
11 Is healthcare data highly inclusive within its facility that the newly developed strategies/software cannot be generalised to other hospitals/ medical facilities?
12 What can be done to improve the quality of data being collected currently?

3.6 Qualitative Data Analysis

Qualitative research is fundamental to enhancing the quality of work and getting measurable data from direct sources as it allows gathering information from many different perspectives. The two main approaches are Thematic and Narrative analysis.

The inductive method can be broken down further into – (ATLAS.ti,.)

Thematic Analysis - data analysis methods that involve identifying, analysing, and interpreting patterns of meaning or themes in qualitative data. It has a flexible and valuable research tool that can provide a rich and detailed yet complex data account. A defining feature of thematic analysis is its ability to highlight similarities and differences across the data set, which can be separated into six stages. Thematic analysis is critical as it provides a step-by-step procedure to analyse data. Moreover, it derives meaningful insights to answer research questions. The method is highly flexible in highlighting important themes that might directly emerge from the dataset. It can be applied as a bridge between raw data and meaningful insights. (Rev, 2023)

- Familiarisation with the data -Reading and listening to the interview to understand the interview's concept and see if any trends or patterns can be drawn.
- Generating initial codes - Identifying significant information from the interviews and sectioning them into meaningful segments.
- Searching for themes - Identifying broader patterns and themes within the data and their responses to find meaning within the dataset.
- Reviewing for themes - checking the themes against the coded extracts and the entire data set to ensure they tell a compelling story about the data.
- Defining and naming themes - Thoroughly define the major themes within the data, ensure that these teams are captured in the storytelling, and further refine the story.
- Producing the report - Provide evidence of the interview, the themes and all the Explanations in the form of a report.

Narrative Analysis – Making sense of the individual stories from the interview. This is used to derive qualitative analysis and highlight the essential aspects of the data.

The thematic analysis method was used to analyse the interviews and gain the maximum level of information.
Some of the Key information from the interviews were:

- Issues within the way the current data is handled.
- Problems with routinely collected data in the primary care dataset.
- How different practices use different coding systems, and how this impacts large-scale datasets.
- How is data processed at the GP surgery?

Design study methodology:
Stage 5– Data Abstraction, Visual Encoding & Interaction
Stage 6 – Implement: Prototypes, Tools and Usability are used to design effective visualisation.

Both stages are highly important as design decisions, implementation, and changes are mainly carried out during this project stage. These stages allow us to get feedback on the process and visualisations. The paper suggests that to avoid any pitfalls, it is essential to have a "Broad consideration space of possible solutions, then narrow proposal space based on design principles and guidelines."
Design mock-ups and low-level prototypes for the dashboard and the visualisations were created during this coursework stage. Furthermore, these stages are also referred to when selecting the types of disorders and codes to explore in the results section. Initially, the inspection of the dataset started on a broader scale, with many spectrum/ large disorder codes that could be used to answer the question. However, based on the selection criterion (Explained further in Chapter 4), the ideas on how to answer the research question were narrowed down to use specific SNOMED codes for Coronavirus and Influenza. The visualisation designs dashboard was changed based on the requirement of the scenario.

3.7 Interview Details

Some of the main issues discovered during the interview are:

- The primary care has 61 GP Surgeries, and the information is collected routinely from multiple practices.
- The admin staff deal with coding the information on the backend - who are not clinical experts.
- Data missingness is a complex issue that is difficult to notice daily - capturing the trends of missing data is highly beneficial.
- Multiple columns in the dataset have issues - pointing trends will be helpful.
- Adding geographic logic could help see if particular practices have greater issues than others.
- There are issues with source-level understanding of the data.

3.8 Initial Dashboard Design

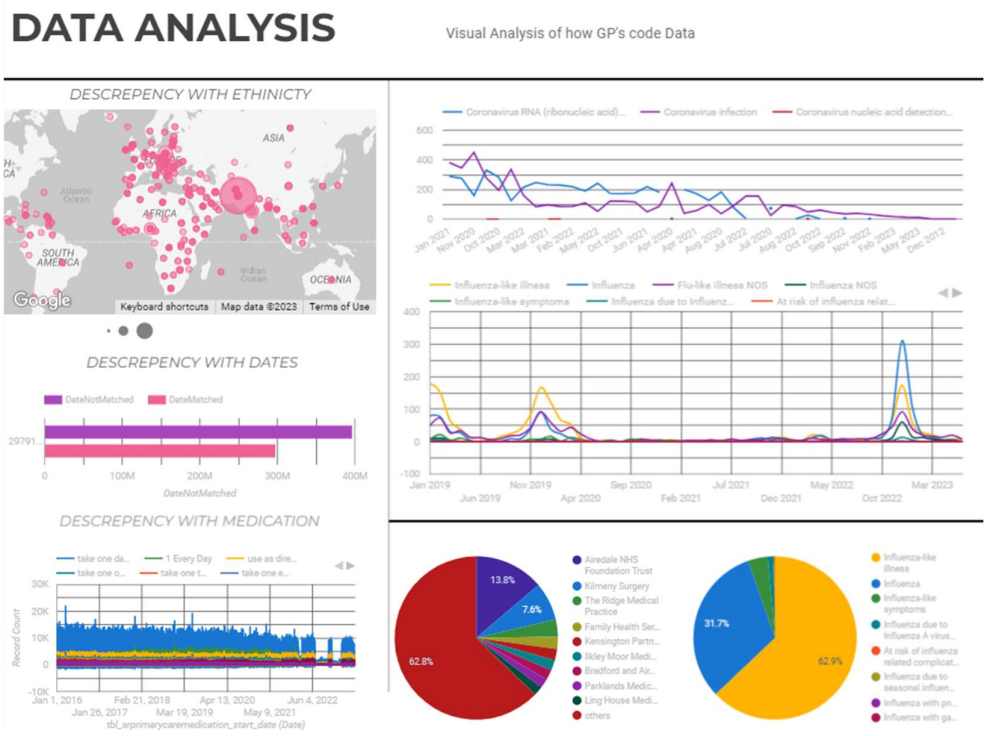


Figure 4: Initial ideas of the dashboard, the basic prototypes of the initial version of the Dashboard ideas.

The following stages after the implementation are regarding the release of the product to gather feedback and the analysis phase, which requires reflecting on the value of the research, the findings, and the analysis of the proposed guidelines. These stages are detailed in chapters 5 and 6 of the report.

Chapter 4 - Results

4.1 Preliminary Analysis of the Dataset

The graphs created using Google Looker Studio below show the initial issues within the GP practices that lower the data quality; the main issue highlighted in Figure 5 is the variations used to enter the "Birthplace" category. Some places are entered as Countries, whereas others are entered as "Hospitals". The issue, if further amplified, is that each minor spelling change creates its category, such as "Pakistan"," pakistan", and "PAKISTAN", all having separate categories because of the way the data is typed. This issue requires minor changes within the database system, such as introducing a picklist and using a standardised approach of entering the country of birth or the name of the hospitals. Where the name of the hospitals is not available, introduce a second category as such name of the Country → District / Locality → Hospital. Even if all the data is standardised for the country section, that is one step closer for cleaner data.

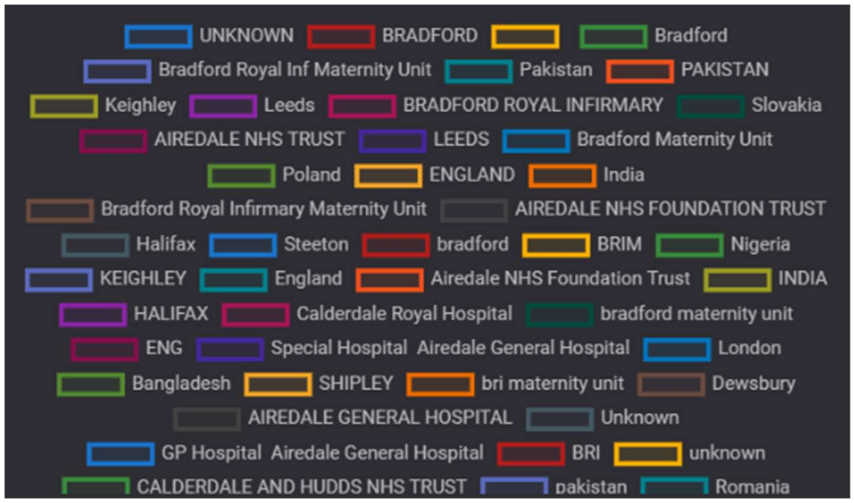


Figure 5: Data issues in the "Birthplace" column for registered Patients

	birthplace	Record Count ▾
1.	UNKNOWN	789,301
2.	BRADFORD	401,902
3.	null	242,234
4.	Bradford	143,616
5.	Bradford Royal Inf Maternity Unit	136,445
6.	Pakistan	83,600
7.	PAKISTAN	76,411
8.	Keighley	49,517
9.	Leeds	38,009
10.	BRADFORD ROYAL INFIRMARY	36,356
11.	Slovakia	31,020
12.	AIREDALE NHS TRUST	26,649
13.	LEEDS	25,848
14.	Bradford Maternity Unit	23,992

Figure 6: Table showing the record count of each Birthplace.

Untitled 6 RUN SAVE SHARE SCHEDULE MORE

```
1  SELECT birthplace, COUNT(*) AS NumberOfRecords
2  FROM `CB_FDM_PrimaryCare_V8.tbl_srpatient`
3  GROUP BY birthplace;
```

Query results

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS | CHART PREVIEW

Row	birthplace	NumberOfRecords
1	*null*	242234
2	Nowy Sacz	1
3	Romford	1
4	,achester	1
5	-	35
6	--	6
7	--SUNDERLAND	1
8	-Krallovsky Chlmec slovakia	6
9	-Warrington	2
10	-in toilet!.	1
11	-kralovsky Chlmec	6
12	0	2
13	0024968	11
14	01 02 1972	1
15	01 03 2002	1
16	01 10 1971	2
17	01011948	1
18	01061984	1
19	01101953	1

Figure 7 - The query showing all the details for Birthplace.

Figure 7 shows that all characters are accepted to be entered into the dataset. In this instance, multiple records have just numbers entered as places of birth. In some cases, it could be assumed that the Date of birth has been included instead of the actual birthplace. Furthermore, one incident even recorded the birthplace as “In toilet!” This shows the quality of the data. New regulations must be introduced to maintain the quality of the data.

Furthermore, this issue is further seen in other categories with a much higher impact on the medical industry. The "*medicationdosage"* column on the dataset, has multiple instances of the data description repeated on several instances. The images below give a small snippet of the seriousness of the issue faced within healthcare practices, "ONE TO BE TAKEN DAILY" and "TAKE ONE DAILY" are essentially the same description, worded slightly differently. However, "ONE TO BE TAKEN DAILY" has over eleven million records and "TAKE ONE DAILY" has over thirty-five million records.

The medical staff or the admin staff mainly look at the easiest and quickest way of entering the data, and with thousands of people entering similar data into a single system that allows open-coded data, these issues are bound to arise.

	medicationdosage	Record ...
1.	take one daily	35,655,092
2.	ONE TO BE TAKEN DAILY	11,809,799
3.	use as directed	11,705,706
4.	1 Every Day	7,269,844
5.	take one each morning	7,121,763
6.	take one once daily	7,061,336
7.	take one twice daily	6,746,359
8.	take one at night	6,410,469
9.	take one 3 times/day	4,760,192
10.	AS DIRECTED	4,096,794
11.	use As directed	3,027,135
12.	TAKE ONE DAILY	3,018,769
13.	ONE TO BE TAKEN TWICE A DAY	2,747,668

Figure 8: Table of "MedicationDosage"

Figure 9: A time series chart showing the use each category.

4.2 Inclusion Criteria

Establishing inclusion and exclusion criteria is extremely important for research studies. Inclusion criteria are key features of the target population that investigators will use to answer their research questions (Patino and Ferreira, 2018) Standard Inclusion/Exclusion criteria are Dates, Participants, Exposure to interest, and Geography Locations.

In this study, the inclusion criteria for each graph varies depending on the type of information that needs to be retrieved to show and analyse the trend.

Inclusion criteria make a particular difference in query writing, where the conditions' names depend on small things, such as the capitalisation of the characters. "Coronavirus" and "coronavirus" are two separate entities; therefore, it is essential to understand that the type of data retrieved is specific to the types of queries written.

```
SELECT DISTINCT (ctv3text)
FROM `CB_FDM_PrimaryCare_V8.tbl_srcode`
WHERE (ctv3text LIKE '%COVID%' OR ctv3text LIKE '%Coronavirus%'OR ctv3text LIKE
'%coronavirus%')
```

```
SELECT *
FROM `CB_FDM_PrimaryCare_V8.tbl_srcode`
WHERE (ctv3text LIKE '%COVID%' OR ctv3text LIKE '%Coronavirus%'OR ctv3text LIKE '%Long
COVID%')
```

Above is the example of two queries that would retrieve very different results based on the inclusivity criteria set in the ctv3text section.

4.3 SNOMEDCT Code Errors

```
CREATE TABLE `CB_2107.SNOMED` AS
SELECT * FROM `CB_LOOKUPS.tbl_EFI2_Codelist`
WHERE (Codedescription LIKE "%bronchitis")
```

DETAILS | PREVIEW | LINEAGE

SNOMEDCT_CONCEPTID	CTV3	Provenance	Codedescription
52571006	H31y1	*null*	Chronic tracheobronchitis
49691004	X101j	*null*	Industrial bronchitis
195949008	H312000	mar-copd	Chronic asthmatic bronchitis
866901000000103	Yaw0N	qof-copd	Eosinophilic bronchitis
63480004	H31..	efi-asthma_copd	Chronic bronchitis
61937009	H310.	efi-asthma_copd	Simple chronic bronchitis
360470001	H310.	efi-asthma_copd	Simple chronic bronchitis
89549007	H310.	efi-asthma_copd	Simple chronic bronchitis
74417001	H311.	efi-asthma_copd	Mucopurulent chronic bronchitis
84409004	XE0YM	efi-asthma_copd	Purulent chronic bronchitis
405720007	Xa0IZ	efi-asthma_copd	Asthmatic bronchitis

Figure 10 – Screenshot of the CTV3 code and their descriptions

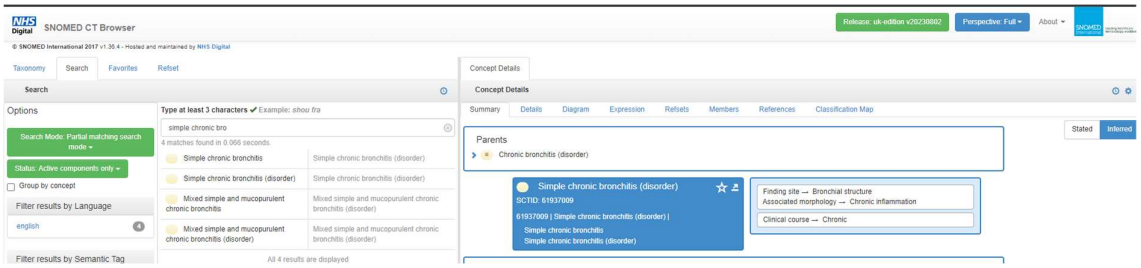


Figure 11 – Screenshot from the SNOMEDCT Browser – Definition of the codes (termbrowser.nhs.uk)

Figure 10 highlights a few initial issues with the medical codes. According to research and as seen above, a single CTV3 code has been mapped into multiple SNOMED_CONCEPTID due to minor detail differences, where SNOMED_CT may have a more general definition.

In a broader context, since SNOMEDCT and CTV3 codes often get mapped into each other, it could be expected that the explanations would be similar. However, in some instances, this is different. In recent

years, specifically in 2020, TTP (Clinical software) has started to transition from CTV3 to SNOMEDCT; during the first phase, users can input data based on either of the two codes (SNOMED CT in SystmOne Guidance for users, n.d.). This creates an issue when there are slight mismatches in the definitions, leading to admin staff selecting codes with a broader meaning to avoid significant miscommunications. Almost all disorders have a "parent" code that describes the primary disorder and additional "child" codes that differ based on the symptoms. (www.opencodelists.org) If definitions for the child codes are more complex in SNOMED_CT, the parent code with broader definitions may be preferred. This means that patient records are assigned to similar but wrong codes, minimising the authenticity of the data. From a daily clinical perspective, this issue will not be noticed instantly, but when the data is collated and extracted for research purposes, the lack of use of specific codes can be noticed. This issue is discussed in detail further in the report.

When inspected closer using the SNOMED CT Browser for this specific instance, it can be deduced that 2/3 of the SNOMED codes are inactive as shown below. One of the issues faced is if the inactive codes have patient records assigned to them, which could lead to two problems. The records under the old codes will be missed if those codes are not included in the Query selection criteria or if the records are updated to the new SNOMED code; there is a loss of definition and granularity in the converted records.

The main SNOMEDCT code that is used in the online dictionary (Image) is 61937009 for Simple Chronic Bronchitis, whereas SNOMEDCT_CONCEPTID 89549007 has a code description of "Catarrhal bronchitis", which falls under the invalid category of SNOMEDCT.

It is difficult to understand the concept of these codes and in which situation they have been used in GP surgeries.

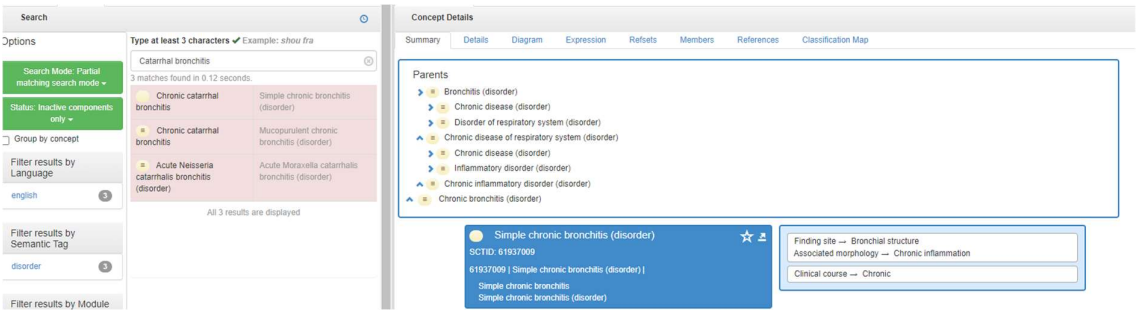


Figure 12 -– Screenshot from the SNOMEDCT Browser – Definition of the codes (termbrowser.nhs.uk)

The issue in this case is the use of inactive SNOMEDCT codes. This could be a major issue in large datasets if the concept codes are not updated in the dataset accordingly (SNOMED CT Fact Sheet Inactive Codes, 2019), (confluence.ihtsdotools.org). The SNOMEDCT documentation states that the codes are inactive if the "description is outdated, ambiguous or duplicate". For GP Practices, once the code is declared inactive, it is no longer available to select from the database. However, the issue begins when old records are not replaced with new codes within the organizations, so in this case, if a SQL search is run to only include instances where code – 61937009 is needed since it is the active code for Bronchitis, all instances recorded under the inactive code 89549007 will be missed. This created a huge gap in research if not researched properly. The SNOMEDCT document states, "There is no mandate to recode existing entries that now have inactive SNOMED CT codes. Historical inactive codes should still be visible to users; must be searchable in the system but identifiable as inactive".

One suggestion to deal with this issue could be to make it mandatory to update the old codes to new SNOMED codes according to the latest data release; this is a very time-consuming and costly process to maintain the upkeep on such a large scale of data. However, it currently seems like the most valid option to ensure no data is lost in the process.

Furthermore, the guidance for SNOMED CT in SystmOne states, "*A single SNOMED CT Concept ID can* be mapped to multiple CTV3 codes". (SNOMED CT in SystmOne Guidance for users, n.d.) The document also suggests that The CTV3 directory will be available on the side panel to cross-check the CTV3 code description with the SNOMED Code. This makes the transition relatively straightforward; however, this is only effective if the transition from CTV3 to SNOMED is made relatively fast. Otherwise, the updates to the description of SNOMED may cause confusion and errors while selecting code, especially during the admin stage, where staff may lack knowledge of clinical terminology. One of the significant issues with multiple code mapping is the loss of granularity in the definitions from CTV3, which may be more defined. In contrast, SNOMED may have a general concept, leading to semantic ambiguity, as subtle changes in the description may cause a loss of minor symptoms.

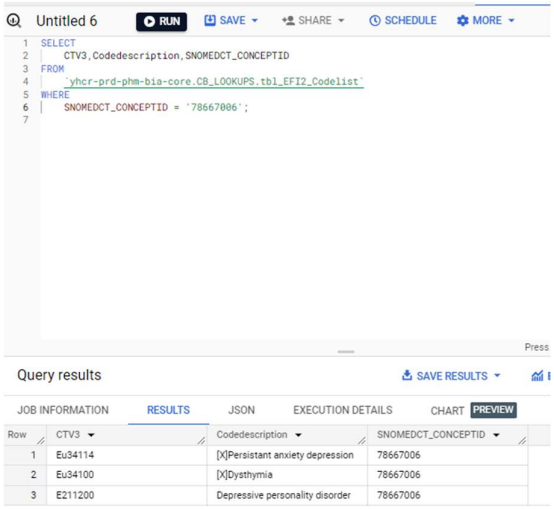


Figure 13: Showing the multiple mapping of the CTV3 codes to SNOMED.

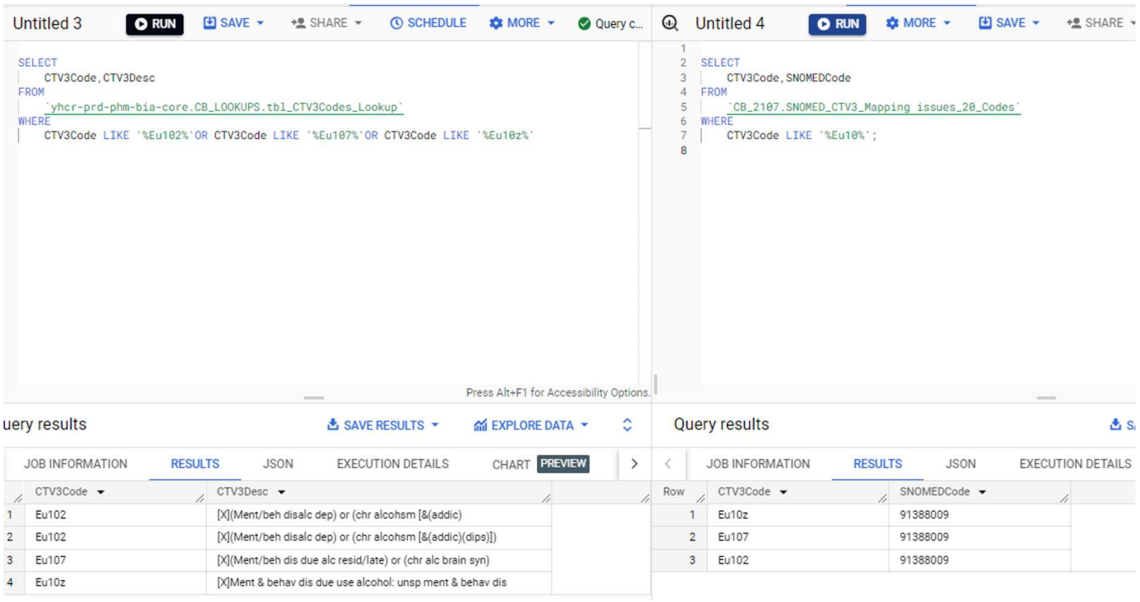


Figure 14: Showing the Subtle changes in the definition of the Code description in CTV3, that are mapped into single SNOMED.

4.4 Different Codes Used for the Merging of the Data.

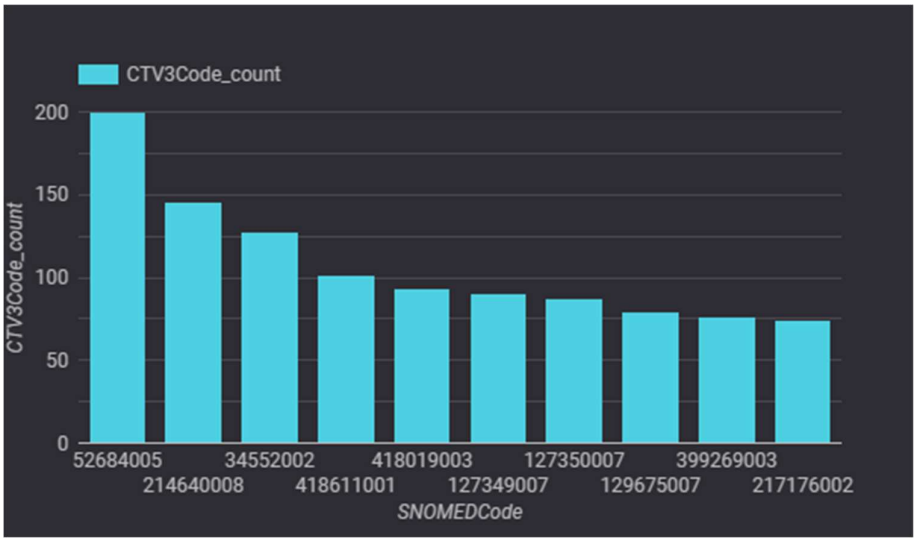


Figure 15: Shows the top ten mappings of CTV3 codes to SNOMED Codes. (Includes terminology of all categories)

One of the other issues discovered during the dataset analysis is that read code (CTV3 Codes) read version two and clinical terms version three were discontinued after April 2018. There exist two versions of CTV codes since 1985: version two and version three. Both versions have provided a standard vocabulary for clinicians to record patient findings and procedures for primary and secondary care healthcare systems. Updated codes are released multiple times a year; however, the last updated version two for CTV codes was released in 2016, and version three was released in 2018, (NHS Digital) with SNOMED being introduced in April 2018 in a phased release. (www.england.nhs.uk) This is a significant issue as some surgeries use multiple coding systems, whereas some GP surgeries rely solely on the CTV3 codes. (NHS Digital) In order to move on to the new system, these CTV3 codes need to be mapped into the new SNOMED coding system. Since the replacement of SNOMED is a slow process and cannot be implemented overnight, in many cases, multiple coding systems are being used for long periods, leading to some discrepancies within the code selection. Furthermore, since medical datasets are large and new information is recorded daily, switching from an old coding system to a new one takes time to get used to and switching over can be a significant task that requires precision. The analysis of the dataset from SNOMED to CTV3 mapping also showed that often, there are 1 to many relationships between the SNOMED codes and the CTV3 codes, in which cases, many CTV3 codes have been mapped into a single SNOMED code. The graph above shows the SNOMED code with the highest number of mappings from CTV3 codes.

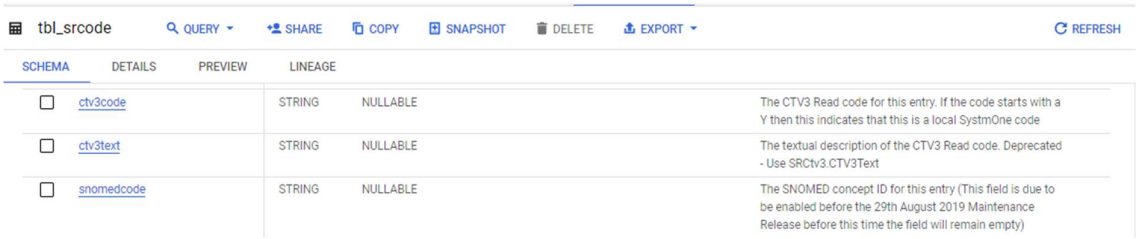


Figure 16: Shows the description provided by Connected Bradford primary care Dataset regarding the SNOMED Coding switchover.

The datasets state that the SNOMED concept IDs will be enabled before the 29th of August 2019. It is important to notice that each practice is responsible for its data; hence, any missing data is not the responsibility of the staff at Connected Bradford.
When a query was run on the primary care dataset, data suggests multiple entries missed the SNOMED code entries. It is difficult to know whether this issue is explicitly caused by the delay in mapping the CTV3 codes to the SNOMED codes.
The query below was run specifically to include all years in the dataset, and the data was extracted from the "date event recorded" section to ensure all data was accounted for. The graph below shows that their mapping issue/ missing SNOMED codes specifically became an issue during 2019 up until 2023, when the current data is available. The issue is highly noticeable in "Airedale NHS Foundation Trust" and "Kilmeny Surgery", specifically during the "COVID Years".

```
CREATE TABLE `CB_2107.NULL_SNOMED_CODE_PER_YEAR_SITE` AS
SELECT
    care_site_id,
    EXTRACT(YEAR FROM dateeventrecorded) AS Year,
    COUNT(*) AS NULL_SNOMED_CODE
FROM `CB_FDM_PrimaryCare_V8.tbl_srcode`
WHERE care_site_id IS NOT NULL AND snomedcode IS NULL
GROUP BY care_site_id, Year
ORDER BY care_site_id, Year;
```

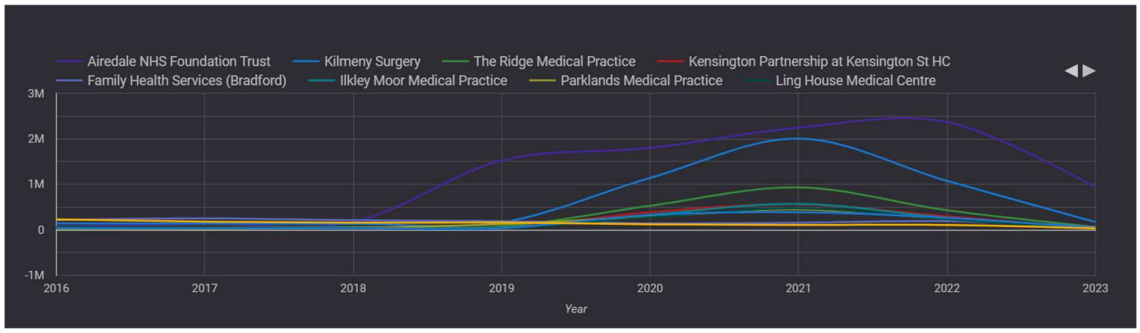


Figure 17: Showing the dataset with patients records that are missing code entry in the SNOMED (Code) Column.

On the contrary, a query was run to check if CTV3 codes were missing, and none were retrieved; this suggests that the issue lies within the mapping of the CTV3 codes to SNOMED. It is still unclear at this instance whether the missing codes are because some practices are still using CTV3 or if there is a delay in mapping the CTV3 codes to the SNOMED codes. If the latter is indeed the issue, they need to be sorted out at the earliest opportunity, as CTV3 codes have not been updated since 2018, and SNOMED codes are updated regularly, with some invalid codes. Mapping issues may arise, causing loss of data and information due to changes in definitions and descriptions.

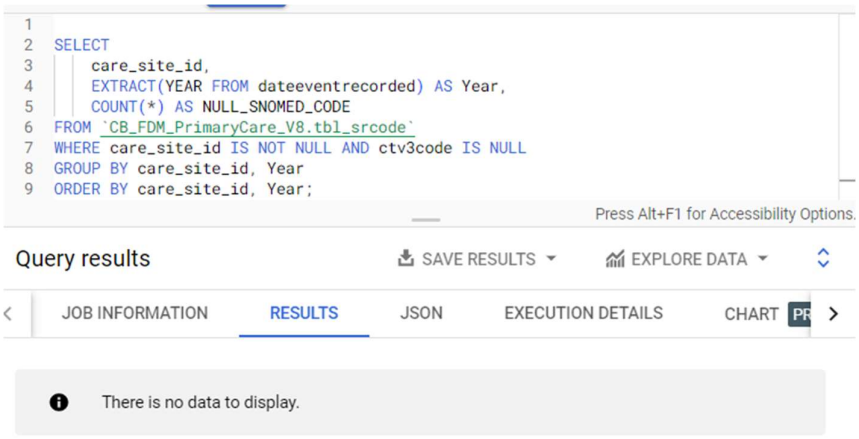


Figure 18: Showing no entries found in the same patients record table with no missing CTV3 codes.

4.5 Data Not being Recorded at the Correct Time.

Data not being recorded on the correct date, is the most significant problem faced. Often, the data is not recorded at the correct time, so the data is sent to a team of coders or admin team members who code the data based on conversation recordings or some other information format. The graph below is simplified to include only data from Acute Bronchitis diagnosed in 2020. This has been done specifically to show the scale of problems with the smallest of the categories in the most recent years. Recording data at a later event always has its issues since it leads to wrong codes being selected or the admin staff not having enough clinical knowledge to make the correct choices. The data has also been filtered after 2021 because the original dataset contains data greater than 1950, during which most records were on paper, so all data is accounted for in Figure 20**.** Using the visualisation helps them see the extent of the problem, which is otherwise not possible on large-scale data.

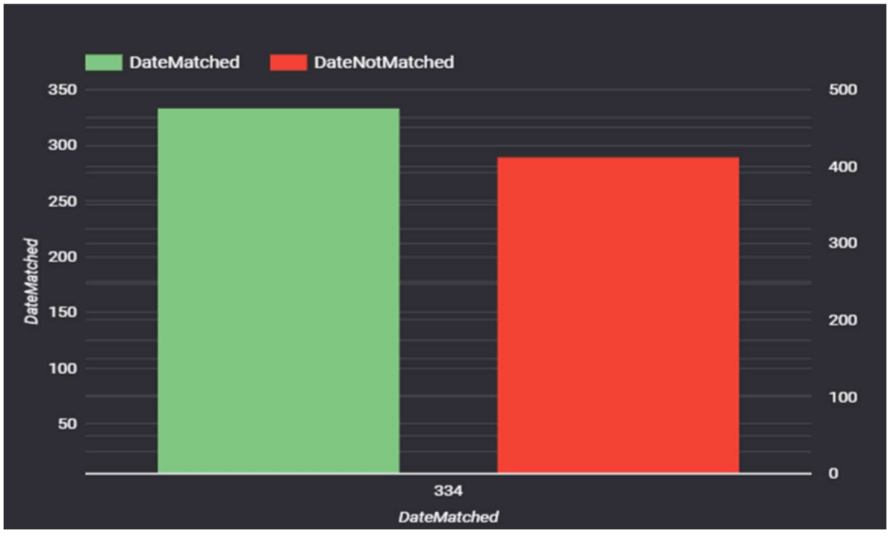


Figure 19 – Mismatching in the data for acute bronchitis in the year 2021 - Connected Bradford primary care dataset.

Figure 20 - mismatch in dates for the entire data set across all healthcare practises - Connected Bradford primary care dataset.

The figure above shows data with records over 3.9 million, where at least 1 million are not matched simultaneously. This graph includes all the data that have been transferred from paperwork to digital systems; therefore, there may be some skewing in the data, hence the use of Figure 19 to show the extent of mismatch within the most recent cases of 2020.

Figure 20 shows that almost 3.9 million cases are not recorded on the event date. The main aim of the graph is to show that if events are not recorded on the same date as they occur, discrepancies are bound to happen. This mainly happens when selecting codes for the diagnosis of the patients. Since the events are not recorded on the same day, there is no guaranteed proof that the exact current diagnosis code has been chosen as there are multiple codes with similar diagnosis descriptions for disorders such as spectrum disorders where each symptom changes the type of codes and descriptions selected. In other cases, multiple coding systems, such as ICD 10 and CTV3, are used in smaller GP practices. When the data are merged onto a larger system, they need to be merged on a single coding system to extract the data easily. Discrepancies occur when these datasets are merged if the correct codes are not selected from SNOMED that match the ICD 10 and CTV3 descriptions. As seen in the above cases where multiple CTV3 codes are mapped into single SNOMED codes, the variation in the descriptions and the level of depth in the knowledge is lost during conversion.

4.6 SNOMEDCT Codes Explored using Coronavirus and Influenza

Introduction of New Codes and their uses explored using the Coronavirus and Influenza as Examples. Often, new codes are introduced to accommodate new viruses' bacteria and other medical issues. The National Library of Medicine website mentions that "The IHTSDO (SNOMED CT International Edition) releases (and NLM makes it available) on January 31st and July 31st of every year. The NLM then produces the SNOMED CT US Edition and makes it available approximately one month following the international release in March and September each year." (www.nlm.nih.gov)

However, in extreme cases such as coronavirus, new codes are introduced quickly to accommodate changes in the logs to the datasets. Research into the SNOMED concept table from the Connected-Bradford dataset showed multiple codes for coronavirus in different subcategories for measurements and metadata.

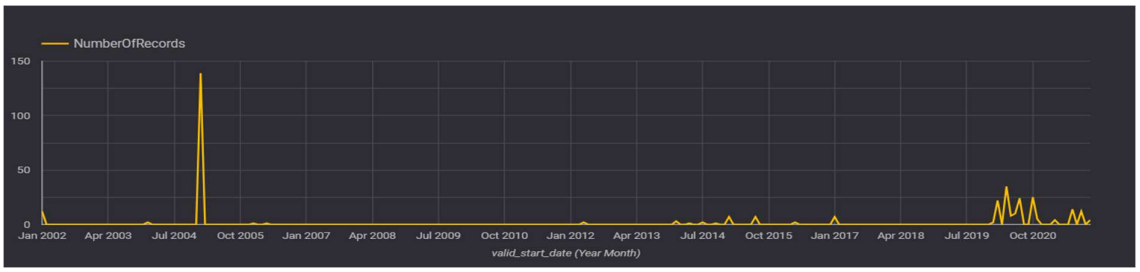


Figure 21 – Shows all the SNOMED codes and when they were released as a time series.

The above graph shows the time series when the coronavirus code was released. The codes above were extracted from the dataset using the Query below. The code concepts were limited to SNOMED, and a variation of the disease name was used to capture all the variations. The graph shows that most of the codes were released in 2005.

```
CREATE TABLE `CB_2107.NEW_COVID_SNOMED` AS
SELECT *
FROM `CB_CDM_VOCAB.concept`
WHERE vocabulary_id = 'SNOMED'
  AND (concept_name LIKE '%Coronavirus %' OR concept_name LIKE '%COVID %'OR
concept_name LIKE '%coronavirus %'OR concept_name LIKE '%corona %' )
  AND Invalid_reason  IS NULL ;
```

Row	ctv3code	ctv3text	snomedcode	CountOfRecords
1	A795.	Coronavirus infection	186747009	2276
2	AyuDC	[X]Coronavirus infection unspecified	186747009	25
3	X73lE	Coronavirus	697933000	282
4	XaaNr	Coronavirus contact	702547000	1001
5	Xab7a	Coronavirus nucleic acid detection assay	1029481000000103	5
6	Xaboe	Coronavirus ribonucleic acid detection assay	1008541000000105	4849
7	Y20fa	Coronavirus disease 19 caused by severe acute respiratory syndrome coronavirus 2 (disorder)	null	1143
8	Y20fa	Coronavirus disease 19 caused by severe acute resp	null	1143
9	Y228d	Coronavirus disease 19 caused by severe acute resp	null	1998
10	Y228d	Coronavirus disease 19 caused by severe acute respiratory syndrome coronavirus 2 confirmed by laboratory test (situation)	null	1998
11	Y228e	Coronavirus disease 19 caused by severe acute respiratory syndrome coronavirus 2 confirmed using clinical diagnostic criteria (situation)	null	8969
12	Y228e	Coronavirus disease 19 caused by severe acute resp	null	8969
13	Y229c	Coronavirus disease 19 severity score (observable entity)	null	42
14	Y229c	Coronavirus disease 19 severity score (observable	null	42
15	Y22a1	Coronavirus disease 19 caused by severe acute resp	null	1196
16	Y22a1	Coronavirus disease 19 caused by severe acute respiratory syndrome coronavirus 2 excluded (situation)	null	1196
17	Y22a2	Coronavirus disease 19 caused by severe acute resp	null	7646
18	Y22a2	Coronavirus disease 19 caused by severe acute respiratory syndrome coronavirus 2 excluded by laboratory test (situation)	null	7646
19	Y22a3	Coronavirus disease 19 caused by severe acute resp	null	32819
20	Y22a3	Coronavirus disease 19 caused by severe acute respiratory syndrome coronavirus 2 excluded using clinical diagnostic criteria (situation)	null	32819
21	Y22a5	Coronavirus disease 19 severity scale (assessment scale)	null	12
22	Y22aa	Signposting to CHMS (COVID-19 Home Management Service) (procedure)	null	145
23	Y22aa	Signposting to CHMS (COVID-19 Home Management Serv	null	145
24	Y26b0	PC19Q I have had a test for COVID-19	null	16
25	Y26b2	PC19Q My test for COVID-19 was positive	null	115
26	Y26b3	PC19Q My test for COVID-19 was negative	null	400
27	Y26b5	PC19Q Someone in my household has had a positive test for COVID-19	null	2

Figure 22 – Shows the CTV3 Codes and their definitions and SNOMED.

Most of these codes are released as vaccinations, new medication dosages, or to determine the PCR test or refusal to take vaccination. The codes also have information related to the SARS coronavirus and many more entries that are not relevant to this study's research purpose. Hence, the study has been limited to only include the codes that record Coronavirus as a condition and Observations.
A query was run to match all the CTV3 codes to match the SNOMED mappings to confirm the above selection of SNOMED codes.

```
CREATE TABLE `CB_2107.NEW_ALL_COVID_CASES_Grouped_CTV3` AS
SELECT ctv3code, ctv3text,snomedcode, COUNT(*) AS CountOfRecords
FROM `CB_2107.NEW_ALL_COVID_CASES`
GROUP BY ctv3code, ctv3text,snomedcode;
```

This step alone comes with much ambiguity, as filtering data further to match the SNOMED codes. The query below was run to categorise the codes to understand the concepts better – The reason this query was run specifically to confirm that the missing fields in Figure 22 are not missing SNOMED field mapping. The above query confirmed the results.

Untitled 2 RUN SAVE SHARE SCHEDULE MORE

```
1
2  SELECT A.ctv3code,CountOfRecords,ctv3text, B.SNOMEDCode
3  FROM `CB_2107.NEW_ALL_COVID_CASES_Grouped_CTV3` A
4  INNER JOIN `CB_LOOKUPS.tbl_CTV3ToSnomed_Map` B ON A.ctv3code = B.CTV3Code;
```

Processing location: europe-west2

Query results

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS | CHART PREVIEW | EXECUTION GRAPH

Row	ctv3code	CountOfRecords	ctv3text	SNOMEDCode
1	AyuDC	25	[X]Coronavirus infection unspecified	186747009
2	A795.	2276	Coronavirus infection	186747009
3	X73lE	282	Coronavirus	697933000
4	XaaNr	1001	Coronavirus contact	702547000
5	Xaboe	4849	Coronavirus ribonucleic acid detection assay	1008541000000105
6	Xab7a	5	Coronavirus nucleic acid detection assay	1029481000000103

Figure 23- Query showing the Matched SNOMED code retrieved from the matching table created for patients record.

These were the results that are extracted back.

When Figure 22 is inspected closely, the CTV3 codes without the SNOMED mappings are due to the other coding systems, such as “Nebraska Lexicon”. Hence the duplication of the CTV3 codes and the slight variation of the description. Majority of the codes are recorded under the “coronavirus infection” and “Coronavirus ribonucleic acid detection assay”.

Untitled 4 RUN SAVE SHARE SCHEDULE MORE

```
1  SELECT *
2  FROM `CB_2107.NEW_ALL_COVID_CASES_2017`
3  WHERE ctv3code = 'X731E';
```

Press Alt+F1 for Accessibility Options

Query results SAVE RESULTS EXPLORE DATA

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS | CHART PREVIEW

There is no data to display.

Figure 24 - Multiple Checking to ensure that cases are not missed.

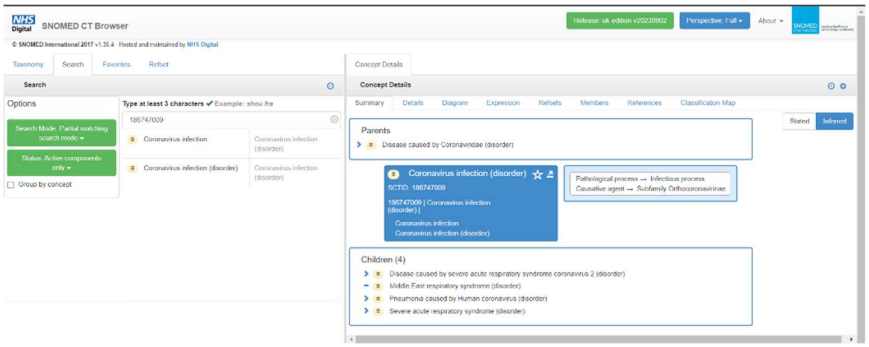


Figure 25 – Definition of Coronavirus Code on the SNOMEDCT website (NHS Digital, 2018)

The graphs below show that "Coronavirus infection" and "Coronavirus RNA (ribonucleic acid) detection assay" were introduced to the SNOMED category in the year 2002 and 2016, respectively. The code used the most for recording Coronavirus infection was 186747009 (Coronavirus Infection), introduced in 2002. The image above shows the definition of the code. It is already noted that the descriptions and definitions of SNOMED codes are one of the biggest problems that lead to discrepancies, and some of the codes are not being used as regularly as the descriptions are confusing or lack depth. In this case, the code does not have an extensive definition of the disorder's symptoms, just as coronavirus infection.

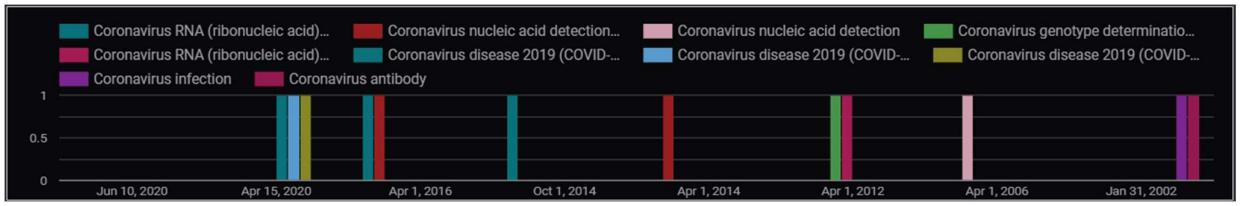


Figure 26 – Coronavirus – Timeseries of simplified codes.

The above graph shows a simplified version of the time series that shows when the most used type of coronavirus codes were released that were used to record the condition while it was happening. The graph also shows some codes being released twice with the same definition. Some codes are released under the concept_class_id of Model Command, Substance, Observable Entity and Procedure. One of the issues is if the wrong domain ideas are selected based on their description, cases can be recorded under the wrong SNOMED code. This issue is most likely to occur during the later stages, where the data admin staff are entering the codes later.

SCHEMA DETAILS PREVIEW LINEAGE

Row	concept_id	concept_name	domain_id	vocabula	concept_class_id	st	concept_code	valid_start_date	valid_end_date
4	44811805	Coronavirus nucleic acid detection assay	Measurement	SNO...	Procedure		906711000000107	2014-04-01	2099-12-31
5	37394268	Coronavirus nucleic acid detection assay	Measurement	SNO...	Observable Entity		1029481000000103	2016-04-01	2099-12-31
6	37392788	Coronavirus RNA (ribonucleic acid) detection assay	Measurement	SNO...	Observable Entity		1008541000000105	2016-04-01	2099-12-31
7	45770687	Coronavirus RNA (ribonucleic acid) detection assay	Measurement	SNO...	Procedure		933791000000101	2014-10-01	2099-12-31

Figure 27 – Showing the duplication of the Concept Names under multiple concept_class_id

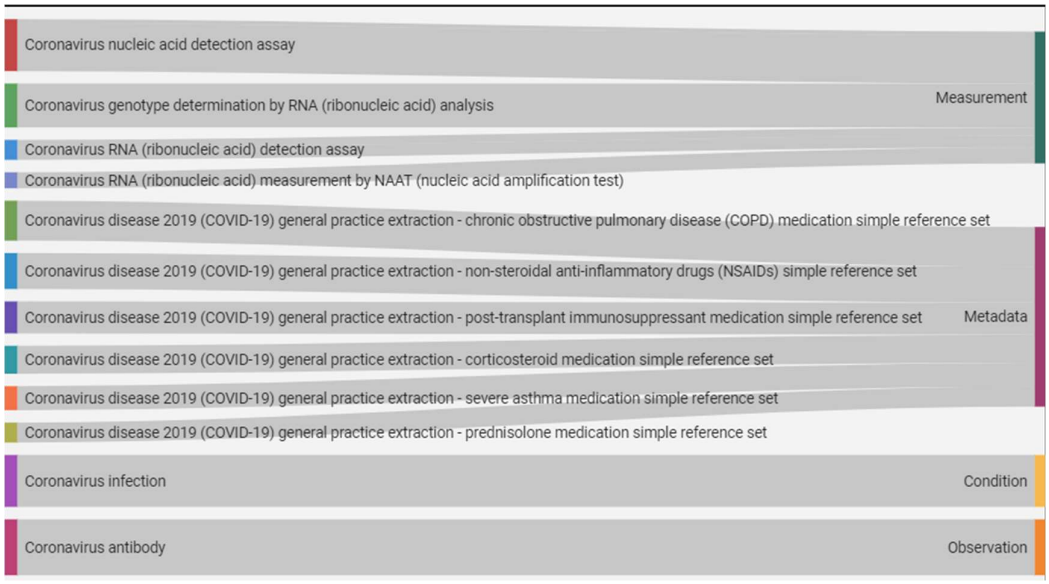


Figure 28 – Coronavirus SNOMED codes distributed into their Domain Categories.

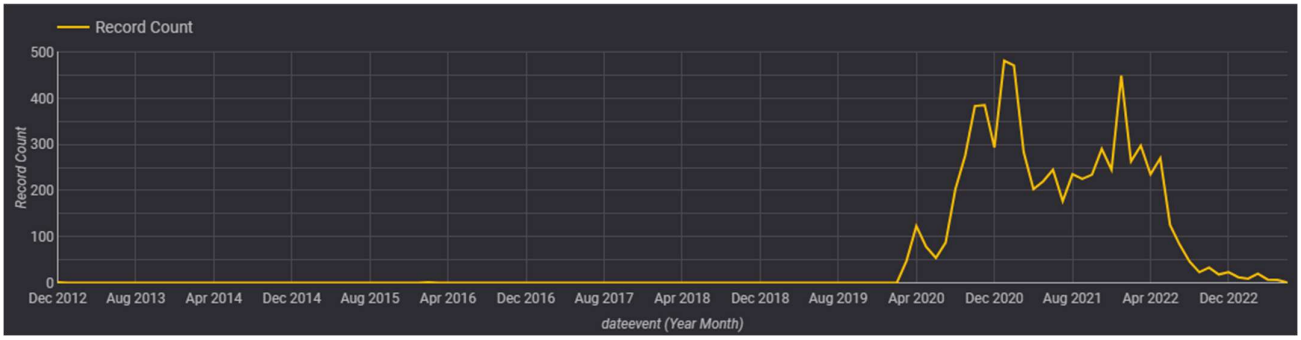


Figure 29 – Showing the rise in Coronavirus cases and when the SNOMED codes for covid were started to be used to record people's diagnosis.

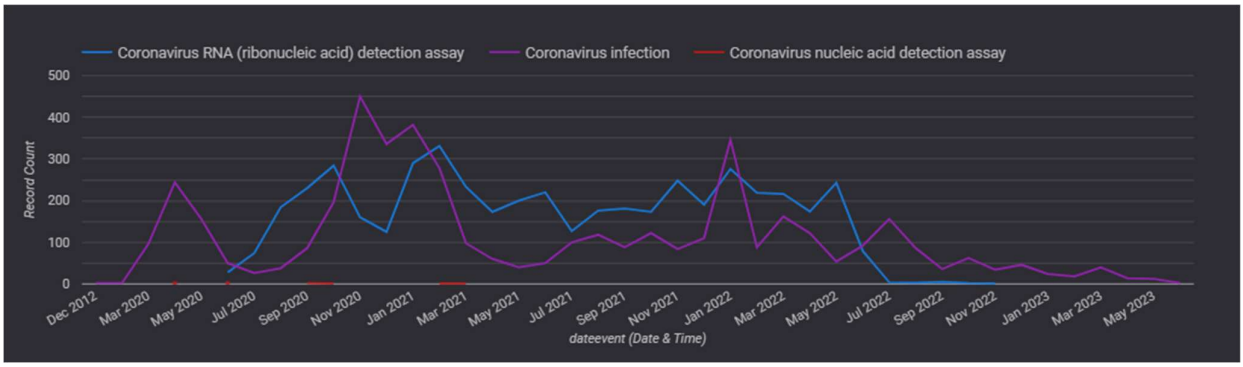


Figure 30 – Showing a detailed time series graph of the three most used SNOMED codes and how their uses varied across the peak covid times.

Figure 30 suggests a lack of definition within the description of the SNOMED codes. This means that administrative staff with less clinical knowledge are often more likely to enter codes based on a loose description of the incident rather than provide a code with a thorough definition. This means that data is recorded under the wrong entities or a loose definition that cannot be examined in detail. Coronavirus is a heightened problem as multiple variants and other symptoms were added and changed quickly to accommodate the changes in the diagnosis, which may lead to general codes being entered into the diagnosis list before detailed definitions are released. Furthermore, since during COVID, most admin staff were working from home, the information and the selection of codes is bound to happen on a later date than the date of the event, leading to missing information and, therefore, entering wrong SNOMED codes that match the general description but may not have an accurate definition of the exact diagnosis of the patient.

Figure 30 shows that "coronavirus infection" and "coronavirus RNA (ribonucleic acid) detection assay" are the two most frequently used codes to diagnose people. Figure 31 suggests that only a few practices used the "coronavirus RNA (ribonucleic acid) detection assay" code. However, most of the recorded cases are from Westcliff Health Innovations. The lack of use of a particular code or one particular code being used by certain practices will cause the data and research to skew if the research codes are not appropriately retrieved. The data suggests that NHS Bradford & Airedale palliative care services is one of the top medical practises to diagnose people with coronavirus infection code; this could be because some of these more minor GP surgeries may have been shut during the coronavirus.

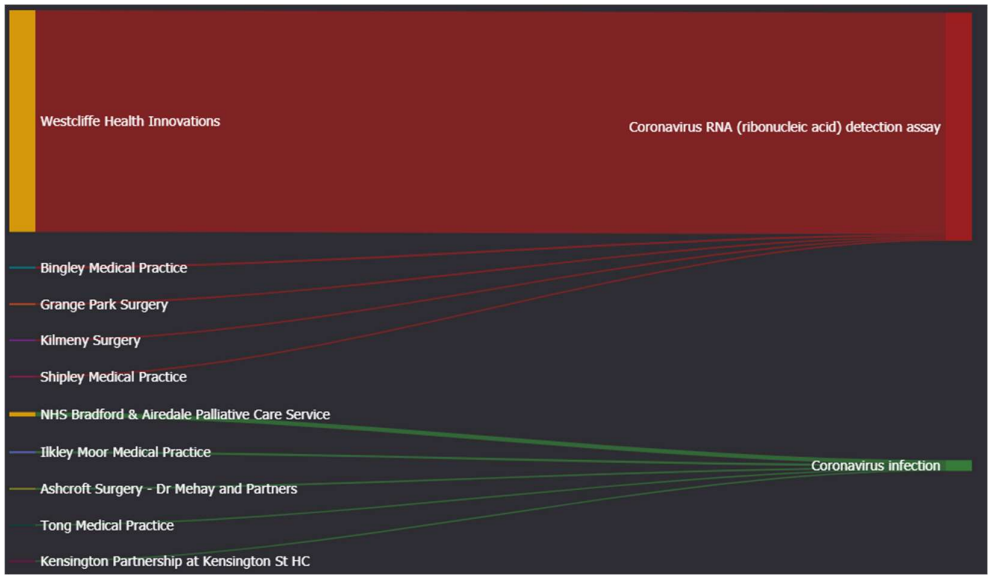


Figure 31- A Sankey diagram showing the distribution of codes for healthcare centres.

4.6.1 Influenza

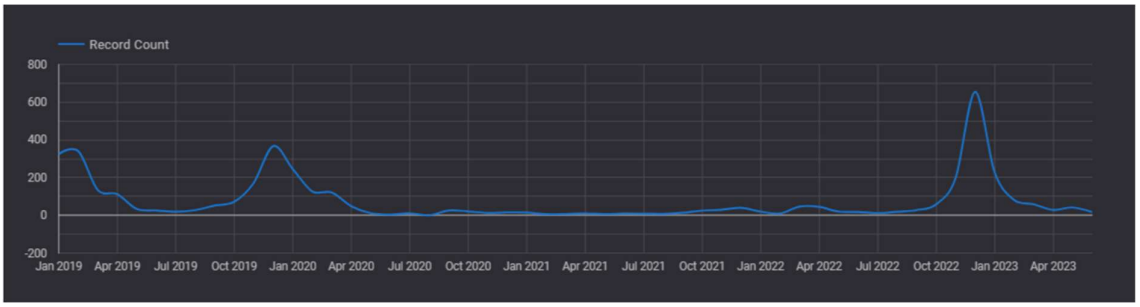


Figure 32- A graph showing the total record count of the influenza cases.

The graph above shows the fluctuations and the changes in the cases of influenza over the three years. In which coronavirus was the most dominant form of disorder across all healthcare practises. This query was run to check if there has been a discrepancy within the use of influenza codes over the three years, as most influenza symptoms matched with Coronavirus symptoms. Figure 32 - suggests that as soon as coronavirus cases started in March 2020, influenza cases almost went down to null. The only possible explanation is that since it was difficult in the beginning to distinguish if a patient genuinely had coronavirus due to a lack of testing kits, all cases with any symptoms listed under coronavirus were recorded under the SNOMED codes for COVID-19.
Figure 32 shows no peaks in January 2021, as shown in all previous years and 2023. This suggests that all possible influenza cases have either been unrecorded or recorded under coronavirus SNOMED codes.

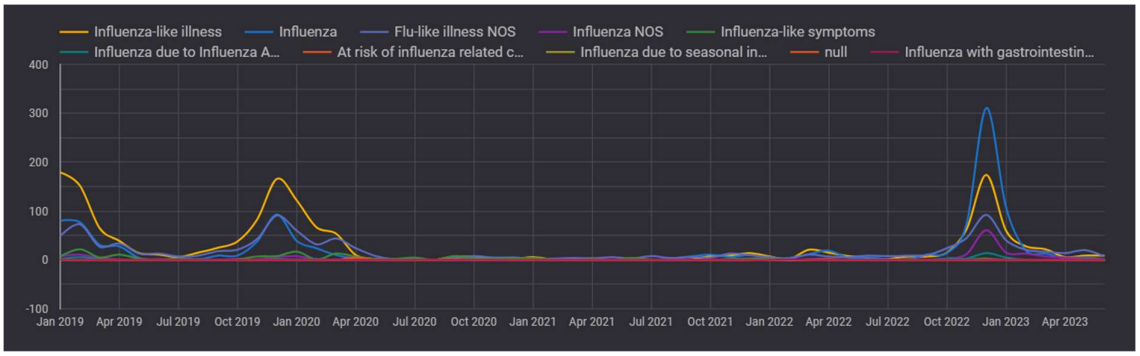


Figure 33 – A graph showing a detailed breakdown of the different types of SNOMED codes used to record the influenza condition.

Figure 33 shows that the SNOMED code "Influenza-like disease" was the most frequently used code. Many different GP surgeries use various coding systems, which have their own breakdown in the definition of the codes. The meaning of the other codes is most likely to get lost while the records are converted. Furthermore, the lack of definition, "Influenza-like disease", "Influenza", and "Flu-like illness" are very similar in descriptions, leading to wrong codes being chosen due to lack of clarity when the event is recorded later. This could be one of the reasons why the code "Influenza-like disease" is the most dominant type of code across the dataset. Codes "Influenza-like illness" and "Influenza-like Symptoms" can be easily mistaken by a person without a strong clinical background entering the data. On further analysis, it can also be noticed that in December 2022, there was a sudden spike in the "Influenza" code, while all the past peaks have been related to "Influenza-like illness." the sudden switch in the use of codes is difficult to understand, especially when the peak is very sharp and rises almost simultaneously with the use of "coronavirus" codes. In contrast, this switch/use of this particular code was not noticed in 2021, when almost no cases of influenza were recorded under any Influenza codes. The difference is highly contrasted from the "Expected codes" to the "Recoded codes".

The Pie chart below shows the different SNOMED code concept names for "Influenzas" as the main category. The graph shows that "Influenza-like illness" is the most used type of code to record the condition across all practices. The issue with this is the simplification of the code. Ambiguity is extremely high while recording illness under codes that have vague definitions. There is no particular way of knowing if the diagnosis was recorded under the correct condition unless the patient is later diagnosed under a different SNOMED code for a similar condition at a secondary care centre or the same healthcare centre. Furthermore, new questions are raised, such as "What are the differences that help to distinguish the selection of SNOMED code from "Influenza-like illness" or "Influenza-like symptoms"?

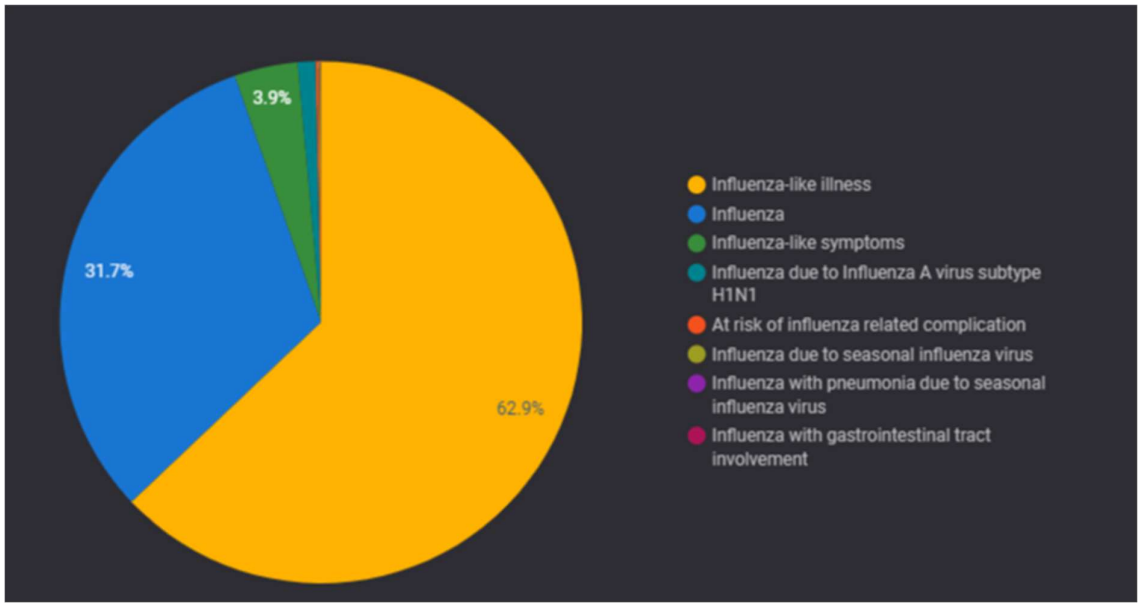


Figure 34 – Shows Acute Bronchitis data from the Year 2020

4.7 Issues regarding the "Expected Versus" VS "Observed Codes"

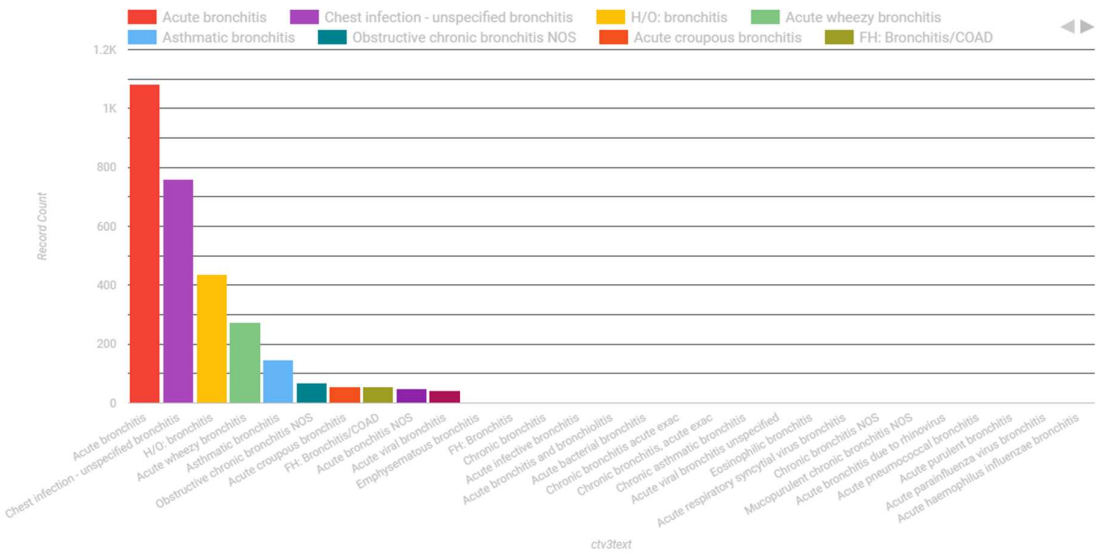


Figure 35– A graph showing all variations of the CTV3 codes available for Bronchitis.

Figure 35 Shows all the CTV3 codes available for Bronchitis to record the patient's diagnosis. As shown, "Acute Bronchitis" was the most used code. The search was run on all records with "Date Recorded from 2020-01-01. For this research, further queries and graphs have been exclusively focused on the "Acute Bronchitis" codes.

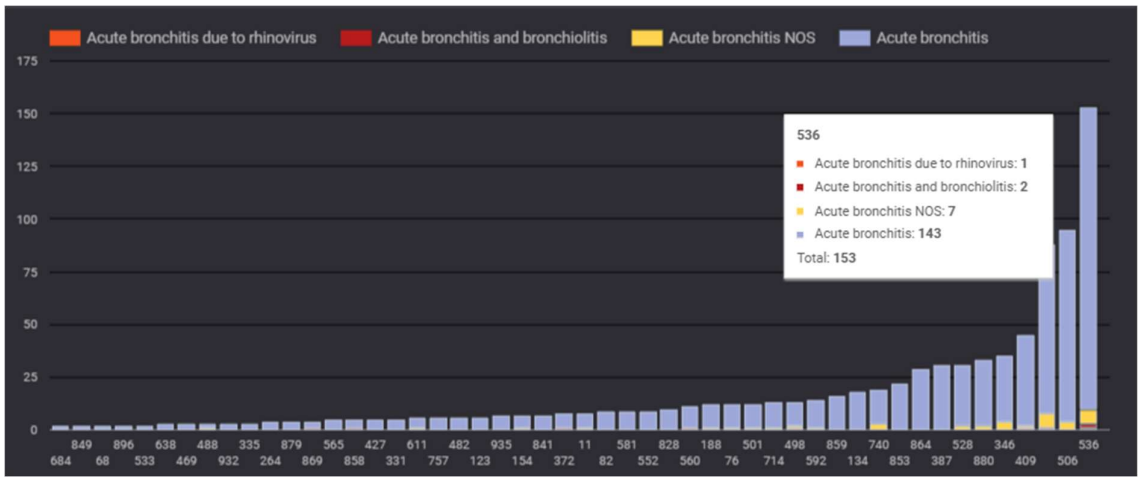


Figure 36 – The graph above shows the usage of different types of SNOMED codes.

Figure 36 shows the difference in the usage of the SNOMED codes. Acute Bronchitis is more commonly known as chest cold, which lasts less than three weeks CDC (2022). Figure 36 has multiple codes that can be used to diagnose Acute Bronchitis under CTV3 text, which matches the SNOMED concept codes. Figures 34 and 35 show a slight variation in the description of the codes from CTV3 to SNOMED. Although the difference may not be significant, it is still noticeable between the two categories. "Acute Bronchitis" and "Acute Bronchitis NOS" are codes under the CTV3 regulation mapped into the same SNOMED code. Figure 35 shows that multiple clinic cases are registered under the "Acute bronchitis NOS." When converted to SNOMED Concept codes, The granularity of these cases will be lost. "NOS" stands for "Not Otherwise Specified" and is a subcategory in disease/disorder classification systems. It is used to note the presence of a condition where the symptoms presented indicate a general diagnosis within a family of disorders but do not meet criteria established for specific diagnoses within that family." (Mental Health America)

The difference in this case is not of significant concern. However, Figures 34 and 35 show the need for more terminology and diagnosis descriptions when moving from one coding practice to another.

ctv3code	ctv3text	care_site_id	SNOMEDCode	concept_name
XE0Xr	Acute bronchitis	188	10509002	Acute bronchitis
XE0Xr	Acute bronchitis	188	10509002	Acute bronchitis
H060z	Acute bronchitis NOS	188	10509002	Acute bronchitis
H060z	Acute bronchitis NOS	188	10509002	Acute bronchitis
XE0Xr	Acute bronchitis	188	10509002	Acute bronchitis

Figure 37 – Variation in CTV3 and SNOMED code description

H06..	Acute bronchitis and bronchiolitis	372	195712009	Acute bronchitis and/or bronchiolitis
H06..	Acute bronchitis and bronchiolitis	372	195712009	Acute bronchitis and/or bronchiolitis
XE0Xr	Acute bronchitis	372	10509002	Acute bronchitis
XE0Xr	Acute bronchitis	372	10509002	Acute bronchitis

Figure 38 – Variation in CTV3 and SNOMED code description

Furthermore, Figure 35 shows that the Acute Bronchitis code is the most used code across all practices in the primary care dataset. Almost all cases from 2020 have been recorded under “Acute Bronchitis”. “Acute Bronchitis and Bronchiolitis” is rarely used – One of the main reasons behind this could be because Bronchitis and Bronchiolitis both have very similar symptom (Healthline, 2017) therefore, it is difficult to diagnose, in which case the clinicians and admin staff rely on selection the code that is acceptable for the general conditions. The more precise subcategories of codes go unused, leading to possible misdiagnosis. The issue is further carried on when the data is used for research, and the discrepancy in the selection of the codes may lead to inconsistent data from some practices, which will skew the data and its analysis.

4.8 Dashboard Screenshots

The dashboard is found on Page 9,10 &11

There may be issues accessing the dashboard due the Data sharing Agreement hence the screenshots are included in the document.

Link to the Dashboard - https://lookerstudio.google.com/s/qBzLdITeCuE

Link to the Dashboard - https://lookerstudio.google.com/reporting/6c049aef-f546-421d-ab14-778373d35fc5

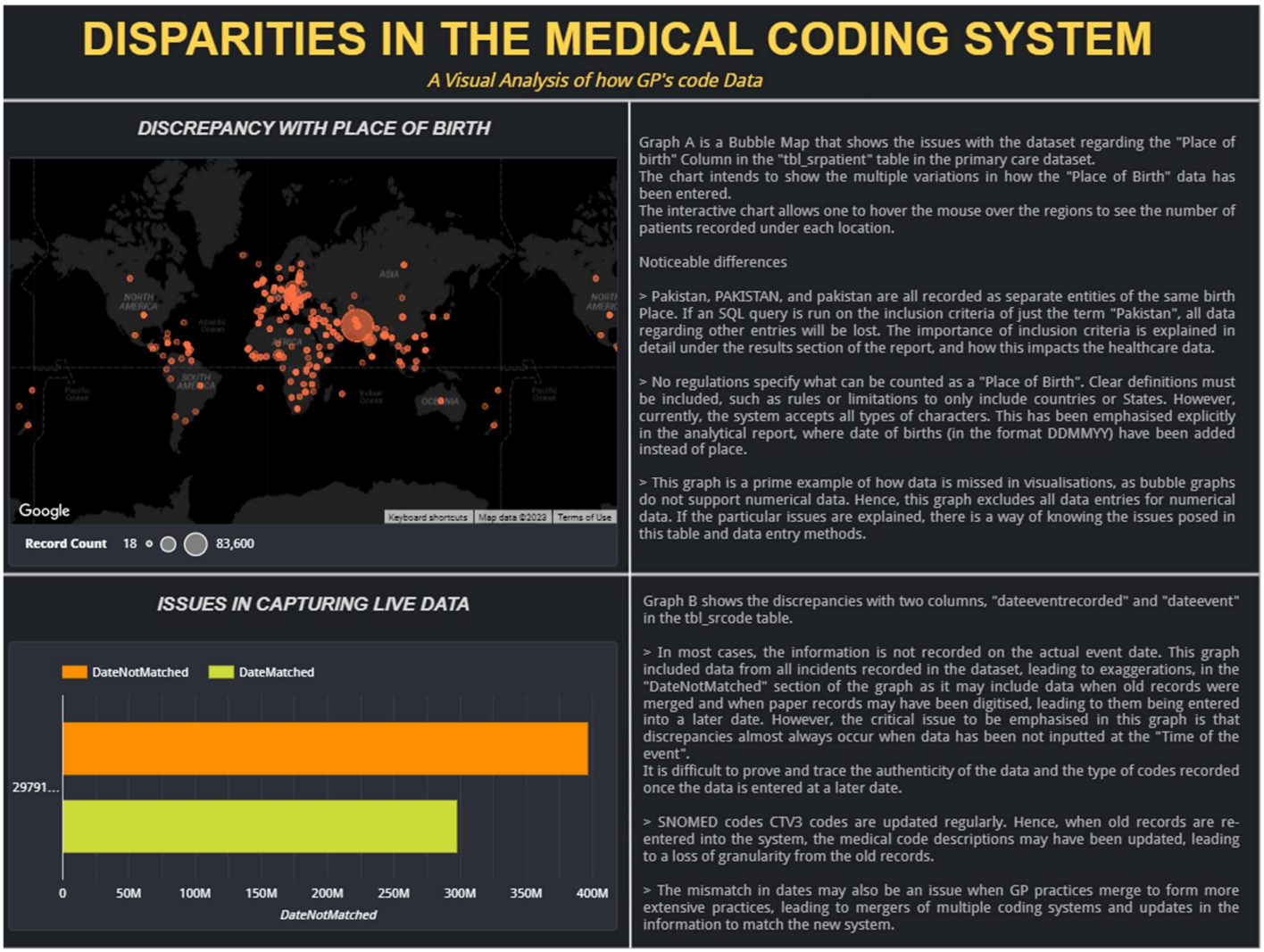


Figure 39 – Dashboard Screen 1

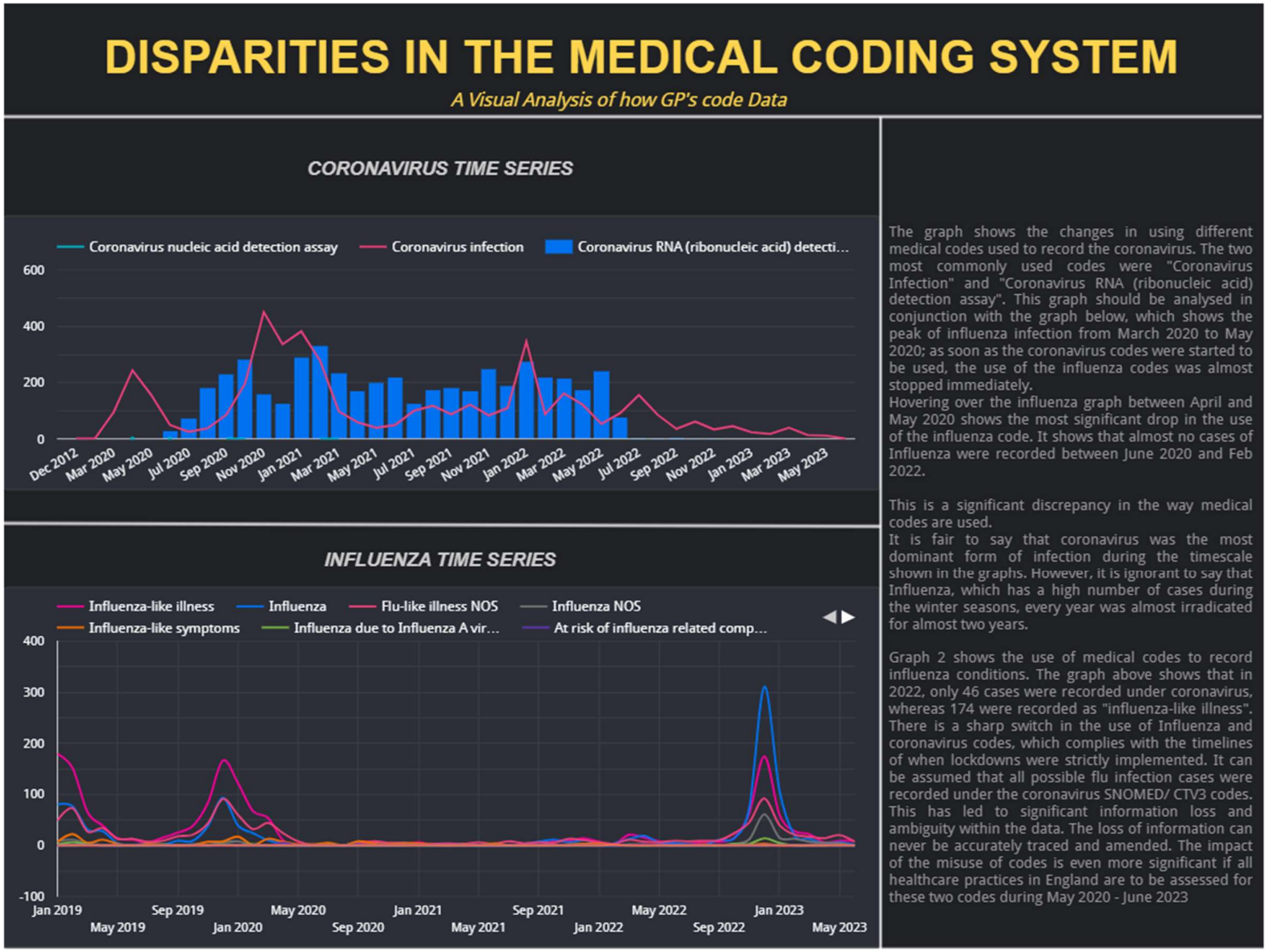


Figure 40 - Dashboard Screen 2

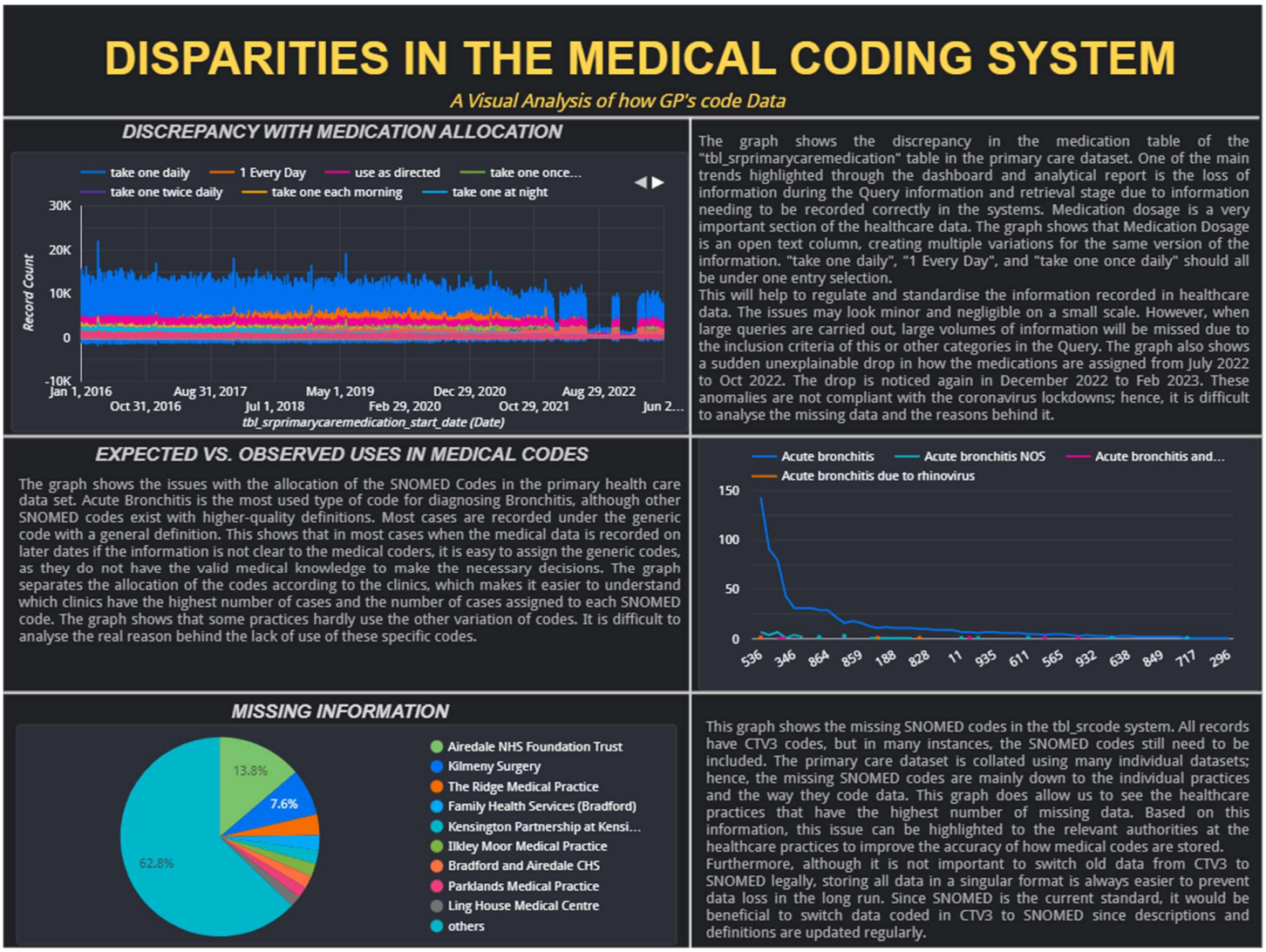


Figure 41 - Dashboard Screen 3

Chapter 5- Discussion

5.1 Objective Fulfilment

The project aimed to deliver an analytical workflow and a dashboard showing how GP's code data, specifically how the codes are used, cause discrepancies within healthcare data. That dashboard will allow the connected Bradford team members to check if specific issues in how the SNOMED codes are being used have a more significant impact on the healthcare system than initially thought and if there is a way to address these issues.

5.2 Existing Conversation Tools and Software

Existing tools and software have been used to showcase the issues within the dataset using visualisation created using Google Looker Studio. Google Looker Studio was the most accessible tool available to design the visualisations, as according to the data sharing agreement signed with connected Bradford, the data was not allowed to be downloaded and used in other platforms such as Tableau or Power BI. The advantages and disadvantages of the software used during the coursework are mentioned in the methodology section.

Initial research has shown some work being done regarding the uses of medical codes and how missing information affects healthcare data in general. However, no specific case studies showed the use of SNOMED code and how switching the codes from CTV3 to SNOMED could affect the medical data and its uses in the future.

The main limitation in terms of the tools being used is that the dataset is non-transferable from the Google Cloud platform. All features and tools used are from Google that could accommodate the user platform set up by the team at Connected Bradford.

5.3 Examining the data and Query.

The data was primarily inspected using SQL queries to create tables, which were then connected to Google Looker Studio to create meaningful graphs. The data queries are particular to the type of research conducted on this project; hence, it is very difficult to compare the visualisation results to other research studies. However, the general idea of the research carried out in this project can be similar to the research paper by Ruddle et al., which discusses the patterns of missing data from the NHS dataset using missing values. This report concentrates more on the effects of SNOMED codes and the issues faced with their descriptions and selection.

Inclusion criteria, discussed in Chapter 4, are highly important in the type of data selection and the results produced because of it. Although the initial research in Chapter 2 does discuss the issues with missing data and the validity of the healthcare datasets based on discrepancies, it is highly difficult to find examples that specifically concentrate on the issues faced with the SNOMED medical coding system and their effects on the NHS system.

The results section uses the connected Bradford dataset to show the specific issues within the medical coding system in the NHS. In contrast, other research studies were conducted on much larger datasets and focused on a broader range of issues than the one outlined as part of this research. Although the research follows a general trend of showing the issues within healthcare data, it can only be generalised as with any other research project if the inclusion criteria selected are the same.

5.4 Generating the Visualisation

All visualisations were generated through Google Looker Studio, which was easy to implement as Google Looker Studio has pre-defined features which allow one to connect data from big Query and create meaningful visualisation. There was an additional feature, such as filter selections, that could have been applied to the graphs, making it easy to create the graphs and Dashboard. Other software could have been explored to attempt other visualisations. However, the restriction with the dataset caused limitations.

5.5 Generating the Query and Story

It was a long process to research the type of disorders to investigate. The Athena dataset library and the Open code website helped recognise the types of medical codes to select and how the medical codes work. The variations in the descriptions of the codes vary, and since the changes were implemented to stop the use of CTV3 codes, it was more important to curate the queries in a way that showed the variations in the descriptions and however slight the changes were, it was vital to show in the narrative the difference it would make on a large scale on data traceability and issues that come with it. The selection of the disorders to investigate was fully understood after the interview with the Connected Bradford team, who helped to understand better the data collected and the backend process of how the data is coded. Analysing the interview in detail allowed us to identify minor details that look trivial on a small scale; however, when looked at from a broader perspective, it can substantially affect how medical data is used.

5.6 Interview Process

The selection of interview participants and conducting the interview itself was an important data collection and analysis process. The responses from the participants were very important in selecting the medical codes to show the discrepancies and to understand the backend process.

5.7 Results Validity and generalisability

The results obtained from this research should not be generalised to other medical coding systems, as the inclusivity criteria vary for every research. Furthermore, the data used for the research is a tiny sample of the healthcare data available. Although the primary care dataset from Connected Bradford has many records, the primary care dataset is very small compared to the datasets from NHS Digital and other larger datasets. On a bigger scale, the same inclusion criteria used in the study should be applied to larger datasets to obtain similar patterns and trends.

Each dataset research method has its variations. The type of visualisation depends specifically on the type of medical codes chosen to explore. However, the discrepancies in how medical data are coded are not inclusive to the dataset used in this research. This occurrence happens at every healthcare facility; hence, the general concept of the research can be applied to any healthcare dataset to check if similar results can be obtained.

5.8 Recommendations

The research showed that there are many discrepancies in healthcare datasets. To better understand the issue on a larger scale, it is recommended to use a larger dataset or datasets from a new healthcare trust that will help explore the issues on a broader scale or provide data for comparison. The queries should be designed to address broader disorders, where assigning a diagnosis can be challenging; this might highlight the medical code selection issues in a greater sense, as according to the research conducted above, the codes with general/ broad descriptions are often used more than the ones with greater depth in definition.

5.9 New Learning

The entire research allowed me to learn how to conduct data analysis in-depth, carry out a literature review to match the quality of new research and find data patterns on a large-scale dataset. Working with large-scale datasets comes with its challenges; in this instance, it allowed me to strengthen my knowledge of navigating around the Google Cloud platform and using Google Looker Studio.

While writing an analytical report comes with challenges, deviating from the research topic is often easy if other discoveries are made. It is essential to acknowledge which data is important in answering the question proposed at the beginning of the research.

This report helped me understand how complex research questions can be broken down into multiple stages to explain the method, the reasoning behind the methods taken, and how this helped answer the specific question.

5.10 Answering the Research Question

The research question proposed at the beginning of the report has been answered successfully, as the multiple disparities in the dataset are highlighted, with particular emphasis on the use of SNOMED codes and the discrepancies that occur based on the lack of descriptions and the use of multiple coding systems.

Chapter 6: Evaluation, Reflections and Conclusions

6.1 Choice of objectives

The project's main objective is to achieve an analytical report with visualisations demonstrating discrepancies in at least one stage of the healthcare data system. The report provides strong evidence about the issues within the primary care dataset provided by the connected Bradford team, which comprises data from 61 healthcare practices and how these issues may be prevalent in other healthcare practices.

The two main objectives were to design a dashboard of graphs and write an analytical workflow by using SNOMED codes to show discrepancies that occur in the current dataset and how this impacts the healthcare data in general. Through the report, the Connected Bradford team should be able to better understand the scale of these issues and how some of the issues can be minimised. Both these objectives were achieved successfully, with the report and dashboard representing some significant issues in the selection of medical codes – specifically SNOMED and CTV3.

The data researching, interviews, analysis and writing of the Query were very interesting parts of the project, as they allowed us to gain very specific information about a very niche subject related to healthcare data, which needs to be more in-depth investigated. Therefore, discovering new information, trends and patterns in the dataset was highly valuable.

The report shows the minor and major issues within the dataset. Some of the issues highlighted can be sorted quickly by changing how information is recorded. In contrast, issues regarding SNOMED can only be monitored and reported since the NHS monitors the coding system, and it is challenging to suggest or implement major changes based on the small-scale research conducted in this study.

6.2 Limitations of the Project

The analytical report is aimed to be a structured report that shows the issues within the current dataset. Hence, one of the main limitations is the generalisability of other healthcare trust datasets. The types of coding systems vary from each healthcare practice. Therefore, new strategies cannot be derived based on a single dataset to be generalised across other healthcare practices, as it will not be an effective solution.

6.3 Future Work

To improve the credibility of the overall research, it is essential to carry out the research on a much larger scale. Conducting the research on a wider scale will help to show if the issues investigated in this report are specific to certain demographical areas or if the general trend is followed across all primary healthcare datasets in terms of discrepancies in the use of the medical codes.

Furthermore, one of the limitations faced by this project is that the data sets could not be retrieved from the cloud platform due to the data protection statement; hence, the creation of visualizations could not be experimented with using other platforms. Therefore, for future projects, if there is a potential to download the datasets and experiment further with other platforms, it would be interesting to check the results in comparison.

Moreover, this report highlights the discrepancies that occur during the admin stage of the work. However, in future work, other stages, such as processing and retrieval, can be investigated to see if a wider range of results can be collected.

More interview participants, including clinicians and medical coders, can be interviewed to understand their perspective on how the data is entered and dealt with.

6.4 Conclusion

This project was a good starting point in investigating a small section of a large-scale problem. Healthcare data is highly valuable and challenging to maintain and upkeep due to the daily volume of data collected. As mentioned in the report, if some minor changes are applied to the data entry system, it would improve the data quality and make a significant difference in the future uses of the data. Visualisation helps to notice large-scale trends and patterns in the data, which can be studied in depth based on the impact it could have on the community. Therefore, this report is a starting step in much more comprehensive research that should be carried out on a broader scale to improve the authenticity of the data and derive new strategies.

Glossary

- **SNOMED CT –** Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) is a structured clinical vocabulary for the use in an electronic record.
  Throughout the report the SMOMED CT codes have been referred as SNOMED codes.

- **CTV3 Codes** – The CTV3 codes also knows as Read coded clinical Terms are a comprehensive computerised coded thesaurus for use by clinicians. They are available in two formats Version2 and Version 3. Version – the most updated version of CTV codes is known as CTV3 Codes. The CTV3 codes are refers throughout the document as CTV3 codes or Read codes.

- **ICD10** – International Classification od Disease, Tenth Revision (ICD-10) The international Classification of the Diseases.
-
- **Primary care Dataset** – Referred throughout the report as “dataset”, “Connected Bradford Dataset”.

References

- Ruddle, R.A., Adnan, M. and Hall, M. (2022). Using set visualisation to find and explain patterns of missing values: a case study with NHS hospital episode statistics data. *BMJ Open*, 12(11), p.e064887.

- Khare, R., Utidjian, L., Ruth, B.J., Kahn, M.G., Burrows, E., Marsolo, K., Patibandla, N., Razzaghi, H., Colvin, R., Ranade, D., Kitzmiller, M., Eckrich, D. and Bailey, L.C. (2017). A longitudinal analysis of data quality in a large pediatric data research network. *Journal of the American Medical Informatics Association*, 24(6), pp.1072–1079.

- Bacon, S. and Goldacre, B. (2019). Overcoming Barriers to Working With NHS Open Data (Preprint). *Journal of Medical Internet Research*.

- Ghafur, S., Fontana, G., Halligan, J., O'shaughnessy, J. and Darzi, A. (No Date). Maximising its impact on the health and wealth of the United Kingdom.

- NHS Digital (2018). *SNOMED CT - NHS Digital*. NHS Digital. Available at: https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct. (Accessed 10 Sept. 2023).

- Petersen, I., Welch, C.A., Nazareth, I., Walters, K., Marston, L., Morris, R.W., Carpenter, J.R., Morris, T.P. and Pham, T.M. (2019). Health indicator recording in UK primary care electronic health records: key implications for handling missing data. *Clinical Epidemiology*, Volume 11, pp.157–167.

- Fung, K.W., Xu, J., Ameye, F., Gutierrez, A.R. and D'Have, A. (2018). Achieving Logical Equivalence between SNOMED CT and ICD-10-PCS Surgical Procedures. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2017, pp.724–733. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977651/.

- Hammad, M. (2020). *RUP and its Phases*. GeeksforGeeks. Available at: https://www.geeksforgeeks.org/rup-and-its-phases/. (Accessed 10 Sept. 2023).

- Unwin, A. (2020). Why Is Data Visualization Important? What Is Important in Data Visualization? *Harvard Data Science Review*, 2(1).

- Sedlmair, M., Meyer, M. and Munzner, T. (2012). Design Study Methodology: Reflections from the Trenches and the Stacks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12), pp.2431–2440.

- Midway, S.R. (2020). Principles of Effective Data Visualization. *Patterns*, [online] 1(9), p.100141.

- SNOMED CT in SystmOne Guidance for users. (No Date.). Available at: https://customerportal.notts-his.nhs.uk/uploads/899/SNOMED_CT_in_SystmOne_2.0_Nov_2018.pdf (Accessed 1 Sept. 2023).

- termbrowser.nhs.uk. (No Date.). *NHSDigital SNOMED CT Browser*. [online] Available at: https://termbrowser.nhs.uk/?perspective=full&conceptId1=61937009&edition=uk-edition&release=v20230802&server=https://termbrowser.nhs.uk/sct-browser-api/snomed&langRefset=999001261000000100 (Accessed 5 Sept. 2023).

- SNOMED CT Fact Sheet Inactive Codes. (2019). Available at: https://confluence.ihtsdotools.org/download/attachments/96799441/SNOMED_CT_Inactives_Fact_sheet_v2.0.pdf?api=v2#:~:text=Concepts%20are%20made%20inactive%20if (Accessed 12 Sept. 2023).

- confluence.ihtsdotools.org. (No Date). *Concept Inactivation - SNOMED CT Editorial Guide - SNOMED Confluence*. [online] Available at: https://confluence.ihtsdotools.org/display/DOCEG/Concept+Inactivation (Accessed 12 Sept. 2023).

- W3Schools (2019). *SQL Joins*. [online] W3schools.com. Available at: https://www.w3schools.com/sql/sql_join.asp. (Accessed 25 Jul. 2023).

- NHS Digital. (No Date.). *Read Codes*. [online] Available at: https://digital.nhs.uk/services/terminology-and-classifications/read-codes. (Accessed 2 Sept. 2023).

- www.england.nhs.uk. (No Date.). *NHS England» SNOMED CT*. [online] Available at: https://www.england.nhs.uk/digitaltechnology/digital-primary-care/snomed-ct/. (Accessed 15 Sept. 2023).

- SNOMED CT in SystmOne Guidance for users. (No Date). Available at: https://customerportal.notts-his.nhs.uk/uploads/899/SNOMED_CT_in_SystmOne_2.0_Nov_2018.pdf. (Accessed 16 Sept. 2023).

- CDC (2022). *Suffering from a chest cold?* [online] Centers for Disease Control and Prevention. Available at: https://www.cdc.gov/antibiotic-use/bronchitis.html#:~:text=A%20chest%20cold%2C%20often%20called. (Accessed 21 Sept. 2023).

- Mental Health America. (No Date). *Not Otherwise Specified, Other Specified Disorder or Unspecified Disorder*. [online] Available at: https://mhanational.org/conditions/not-otherwise-specified-other-specified-disorder-or-unspecified-disorder. (Accessed 15 Sept. 2023).

- Healthline. (2017). *Bronchiolitis vs. Bronchitis: Symptoms, Causes, and Treatments*. [online] Available at: https://www.healthline.com/health/bronchiolitis-vs-bronchitis#symptoms. (Accessed 14 Sept. 2023).

- Google Cloud (2023). *Advantages Of Cloud Computing*. [online] Google Cloud. Available at: https://cloud.google.com/learn/advantages-of-cloud-computing. (Accessed 19 Sept. 2023).

- Google Cloud. (No Date). *What is BigQuery?* [online] Available at: https://cloud.google.com/bigquery/docs/introduction#bigquery-analytics (Accessed 19 Sept. 2023).

- support.google.com. (No Date). *Welcome to Looker Studio! - Looker Studio Help*. [online] Available at: https://support.google.com/looker-studio/answer/6283323?hl=en. (Accessed 19 Sept. 2023).

- Jupyter (2019). *Project Jupyter*. [online] Jupyter.org. Available at: https://jupyter.org/. (Accessed 19 Sept. 2023).

- www.nobledesktop.com. (No Date). *Top 5 Uses for Jupyter Notebook | Classes Near Me Blog*. [online] Available at: https://www.nobledesktop.com/classes-near-me/blog/top-uses-for-jupyter-noteboook#:~:text=Jupyter%20Notebook%20is%20widely%20used (Accessed 19 Sept. 2023).

- ATLAS.ti. (No Date). *Differences between Thematic & Content Analysis*. [online] Available at: https://atlasti.com/guides/qualitative-research-guide-part-2/content-analysis-vs-thematic-analysis. (Accessed 21 Sept. 2023).

- Rev (2023). *How to Analyze Interview Transcripts in Qualitative Research*. [online] Rev. Available at: https://www.rev.com/blog/transcription-blog/analyze-interview-transcripts-in-qualitative-research. (Accessed 21 Sept. 2023).

- Sewell, M. (No Date). *QUALITATIVE INTERVIEWS IN EVALUATION*. [online] ag.arizona.edu. Available at: https://ag.arizona.edu/sfcs/cyfernet/cyfar/Intervu5.htm. (Accessed 21 Sept. 2023).

- www.nlm.nih.gov. (No Date). *SNOMED CT FAQs*. [online] Available at: https://www.nlm.nih.gov/healthit/snomedct/faq.html#:~:text=The%20IHTSDO%20(SNOMED%20CT%20International. (Accessed 9 Sept. 2023).

- Patino, C.M. and Ferreira, J.C. (2018). Inclusion and Exclusion Criteria in Research studies: Definitions and Why They Matter. *Jornal Brasileiro De Pneumologia*, [online] 44(2), p.84.

- www.opencodelists.org. (No Date). *OpenCodelists*. [online] Available at: https://www.opencodelists.org/ (Accessed 12 Sept. 2023).

- Muralidhar, K.S.V. (2021). *Merging tables using SQL*. [online] Medium. Available at: https://towardsdatascience.com/merging-tables-using-sql-a2e60ff687e9.
- Mode Resources. (2016). *SQL UNION | Intermediate SQL - Mode*. [online] Available at: https://mode.com/sql-tutorial/sql-union/.

Reference list included references used for the SLQ query.

Appendix I

SQL Queries

Visualisation from Google Looker studio

SQL Querry

Initital Test Query to Examin the Dataset

In [ ]:

```
CREATE TABLE `CB_2107.SNOMED` AS
SELECT * FROM `CB_LOOKUPS.tbl_EFI2_Codelist`
WHERE (Codedescription LIKE "%bronchitis")
#Querry 1 - was run on simple dissorder to see code dupliacation exist
```

In [ ]:

```
CREATE TABLE `CB_2107.SNOMED_FLU` AS
SELECT * FROM `CB_LOOKUPS.tbl_EFI2_Codelist`
WHERE (SNOMEDCT_CONCEPTID LIKE "%6142004 ")
##### This code was run to find all the types of flue codes that are avaialble on the SNOMED data list for FLU
```

In [ ]:

```
CREATE TABLE `CB_2107.People_Details` AS
SELECT * FROM `CB_FDM_PrimaryCare_V8.person`

WHERE (year_of_birth) > 1950
  AND death_datetime IS NULL;
### This code is run to create a table from the entire dataset, that shows only people born after the year 1950,
#the dataset contains data that is as Old as people born in the 1900, however for the purpose if this reserch
#I have decided to unly include the poeple from 1950
```

In [ ]:

```
CREATE TABLE `CB_2107.death_Count_null_per_Year` AS
SELECT (year_of_birth) AS birth_year,
  COUNT(*) AS null_death_count
FROM
   `CB_FDM_PrimaryCare_V8.person`
WHERE
  death_datetime IS NULL
GROUP BY (year_of_birth)
ORDER BY (year_of_birth);
### This Query was run to see how many death data was captured and how many null
#values per data per year, this was done to check the missinging value feature
```

In [ ]:

```
SELECT (year_of_birth) AS birth_year,
  COUNT(*) AS null_birth_count
FROM
   `CB_FDM_PrimaryCare_V8.person`
WHERE
  year_of_birth IS NULL
GROUP BY (year_of_birth)
ORDER BY (year_of_birth);
#### There are no null values for Date of birth - That data is captured perfectly
```

In [ ]:

```
SELECT
  COUNT(*) AS ethnicity_source_value
FROM
  `CB_FDM_PrimaryCare_V8.person`
WHERE
  ethnicity_source_value = 'Unknown/Refuse to say';
#This category should have been split into two - to analyse if people did not know their ethnicity or
# if they refused to say? There is no way to distinguish these details. Also, was this data recorded on the day the patient was
# registered? or was it captured later and only entered as "Unknow/refuse to say" because the people who recorded the
#did not know then? There should be two separate sections, and this field must be made mandatory to
#understand the origins of people, especially considering how diverse the country is and how much this data could mean.
#Also, when this data gets carried forward, there is a possibility that there are two separate categories.
#but since this data is not separated.. which category will this fit into? If it is placed in the wrong category, will that miss
#Lead research?
```

In [ ]:

```
SELECT
p.person_id,
a.care_site_id,
m.nameofmedication,
c.care_site_name
FROM `yhcr-prd-phm-bia-core.CB_FDM_PrimaryCare_V8.person` AS p
LEFT JOIN `yhcr-prd-phm-bia-core.CB_FDM_PrimaryCare_V8.tbl_srappointment` AS a ON p.person_id = a.person_id
LEFT JOIN `yhcr-prd-phm-bia-core.CB_FDM_PrimaryCare_V8.tbl_srprimarycaremedication` AS m ON p.person_id = m.person_id
LEFT JOIN `yhcr-prd-phm-bia-core.CB_FDM_PrimaryCare_V8.care_site` AS c ON a.care_site_id = c.care_site_id
LIMIT 1000;
```

```
Find out the count CTV3 codes that are assigned according to thier care site id
number of record per CTV3 code per Care site ID -
Then find the code for Broncitis through open code list
see which GP practice used which of the codes the most often..
see if there is a problem? - If a certain code is used more often then others?
```

In [ ]:

```
CREATE TABLE `CB_2107.MassTest` AS
  SELECT
  P.person_id,
  C.person_id AS srcode_person_id,
  C.care_site_id,
  Ca.care_site_name,
  C.ctv3code,
  C.ctv3text,
  C.dateevent

FROM `yhcr-prd-phm-bia-core.CB_FDM_PrimaryCare_V8.person` AS P
JOIN
  `yhcr-prd-phm-bia-core.CB_FDM_PrimaryCare_V8.tbl_srcode` AS C ON P.person_id = C.person_id
JOIN
  `yhcr-prd-phm-bia-core.CB_FDM_PrimaryCare_V8.care_site`  AS Ca ON C.care_site_id = Ca.care_site_id;
# Testing to join the tables based on Primary Key
```

In [ ]:

```
CREATE TABLE `CB_2107.Aggregate_Data_MassTest` AS
SELECT
  C.care_site_id,
  C.ctv3code,
  C.ctv3text,
  COUNT(*) AS RecordCount
FROM
  `yhcr-prd-phm-bia-core.CB_FDM_PrimaryCare_V8.person` AS P
JOIN
   `yhcr-prd-phm-bia-core.CB_FDM_PrimaryCare_V8.tbl_srcode` AS C ON P.person_id = C.person_id
JOIN
  `yhcr-prd-phm-bia-core.CB_FDM_PrimaryCare_V8.care_site`  AS Ca ON C.care_site_id = Ca.care_site_id
GROUP BY
  C.care_site_id,
  C.ctv3code,
  C.ctv3text
ORDER BY
  C.care_site_id,
  C.ctv3code;

#### group data in the above table by  record count  CareSiteID and CTV3 Code - This shows the care site ID, the CTV Code
# and the number of patients per code per care site - This shows if a certain practice uses a cerain code more regularly
# than the other practices THIS IS FOR ALL ISSUES
```

In [ ]:

```
CREATE TABLE `CB_2107.Bronchitis` AS
SELECT ctv3code,ctv3text,person_id,care_site_id
FROM `CB_FDM_PrimaryCare_V8.tbl_srcode`
WHERE ctv3text LIKE '%bronchitis%';
# Query to get all the patients that have bronchitis
```

In [ ]:

```
CREATE TABLE `CB_2107.Bronchitis_Test` AS
SELECT ctv3code,
ctv3text,
person_id,
care_site_id,
FROM `CB_FDM_PrimaryCare_V8.tbl_srcode`
WHERE ctv3text LIKE '%bronchitis%'
GROUP BY
  care_site_id,
  ctv3code,
  ctv3text,
  person_id;
#Table created - May not be a useful code - coz nothing is aggregated - it just pulls each indevidual person that has
#bronchitis per caresite id
```

In [ ]:

```
CREATE TABLE `CB_2107.Bronchitis_Aggregated_by_CareSite` AS
SELECT
  care_site_id,
  ctv3code,
  ctv3text,
  COUNT(DISTINCT person_id) AS PersonCount
FROM
  `CB_FDM_PrimaryCare_V8.tbl_srcode`

WHERE
  ctv3text LIKE '%bronchitis%'

GROUP BY
  care_site_id,
  ctv3code,
  ctv3text
ORDER BY
  care_site_id,
  ctv3text;

# Get the numebr of records at each site, per type of bronchitis
#A table has ctv3code, ctv3text, person_id, care_site_id - i want to know the number of people
#diagnosed with each disease in each care_site_id , where the code meaning contains the term "bronchitis"
```

In [ ]:

```
CREATE TABLE `CB_2107.Dates_notMatch_All` AS
SELECT care_site_id,
  CASE WHEN dateeventrecorded = dateevent THEN 'Match'
       ELSE 'DontMatch' END AS RecordCategory,
  COUNT(*) AS RecordCount
FROM
  `CB_FDM_PrimaryCare_V8.tbl_srcode`
GROUP BY
  dateevent,
  care_site_id,
  dateeventrecorded;

    ## A table has Table had incidenthappened and incidentrecorded, for each record these two dates may not be exactly the same,
    #Write SQL to get the number of records that have matching and mismatching records
```

In [ ]:

```
CREATE TABLE `CB_2107.Dates_notMatch_All_groupedbyCaresite` AS
SELECT
  care_site_id,
  SUM(CASE WHEN dateevent = dateeventrecorded THEN 1 ELSE 0 END) AS matching_records,
  SUM(CASE WHEN dateevent <> dateeventrecorded THEN 1 ELSE 0 END) AS mismatching_records
FROM
  `CB_FDM_PrimaryCare_V8.tbl_srcode`
GROUP BY
  care_site_id;
    # Total number of Matched and not matched number of Date event happend and date recorded per care_site_id
    # This needs to be matched with Care_site table to get the care site Name
```

In [ ]:

```
CREATE TABLE `CB_2107.Bronchitis_Care_Site_Name` AS
SELECT x.*, y.care_site_name
FROM `CB_FDM_PrimaryCare_V8.care_site` y
INNER JOIN `CB_2107.Bronchitis_Aggregated_by_CareSite`x ON
 x.care_site_id = y.care_site_id
#Inner join to join get the care_site Name and id Matched
```

In [ ]:

```
CREATE TABLE `CB_2107.Bronchitis_Care_Site_Acute_Bron` AS
SELECT *
FROM `CB_2107.Bronchitis_Care_Site_Name`
WHERE ctv3text LIKE '%Acute bronchitis%';
# Only care sites with Accute Bronchitis were selected to see if codes are selcted more often then others
```

Null SNOMED VALUES

In [ ]:

```
CREATE TABLE `CB_2107.NULL_SNOMED_CODE_PER_YEAR_SITE` AS
SELECT
    care_site_id,
    EXTRACT(YEAR FROM dateeventrecorded) AS Year,
    COUNT(*) AS NULL_SNOMED_CODE
FROM `CB_FDM_PrimaryCare_V8.tbl_srcode`
WHERE care_site_id IS NOT NULL AND snomedcode IS NULL
GROUP BY care_site_id, Year
ORDER BY care_site_id, Year;
```

In [ ]:

```
CREATE TABLE `CB_2107.NULL_SNOMED_CODE_PER_YEAR_SITE_JOINED` AS

SELECT A.*, B.care_site_name
FROM `yhcr-prd-phm-bia-core.CB_2107.NULL_SNOMED_CODE_PER_YEAR_SITE` A
INNER JOIN `CB_FDM_PrimaryCare_V8.care_site`B ON A.care_site_id = B.care_site_id;
```

Bronchitis Test Codes

In [ ]:

```
CREATE TABLE `CB_2107.ALL_Bronchitis` AS
SELECT person_id,dateeventrecorded,dateevent,ctv3code,ctv3text,care_site_id
FROM `CB_FDM_PrimaryCare_V8.tbl_srcode`
WHERE (dateeventrecorded > CAST('2020-01-01' AS DATE))
AND (ctv3text LIKE '% bronchitis%' OR ctv3text LIKE '% Bronchitis%');
```

In [ ]:

```
# Get all data for bronc with all details and create a new table where the date is greater than 2020 Jan
#and only conatins Acute Bron
CREATE TABLE `CB_2107.Acute Bronchitis` AS
SELECT *
FROM `CB_FDM_PrimaryCare_V8.tbl_srcode`
WHERE (dateeventrecorded > CAST('2020-01-01' AS DATE))
AND (ctv3text LIKE '%Acute bronchitis%');
```

In [ ]:

```
CREATE TABLE `CB_2107.Acute Bronchitis_Simplified` AS
SELECT person_id,dateeventrecorded,dateevent,ctv3code,ctv3text,care_site_id
FROM `CB_FDM_PrimaryCare_V8.tbl_srcode`
WHERE (dateeventrecorded > CAST('2020-01-01' AS DATE))
AND (ctv3text LIKE '%Acute bronchitis%');
# Simplified to only include certain fields
```

In [ ]:

```
CREATE TABLE `CB_2107.Acute Bronchitis_Simplified_SNOMED` AS
SELECT a.*, b.SNOMEDCode
FROM `CB_2107.Acute Bronchitis_Simplified`a
INNER JOIN  `CB_LOOKUPS.tbl_CTV3ToSnomed_Map`b ON a.CTV3Code = b.CTV3Code;
# SNOMED Codes are mapped
```

In [ ]:

```
CREATE TABLE `CB_2107.Acute Bronchitis_Simplified_Match` AS
SELECT
    SUM(CASE WHEN dateevent = dateeventrecorded THEN 1 ELSE 0 END) AS DateMatched,
    SUM(CASE WHEN dateevent <> dateeventrecorded THEN 1 ELSE 0 END) AS DateNotMatched
FROM `CB_2107.Acute Bronchitis_Simplified_SNOMED`;
# Query to check the match and do not Match field
```

In [ ]:

```
CREATE TABLE `CB_2107.Acute Bronchitis_Simplified_2021` AS
SELECT person_id,dateeventrecorded,dateevent,ctv3code,ctv3text,care_site_id
FROM `CB_FDM_PrimaryCare_V8.tbl_srcode`
WHERE (dateeventrecorded > CAST('2021-01-01' AS DATE))
AND (ctv3text LIKE '%Acute bronchitis%');
```

In [ ]:

```
CREATE TABLE `CB_2107.Acute Bronchitis_Simplified_Match_2021` AS
SELECT
    SUM(CASE WHEN dateevent = dateeventrecorded THEN 1 ELSE 0 END) AS DateMatched,
    SUM(CASE WHEN dateevent <> dateeventrecorded THEN 1 ELSE 0 END) AS DateNotMatched
FROM `CB_2107.Acute Bronchitis_Simplified_2021`;
# Query to check the match and do not Match field
```

In [ ]:

```
SELECT ctv3code,ctv3text, COUNT(*) AS NumberOfRecords
FROM `CB_2107.ALL_Bronchitis`
GROUP BY ctv3code,ctv3text;
# Query to check all the CTV3 codes for Bronchitis and thier number of records
```

In [ ]:

```
CREATE TABLE `CB_2107.All Not Match` AS
SELECT
    SUM(CASE WHEN dateevent = dateeventrecorded THEN 1 ELSE 0 END) AS DateMatched,
    SUM(CASE WHEN dateevent <> dateeventrecorded THEN 1 ELSE 0 END) AS DateNotMatched
FROM `CB_FDM_PrimaryCare_V8.tbl_srcode`;
# Match / Not Match for All Data
```

In [ ]:

```
CREATE TABLE `CB_2107.Acute Bronc agg by site` AS
SELECT
  care_site_id,
  ctv3code,
  ctv3text,
  COUNT(DISTINCT person_id) AS PersonCount
FROM
  `CB_2107.Acute Bronchitis_Simplified_SNOMED`

GROUP BY
  care_site_id,
  ctv3code,
  ctv3text
ORDER BY
  care_site_id,
  ctv3text;
# Acute Bronc aggregated by Site
```

In [ ]:

```
CREATE TABLE `CB_2107.SNOMED_CTV3_Mapping issues` AS
SELECT SNOMEDCode,COUNT(CTV3Code) AS CTV3CodeCount
FROM `CB_LOOKUPS.tbl_CTV3ToSnomed_Map`
GROUP BY SNOMEDCode
HAVING COUNT(DISTINCT CTV3Code) > 1;
#Table A has Ctcode and medCode - Multiple Ctcode can be mapped into a single medCode ,
#SQL query to  Print all the Medcode where a more than 1 ctv code has been assigned to,
#alsong with a recod count of the number of ctv code assigned to each code
```

In [ ]:

```
SELECT
    SNOMEDCode,
    COUNT(CTV3Code) AS CTV3Code_count
FROM
    `CB_LOOKUPS.tbl_CTV3ToSnomed_Map`
GROUP BY
    SNOMEDCode
HAVING
    COUNT(CTV3Code) > 1
ORDER BY
    CTV3Code_count DESC, SNOMEDCode DESC
LIMIT 20;
# Same things as above, but only the first 20 recods with the highest count
```

In [ ]:

```
CREATE TABLE `CB_2107.SNOMED_CTV3_Mapping issues_20_Codes` AS
SELECT
    SNOMEDCode,CTV3Code,

FROM
    `CB_LOOKUPS.tbl_CTV3ToSnomed_Map`
WHERE
    SNOMEDCode IN ('52684005', '214640008', '34552002','3160009','3160009','418019003', '127349007','127350007','129675007','399
GROUP BY
    SNOMEDCode,CTV3Code;
# SNOMED and CTV3 codes for the top 20 codes with highest CTV3 Mapping
# there is an issue with dataset, this table ot the other tables in dataset that provides similar details which one is correct?
```

In [ ]:

```
CREATE TABLE `CB_2107.SNOMED_CTV3_Mapping issues_concept code-ID` AS
SELECT SNOMEDCT_CONCEPTID,COUNT(CTV3) AS CTV3CodeCount
FROM `yhcr-prd-phm-bia-core.CB_LOOKUPS.tbl_EFI2_Codelist`
GROUP BY SNOMEDCT_CONCEPTID
HAVING COUNT(DISTINCT CTV3) > 1;
# Using the EF1_2 Codelist to draw codes that are matched to multiple snomed/CTV3
```

In [ ]:

```
SELECT
    SNOMEDCT_CONCEPTID,
    STRING_AGG(CTV3, ', ') AS CTV3,
    STRING_AGG(Codedescription, ', ') AS Codedescription
FROM
    `yhcr-prd-phm-bia-core.CB_LOOKUPS.tbl_EFI2_Codelist`
GROUP BY
    SNOMEDCT_CONCEPTID;
 #NOT THE DESIRED RESULT
```

In [ ]:

```
SELECT
    CTV3Code,CTV3Desc
FROM
    `yhcr-prd-phm-bia-core.CB_LOOKUPS.tbl_CTV3Codes_Lookup`
WHERE
    CTV3Code LIKE '%Eu102%'OR CTV3Code LIKE '%Eu107%'OR CTV3Code LIKE '%Eu10z%'

#Shows the CTV3 code and thier decription
#This code and the code below shows that multiple CTV3 codes such as shows int he document have been mapped into single CTV3.
# Anothe issue some use SNOMED some use CTV3 - the ones that used CTV3, is all code is not accounted for
# some info will be missed.
```

In [ ]:

```
SELECT
    CTV3Code,SNOMEDCode
FROM
    `CB_2107.SNOMED_CTV3_Mapping issues_20_Codes`
WHERE
    CTV3Code LIKE '%Eu10%';
    #Shows the SNOMED CODE AND CTV3 Code- Check Desc Above
```

In [ ]:

```
SELECT
    CTV3,Codedescription,SNOMEDCT_CONCEPTID
FROM
    `yhcr-prd-phm-bia-core.CB_LOOKUPS.tbl_EFI2_Codelist`
WHERE
    SNOMEDCT_CONCEPTID = '191448002';
# This code tests for all the codes that have multiple CTV3 codes mapped into them. this table is found SNOMED_CTV3_Mapping
#iSSUES cONCEPT CODE_id
```

Covid SQL

In [ ]:

```
CREATE TABLE `CB_2107.COVID_CODE_INTRO` AS
SELECT * FROM `CB_CDM_VOCAB.concept`
WHERE (valid_end_date > CAST('1950-01-01' AS DATE)) AND
(concept_name LIKE '%Coronavirus%' OR concept_name LIKE '%Coronavirus%')
ORDER BY valid_end_date ASC

# Shows all the people that have Coronavirus in thier GP records
# Multiple different codeing systems are used at each GP
# THIS CODE IN SNOMED WAS INTODUCED IN 2020 MAY
# what was this before- Covid stated in early 2020 ? Many records have been missed?
```

In [ ]:

```
CREATE TABLE `CB_2107.COVID_CODE_DIFFERENT` AS
SELECT vocabulary_id, COUNT(*) AS count
FROM `CB_2107.COVID_CODE_INTRO`
GROUP BY vocabulary_id;
# This query has all the Coding systems, and how many codes under each system since 1950, with SNOMED having the second
#highest - this includes the Obervation and medication and all categories - Just to get an Idea (Initial analysis)
```

In [ ]:

```
CREATE TABLE `CB_2107.COVID_SNOMED` AS
SELECT *
FROM `CB_CDM_VOCAB.concept`
WHERE vocabulary_id = 'SNOMED'
  AND concept_name LIKE '%Coronavirus %'
  AND valid_end_date > CAST('1950-01-01' AS DATE);

# This query only gets the Concept code for COVID that have SNOMED id
```

In [ ]:

```
CREATE TABLE `CB_2107.COVID_SNOMED_INVALID` AS
SELECT *
FROM `CB_2107.COVID_SNOMED`
WHERE invalid_reason != 'null';
#Creates A LIST of invalid codes for Covid for SNOMED
```

In [ ]:

```
CREATE TABLE `CB_2107.COVID_SNOMED_VALID` AS
SELECT *
FROM `CB_2107.COVID_SNOMED`
WHERE invalid_reason IS NULL ;
#Creates A LIST of valid codes for Covid for SNOMED
```

In [ ]:

```
CREATE TABLE `CB_2107.COVID_Cases` AS
SELECT
    *

FROM
    `CB_FDM_PrimaryCare_V8.tbl_srcode`
WHERE
    SNOMEDCode IN ('120814005', '186747009', '204351000000100','906711000000107',
                   '1029481000000103','1008541000000105', '933791000000101','817221000000104',
                   '817111000000102','225631000001109','225591000001109','225581000001107','225681000001108',
                   '226631000001102','225571000001105');
#List of people with Covid - Only checking for SNOMED Codes
```

In [ ]:

```
CREATE TABLE `CB_2107.COVID_SNOMED_JOINED` AS
SELECT *
FROM `CB_2107.COVID_SNOMED_VALID`
INNER JOIN `CB_2107.COVID_Cases`  ON CAST(`CB_2107.COVID_SNOMED_VALID`.concept_code AS STRING) = `CB_2107.COVID_Cases` .snomedco
# The SNOMED codes have noe been Matched for thier description
```

In [ ]:

```
CREATE TABLE `CB_2107.COVID_CODE_START_Date` AS
SELECT
    snomedcode,concept_name,valid_start_date,
    EXTRACT(YEAR FROM dateevent) AS Year,
    EXTRACT(MONTH FROM dateevent) AS Month,
    COUNT(*) AS NumberOfRecords
FROM `CB_2107.COVID_SNOMED_JOINED`
GROUP BY snomedcode, Year, Month, concept_name,valid_start_date
ORDER BY snomedcode, Year, Month,concept_name,valid_start_date;
# Start Date of all Covid Codes, and when they were used the most
# Although the Codes
```

In [ ]:

```
CREATE TABLE `CB_2107.COVID_BY_CARESITE_BY_MMONTH_YEAR` AS

SELECT
    care_site_id,
    EXTRACT(YEAR FROM dateevent) AS Year,
    EXTRACT(MONTH FROM dateevent) AS Month,
    COUNT(*) AS NumberOfRecords
FROM `CB_2107.SNOMED_NAME_MAPPED`
GROUP BY care_site_id, Year, Month
ORDER BY care_site_id, Year, Month;


#Covid case aggregated by Year and Month per Caresite ID
```

In [ ]:

```
CREATE TABLE `CB_2107.COVID_BY_CARESITE_name_BY_MONTH_YEAR` AS
SELECT A.*, B.care_site_name,location_id,place_of_service_concept_id
FROM `CB_2107.COVID_BY_CARESITE_BY_MMONTH_YEAR` A
INNER JOIN `CB_FDM_PrimaryCare_V8.care_site` B
ON A.care_site_id = B.care_site_id;
#CareSite Name joined from Primarycare_CareSite
```

In [ ]:

```
CREATE TABLE `CB_2107.COVID_BY_CARESITE_name_Aggrigated` AS
SELECT
    care_site_name,
    concept_name,
    EXTRACT(DATE FROM dateevent) AS Year,
    COUNT(*) AS NumberOfCases
FROM `yhcr-prd-phm-bia-core.CB_2107.SNOMED_NAME_MAPPED_CareSiteName`
GROUP BY care_site_name, concept_name, Year
ORDER BY care_site_name, Year, concept_name;
```

COVID Codes Part II

In [ ]:

```
CREATE TABLE `CB_2107.NEW_ALL_COVID_CASES` AS

SELECT *
FROM `CB_FDM_PrimaryCare_V8.tbl_srcode`
WHERE (ctv3text LIKE '%COVID%' OR ctv3text LIKE '%Coronavirus%'OR ctv3text LIKE '%Long COVID%')
#table of all cases of Covid from the Peopl Table
```

In [ ]:

```
CREATE TABLE `CB_2107.NEW_ALL_COVID_CASES_2017` AS
SELECT * FROM `CB_2107.NEW_ALL_COVID_CASES`
WHERE (dateevent > CAST('2017-12-30' AS DATE))
#cASES FROM 2017
```

In [ ]:

```
CREATE TABLE `CB_2107.NEW_ALL_COVID_CASES_Grouped_CTV3` AS
SELECT ctv3code, ctv3text,snomedcode, COUNT(*) AS CountOfRecords
FROM `CB_2107.NEW_ALL_COVID_CASES`
GROUP BY ctv3code, ctv3text,snomedcode;
# Group the table by CTV3 codes
```

In [ ]:

```
CREATE TABLE `CB_2107.NEW_COVID_SNOMED` AS
SELECT *
FROM `CB_CDM_VOCAB.concept`
WHERE vocabulary_id = 'SNOMED'
  AND (concept_name LIKE '%Coronavirus %' OR concept_name LIKE '%COVID %'OR concept_name LIKE '%coronavirus %'OR concept_name LI
  AND Invalid_reason  IS NULL ;
# A list of all Covid Codes from SNOMED Concept
```

In [ ]:

```
CREATE TABLE `CB_2107.NEW_COVID_SNOMED_GROUPED` AS

SELECT valid_start_date, COUNT(*) AS NumberOfRecords
FROM `CB_2107.NEW_COVID_SNOMED`
GROUP BY valid_start_date;
# Number of codes grouped by thier release year
```

In [ ]:

```
SELECT *
FROM `CB_2107.NEW_ALL_COVID_CASES_2017`
WHERE ctv3code = 'X731E';
```

Influenza SQL

In [ ]:

```
#CREATE TABLE `CB_2107.Influenza_Cases` AS
SELECT
    *

FROM
    `CB_FDM_PrimaryCare_V8.tbl_srcode`
WHERE
    SNOMEDCode IN ('10629191000119100', '719865001', '719590007','906711000000107',
                   '408687004','408685007', '407496005','441119003',
                   '195726000','27475006','230188005','59221008','30270006',
                   '64880000','64917006','68949000','78046005','78911000','85832003',
                   '70233007', '35377009', '192721000','10809006','932221000000103',
                   '1033091000000109','1033111000000104','1033051000000101','1033071000000105');
    # Creating a list of people who has Influenza using the valid codes
    #Table deleted
```

In [ ]:

```
CREATE TABLE `CB_2107.Influenza_Cases_2019` AS

SELECT
    *

FROM
    `CB_FDM_PrimaryCare_V8.tbl_srcode`

WHERE  (dateevent > CAST('2019-01-01' AS DATE)) AND
ctv3text  LIKE '%Influenza%' AND ctv3text  LIKE '%flu%' AND
  (ctv3text NOT LIKE '%vacc%' AND ctv3text NOT LIKE '%SMS%' AND ctv3text NOT LIKE '%immunisation%'
  AND ctv3text NOT LIKE '%invite %'AND ctv3text NOT LIKE '%Vaccine%');

    # All the cases with inlfuenza from 2019 and Above
```

In [ ]:

```
CREATE TABLE `CB_2107.Influenza_Cases_2019_Record_Count` AS
SELECT snomedcode, COUNT(*) AS RecordCount
FROM `CB_2107.Influenza_Cases_2019`
GROUP BY snomedcode
ORDER BY snomedcode;
# aLL THE RECORD FOR Inlfuelza code FROM pERSON TO snomed
```

In [ ]:

```
CREATE TABLE `CB_2107.Influenza_Cases_2019_Record_Count_joined` AS
SELECT A.*,B.*
FROM `CB_2107.Influenza_Cases_2019_Record_Count` A
INNER JOIN `CB_CDM_VOCAB.concept`B
ON A.snomedcode = B.concept_code;
# Joined table for all the codes extracted using the srcode first - where condidtion and everything is mixed -
#From person to SNOMED
```

In [ ]:

```
SELECT
    *

FROM
    `CB_CDM_V1_50k_Random.concept`
WHERE
    concept_code IN ('1033091000000109', '195878008', '195929004','315642008',
                     '359351000000100','407479009', '407480007','441049004',
                     '441345003','442696006','444426005','511781000000101','55014007',
                     '6142004','711330007','95891005');

    # All the matching codes from the cases table to match the Concept table from the rocord count table
    # Checking from pERSON TABLE TO snomed
```

Influenza Code Part II - Testing inclusion criteria

In [ ]:

```
CREATE TABLE `CB_2107.Influenza_Active_Codes` AS

     SELECT *
FROM `CB_CDM_VOCAB.concept`

  WHERE concept_name LIKE '%Influenza%'
 AND domain_id = 'Condition' AND vocabulary_id = 'SNOMED'
  AND invalid_reason IS  NULL;
# Checking for valid SNOMED cases with Influenza - Has many codes
```

In [ ]:

```
#Checing from SNOMED to person
CREATE TABLE `CB_2107.Influenza_Cases_snomed_Cases` AS
SELECT
    *

FROM
    `CB_FDM_PrimaryCare_V8.tbl_srcode`
WHERE
    SNOMEDCode IN ('10628911000119103', '10674911000119108', '142951000119106','10629351000119108',
                   '711334003','10685111000119102', '328531000119104','10677711000119101',
                   '772839003','772810003','772828001','866126000','719865001',
                   '719590007','453931000124108','434931000124106','435051000124104','707448003','961000124101',
                   '70233007', '35377009', '192721000','10809006','932221000000103',
                   '941000124100','951000124103','921000124107','931000124105',

                   '442438000','442696006','1149091008',
                   '738276008','471361000124100','713083002','450716003','450715004','74644004',
                   '195923003', '43692000', '41269000','195929004','24662006',
                   '315642008','95891005','195924009','81524006',

                   '6142004','61700007','1033091000000109',
                   '1033051000000101','1033071000000105','1033111000000104') AND (dateevent > CAST('2019-01-01' AS DATE));
    # tHIS METHOD WAS ALSO TRIED - BUT DID NOT GIVE A LOT OF VARIATION IN THE SNIMED CODE
    #So this table only gives cases from SRCode that are measured under Condition
```

In [ ]:

```
CREATE TABLE `CB_2107.Influenza_Cases_snomed_Cases_Record` AS
SELECT snomedcode, COUNT(*) AS RecordCount
FROM `CB_2107.Influenza_Cases_snomed_Cases`
GROUP BY snomedcode
ORDER BY snomedcode;
# This has SNOMED to person - ONLY Cases under Condiditon
```

In [ ]:

```
CREATE TABLE `CB_2107.Influenza_Cases_snomed_Cases_Record_joined` AS
SELECT A.*,B.*
FROM `CB_2107.Influenza_Cases_snomed_Cases_Record` A
INNER JOIN `CB_CDM_VOCAB.concept`B
ON A.snomedcode = B.concept_code;
# SNOMED TO person Mapping only condidtion = But includes SNOMED and Nebraska?
```

In [ ]:

```
CREATE TABLE `CB_2107.Influenza_Cases_snomed_Cases_Record_joined_SNOMED_ONLY` AS
SELECT *
FROM `CB_2107.Influenza_Cases_snomed_Cases_Record_joined`
WHERE (vocabulary_id = 'SNOMED' )
# This removed the dupliactions of SNOMED and Nabraska = Only SNOMED Kept
```

In [ ]:

```
    SELECT *
FROM `CB_CDM_VOCAB.concept`
WHERE (domain_id = 'Condition' or domain_id = 'Observation')

#Checking both Observation and Condition
```

In [ ]:

```
SELECT DISTINCT (ctv3text)
FROM `CB_FDM_PrimaryCare_V8.tbl_srcode`
WHERE (ctv3text LIKE '%COVID%' OR ctv3text LIKE '%Coronavirus%'OR ctv3text LIKE '%coronavirus%')
# Inclusion Criteria Tested
```

DISPARITIES IN THE MEDICAL CODING SYSTEM

A Visual Analysis of how GP's code Data

DISCREPANCY WITH PLACE OF BIRTH

Graph A is a Bubble Map that shows the issues with the dataset regarding the "Place of birth" Column in the "tbl_srpatient" table in the primary care dataset.
The chart intends to show the multiple variations in how the "Place of Birth" data has been entered.
The interactive chart allows one to hover the mouse over the regions to see the number of patients recorded under each location.

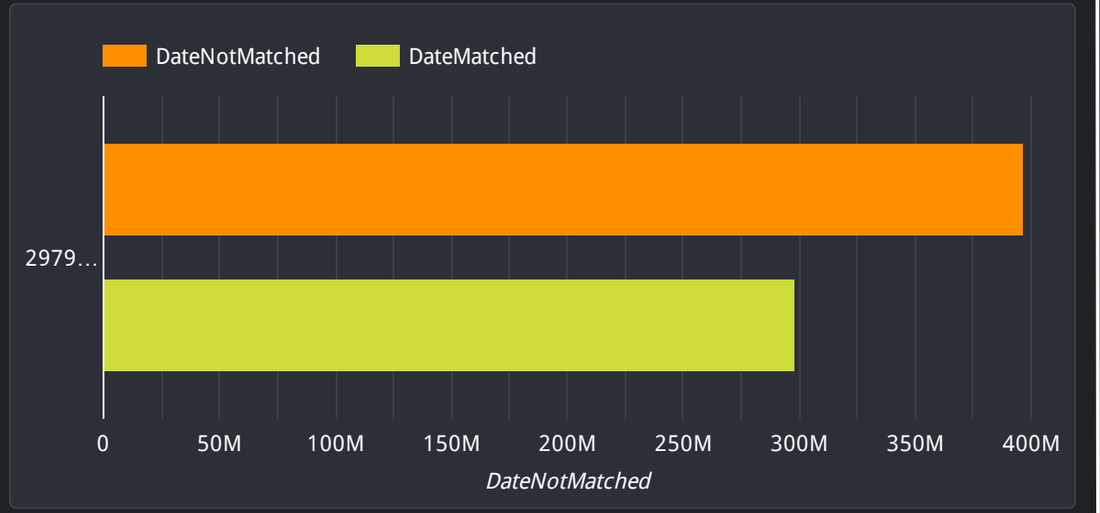
Noticeable differences

> Pakistan, PAKISTAN, and pakistan are all recorded as separate entities of the same birth Place. If an SQL query is run on the inclusion criteria of just the term "Pakistan", all data regarding other entries will be lost. The importance of inclusion criteria is explained in detail under the results section of the report, and how this impacts the healthcare data.

> No regulations specify what can be counted as a "Place of Birth". Clear definitions must be included, such as rules or limitations to only include countries or States. However, currently, the system accepts all types of characters. This has been emphasised explicitly in the analytical report, where date of births (in the format DDMMYY) have been added instead of place.

> This graph is a prime example of how data is missed in visualisations, as bubble graphs do not support numerical data. Hence, this graph excludes all data entries for numerical data. If the particular issues are explained, there is a way of knowing the issues posed in this table and data entry methods.

ISSUES IN CAPTURING LIVE DATA



Graph B shows the discrepancies with two columns, "dateeventrecorded" and "dateevent" in the tbl_srcode table.

> In most cases, the information is not recorded on the actual event date. This graph included data from all incidents recorded in the dataset, leading to exaggerations, in the "DateNotMatched" section of the graph as it may include data when old records were merged and when paper records may have been digitised, leading to them being entered into a later date. However, the critical issue to be emphasised in this graph is that discrepancies almost always occur when data has been not inputted at the "Time of the event".
It is difficult to prove and trace the authenticity of the data and the type of codes recorded once the data is entered at a later date.

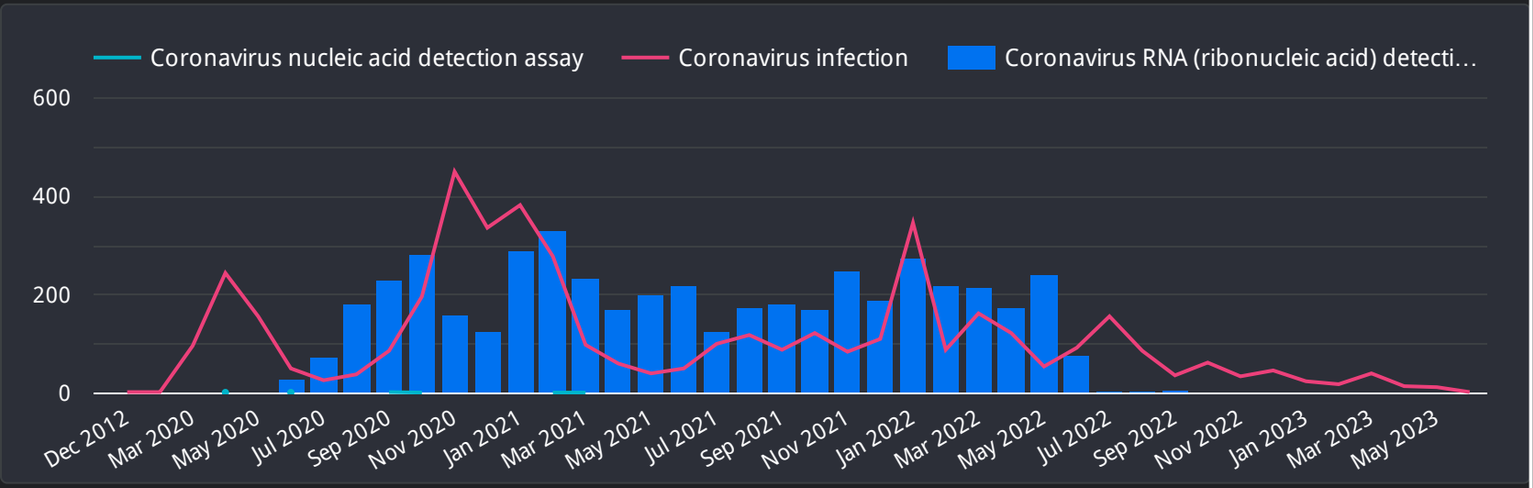
> SNOMED codes and CTV3 codes are updated regularly. Hence, when old records are re-entered into the system, the medical code descriptions may have been updated, leading to a loss of granularity from the old records.

> The mismatch in dates may also be an issue when GP practices merge to form more extensive practices, leading to mergers of multiple coding systems and updates in the information to match the new system.

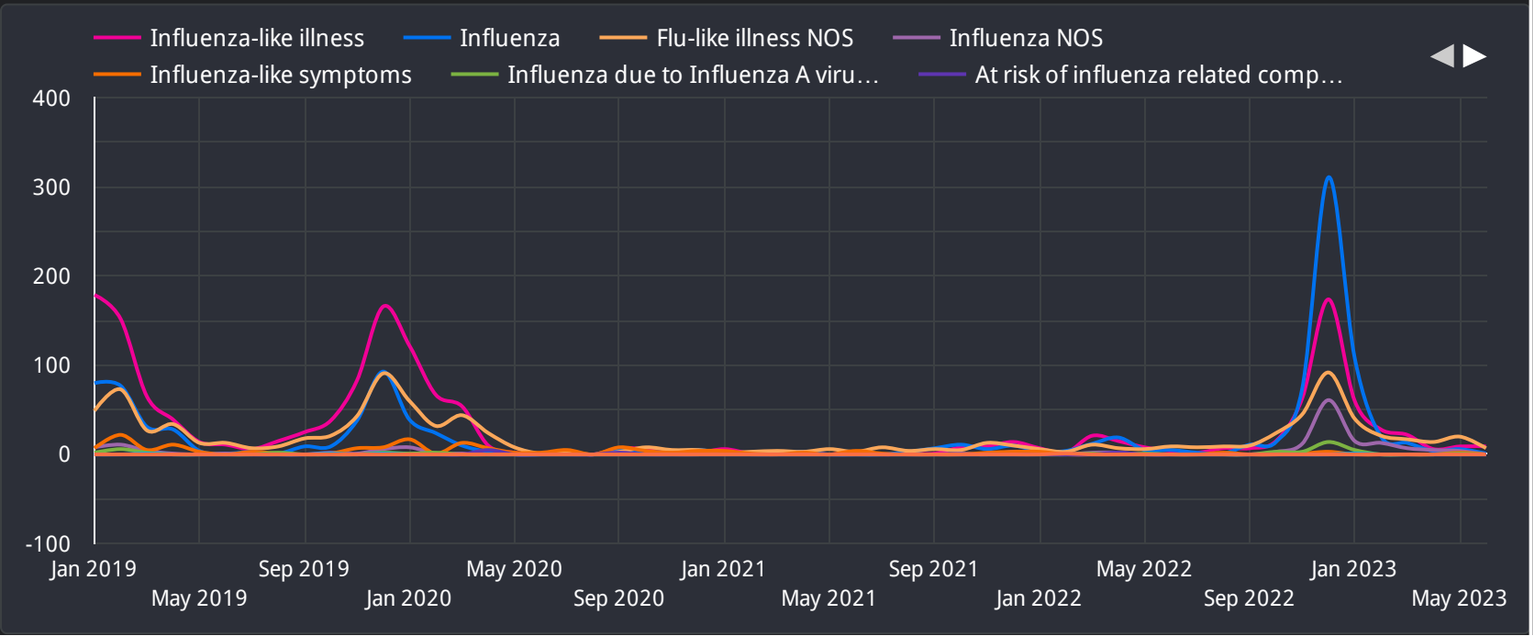
DISPARITIES IN THE MEDICAL CODING SYSTEM

A Visual Analysis of how GP's code Data

CORONAVIRUS TIME SERIES



INFLUENZA TIME SERIES



The graph shows the changes in using different medical codes used to record the coronavirus. The two most commonly used codes were "Coronavirus Infection" and "Coronavirus RNA (ribonucleic acid) detection assay". This graph should be analysed in conjunction with the graph below, which shows the peak of influenza infection from March 2020 to May 2020; as soon as the coronavirus codes were started to be used, the use of the influenza codes was almost stopped immediately.
Hovering over the influenza graph between April and May 2020 shows the most significant drop in the use of the influenza code. It shows that almost no cases of Influenza were recorded between June 2020 and Feb 2022.

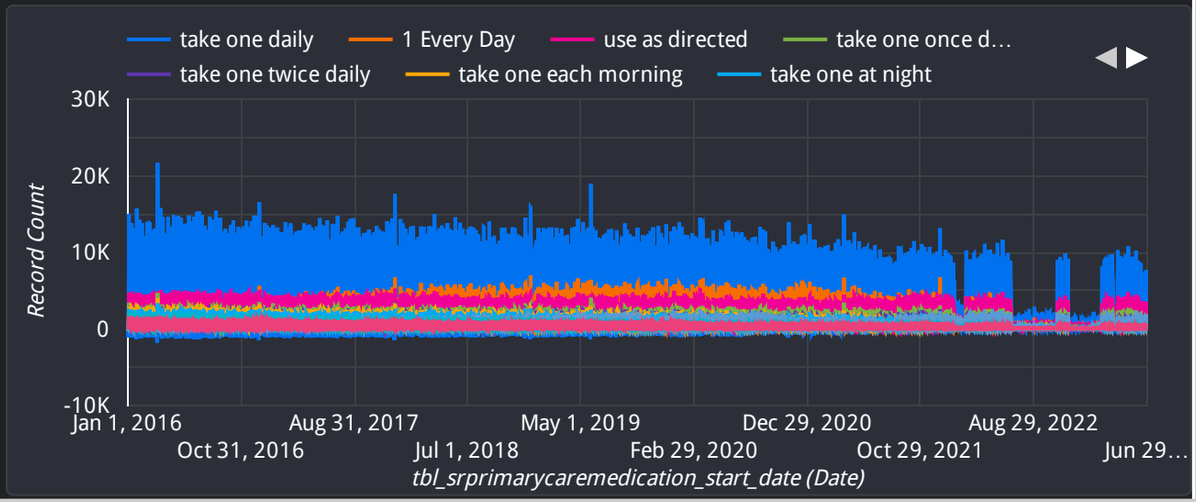
This is a significant discrepancy in the way medical codes are used.
It is fair to say that coronavirus was the most dominant form of infection during the timescale shown in the graphs. However, it is ignorant to say that Influenza, which has a high number of cases during the winter seasons, every year was almost irradicated for almost two years.

Graph 2 shows the use of medical codes to record influenza conditions. The graph above shows that in 2022, only 46 cases were recorded under coronavirus, whereas 174 were recorded as "influenza-like illness". There is a sharp switch in the use of Influenza and coronavirus codes, which complies with the timelines of when lockdowns were strictly implemented. It can be assumed that all possible flu infection cases were recorded under the coronavirus SNOMED/ CTV3 codes. This has led to significant information loss and ambiguity within the data. The loss of information can never be accurately traced and amended. The impact of the misuse of codes is even more significant if all healthcare practices in England are to be assessed for these two codes during May 2020 - June 2023

DISPARITIES IN THE MEDICAL CODING SYSTEM

A Visual Analysis of how GP's code Data

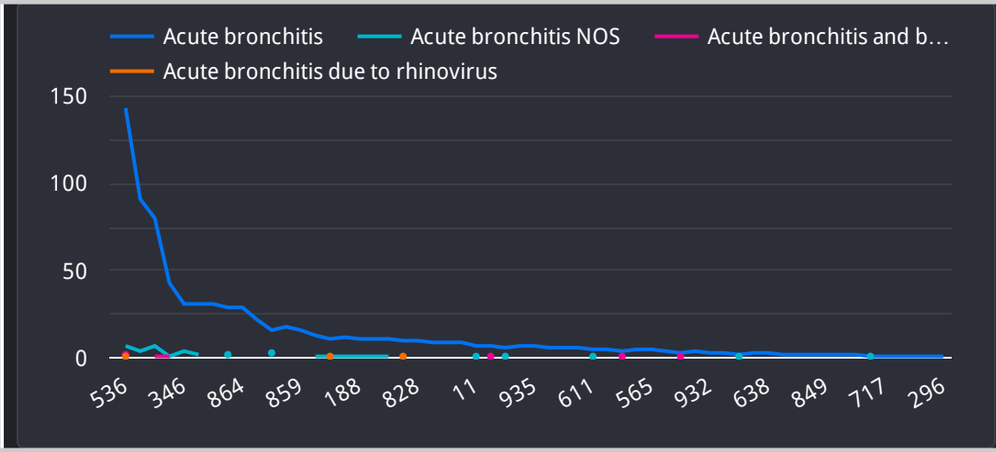
DISCREPANCY WITH MEDICATION ALLOCATION



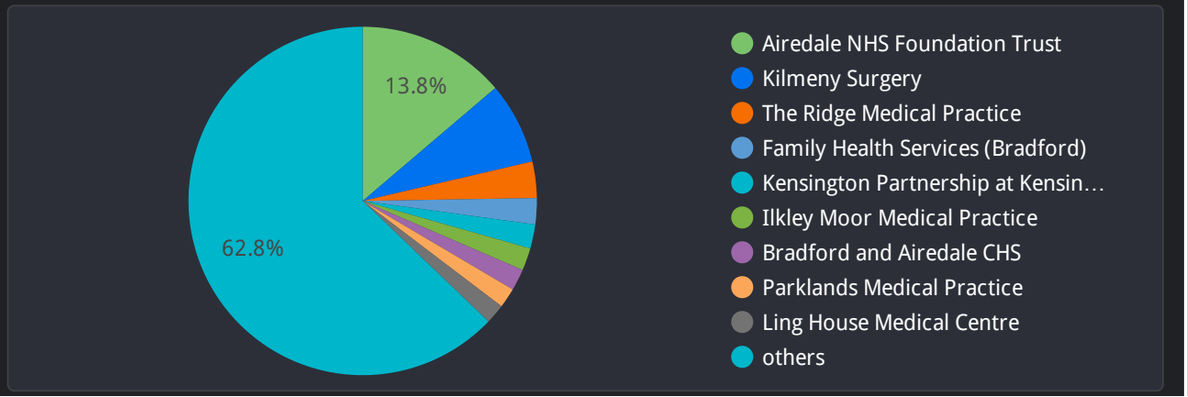
The graph shows the discrepancy in the medication table of the "tbl_srprimarycaremedication" table in the primary care dataset. One of the main trends highlighted through the dashboard and analytical report is the loss of information during the Query information and retrieval stage due to information needing to be recorded correctly in the systems. Medication dosage is a very important section of the healthcare data. The graph shows that Medication Dosage is an open text column, creating multiple variations for the same version of the information. "take one daily", "1 Every Day", and "take one once daily" should all be under one entry selection.
This will help to regulate and standardise the information recorded in healthcare data. The issues may look minor and negligible on a small scale. However, when large queries are carried out, large volumes of information will be missed due to the inclusion criteria of this or other categories in the Query. The graph also shows a sudden unexplainable drop in how the medications are assigned from July 2022 to Oct 2022. The drop is noticed again in December 2022 to Feb 2023. These anomalies are not compliant with the coronavirus lockdowns; hence, it is difficult to analyse the missing data and the reasons behind it.

EXPECTED VS. OBSERVED USES IN MEDICAL CODES

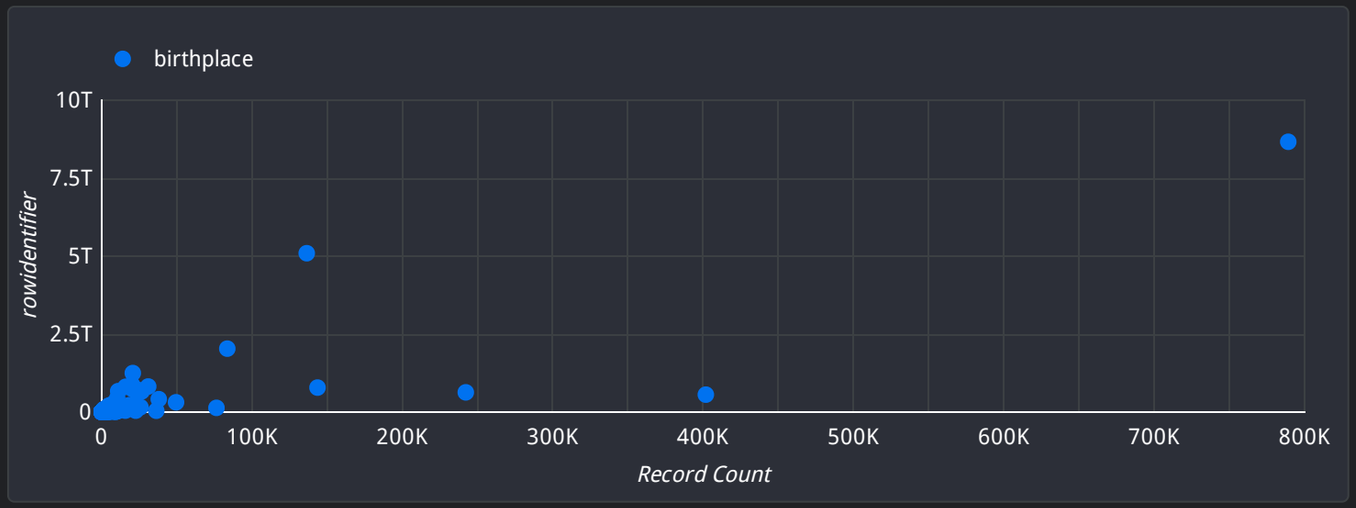
The graph shows the issues with the allocation of the SNOMED Codes in the primary health care data set. Acute Bronchitis is the most used type of code for diagnosing Bronchitis, although other SNOMED codes exist with higher-quality definitions. Most cases are recorded under the generic code with a general definition. This shows that in most cases when the medical data is recorded on later dates if the information is not clear to the medical coders, it is easy to assign the generic codes, as they do not have the valid medical knowledge to make the necessary decisions. The graph separates the allocation of the codes according to the clinics, which makes it easier to understand which clinics have the highest number of cases and the number of cases assigned to each SNOMED code. The graph shows that some practices hardly use the other variation of codes. It is difficult to analyse the real reason behind the lack of use of these specific codes.

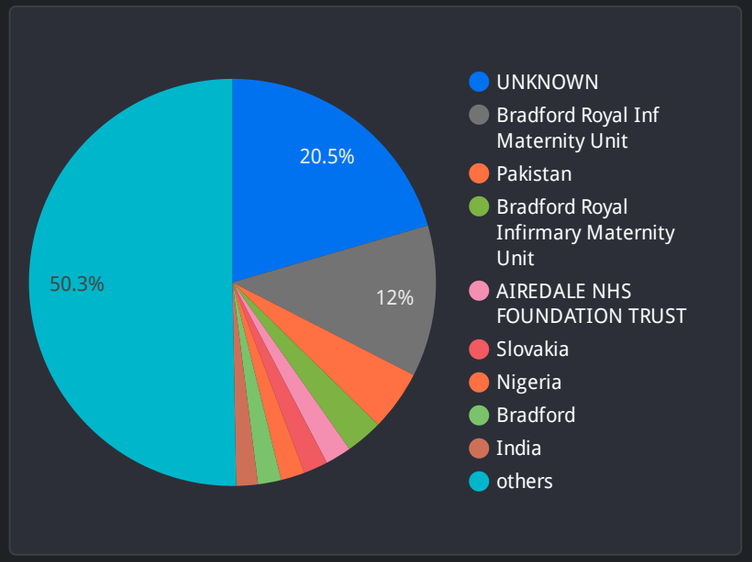


MISSING INFORMATION



This graph shows the missing SNOMED codes in the tbl_srcode system. All records have CTV3 codes, but in many instances, the SNOMED codes still need to be included. The primary care dataset is collated using many individual datasets; hence, the missing SNOMED codes are mainly down to the individual practices and the way they code data. This graph does allow us to see the healthcare practices that have the highest number of missing data. Based on this information, this issue can be highlighted to the relevant authorities at the healthcare practices to improve the accuracy of how medical codes are stored.
Furthermore, although it is not important to switch old data from CTV3 to SNOMED legally, storing all data in a singular format is always easier to prevent data loss in the long run. Since SNOMED is the current standard, it would be beneficial to switch data coded in CTV3 to SNOMED since descriptions and definitions are updated regularly.



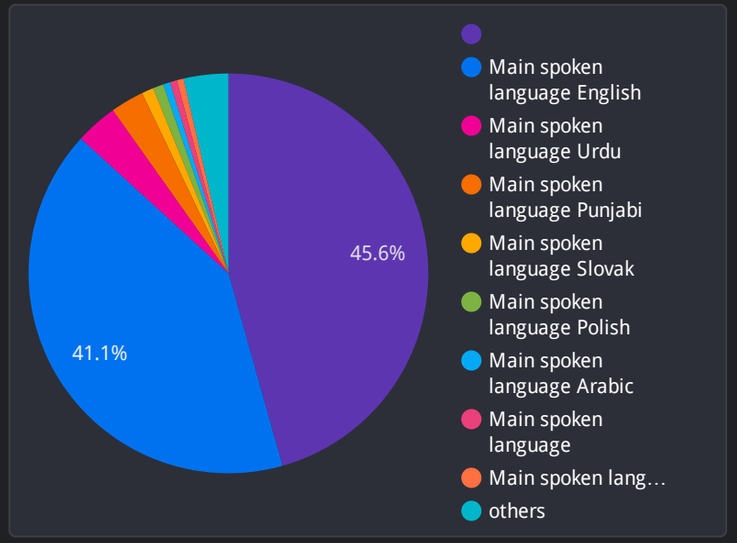


	table_name	Record Count
1.	tbl_srreferralin	61
2.	tbl_srprimarycare...	33
3.	tbl_srappointment	33
4.	tbl_srreferralout	30
5.	tbl_srcode	28
6.	tbl_srimmunisation	28
7.	tbl_srpatientaddres...	24

1 - 17 / 17

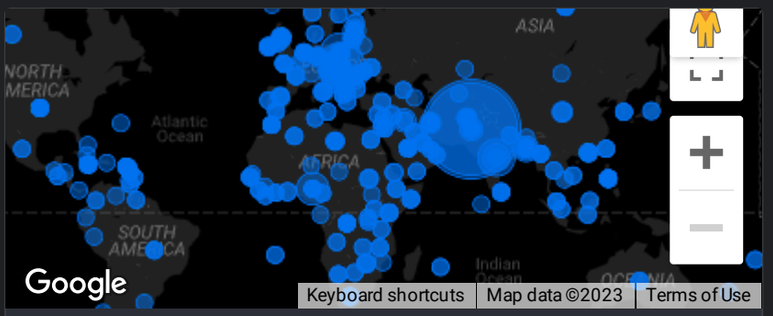
	ethnici...	birthpl...	lan...	Record...
1.	Not speci...	UNKNO...	null	149,080
2.	Ethnic gr...	UNKNO...	null	114,431
3.	British or...	UNKNO...	Main...	110,820
4.	White Bri...	UNKNO...	Main...	75,481
5.	British or...	BRADFORD	Main...	67,839
6.	Ethnic gr...	null	null	66,654

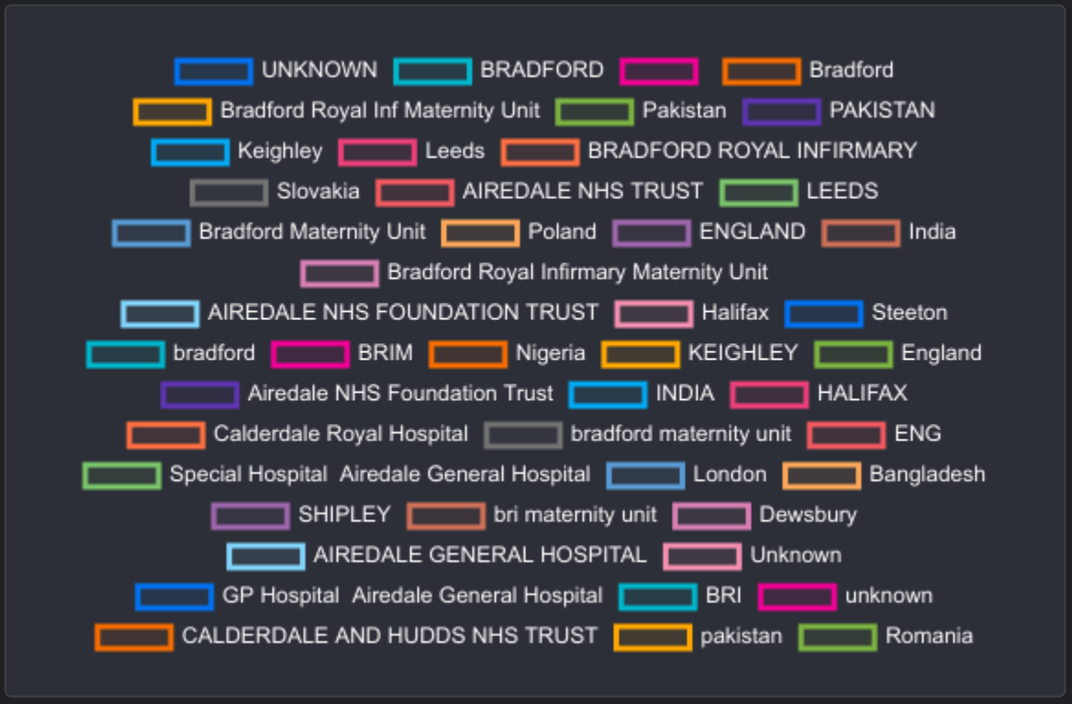
1 - 100 / 102952



	birthplace	Record Count
1.	UNKNOWN	789,301
2.	BRADFORD	401,902
3.	null	242,234
4.	Bradford	143,616
5.	Bradford Royal Inf M...	136,445
6.	Pakistan	83,600
7.	PAKISTAN	76,411
8.	Keighley	49,517
9.	Leeds	38,009

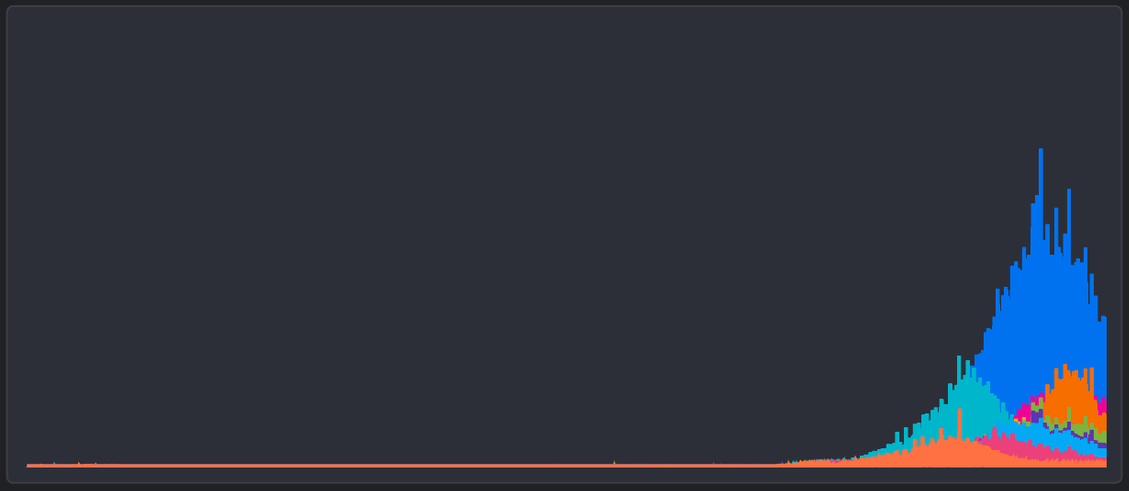
1 - 100 / 34386





	medicationdosage	Re...
1.	take one daily	35,...
2.	ONE TO BE TAKEN DAILY	11,...
3.	use as directed	11,...
4.	1 Every Day	7,2...
5.	take one each morning	7,1...
6.	take one once daily	7,0...
7.	take one twice daily	6,7...

1 - 100 / 3077216

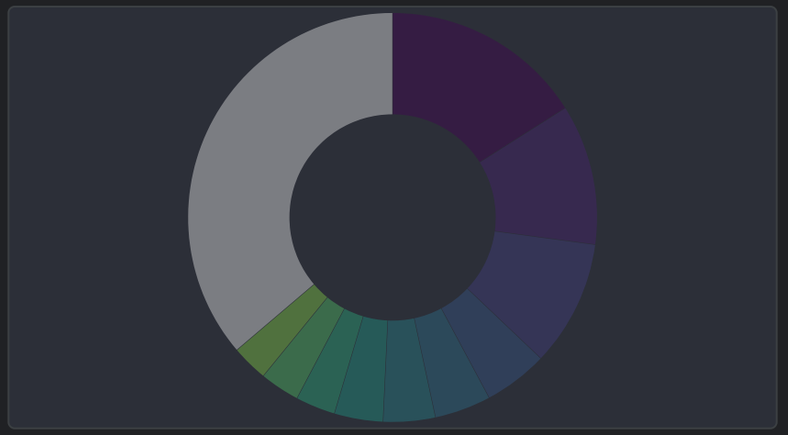


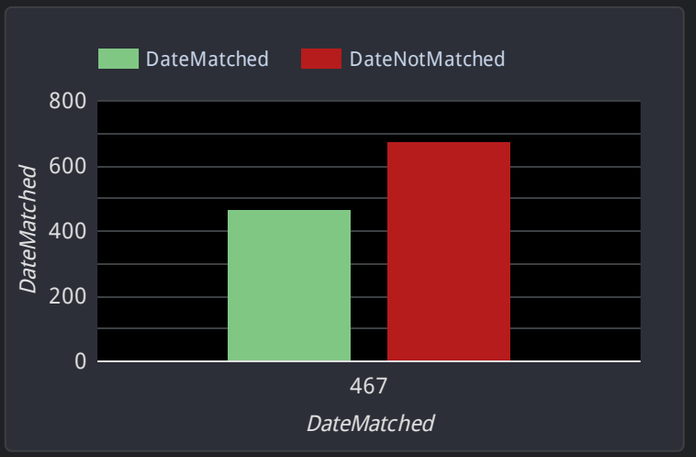
	care_site_name	care_site_id	locat...
1.	Dr Akbar's Surgery	38588	38588
2.	Moorside District Nurses	37122	37122
3.	Hillside Bridge Surgery	38609	38609
4.	Family Health Services (Bradf...	64975	64975
5.	Stop Smoking Service - Bradf...	29061	29061
6.	Dr de Haar & Partners	555916	555916
7.	Primrose Surgery	555902	555902
8.	Low Moor Medical Practice	38691	38691
9.	Avicenna Medical Practice	560	560
10.	Park Grange Medical Centre	38582	38582

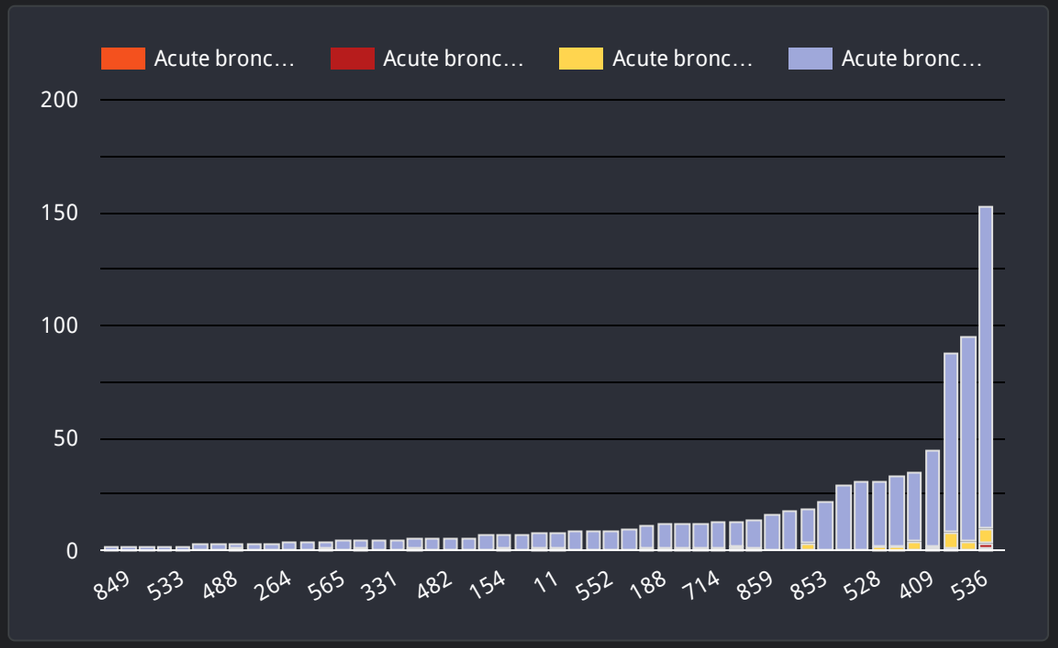
1 - 100 / 119

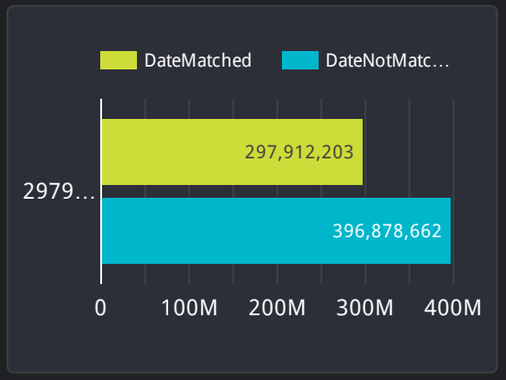
	care_site_i...	matching_...	mismatchi...
1...	961	58	57988
2...	958	12002	29
3...	957	267	80
4...	956	7732	11829
5...	955	2857	60
6...	944	360	1225

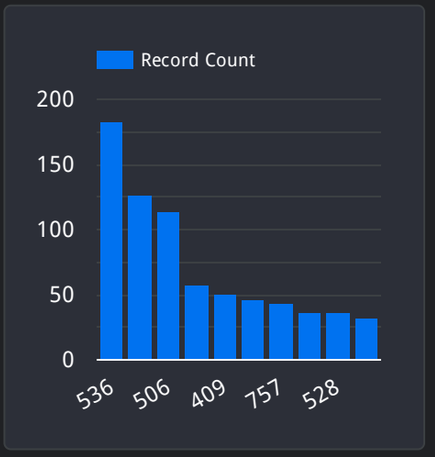
1 - 100 / 348

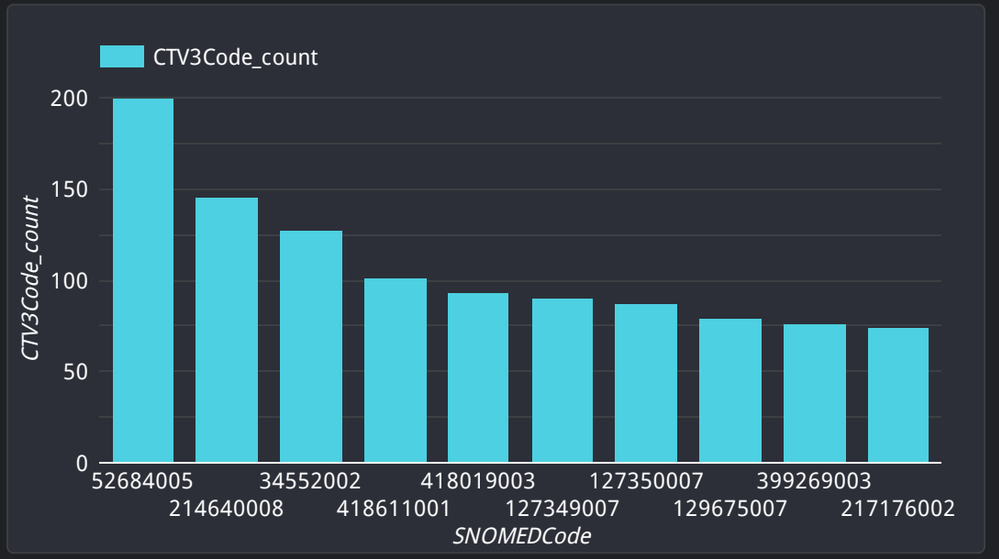


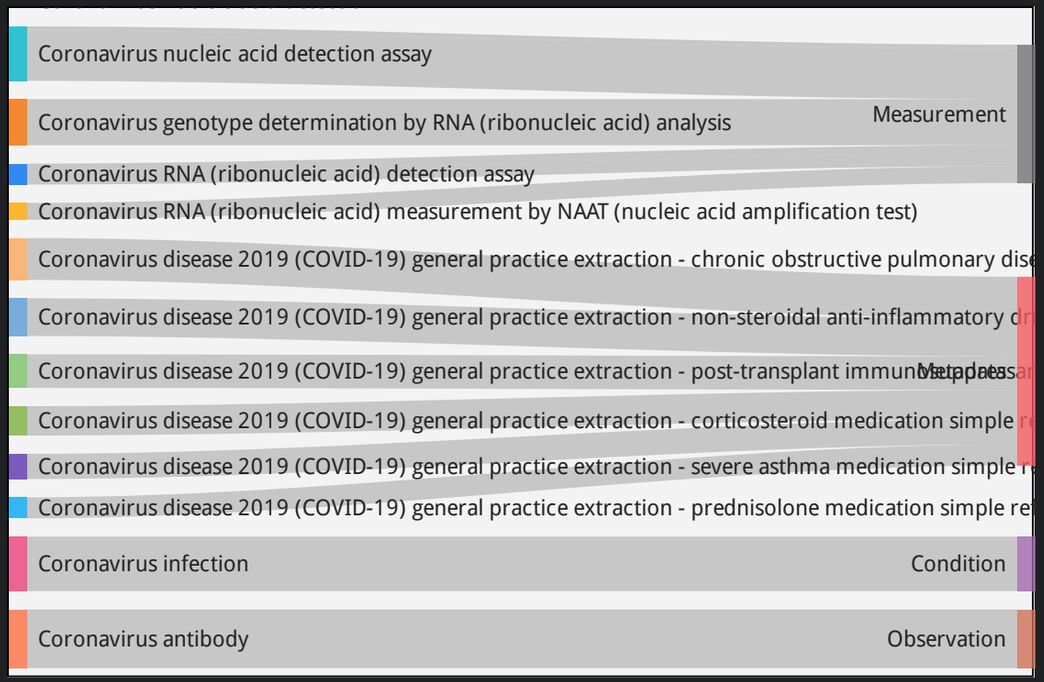


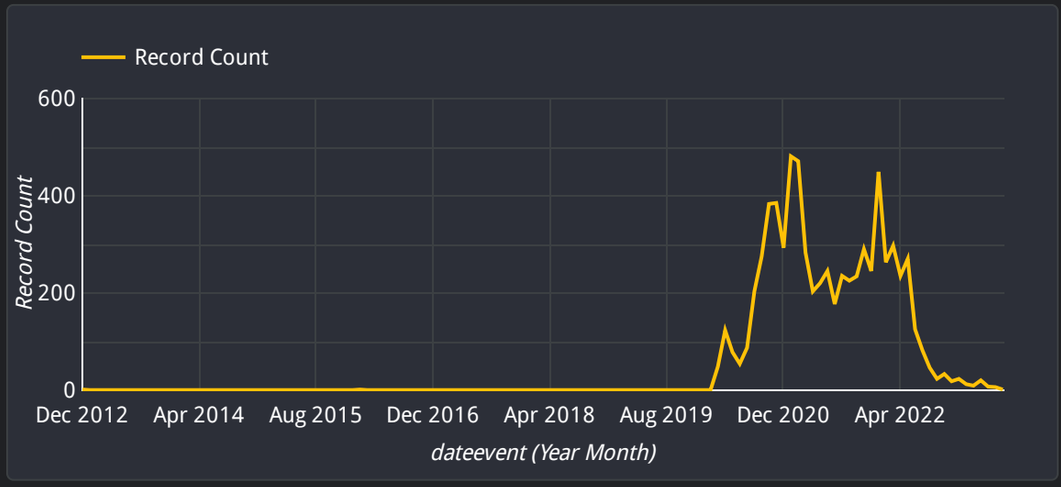


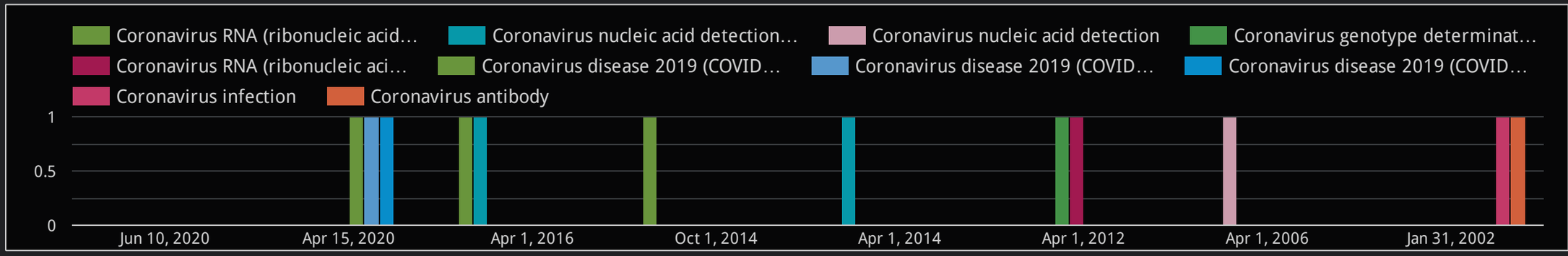


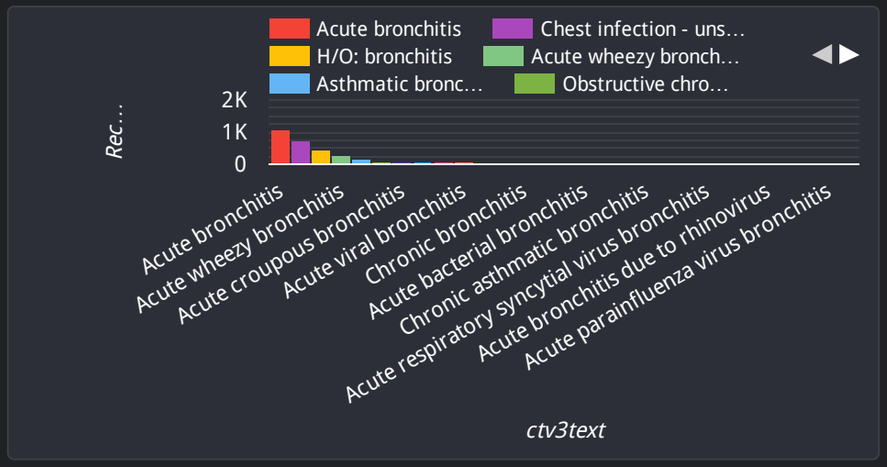


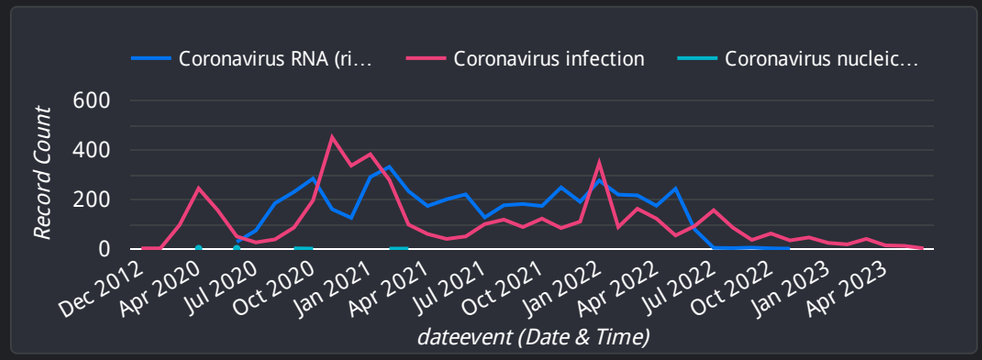


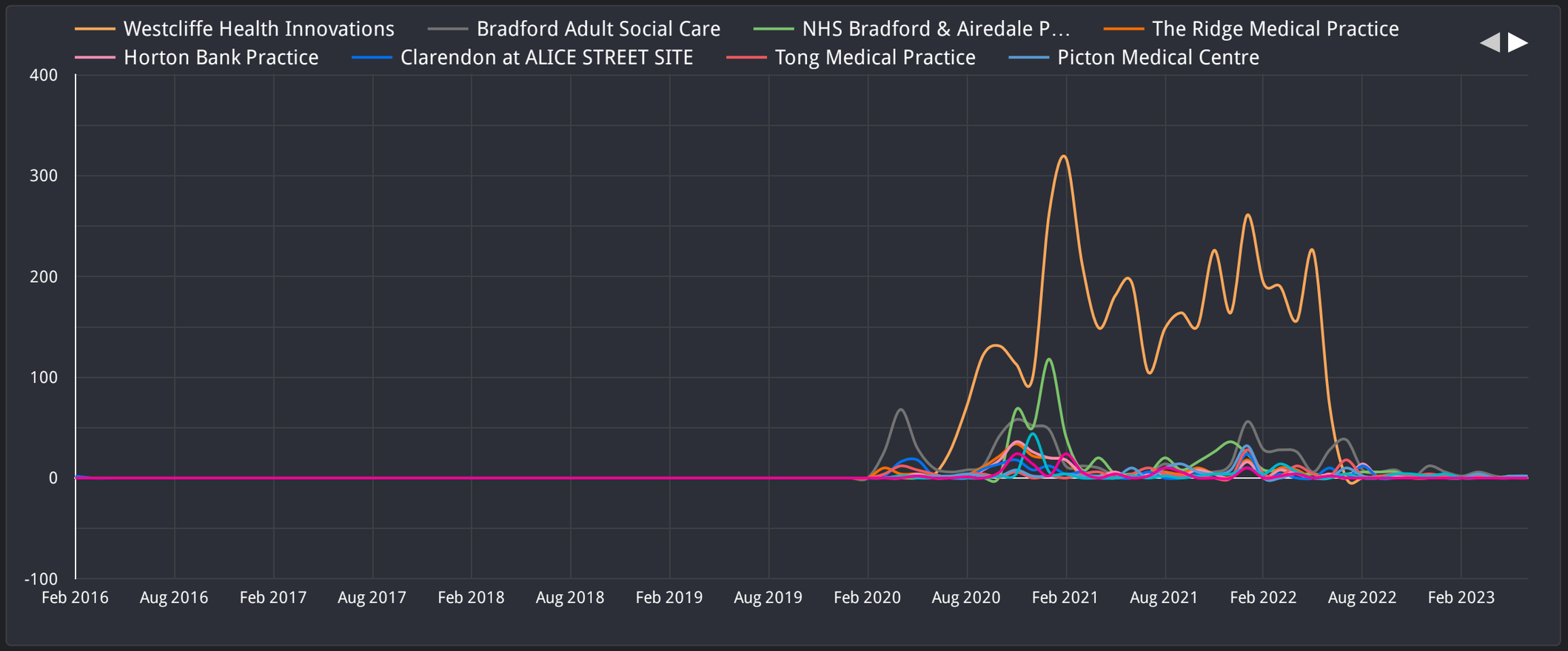

Coronavirus nucleic acid detection assay
Coronavirus genotype determination by RNA (ribonucleic acid) analysis
Coronavirus RNA (ribonucleic acid) detection assay
Coronavirus RNA (ribonucleic acid) measurement by NAAT (nucleic acid amplification test)
Coronavirus disease 2019 (COVID-19) general practice extraction - chronic obstructive pulmonary dise
Coronavirus disease 2019 (COVID-19) general practice extraction - non-steroidal anti-inflammatory dr
Coronavirus disease 2019 (COVID-19) general practice extraction - post-transplant immunosuppressa
Coronavirus disease 2019 (COVID-19) general practice extraction - corticosteroid medication simple re
Coronavirus disease 2019 (COVID-19) general practice extraction - severe asthma medication simple re
Coronavirus disease 2019 (COVID-19) general practice extraction - prednisolone medication simple re
Coronavirus infection
Coronavirus antibody
Measurement
Condition
Observation


Record Count
600
400
200
0
Record Count
Dec 2012
Apr 2014
Aug 2015
Dec 2016
Apr 2018
Aug 2019
Dec 2020
Apr 2022
dateevent (Year Month)

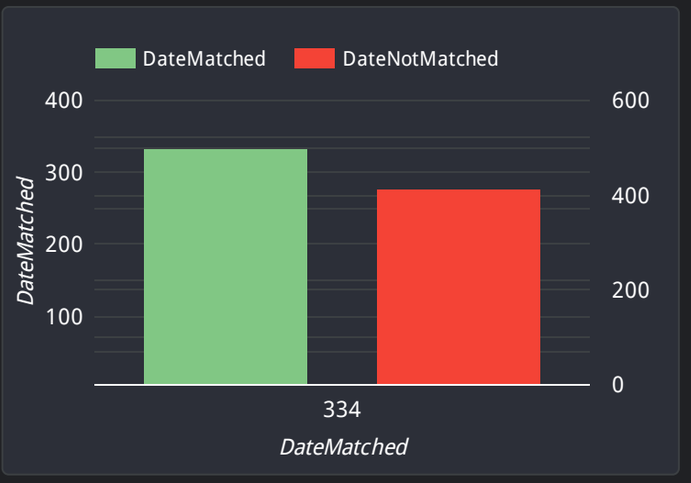

Coronavirus RNA (ribonucleic acid…
Coronavirus nucleic acid detection…
Coronavirus nucleic acid detection
Coronavirus genotype determinat…
Coronavirus RNA (ribonucleic aci…
Coronavirus disease 2019 (COVID…
Coronavirus disease 2019 (COVID…
Coronavirus disease 2019 (COVID…
Coronavirus infection
Coronavirus antibody
1
0.5
0
Jun 10, 2020
Apr 15, 2020
Apr 1, 2016
Oct 1, 2014
Apr 1, 2014
Apr 1, 2012
Apr 1, 2006
Jan 31, 2002

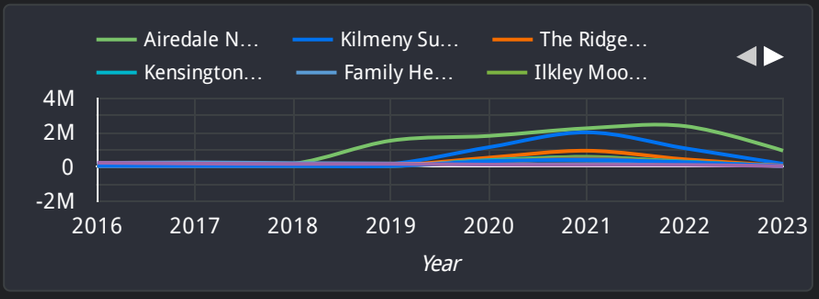

Acute bronchitis
Chest infection - uns…
H/O: bronchitis
Acute wheezy bronch…
Asthmatic bronc…
Obstructive chro…
Rec…
2K
1K
0
Acute bronchitis
Acute wheezy bronchitis
Acute croupous bronchitis
Acute viral bronchitis
Chronic bronchitis
Acute bacterial bronchitis
Chronic asthmatic bronchitis
Acute respiratory syncytial virus bronchitis
Acute bronchitis due to rhinovirus
Acute parainfluenza virus bronchitis
ctv3text

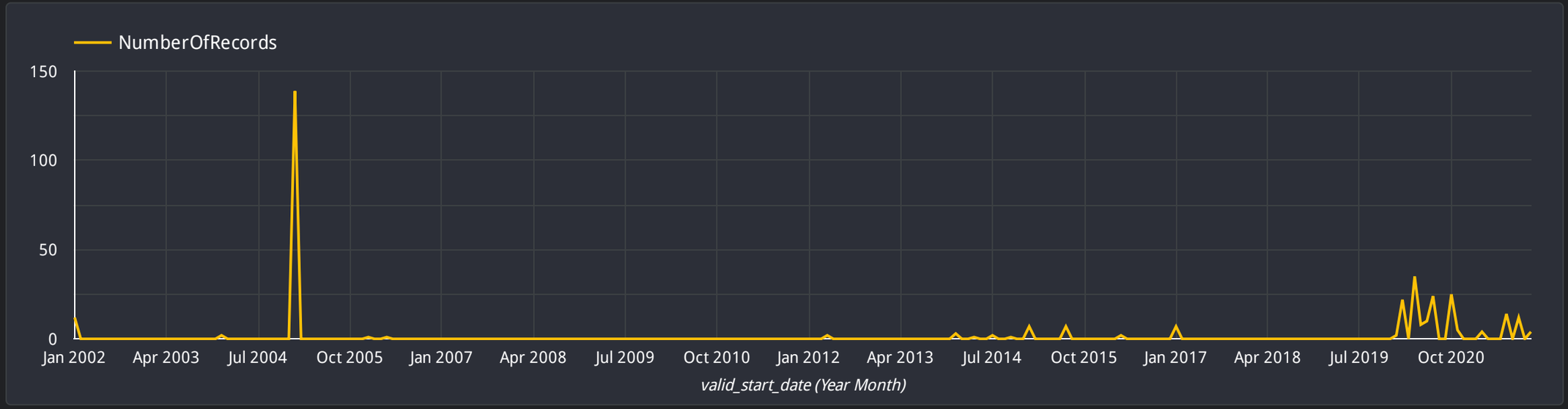

Coronavirus RNA (ri…
Coronavirus infection
Coronavirus nucleic…
Record Count
600
400
200
0
Dec 2012
Apr 2020
Jul 2020
Oct 2020
Jan 2021
Apr 2021
Jul 2021
Oct 2021
Jan 2022
Apr 2022
Jul 2022
Oct 2022
Jan 2023
Apr 2023
dateevent (Date & Time)

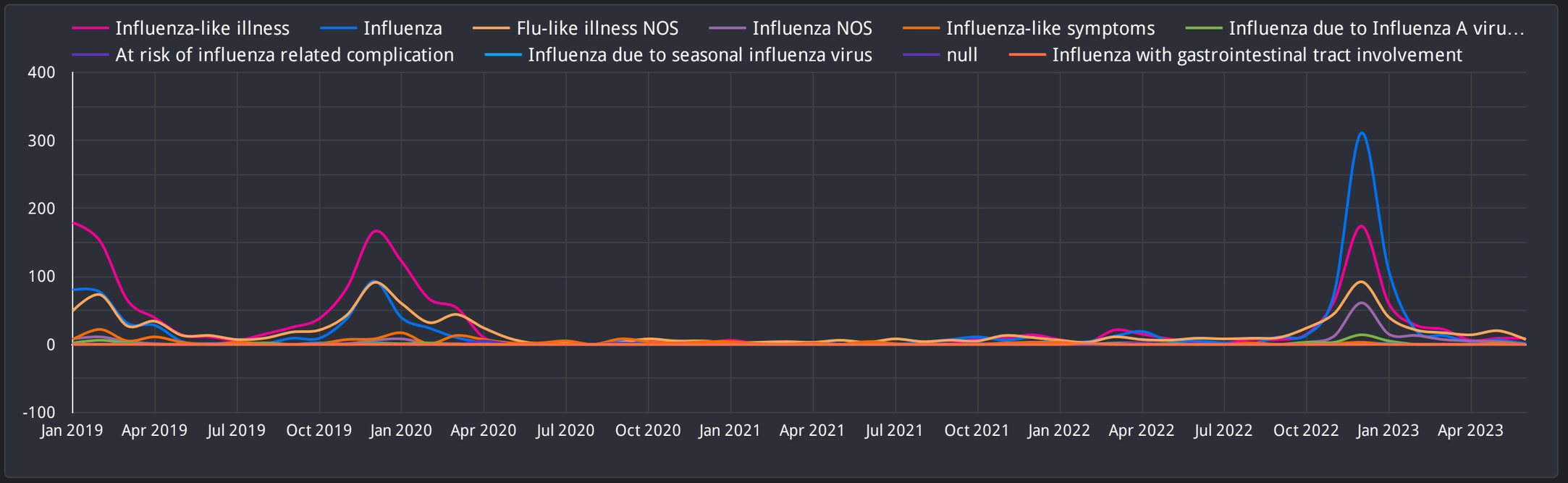

Westcliffe Health Innovations
Bradford Adult Social Care
NHS Bradford & Airedale P…
The Ridge Medical Practice
Horton Bank Practice
Clarendon at ALICE STREET SITE
Tong Medical Practice
Picton Medical Centre
400
300
200
100
0
-100
Feb 2016
Aug 2016
Feb 2017
Aug 2017
Feb 2018
Aug 2018
Feb 2019
Aug 2019
Feb 2020
Aug 2020
Feb 2021
Aug 2021
Feb 2022
Aug 2022
Feb 2023

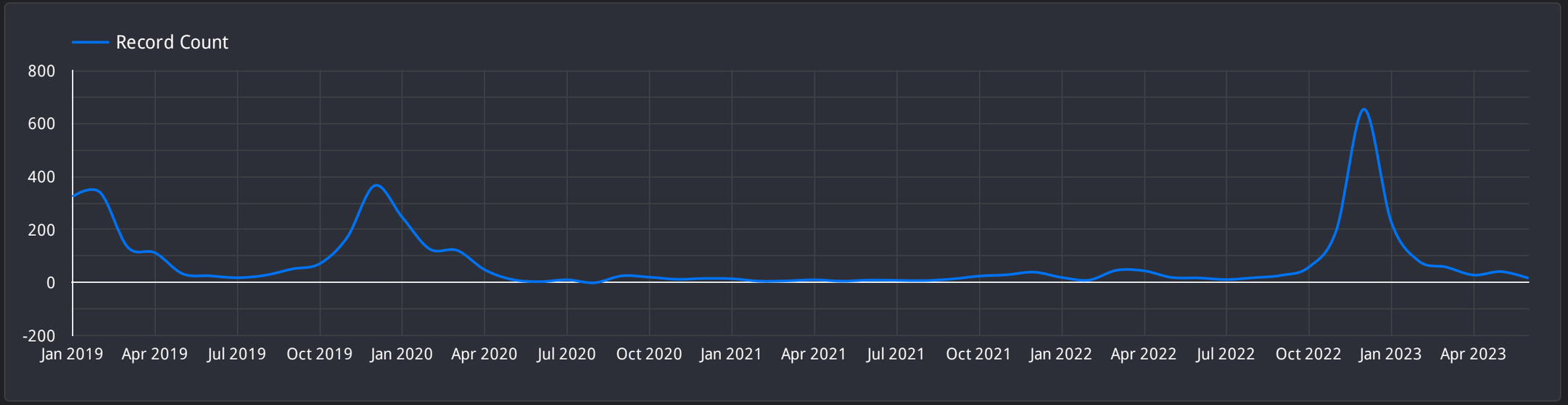

Westcliffe Health Innovations
Coronavirus RNA (ribonucleic acid) detection assay
Bingley Medical Practice
Grange Park Surgery
Kilmeny Surgery
Shipley Medical Practice
NHS Bradford & Airedale Palliative Care Service
Ilkley Moor Medical Practice
Ashcroft Surgery - Dr Mehay and Partners
Tong Medical Practice
Kensington Partnership at Kensington St HC
Coronavirus infection


DateMatched
DateNotMatched
400
300
200
100
600
400
200
0
DateMatched
334
DateMatched


Airedale N...
Kilmeny Su...
The Ridge...
Kensington...
Family He...
Ilkley Moo...
4M
2M
0
-2M
2016
2017
2018
2019
2020
2021
2022
2023
Year


NumberOfRecords
150
100
50
0
Jan 2002
Apr 2003
Jul 2004
Oct 2005
Jan 2007
Apr 2008
Jul 2009
Oct 2010
Jan 2012
Apr 2013
Jul 2014
Oct 2015
Jan 2017
Apr 2018
Jul 2019
Oct 2020
valid_start_date (Year Month)


Influenza-like illness
Influenza
Flu-like illness NOS
Influenza NOS
Influenza-like symptoms
Influenza due to Influenza A viru…
At risk of influenza related complication
Influenza due to seasonal influenza virus
null
Influenza with gastrointestinal tract involvement
400
300
200
100
0
-100
Jan 2019
Apr 2019
Jul 2019
Oct 2019
Jan 2020
Apr 2020
Jul 2020
Oct 2020
Jan 2021
Apr 2021
Jul 2021
Oct 2021
Jan 2022
Apr 2022
Jul 2022
Oct 2022
Jan 2023
Apr 2023


Record Count
800
600
400
200
0
-200
Jan 2019
Apr 2019
Jul 2019
Oct 2019
Jan 2020
Apr 2020
Jul 2020
Oct 2020
Jan 2021
Apr 2021
Jul 2021
Oct 2021
Jan 2022
Apr 2022
Jul 2022
Oct 2022
Jan 2023
Apr 2023

MSc Data Science Research Project Proposal

Author: Manisha Mendonca

Supervisor: Dr Mai Elshehaly

Research Question - Using Visual Analytic workflow design to analyse variations and their implications in the Use of SNOMED by GPs Across England.

1. Introduction

Healthcare professionals across England use Electronic Healthcare records (EHRs) to record the "incident" when a patient visits the Doctor; each disorder is subdivided into multiple small categories to distinguish them based on their specific symptoms; this can have a varied impact on the type of diagnosis and further treatments received by the patients. Before the diagnosis is made, the data goes through a pipeline, and depending on the kind of practices, the codes vary, resulting in discrepancies. Systematised Nomenclature of Medicine Clinical Terms (SNOMED CT) is a structured clinical vocabulary for use in an electronic health record; it helps ensure that data is consistent across the entire NHS system and helps prevent loss of information through human error. SNOMED CT is used by all healthcare professionals in England and other countries such as Sweden for breast cancer treatment management, electronic referral services and storing summary care records (NHS Digital 2018).

1.1 Aim and Objective

The project aims to use a web-enabled visual analytics workflow that can demonstrate how the "selection" of code leads to discrepancies and how these issues can be streamlined, leading to a more accurate diagnosis. The connected Bradford dataset will be used to analyse the data quality and how the codes are used differently across the county. The connected Bradford analysis team will mainly use the workflow to identify data quality and challenges. The successful completion of the project aim will lead to creation of an interactive dashboard to demonstrate key challenges and quality of healthcare data. The Athena vocabularies repository will be used with the Bradford dataset to retrieve semantic relationships between SNOMED codes.

2. Critical Context

Missing data is the most common data quality issue in EHRs (electronic health records). Visualisation targeting specific patterns of missing data can help researchers and organisations understand how multifield missing data patterns can affect data. Interactive visualisations such as heat maps and histograms can help to reveal unexpected gaps in diagnosis and help to gain ratio and entropy calculations to identify the origins of unexpected patterns in data in terms of values and other fields. Investigating patterns of missing value is highly important as it helps to understand the bias within data. Finding such patterns will continuously help improve their data quality and make informed choices. Visualisation dashboards are widely used in healthcare, however, need more focus on displaying high-level metrics such as number values for missing fields in large, collated datasets. Due to the lack of high-level features, analysing multifield patterns of missing values in detail is challenging. The study conducted by Ruddle et al. shows the use of bar charts and histograms to uncover patterns in missing data in over 16 million records from the national admitted patient care data set. Although bar charts and histograms are standard visualisation techniques, using them in combination with interactive tools helps to demonstrate results that show many discrepancies and many gaps in diagnosis and operation codes. The study helps the analysts and researchers by making it easy to check the data they have received by pinpointing specific data in certain timeframes to see trends in codes and diagnoses. (Ruddle et al., 2022)

The design study methodology paper reflects on how visualisation has become an increasingly popular approach to show anomalies within large data sets and to draw analogies. The article also indicates that problem-driven research aims to work with real users to solve real-life problems and use visualisations specific to real-world issues faced by domain experts that can validate the design and reflect the lessons learned to refine the visualisation design guidelines. Task clarity is essential when deciding to make a visual approach. The complexity of the evaluation and the task analogy chooses the type of visualisation techniques that need to be applied. (Sedlmair et al., 2012)

A study conducted by Petersen et al. in 2019 shows the critical implications of handling missing data in electronic health records. The report shows that missing data will affect statistical analysis in health and research studies. This could cause misleading information and involve a long-term clinical examination. Demographic characteristics and disease status must also be considered during the implementation of visual techniques while validating the discrepancy in data. The THIN primary care database was used to analyse the quality of the data sets and see how information and diagnosis were recorded by practice staff in different areas based on the hierarchical coding system. The study concentrated on general factors such as age, gender, and disease status to show the importance of how missing data and handling of it could affect further research in its interpretation. (Petersen et al., 2019)

There are multiple stages of data entry where information could be misinterpreted, leading to discrepancies. The pipeline stages involve an "entry stage", where healthcare staff record diseases and procedures by selecting the SNOMED codes, followed by the "retrieval stage", where analysts select specific codes to retrieve information from the database relating to that specific disease, and finally a "processing stage", where an analyst may run SQL join statements to retrieve information and form analysis. During any of these stages, if the correct types of codes are not used it could lead to a wrong interpretation of how certain diseases are diagnosed more commonly or less commonly across England.

SNOMED CT is now mandatory to use in England; it comes with its disadvantages of being a multi-classification complex system. Healthcare language is complex; therefore, simplifying it requires multiple stages of analysis and interpretation to preserve data. Representation of clinical information in SNOMED CT is complex; therefore, the selection of language used to make the diagnosis is crucial to ensure it is accurate in diagnosing the patient. Lack of transparency and use of different codes to interpret a "comprehensive" disorder such as "autistic spectrum disorder" is one of the most significant disadvantages. Based on the practitioner's discretion, some "codes" could be chosen more often than others, this classification could lead to missing significant details as SNOMED CT has problems with a lack of terminological clarity, and synonyms are not always very clear in terms of context and language. This problem is further amplified during the "retrieval stage", where there could be failures in finding appropriate concepts despite relevant searches. (Rossander et al., 2021)

Analysing the use of SNOMED CT codes across England can help us to give an idea of the types of code that are most used across England and the least used across England for disorders such as autistic spectrum disorder, which have multiple types of diagnosis. The "Athena vocabulary Repository" will be used with the "Bradford Connected Dataset" to reveal discrepancies in the pipeline. This could further help to simplify the pipeline to see if some codes lack valuable descriptions and hence are chosen less often than others across some practices (Jones, 2022). Visualisation techniques such as interactive dashboards to identify patterns for analysts can be very valuable to recognise patterns in datasets that can reveal data quality issues that were previously unknown. This data can then target/ locate locations in England where the discrepancies are at their highest (Ruddle et al., 2022).

Data Visualisation usually helps to identify limitations within the data, find patterns that can use for the critical development of future resources and evaluate key concerns. Data can be explored by either analysing the "missingness" or the "completeness" of information (Hengesbach et al., 2022). The visualisations maybe sometimes limited due to the quality and accuracy of the data; however, these limitations must be addressed in the visualisation, so these measures do not influence the perspective of the design. The purpose of workflow visualisation should be to enhance and express the data in a graphical manner that is easy to understand and interpret. Both techniques can be applied to the Bradford-connected data set to enhance the visualisation and find patterns in data across England.

3. Approaches: Methods & Tools for Design, Analysis & Evaluation

The connected Bradford data set provides access to a large data set that can be used to analyse the variations in the uses of SNOMED CT codes across England. The "Bradford Institute for Health research: data sharing agreement" has been completed to gain access to the data set and design visualisations to answer the research question. The data will be accessed via Google cloud and not be shared with any third parties. Data will also not be stored locally or downloaded from the secure platform during any project stage. The data set is already shared in the anonymous form, and the project will not involve any re identification of the individuals. Once the dataset is obtained, the likely results will be to generate insights that the analysts at the Bradford data centre can generate based on the visualisation workflow tool generated as the project outcome.

The questions will be addressed using a visualisation workflow technique created through tools such as Google looker studio, which helps create interactive dashboards and help analyse large scale data sets (Looker Studio Overview, 2023). The data analyst will then be able to navigate through the workflow to address and find insights into any discrepancies in the use of SNOMED CT codes across England healthcare practises. These insights could be further analysed to get in depth analysis onto changing certain definitions and helping streamlining codes that are no longer used or no longer valid according to the current healthcare requirements.
Google looker studio has built in partner connectors that make it possible to link to any dataset virtually and access data sets and turn them into interactive reports and dashboards to visualise the information to make informed decisions. Since the connected Bradford data set will be stored on Google cloud, it will be easy to connect the data set on to Google looker studio and build visualisations. Visualisations can be represented using highly configurable charts and tables, it easily can be connected into other data sources such as Google cloud share and build interactive charts and graphs that data analysts can filter range and control to examine patterns in data sets that would otherwise be not possible. It could also include clickable images and different styles and colour themes to enhance the visualisation. Google looker studio also allows a big query, which will be used to retrieve and analyse certain SNOMED CT codes (Looker Studio Help, 2023). In conjunction with Google Looker studio, other platforms such as D3 (data-driven documents) are a JavaScript library for producing interactive Stand-alone data visualisation in web browsers. D3 works extremely well with large data sets, is open source, and does not require additional technology or plugins to create visualisations. Like Google looker studio, network nodes and edges can be adjusted in weights and colour to define parameters and enhance visualisations (Taskesen, 2023).

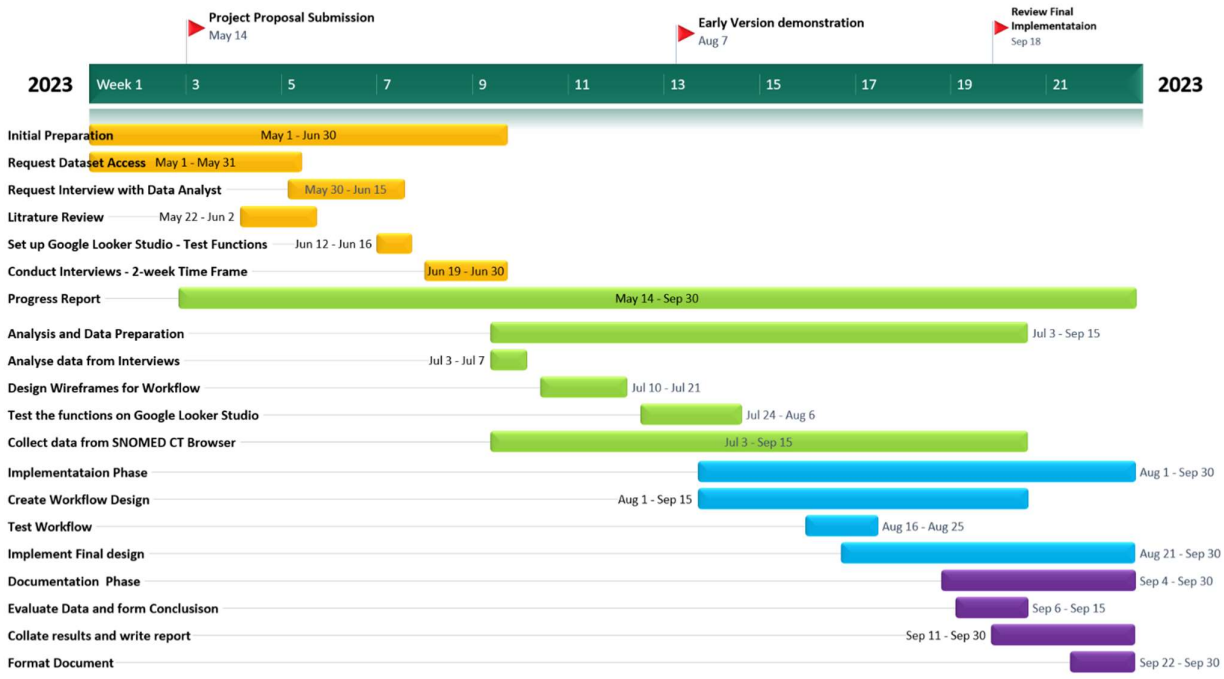
Data Collection: Domain experts who are data analysts from the connected Bradford team will be interviewed to gain further insights into how the visualisations could be constructed to form informed decisions. Interviews will take place online and the sessions will be recorded then transcribed and all data will be anonymised. The original recordings will then be deleted securely so that once the data is published it will not be used to quote any specific individual that has taken part in the interview. The interview will only be used to collect qualitative information to understand the depth of how quality of information can impact at each stage when healthcare professionals add data into the electronic healthcare records and how visualisations could enhance to understand this data and form analysis. The interviews will be conducted through Microsoft Teams or zoom application. The Interviewee will be asked to complete a consent form before proceeding with the interview and will be made aware of the complete description of the coursework and how the data collected in the interview will be used to enhance the coursework.

Method of data collection: The Athena Vocabularies repository will be used to retrieve semantic relationships between SNOMED codes, these codes will then be applied to the SNOMED CT Browser, to gain details on the various disorders and the different types of codes that used for diagnosis. Once all the relevant data is obtained, matching health records from the Bradford healthcare dataset are obtained by running a big Query to retrieve relevant data.
Design wireframes using Power point/ Adobe XD to get initial ideas of how the information can be displayed.

Analysis and Evaluation:

Once the data is collected, it will be analysed to see trends in the data sets, that would otherwise be very difficult to note in raw data in its tabular format. The data will be compared to the entire England to see if specific geographical locations have variations in the use of SNOMED CT codes and how this affects the analysts who then choose sections of data and analyse them. Once the analysis is complete, further insights from the Bradford-connected data team analyst could be asked to check and use the workflow to see if the implementation meets the expected standard in streamlining the process of discovering discrepancies and helping to make better-informed decisions about data and assess its overall quality. Once the workflow visualisation tool is ready, this tool could be used to curate advanced data sets and used by other healthcare professionals to pinpoint certain sections of data and drive conclusions that can help to understand how SNOMED CT codes are used across practitioners in England.

4. Work Plan



5. Risk Analysis

Task No	Risk	Probability	Impact	Plan
1	Scheduling issues with interviews with the Bradford healthcare team.	Medium	High	Schedule the interview well in advance and send all necessary consent documents before the interview. – Allow up to 4 weeks to conduct the interview.
2	Software issues leading to delays in proposed internal deadlines	Low	High	Store the report/ analysis data on Google Cloud and locally (Excluding the dataset) and manage regular version control to minimize data loss.
3	Unable to understand/ retrieve valuable information from the dataset	Low	High	Speak with the supervisor and seek help understanding the dataset and how to proceed with the project.
4	Issues in designing the workflow	Medium	High	Explore all possible platforms that can be used to design the workflow- Discuss any further problems with the supervisor.
5	Scope of the project too complex to complete within the timeframe	Low	High	Discuss the project's scope with the supervisor to see how this can be resolved.
6	Unable to generate valid insights from the dataset	Low	High	Discuss the project's scope with the supervisor to see how this can be resolved.

References

1. NHS Digital (2018). *SNOMED CT - NHS Digital*. [online] NHS Digital. Available at: https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct. (Accessed: 11 May 2023)

2. Jones, M. (2022). *Chapter 22 - Benefits of using SNOMED CT in the UK National Health Service (NHS)*. [online] ScienceDirect. Available at: https://www.sciencedirect.com/science/article/pii/B9780128234136000239 (Accessed: 11 May 2023)

3. Petersen, I., Welch, C.A., Nazareth, I., Walters, K., Marston, L., Morris, R.W., Carpenter, J.R., Morris, T.P. and Pham, T.M. (2019). Health indicator recording in UK primary care electronic health records: key implications for handling missing data. *Clinical Epidemiology*, Volume 11, pp.157–167. Available at: https://pubmed.ncbi.nlm.nih.gov/30809103/ (Accessed: 13 May 2023)

4. Rossander, A., Lindsköld, L., Ranerup, A. and Karlsson, D. (2021). A State-of-the Art Review of SNOMED CT Terminology Binding and Recommendations for Practice and Research. *Methods of Information in Medicine*, 60(S 02), pp. e76–88. Available at: https://www.thieme-connect.com/products/ejournals/pdf/10.1055/s-0041-1735167.pdf (Accessed: 12 May 2023)

5. Ruddle, R.A., Adnan, M. and Hall, M. (2022). Using set visualisation to find and explain patterns of missing values: a case study with NHS hospital episode statistics data. *BMJ Open*, [online] 12(11), p.e064887. Available at: https://bmjopen.bmj.com/content/bmjopen/12/11/e064887.full.pdf (Accessed: 12: May 2023)

6. Hengesbach, N., McInerny, G.J. and Albuquerque, J.P. de (2022). Seeing what is not shown. *Information Design Journal*. Available at: https://www.jbe-platform.com/docserver/fulltext/idj.22006.hen.pdf?expires=1683820367&id=id&accname=guest&checksum=6A522C32D14107FC411F7340667A40DE (Accessed : 12 May 2023)

7. lookerstudio.google.com. (n.d.). *Looker Studio Overview*. [online] Available at: https://lookerstudio.google.com/overview. (Accessed: 12 May 2023)

8. support.google.com. (n.d.). *Welcome to Looker Studio! - Looker Studio Help*. [online] Available at: https://support.google.com/looker-studio/answer/6283323?hl=en. (Accessed: 12 May 2023)

9. Taskesen, E. (n.d.). *d3graph: Python package to create interactive network based on d3js.* [online] PyPI. Available at: https://pypi.org/project/d3graph/ (Accessed: 12 May 2023).

10. Sedlmair, M., Meyer, M. and Munzner, T. (2012). Design Study Methodology: Reflections from the Trenches and the Stacks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12), pp.2431–2440. Available at: https://www.cs.ubc.ca/labs/imager/tr/2012/dsm/dsm.pdf (Accessed: 12 May 2023).

Research Ethics Review Form: BSc, MSc and MA Projects

Computer Science Research Ethics Committee (CSREC)

https://www.city.ac.uk/about/governance/committees/cs-research-ethics

Undergraduate and postgraduate students undertaking their final project in the Department of Computer Science are required to consider the ethics of their project work and to ensure that it complies with research ethics guidelines. In some cases, a project will need approval from an ethics committee before it can proceed. Usually, but not always, this will be because the student is involving other people ("participants") in the project.

In order to ensure that appropriate consideration is given to ethical issues, all students must complete this form and attach it to their project proposal document. There are two parts:

PART A: Ethics Checklist. All students must complete this part. The checklist identifies whether the project requires ethical approval and, if so, where to apply for approval.

PART B: Ethics Proportionate Review Form. Students who have answered "no" to all questions in A1, A2 and A3 and "yes" to question 4 in A4 in the ethics checklist must complete this part. The project supervisor has delegated authority to provide approval in such cases that are considered to involve MINIMAL risk.
The approval may be ***provisional*** – *identifying the planned research as* likely to involve MINIMAL RISK. In such cases you must additionally seek ***full approval*** from the supervisor as the project progresses and details are established. ***Full approval*** must be acquired in writing, before beginning the planned research.

A.1 If you answer YES to any of the questions in this block, you must apply to an appropriate external ethics committee for approval and log this approval as an External Application through Research Ethics Online - https://ethics.city.ac.uk/		*Delete as appropriate*
1.1	Does your research require approval from the National Research Ethics Service (NRES)? *e.g. because you are recruiting current NHS patients or staff?* *If you are unsure try -* https://www.hra.nhs.uk/approvals-amendments/what-approvals-do-i-need/	**NO**
1.2	Will you recruit participants who fall under the auspices of the Mental Capacity Act? *Such research needs to be approved by an external ethics committee such as NRES or the Social Care Research Ethics Committee -* http://www.scie.org.uk/research/ethics-committee/	**NO**
1.3	Will you recruit any participants who are currently under the auspices of the Criminal Justice System, for example, but not limited to, people on remand, prisoners and those on probation? *Such research needs to be authorised by the ethics approval system of the National Offender Management Service.*	**NO**
A.2 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee, you must apply for approval from the Senate Research Ethics Committee (SREC) through Research Ethics Online - https://ethics.city.ac.uk/		*Delete as appropriate*

2.1	Does your research involve participants who are unable to give informed consent? *For example, but not limited to, people who may have a degree of learning disability or mental health problem, that means they are unable to make an informed decision on their own behalf.*	**NO**
2.2	Is there a risk that your research might lead to disclosures from participants concerning their involvement in illegal activities?	**NO**
2.3	Is there a risk that obscene and or illegal material may need to be accessed for your research study (including online content and other material)?	**NO**
2.4	Does your project involve participants disclosing information about special category or sensitive subjects? *For example, but not limited to: racial or ethnic origin; political opinions; religious beliefs; trade union membership; physical or mental health; sexual life; criminal offences and proceedings*	**NO**
2.5	Does your research involve you travelling to another country outside of the UK, where the Foreign & Commonwealth Office has issued a travel warning that affects the area in which you will study? *Please check the latest guidance from the FCO - http://www.fco.gov.uk/en/*	**NO**
2.6	Does your research involve invasive or intrusive procedures? *These may include, but are not limited to, electrical stimulation, heat, cold or bruising.*	**NO**
2.7	Does your research involve animals?	**NO**
2.8	Does your research involve the administration of drugs, placebos or other substances to study participants?	**NO**
A.3 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee or the SREC, you must apply for approval from the Computer Science Research Ethics Committee (CSREC) through Research Ethics Online - https://ethics.city.ac.uk/ **Depending on the level of risk associated with your application, it may be referred to the Senate Research Ethics Committee.**		*Delete as appropriate*
3.1	Does your research involve participants who are under the age of 18?	**NO**
3.2	Does your research involve adults who are vulnerable because of their social, psychological or medical circumstances (vulnerable adults)? *This includes adults with cognitive and / or learning disabilities, adults with physical disabilities and older people.*	**NO**
3.3	Are participants recruited because they are staff or students of City, University of London? *For example, students studying on a particular course or module.* *If yes, then approval is also required from the Head of Department or Programme Director.*	**NO**
3.4	Does your research involve intentional deception of participants?	**NO**

3.5	Does your research involve participants taking part without their informed consent?	**NO**
3.5	Is the risk posed to participants greater than that in normal working life?	**NO**
3.7	Is the risk posed to you, the researcher(s), greater than that in normal working life?	**NO**
A.4 If you answer YES to the following question and your answers to all other questions in sections A1, A2 and A3 are NO, then your project is deemed to be of MINIMAL RISK. **If this is the case, then you can apply for approval through your supervisor under PROPORTIONATE REVIEW. You do so by completing PART B of this form.** **If you have answered NO to all questions on this form, then your project does not require ethical approval. You should submit and retain this form as evidence of this.**		*Delete as appropriate*
4	Does your project involve human participants or their identifiable personal data? *For example, as interviewees, respondents to a survey or participants in testing.*	**YES**

PART B: Ethics Proportionate Review Form

If you answered YES to question 4 and NO to all other questions in sections A1, A2 and A3 in PART A of this form, then you may use PART B of this form to submit an application for a proportionate ethics review of your project. Your project supervisor has delegated authority to review and approve this application under proportionate review. You must receive final approval from your supervisor in writing before beginning the planned research.

However, if you cannot provide all the required attachments (see B.3) with your project proposal (e.g. because you have not yet written the consent forms, interview schedules etc), the approval from your supervisor will be ***provisional***. You **must** submit the missing items to your supervisor for approval prior to commencing these parts of your project. Once again, you must receive written confirmation from your supervisor that any provisional approval has been superseded by with ***full approval*** of the planned activity as detailed in the full documents. **Failure to follow this procedure and demonstrate that final approval has been achieved may result in you failing the project module.**

Your supervisor may ask you to submit a full ethics application through Research Ethics Online, for instance if they are unable to approve your application, if the level of risks associated with your project change, or if you need an approval letter from the CSREC for an external organisation.

B.1 The following questions must be answered fully. All grey instructions must be removed.		*Delete as appropriate*
1.1.	Will you ensure that participants taking part in your project are fully informed about the purpose of the research?	**YES**
1.2	Will you ensure that participants taking part in your project are fully informed about the procedures affecting them or affecting any information collected about them, including information about how the data will be used, to whom it will be disclosed, and how long it will be kept?	**YES**
1.3	When people agree to participate in your project, will it be made clear to them that they may withdraw (i.e. not participate) at any time without any penalty?	**YES**
1.4	Will consent be obtained from the participants in your project? Consent from participants will be necessary if you plan to involve them in your project or if you plan to use identifiable personal data from existing records. "Identifiable personal data" means data relating to a living person who might be identifiable if the record includes their name, username, student id, DNA, fingerprint, address, etc. *If YES, you must attach drafts of the participant information sheet(s) and consent form(s) that you will use in section B.3 or, in the case of an existing dataset, provide details of how consent has been obtained.* *You must also retain the completed forms for subsequent inspection. Failure to provide the completed consent request forms will result in withdrawal of any earlier ethical approval of your project.*	**YES**

1.5	Have you made arrangements to ensure that material and/or private information obtained from or about the participating individuals will remain confidential?	**YES**

B.2 If the answer to the following question (B2) is YES, you must provide details		*Delete as appropriate*
2	Will the research be conducted in the participant's home or other non-University location? *If **YES**, you must provide details of how your safety will be ensured.*	**NO**

B.3 Attachments **ALL of the following documents MUST be provided to supervisors if applicable.** **All must be considered prior to final approval by supervisors.** **A written record of final approval must be provided and retained.**	***YES***	***NO***	***Not Applicable***
Details on how safety will be assured in any non-University location, including risk assessment if required (see B2)			***Not Applicable***
Details of arrangements to ensure that material and/or private information obtained from or about the participating individuals will remain confidential (see B1.5) *Any personal data must be acquired, stored and made accessible in ways that are GDPR compliant.*	**Yes**		
Full protocol for any workshops or interviews**	**Yes**		
Participant information sheet(s)**	**Yes**		
Consent form(s)**	**Yes**		
Questionnaire(s)** *sharing a Qualtrics survey with your supervisor is recommended.*	**Yes**		
Topic guide(s) for interviews and focus groups**	**Yes**		
Permission from external organisations or Head of Department** *e.g. for recruitment of participants*	**Yes**		

***If these items are not available at the time of submitting your project proposal, then **provisional approval** can still be given, under the condition that you must submit the final versions of all items to your supervisor for approval at a later date. **All** such items **must** be seen and approved by your supervisor before the activity for which they are needed begins. Written evidence of **final approval** of your planned activity must be acquired from your supervisor before you commence.*

Changes

If your plans change and any aspects of your research that are documented in the approval process change as a consequence, then any approval acquired is invalid. If issues addressed in Part A (the checklist) are affected, then you must complete the approval process again and establish the kind of approval that is required. If issues addressed in Part B are affected, then you must forward updated documentation to your supervisor and have received written confirmation of approval of the revised activity before proceeding.

Templates for Consent and Information

You must use the templates provided by the University as the basis for your participant information sheets and consent forms. You **must** adapt them according to the needs of your project before you submit them for consideration.

Participant Information Sheets, Consent Forms and Protocols must be consistent. Please ensure that this is the case prior to seeking approval. Failure to do so will slow down the approval process.

We strongly recommend using Qualtrics to produce digital information sheets and consent forms.

Further Information

https://www.city.ac.uk/about/governance/committees/cs-research-ethics

https://www.city.ac.uk/research/ethics/how-to-apply/participant-recruitment

https://www.city.ac.uk/research/ethics