CITY
UNIVERSITY OF LONDON
EST 1894

City, University of London

MSc in Data Science

Project Report

2019

Modelling Depression Recurrence Through Analysis of Electronic Health Records

Lesley Dwyer

Supervised by: Dr Cagatay Turkay

1st October 2019

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:
Lesley Dwyer

Abstract

Using South London and Maudsley (SLaM) NHS Foundation Trust anonymised electronic health records, data from a group of 11,971 patients with depression were used for this study. Within this group, 1,467 (12.25%) experienced a recurrence of depression within two years following discharge from SLaM. The patients' clinical, demographic and area-level deprivation data was collected and analysed. The patient data was reshaped into sequences and divided into training and test sets. Several long short-term memory (LSTM) models were produced to identify which patients from the unseen test set would have a recurring episode of depression within two years. The best model used only male patient data and had an area under the receiver operating characteristic curve (AUC) of 0.56, recall of 54% and accuracy of 58%. A key limitation of the study was having limited data (populated for fewer than 20% of patients) for some data entities. This was possibly due to information being added to free text fields, rather than structured fields used in this project.

Acknowledgements

I would like to thank both the Programme Director, Dr Aidan Slingsby, and my supervisor, Dr Cagatay Turkay, for their support and advice before and during the project. I would also like to thank the clinician for his time and expertise on depression and the data. Lastly, I would like to thank the team who supports the CRIS application and data at SLaM, especially the subject matter expert. They were most helpful in getting hot desks and software set up, helping me to understand the data structure and getting questions answered about mental health and the data.

Table of Contents

Declaration 1

Abstract 2

Acknowledgements 2

Chapter 1 – Introduction and Objectives 6

1.1 Background 6

1.2 Research Question and Objectives 6

1.3 Beneficiaries 7

1.4 Methods and Work Plan 7

1.5 Structure of Report 8

Chapter 2 – Context 9

2.1 Machine Learning 9

2.2 Healthcare Domain 10

2.3 Depression and Recurrence 11

2.4 Electronic Health Records 12

2.5 Previous Research 12

2.4.1 Machine Learning and Depression 12

2.4.2 Machine Learning and Hospital Readmission 13

2.4.3 Machine Learning and EHRs 13

2.4.4 Sequential Modelling and Healthcare 13

2.6 Gaps in the Research 14

2.7 Healthcare Provider and Data 15

Chapter 3 – Methods 17

3.1 Research and Setup Phase 17

3.1.1 Access Data and System 17

3.1.2 Prepare Research Participants 18

3.1.3 Select Project Management Framework 18

3.1.4 Research Algorithms 19

3.1.5 Build Prototypes 20

3.2 Define and Analyse Phase 21

3.2.1 Understand Domain and Data 21

3.2.2 Define Data Requirements 22

3.2.3 Collect Data 25

3.2.4 Prepare Data 25

3.2.5 Analyse Data 26

3.2.6 Select Algorithm 26

3.3 Build and Evaluate Phase 27

3.3.1 Select Evaluation Metrics 27

3.3.2 Build Model 28

3.3.3 Build Additional Models by Gender 29

3.3.4 Build Additional Model using Filtered Data 30

Chapter 4 – Results 31

4.1 Data Specifications 31

4.2 Data Extracts 31

4.3 Prototype Results 34

4.4 Exploratory Data Analysis Results 35

4.5 Model Results 45

4.6 Additional Model by Gender Results 47

4.7 Additional Model using Filtered Data Results 49

4.8 Summary 49

Chapter 5 – Discussion 51

5.1 Objectives 51

5.2 Answering the Research Question 53

5.3 Limitations 54

Chapter 6 – Evaluation, Reflections, and Conclusions 56

6.1 Evaluation and Reflections 56

6.1.1 What went well? 56

6.1.2 What could be improved? 56

6.1.3 Future Work 57

6.2 Conclusions 57

Glossary 59

References 60

Chapter 1 – Introduction and Objectives

1.1 Background

Depression affects over 300 million people worldwide (World Health Organization, 2018), and was estimated to cost £9 billion annually in the UK alone (Thomas and Morris, 2003). Depression has a high rate of recurrence with one study showing 38% of people experiencing at least one recurring episode after an initial episode of depression (Eaton *et al.*, 2008). The severity of depression can also change over time. In a Danish study, Kessing (2008) found that the prevalence of severe depressive episodes increases throughout the course of illness for patients with depression.

The availability of electronic health records (EHRs) combined with machine learning can provide a means to better understanding illness. Machine learning has been used in a number of studies to predict outcomes for physical and mental illness, such as opioid use (Che *et al.*, 2017), cardiovascular disease (Zhao *et al*., 2019), dementia (Mahmoud *et al*., 2013), mood disturbances (Cao *et al*., 2017) and depression (Kessler *et al*., 2016). However, some studies do not take advantage of the longitudinal or sequential nature of EHRs, which could be of value. Zhao *et al.* (2019) state that EHRs 'contain a wealth of detailed clinical information and provide several distinct advantages for clinical research, including cost efficiency, big data scalability, and the ability to analyze data over time.'

This project investigates how sequential machine learning models can be used with anonymised EHRs to determine the likelihood that a patient will experience a recurrence of depression within a given timeframe following remission of the illness. Given the rising costs of healthcare in the UK (Cooper *et al*., 2019) and the prevalence of depression, having the ability to predict which patients will experience a relapse of depression could lead to treatment being altered to try and prevent this.

Other research has been done in this area, but it differs or contains gaps when compared to this project. Generally, these studies used different algorithms or data or were asking a different research question.

1.2 Research Question and Objectives

The overall aim of this project is to determine if a patient, having experienced a previous episode of depression, is likely to have a recurrence. Given the vast amount of data stored in EHRs and the important role this data plays in clinical decision support (Evans, 2016), machine learning could help where human capacity is limited. Programs that can process a patient's full medical profile and highlight something relevant to a clinician, could save time, cost and may find something hidden in the data that a human cannot.

Therefore, the **research question** for this project is:

- To what extent can machine learning be used with anonymised electronic health records to determine the likelihood a patient will experience a relapse/recurrence of depression within a given time frame following remission?

In order to answer this question, several **objectives** have been set:

O1. To research sequence modelling algorithms and related literature
O2. To define attributes and thresholds from the source data to be used for analysis
O3. To cleanse and prepare the data for further analysis and use in modelling
O4. To perform exploratory data analysis (EDA) to better understand the data and define data for modelling
O5. To build a machine learning model that determines if a patient will experience a relapse of depression
O6. To investigate whether building separate machine learning models for different patient groups will improve model performance or provide additional understanding of the illness
O7. To investigate whether placing limits on the data included in the model will improve model performance or provide additional understanding of the illness

The **outcomes** of the project are:

- Machine learning models that, for a new/unseen patient, classes that patient as belonging to the relapse or non-relapse group.
- Knowledge discovered during the EDA about patients with depression in the relapse and non-relapse groups.

1.3 Beneficiaries

The anticipated beneficiaries of this work are clinicians working with patients who have been diagnosed with depression. Anonymised data has been used for this, but a modified version of this work using identifiers to identify the patient could provide clinical decision support. It could indicate to a clinician whether a patient is more likely to experience another episode of depression in a specified timeframe. The EHR data used in this study was from the South London and Maudsley (SLaM) NHS Foundation Trust, so the findings could potentially be used by clinicians or staff within SLaM.

1.4 Methods and Work Plan

The methods used for this project include reviewing the literature, researching algorithms, understanding the domain and data, building prototypes, defining data requirements, collecting and preparing data, performing exploratory data analysis, and building models. The initial work plan was produced in the original research proposal and is available in Appendix A, page A-7. During the project, a more detailed level of planning was needed, so I decided to manage the project using the Scrum framework. This is discussed in more detail in section 3.1.3.

1.5 Structure of Report

The context of the problem, including previous research and gaps, is discussed in Chapter 2. Chapter 3 lays out the methods used during the project. Chapter 4 presents the results, and Chapter 5 discusses those results and how they relate to the objectives and research question. The final chapter includes an evaluation of the project, reflections of the work done and conclusions.

Appendices, which have been submitted as a separate document, are structured as follows:

- Appendix A: Research Proposal
- Appendix B: Additional Ethics Forms
- Appendix C: Interview Questions and Notes
- Appendix D: Data Specifications
- Appendix E: SQL Queries
- Appendix F: Python Code (sample) and R Code

In addition, a separate zip file has been submitted with:

- Readme file
- Prototype data and target files
- Two signed informed consent files
- Full Python code in Jupyter notebooks and html (for better readability)
- Original (Word) version of this report
- Notes of meetings with supervisor
- List of user stories

Raw source data has **not** been submitted. One of the conditions of using the data for this project is that it had to remain within a firewall.

Chapter 2 – Context

This chapter gives a brief overview of machine learning, the healthcare domain, depression and electronic health records before presenting previous research in these areas. It then discusses the gaps in the previous research and how this project aims to fill those gaps. Lastly, it includes information on the specific healthcare provider and data used for this project.

2.1 Machine Learning

Mitchell (1997, xv) states ‘the field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience.’ Machine learning does this using techniques from many other disciplines like statistics, information theory, neurobiology, artificial intelligence and others (Mitchell, 1997, p. 2). Bishop (2006, p. 3) describes machine learning in a more practical way by discussing the use of training data to train the machine learning model, and target data representing the known answer or category. The model that has been trained can then be tested on unseen data, i.e. the test data. A model that produces correct answers or categories for the unseen data is said to generalize well. He goes on to categorise machine learning into three types: supervised, unsupervised and reinforcement learning. Supervised learning is when the targets or labels are provided with the data. Unsupervised learning is when there is no target data provided, and the goal may be to find similar groups within the data; a good example of this is clustering. Reinforcement learning involves discovering the best output using trial and error to maximise a reward. Lastly, he discusses the two main types of machine learning problems: classification and regression. Classification problems involve discrete or categorical target data, whereas regression problems have target data that is continuous.

There are many different machine learning algorithms, and different algorithms suit different types of problems. For example, algorithms that can be used for supervised classification problems include decision trees (Mitchell, p. 52), logistic regression (Bishop, p. 205) and Support Vector Machines (SVMs) (Bishop, p. 338). Modelling data in a time series or a sequence is another type of problem. Algorithms suitable for this include Hidden Markov Models (HMMs) (Bishop, p. 610) and recurrent neural networks (RNNs) (Mitchell, p. 119).

Neural networks, according to Mitchell (1997, p. 81), ‘provide a general, practical method for learning real-valued, discrete-valued, and vector-valued functions from examples.’ They were inspired by the architecture of the brain which contains a vast network of connected neurons, and activity between neurons is activated or inhibited (Mitchell, 1997, p. 82). A neural network has a similar structure with neurons, or units, and weights which play a part in activating the units. The weights are learned during the model training. A multi-layer neural network is made up of different layers of neurons including the input layer, one or many hidden layers and an output layer. An example of a feed-forward neural network is shown below.

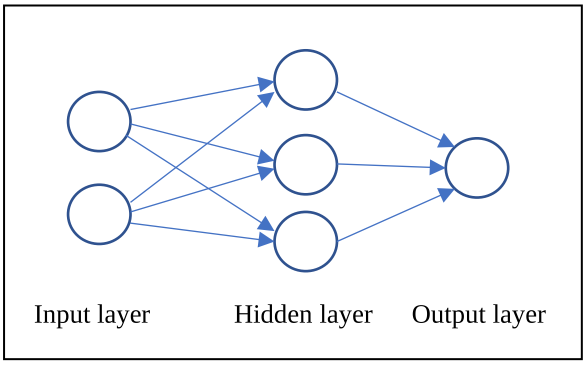


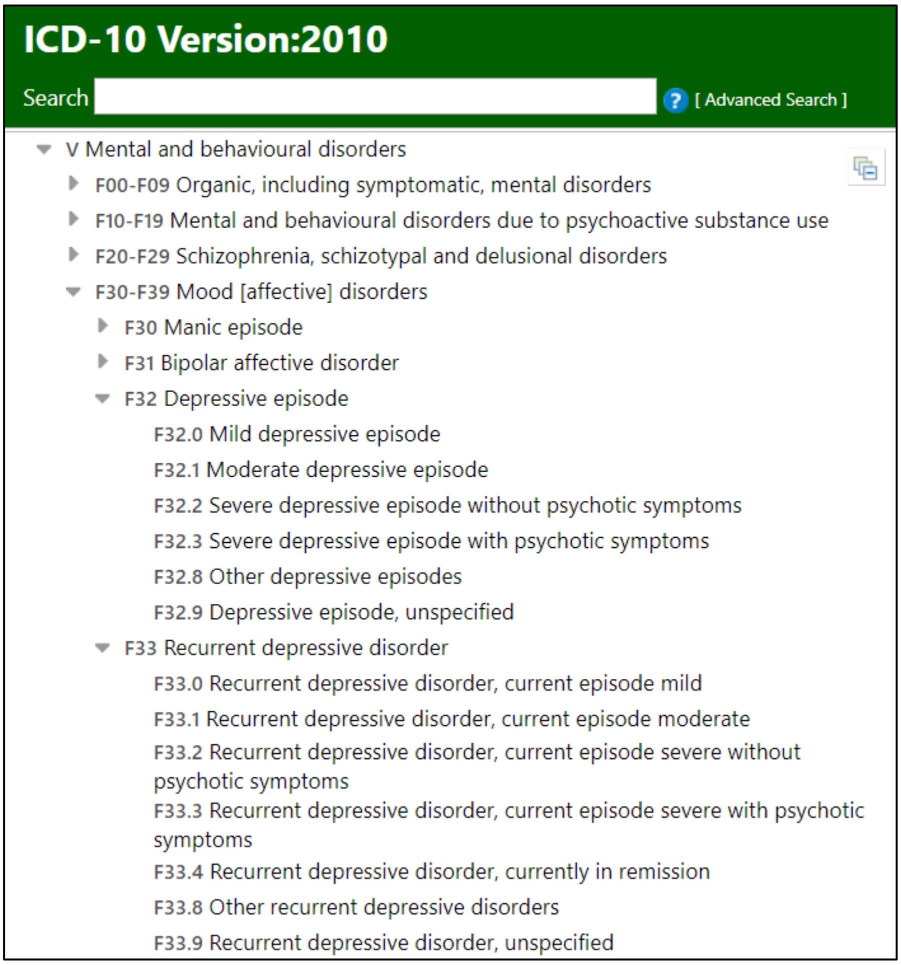
Figure 2.1. Neural network based on example from Bishop (2006, p. 228).

Deep learning describes networks with many hidden layers while recurrent neural networks allow for feedback loops instead of only processing data in one direction. This makes RNNs suitable for modelling sequences (Durstewitz *et al.*, 2019).

2.2 Healthcare Domain

This section explains some terms used in a healthcare setting and specifically within mental health. It includes brief explanations of diagnostic coding standards and scales to measure mental health called HoNOS.

Diagnosis of disease is standardised using a global coding system called the International Classification of Diseases (ICD). The standards get amended periodically with the current version being ICD-10 (World Health Organization, 2019). Each diagnosis is assigned a category, such as F32, which represents a depressive episode. This can be further broken down into sub-categories like F32.1 representing a moderate depressive episode. Figure 2.2 below shows the diagnostic categories and sub-categories of depression from the ICD-10 online application (World Health Organization, 2016).



ICD-10 Version:2010

- V Mental and behavioural disorders
  - F00-F09 Organic, including symptomatic, mental disorders
  - F10-F19 Mental and behavioural disorders due to psychoactive substance use
  - F20-F29 Schizophrenia, schizotypal and delusional disorders
  - F30-F39 Mood [affective] disorders
    - F30 Manic episode
    - F31 Bipolar affective disorder
    - F32 Depressive episode
      - F32.0 Mild depressive episode
      - F32.1 Moderate depressive episode
      - F32.2 Severe depressive episode without psychotic symptoms
      - F32.3 Severe depressive episode with psychotic symptoms
      - F32.8 Other depressive episodes
      - F32.9 Depressive episode, unspecified
    - F33 Recurrent depressive disorder
      - F33.0 Recurrent depressive disorder, current episode mild
      - F33.1 Recurrent depressive disorder, current episode moderate
      - F33.2 Recurrent depressive disorder, current episode severe without psychotic symptoms
      - F33.3 Recurrent depressive disorder, current episode severe with psychotic symptoms
      - F33.4 Recurrent depressive disorder, currently in remission
      - F33.8 Other recurrent depressive disorders
      - F33.9 Recurrent depressive disorder, unspecified

Figure 2.2. ICD-10 online application showing diagnostic categories for mental and behavioural disorders.

The Health of the Nation Outcome Scales (HoNOS) are a set of questions developed following a target set by the Department of Health to improve mental health (Wing *et al*., 1998). The 12 questions measure behaviour, impairment, symptoms and social functioning on a 0-4 rating scale. Additional scales were developed for different groups: HoNOS65+ for older adults, HoNOSCA for Children and Adolescents, HoNOS-secure for use in a secure psychiatric setting like prisons or forensic services, HoNOS-LD for patients with learning disabilities and HoNOS-ABI for patients with Acquired Brain Injury (Royal College of Psychiatrists, 2019). ‘HoNOS is the most widely used routine clinical outcome measure used by English mental health services’ according to the Royal College of Psychiatrists (2019).

2.3 Depression and Recurrence

The World Health Organization (2018) states that depression is the leading cause of disability globally. According to the ICD-10 classification of diseases, patients experiencing a depressive episode typically suffer from low mood, reduced energy, marked tiredness, lack of enjoyment and reduced activity, among other things. Depression can be diagnosed as a depressive episode or recurrent depressive disorder, meaning a patient has repeated episodes of depression. Both types can vary from mild to severe (World Health Organization, 2016). Recurring episodes are sometimes referred to as relapse in the literature.

2.4 Electronic Health Records

Medical records used to be recorded on paper until the technology became available to allow for recoding these details electronically. Today, EHRs are used by different entities like GPs, hospitals and insurance companies, and they include details such as family history, lab results and medication. Natural language processing (NLP) has been implemented in many EHR systems to uncover useful information in the large amount of free text recorded in clinical notes (Evans, 2016).

In addition to their use in patient health care, EHRs are being used for research. Perlis *et al*. (2012) state that research using EHRs 'reflects clinical practice. It also offers far greater efficiency and feasibility than traditional clinical trials, as the data have already been collected and coded.'

Given the sensitive nature of the details stored in EHRs, data security and privacy are vital to their use and acceptance (Evans, 2016). One measure that has been taken when using EHRs for research to ensure data privacy is to anonymise or de-identify the data so it cannot be attributed to a specific patient. Examples where this has been implemented include the Secure Anonymised Information Linkage (SAIL) databank and the Clinical Records Interactive Search (CRIS) system used in this project (McIntosh, 2016).

2.5 Previous Research

2.4.1 Machine Learning and Depression

Previous research has used machine learning techniques to understand or predict illness, and more specifically depression and depression relapse.

In 2014, Wang *et al*. did a study on predicting recurrence of major depression using data from the U.S. National Epidemiological Survey on Alcohol and Related Conditions (NESARC). The surveys were completed by patients who had Major Depressive Disorder (MDD) and had been in remission for at least two months. The researchers used a logistic regression model to predict the recurrence of depression.

A later study by Kessler *et al*. (2016) used machine learning to predict severity and persistence of MDD from baseline self-reports over a period of 10-12 years. The machine learning models used in this case were ensemble regression trees and penalized regression.

Two additional studies from Nie *et al*. (2016) and Sakurai *et al*. (2017) investigated prediction of depression relapse using the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) clinical trial data. Nie *et al*. (2016) used a gradient boosting algorithm and a stochastic dual coordinate ascent algorithm, while Sakurai *et al*. (2017) used a Cox proportional hazard model.

In 2018, Dinga *et al*. used a penalized logistic regression model to predict the naturalistic course of depression. They incorporated 81 attributes from clinical, psychological, and biological data obtained from the Netherlands Study of Depression and Anxiety (NESDA). Patients with MDD or dysthymia were measured at the start of the study and again two years later, but no interventions were applied. The researchers were able to predict rapid remission with 62% accuracy and 0.69 Area Under the Receiver Operating Characteristic curve (AUC) and MDD diagnosis at a 2-year follow up with 66% accuracy and 0.66 AUC. They found symptoms to be the most predictive factors.

2.4.2 Machine Learning and Hospital Readmission

Machine learning has been used to predict hospital readmission, which is a similar type of problem to predicting disease recurrence. One example of this is where Doryab *et al.* (2019) used data collected from Fitbits to determine if patients would be readmitted following pancreatic surgery. They used a variety of machine learning algorithms in this study including random forest, logistic regression, SVMs, Bayesian networks, and boosted logistic regression.

2.4.3 Machine Learning and EHRs

Machine learning has also been combined with EHRs to predict health outcomes.

Using EHRs, Hayes *et al*. (2012) examined associations between mortality and other factors, such as activities of daily living (ADL) impairment, social relationships, living conditions and occupational/recreational activities, for people with serious mental illness. Using a Cox regression model, they found that impairment of ADL was associated with increased mortality.

Another study by Hayes *et al*. (2012) looked at how symptoms were associated with mortality for people with serious mental illness. The data they used from the EHRs included Health of the Nation Outcome Scales (HoNOS), demographic information and area-level deprivation data. This study also used a Cox regression model and found that mortality was associated with physical illness and disability among patients with serious mental illness.

2.4.4 Sequential Modelling and Healthcare

One drawback of the research described above is that it did not consider the temporal nature of healthcare data. However, including the time element can be important. Zhao *et al.* (2019) found that performance improved on several models when incorporating temporal features. The following studies also took this into account.

Hidden Markov Models (HMMs) were used in some examples to model the sequential data found in a healthcare setting. Kawamoto *et al.* (2013) used HMMs in combination with annual health checks to categorise patients into different states of healthy and unhealthy. A study in 2016 by Ayabakan *et al.* used HMMs to model the impact of the hidden status of a patient and its impact on hospital readmission for people with congestive heart failure. Verma *et al*. (2018) conducted a study on patients with chronic obstructive pulmonary disease (COPD). They used continuous time HMMs to model disease progression.

Deep learning algorithms, like recurrent neural networks (RNNs) have also been used within healthcare to model sequential data. According to Shickel, *et al*. (2018), ‘recurrent neural networks (RNNs) are an appropriate choice when data is sequentially ordered (such as time series data or natural language).’

Che *et al*. (2017) used long short-term memory (LSTM), a type of RNN, and deep feedforward networks to predict if a patient was using opioids. Zhao *et al*. (2019) used LSTM and convolutional neural networks (CNN) to predict cardiovascular events. In another study, Pham *et al*. (2016) used EHRs and LSTM to predict health trajectories. They assessed their model on a diabetes cohort and a mental health cohort.

Min *et al*. (2019) used gated recurrent units (GRU), another type of RNN, and LSTM to predict 30-day hospital readmission for patients with COPD. Cao *et al.* (2017) used GRU to detect mood using keystrokes measures from smartphone users, and Mahmoud *et al.* (2013) used RNNs to model and predict the behaviour of dementia patients from sensors in their homes. Also mentioned is a study using data from smartphones in which Suhara *et al*. (2017, as cited in Durstewitz *et al*., 2019) ‘forecast severe depressive states based on individual histories of mood, behavioural logs, and sleep information using a LSTM architecture.’

2.6 Gaps in the Research

Several of the above studies used surveys or clinical trial data and not routine EHRs. Other studies above considered machine learning approaches on EHRs, but not specifically on depression or depression relapse.

The following research used machine learning techniques on EHRs for depression, but the studies differ or contain gaps when compared to this project.

Lin *et al*. (2016) evaluated the following machine learning algorithms on EHR data to predict depression trajectory: k-means clustering, collaborative modelling, individual growth model (IGM), mixed effect model (MEM) and similarity-based CM (SCM). They mainly considered scores from the patient health questionnaire (PHQ-9), a tool used to measure the severity of depression. They were able to identify five main trajectory groups and predict the PHQ-9 scores. They discuss that the main limitation of their study is a lack of demographic and socioeconomic data, as well as additional clinical data.

Lin *et al.* (2018) performed a later study looking at a rule-based strategy to monitor depression from PHQ-9 data in EHRs. They used RuleFit, which is a pruned random forest, to predict depression severity in the next six months. Again, one of the limitations of this study was limited information on socioeconomic and clinical data. They also were not considering patients that were in remission before predicting the future severity.

Perlis *et al.* (2012) used natural language processing (NLP) of clinical notes to classify depressed patients into either symptomatic remission or treatment resistance groups. Their model performed much better than those only using billing diagnostic codes. However, a limitation of their study was a large variation in the quality of the notes, resulting in ambiguity of clinical states.

In another study looking at hospital readmission, Rumshisky *et al.* (2018) used NLP to extract words and identify topics from discharge summaries in EHRs. This was then combined with SVMs to predict psychiatric hospital readmission within 30 days for patients with major depressive disorder.

As far as I am aware, there has not been another study looking specifically at predicting depression relapse in a given timeframe following remission using machine learning and electronic health records.

This project builds on these studies to investigate the use of machine learning to predict depression relapse from anonymised EHRs. The types of data used in previous studies, such as symptoms, HoNOS, demographics and area-level deprivation, informed the data collection in this project. The sequential modelling algorithms in prior studies were also used to inform the prototypes and algorithm selection for this project. Lastly, the project considers different groups of patients and filtering the data as further modelling options.

2.7 Healthcare Provider and Data

The data captured in EHRs is sensitive, personal and protected, and it can therefore be difficult to access for research purposes. However, the South London and Maudsley (SLaM) NHS Foundation Trust, a large mental healthcare provider, has developed the Clinical Records Interactive Search (CRIS) system for mental health research (Perera *et al*., 2016). CRIS contains data from over 250,000 anonymised patient records from the trust's clinical systems capturing routine mental healthcare. A lot of valuable information is recorded in clinical notes on EHRs. To take advantage of this, natural language processing (NLP) was used by Perera *et al.* (2016) to derive and generate records of symptoms, as well as other types of data, from these notes in SLaM. This structured format of symptoms can then be combined with other patient data for analysis.

SLaM data accessed via CRIS was used for this project, including the symptoms from the NLP applications. Figure 2.3 shows a diagram of the CRIS technical architecture. The data was accessed directly from the CRIS SQL database from within the SLaM firewall onsite at the SLaM Biomedical Research Centre (BRC) in the Maudsley Hospital in London.

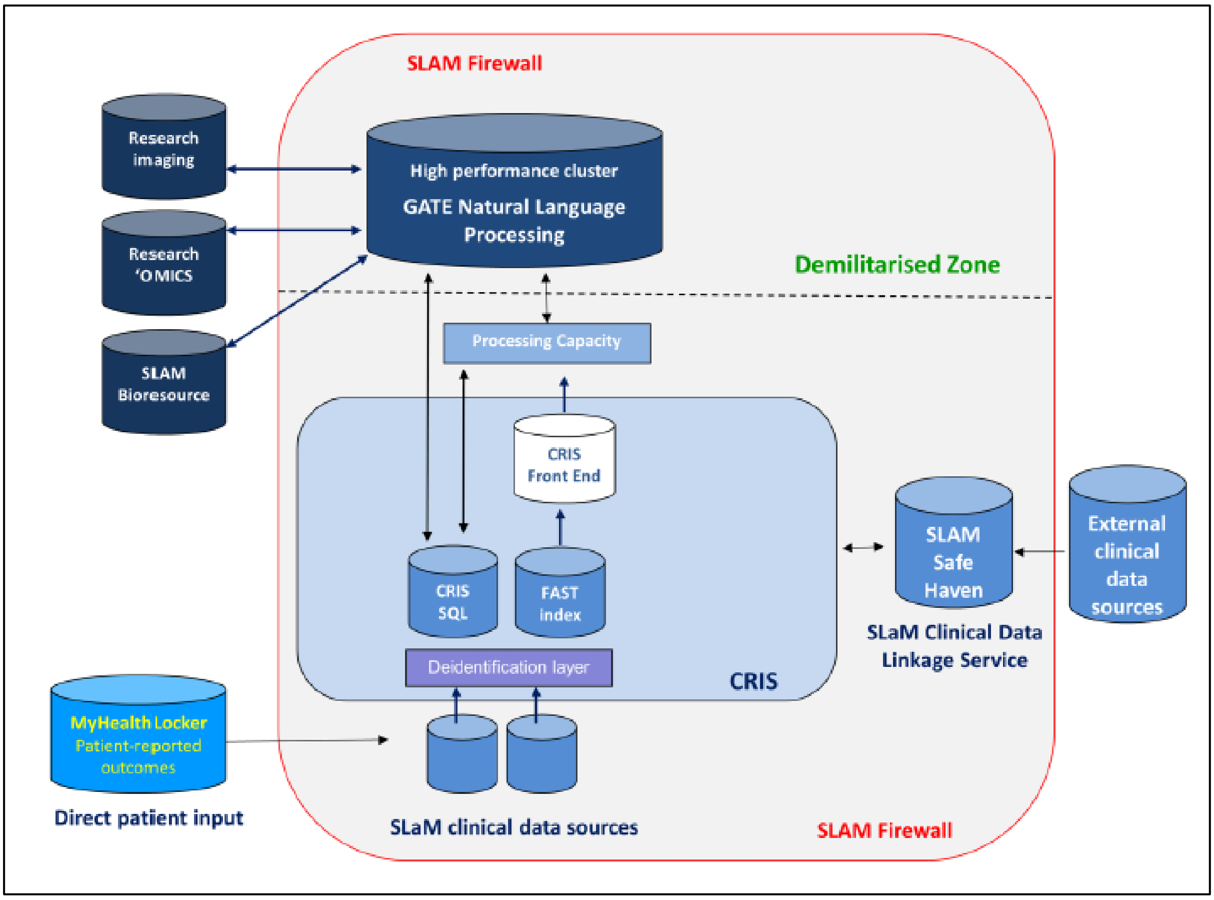
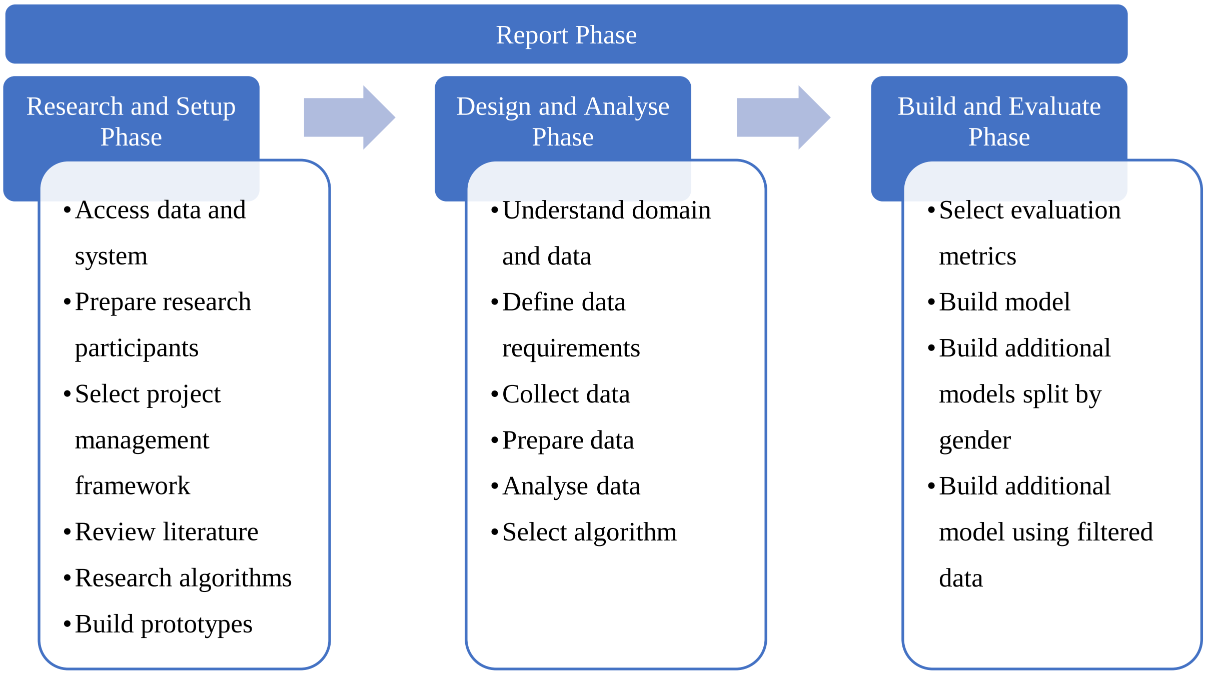


Figure 2.3. CRIS Technical Architecture (Perera *et al*., 2016).

SLaM is considered secondary care and patient admission is usually by referral from a general practitioner (GP). Once a patient has completed treatment with SLaM, they are discharged, usually back to their GP. However, SLaM itself does not hold any GP records on its patients.

Chapter 3 – Methods

This chapter presents the methods used during the project. Methods were split into four phases: research and setup, design and analyse, build and evaluate, and the report. The report phase involved documenting everything from the other three phases and was executed in parallel with these phases. Also, some design and analyse activities were done in parallel with the research and setup activities. Figure 3.1 shows each phase and the methods included in them.



Report Phase

Research and Setup Phase

- Access data and system
- Prepare research participants
- Select project management framework
- Review literature
- Research algorithms
- Build prototypes

Design and Analyse Phase

- Understand domain and data
- Define data requirements
- Collect data
- Prepare data
- Analyse data
- Select algorithm

Build and Evaluate Phase

- Select evaluation metrics
- Build model
- Build additional models split by gender
- Build additional model using filtered data

Figure 3.1. Project phases and methods.

The next sections explain each phase and the methods used, apart from the literature review which was presented in Chapter 2.

3.1 Research and Setup Phase

This phase of the project included getting access to the systems and data, preparing documents for research participants, choosing a method for managing the project, researching algorithms and building prototypes.

3.1.1 Access Data and System

CRIS data access was strictly limited to the SLaM network. Access to the SLaM network was available in person at the SLaM Biomedical Research Centre (BRC) or remotely via the virtual private network (VPN). Access to the network was obtained on the first day of the project, followed by access to the VPN. Access to the CRIS SQL Server database was also provided during the first week.

Software was either provided by SLaM or was installed. R was accessed via a virtual desktop (VDI) from the hot desks. I installed Anaconda directly onto the hot desks in order to use Python. Microsoft SQL Server Management Studio was also installed, with assistance from the CRIS administrator, to access the CRIS SQL Server database directly.

3.1.2 Prepare Research Participants

An advantage of working with the CRIS system and data is the ability to work with the team that created and supports it. A subject matter expert (SME) and clinician were available to assist with questions on the data and the domain. In order to conduct qualitative research with these individuals in the form of semi-structured interviews, a participant information sheet and informed consent form were prepared and sent to both participants. These forms can be found in Appendix B.

3.1.3 Select Project Management Framework

The Scrum Guide (Scrum Guides, 2017) defines Scrum as ‘A framework within which people can address complex adaptive problems, while productively and creatively delivering products of the highest possible value.’ Data science projects are often complex and come with uncertainty, so Scrum was chosen as the framework to manage this project. Scrum follows an iterative approach, where work is continuously inspected and adapted as needed. The figure below presents an overview of the events and artefacts in a Scrum project.

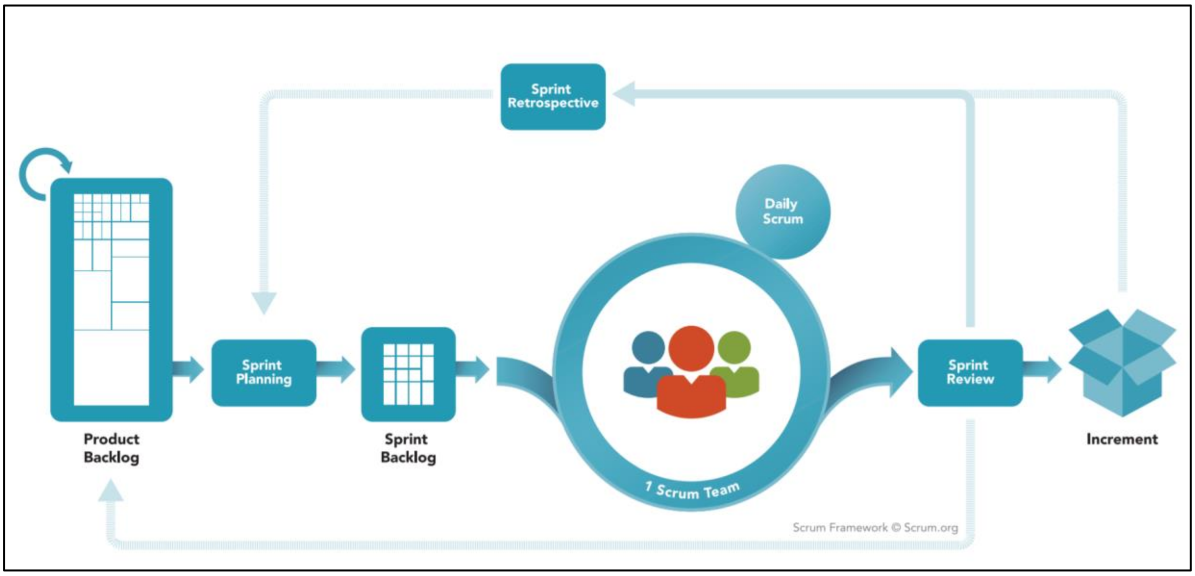


Figure 3.2 The Scrum framework (Scrum.org, 2019)

A Scrum project is split into time increments called sprints in order to manage the work. The work is broken down into product backlog items, or user stories. These are then put in order of priority, so that higher value work is delivered sooner. Items are added from the backlog to a sprint during the sprint planning meeting on the first day of the sprint. Each day, a daily scrum meeting is held, in which the team discusses progress from the previous day, plans for the next day, and any impediments they are facing. At the end of the sprint, a sprint review is held to demonstrate the work achieved to the stakeholders. Lastly, a sprint retrospective is held where the team discusses what went well and what could have been improved during the sprint. Throughout the project, the backlog is refined. This includes adding details to user stories, estimating the size of the stories and prioritising the backlog. The roles on a scrum team include the product owner, who is responsible for managing the product backlog, the development team, who builds the product, and the Scrum master, who helps the team follow the Scrum framework (Scrum Guides, 2017).

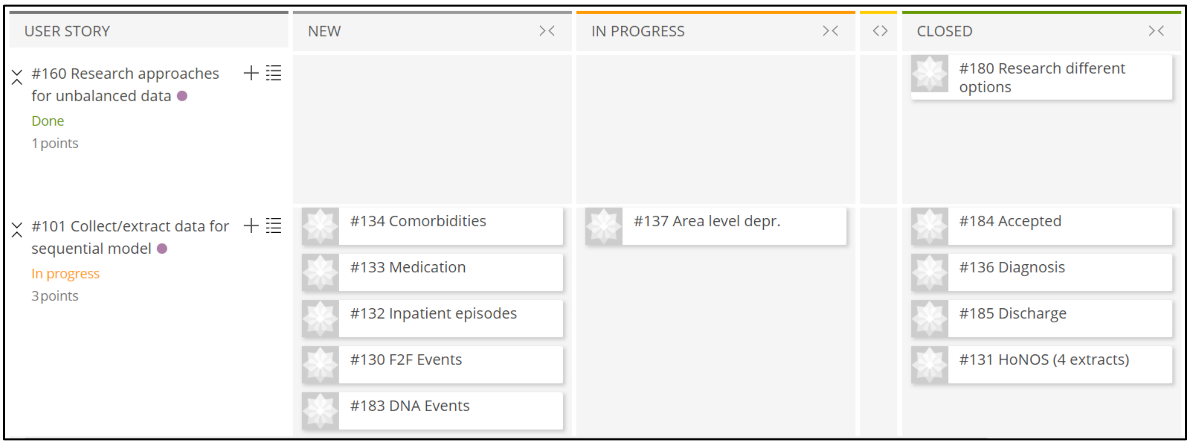
This project was set up with two-week sprints for a total of seven sprints. As this was an individual project, I played all roles on the team, and events were run as individual sessions. The sprint review was not used, as there were no stakeholders.

Table 3.1 shows the events held for a two-week sprint:

Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**
Day 1 Sprint Planning	**Day 2**	**Day 3** Backlog Refinement	**Day 4**	**Day 5**
Day 6	**Day 7**	**Day 8** Backlog Refinement	**Day 9**	**Day 10** Sprint Retrospective

Table 3.1. Project sprint schedule

The project artefacts were managed using a free Software-as-a-Service (SaaS) product called Taiga (Taiga, 2019). This allowed for access from multiple locations. Figure 3.3 shows part of the sprint task board for sprint 4.



USER STORY	NEW	IN PROGRESS	CLOSED
#160 Research approaches for unbalanced data Done 1 points			#180 Research different options
#101 Collect/extract data for sequential model In progress 3points	#134 Comorbidities #133 Medication #132 Inpatient episodes #130 F2F Events #183 DNA Events	#137 Area level depr.	#184 Accepted #136 Diagnosis #185 Discharge #131 HoNOS (4 extracts)

Figure 3.3. Sprint task board (Taiga, 2019)

3.1.4 Research Algorithms

Some objectives of the literature review, discussed in Chapter 2, were to understand how models were used in healthcare, which algorithms were used and the outcomes of the relevant research.

Additionally, it was of interest to understand how the data was prepared to inform my data extract preparation.

Sequential algorithms, such as RNNs and HMMs, were used in several studies to model temporal data. However, some studies also used non-sequential algorithms to model this data by re-shaping the data to accommodate the algorithm. Unfortunately, there was not much detail in the HMM research on data preparation or structure. Therefore, the rest of this section discusses details from the papers using recurrent neural networks.

Studies using RNNs used various methods of structuring temporal data. One technique used to reshape categorical data is one-hot encoding. Using this technique, each categorical value is represented as a column and the columns are populated with a 1 if it is true or a 0 if it is not true for each observation. Figure 3.4 shows a simple example of using one-hot encoding for discharge destination.

Patient	Discharge Destination
100001	GP
100002	Home

⇨

Patient	GP	Home
100001	1	0
100002	0	1

Figure 3.4. Reshaping to one-hot encoded format

Cao *et al.* (2017) considered a sequence of keystrokes within one smartphone session, which usually lasted less than a minute. They used one-hot encoding to represent keystrokes that weren't alpha-numeric characters, e.g. backspace, space, etc. Che *et al.* (2017) created features using one-hot encoding. For the recurrent neural networks, they aggregated the temporal features into one-year windows. For the other models, they aggregated the temporal features for all time. Zhao *et al.* (2019) did something similar, in which they aggregated features across a seven-year observation window and summarised data into one-year windows. For the deep learning models, they used the one-year features, and for the other models, they used both sets of data. Min *et al*. (2019) created medical events per patient based on timestamps. They built deep learning models for different options, such as the sequence of events only, regular time intervals and irregular time intervals. Interestingly, the sequence data alone had the worst performance among the deep learning models, whereas the best performing model used regular monthly intervals.

The activities performed in this section and in the literature discussed in Chapter 2 fulfilled objective O1. The final algorithm selection is discussed in section 3.2.6 taking into account the above research, the prototype results and the other activities performed in the design and analyse phase.

3.1.5 Build Prototypes

Following the literature review and algorithm research, recurrent neural networks were identified as an option for the sequential algorithm. However, there was a risk that the computational power of the

SLaM hot desks would be insufficient to run a deep learning model. In order to test this, two prototypes were built and tested. Sample data was generated, as actual CRIS data was not prepared at this stage.

A prototype LSTM model was created in Python using Keras and Tensorflow libraries. The model was adapted from a blog post by Aishwarya Singh (2019). The sample data contained 89 features and 152 patients with 8 events each.

A prototype model was also created in R using the RNN package. The model was adapted from an R-bloggers post by Mic (2016). The sample data contained 90 features and 15,000 patients with 10 events each.

3.2 Define and Analyse Phase

This phase of the project included understanding the domain and the data, defining the data requirements, collecting the data, preparing the data for analysis, performing exploratory data analysis and selecting an algorithm.

3.2.1 Understand Domain and Data

This phase began with gaining an understanding of the domain and the data. Semi-structured interviews were conducted with the CRIS Subject Matter Expert (SME) and a clinician. The objectives of the interviews were to:

- Define relapse and remission
- Understand the data available and data quality of the attributes used to define relapse and remission
- Define patients with depression
- Confirm the data entities to include in the analysis
- Understand the best data source for these entities
- Understand different dates available and how they were used
- Get answers to other questions about the data
- Gain insight on limiting data to build additional models

The full list of questions and notes from the interviews can be found in Appendix C. The SME provided a data dictionary prior to the project start, and this was used along with the literature research to form an initial list of data entities for analysis. Alongside the interviews, queries of the data using the CRIS SQL database were conducted to review and understand the data.

In addition to the interviews and the literature review, a few other resources were used to gain an understanding of the domain. The ICD-10 website (World Health Organization, 2016) was used to understand the diagnostic codes and details about depression and other illnesses that may be relevant to depression relapse. A paper by Wing *et al.* (1998) was used to better understand the Health of the Nation Outcome Scales (HoNOS). The SME also provided a list of symptoms available as part of the CRIS natural language processing (NLP) applications. This was used in combination with the ICD-10 website to identify depression symptoms that could be captured for patients in CRIS. The Mind website (Mind, 2019) provided information about anti-depressants available in the UK, including a map of the brand names to the generic names.

3.2.2 Define Data Requirements

During interviews with the clinician, the terms remission and relapse were defined. For the purposes of this project, a patient was defined to be in remission once they were discharged from SLaM. A patient was defined to relapse if, following remission, they returned to SLaM, or were 're-referred'. Figure 3.5 shows a basic process flow of the patient including the remission and relapse periods.

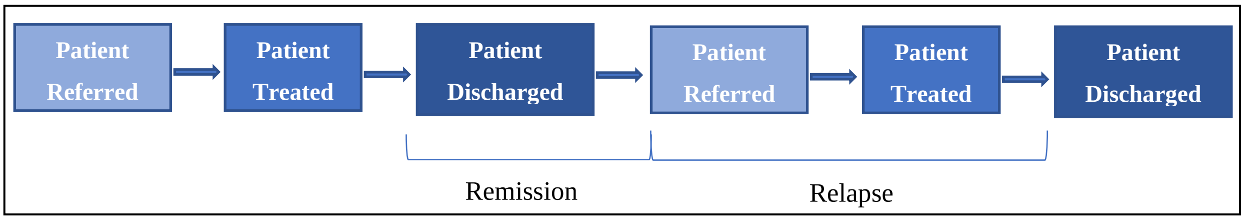


Figure 3.5 Process flow of patient

The tables below show excerpts of the data structure from CRIS, showing synthetic data. The example shows relevant columns for referral, discharge, diagnosis, events and medication for the same patient with BrcId XXXXXXX1.

BrcId	Referral_date	Accepted_date	Discharge_date	Spell_number	Discharge_Destination _ID
XXXXXXX1	21/04/2012	29/04/2012	12/07/2012	1	GP
XXXXXXX1	03/02/2013	15/02/2013	01/08/2013	2	GP

Table 3.2. CRIS Referral table excerpt with synthetic data showing two depressive episodes for same patient.

BrcId	Diagnosis_date	Primary_Diag
XXXXXXX1	05/05/2012	F32.1 – Moderate depressive episode

Table 3.3. CRIS Diagnosis table excerpt with synthetic data.

BrcId	Medication_Start_Date	Gazetteer (*Medication Name*)
XXXXXXX1	09/05/2012	Citalopram

Table 3.4. CRIS Medication tables excerpt with synthetic data.

BrcId	Start_date	Event_Type_Of_Contact_ID	Event_Outcome_ID
XXXXXXX1	07/05/2012	Face To Face	Attended
XXXXXXX1	15/05/2012	Face-to-face	Did not attend
XXXXXXX1	01/06/2012	Face-to-face	Attended

Table 3.5. CRIS Event table excerpt with synthetic data.

The Discharge Destination picklist values shown in Figure 3.6 were extracted and reviewed with the clinician. This list was used to further define remission to include only patients discharged to their GP or ‘Home – No Follow Up Required’.

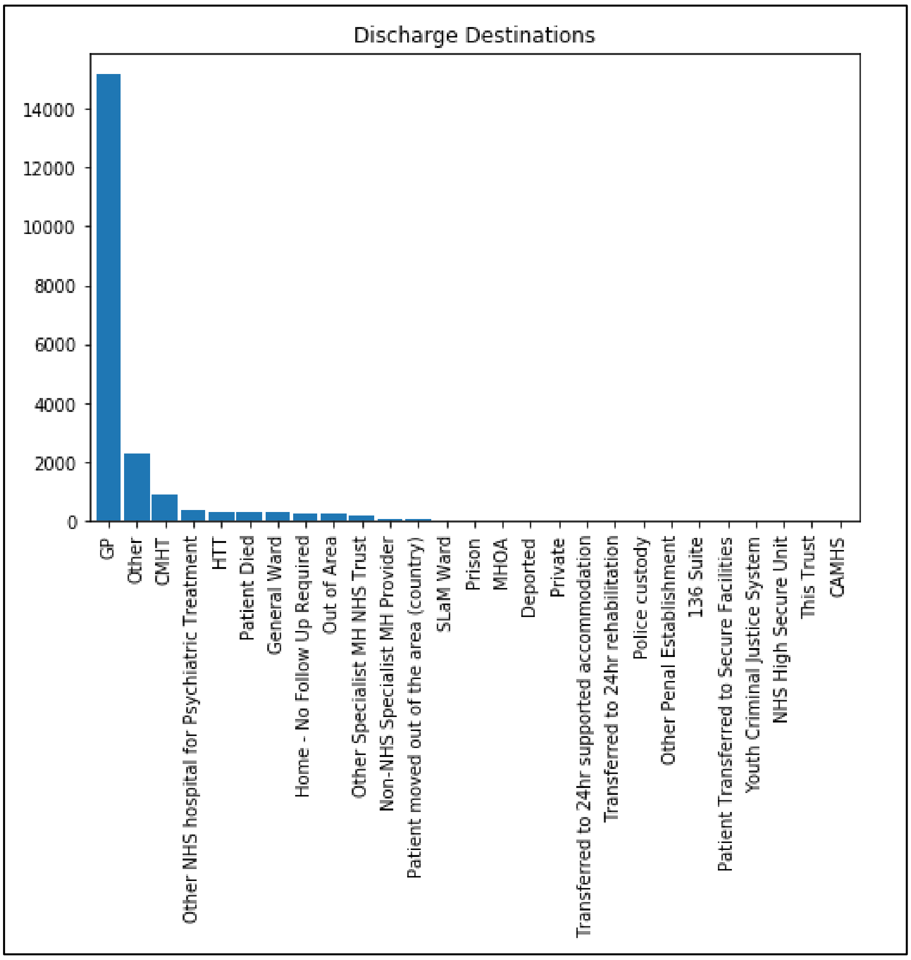


Figure 3.6 Discharge destinations

Some decisions required consultation with the CRIS team:

1) Which date attribute should be used to define the start of a patient episode?
2) Can the overall referral be used to define an episode or do ward stays and inpatient stays need to be considered?
3) Should all patient records be used or only those after a certain date?

The first and second questions were raised with the SME and a CRIS Informatician. The overall referral was deemed appropriate, as it was the entire episode being considered for analysis and not transitions between wards or inpatient stays. Some patients are referred and not accepted. However, they must be accepted before they can receive a diagnosis by SLaM. In addition, there was sometimes a time gap between acceptance and diagnosis. This gap could be interesting to analyse, and it could mean there are other events occurring between acceptance and diagnosis. Considering these points, the accepted date was chosen as the start of an episode.

The third question was also raised with the SME. Generally, 2007 was used as a starting point for CRIS analysis. Data added before that date was migrated to CRIS and is therefore not as reliable. An important point to consider is that initial diagnoses may not actually be the patient's first episode of depression. That may have occurred outside of SLaM or the initial episode may not have been migrated. One option suggested by the SME was to use a later year as a starting point to aim to capture all SLaM-entered first episodes of depression. My analysis showed that most patients (83%) relapsed within three years. Therefore, if data had been missed during a pre-2007 migration, starting three years later should be sufficient to exclude most patients that were relapsing and not having an initial episode of depression. Based on this feedback, 1st January 2010 was used as the starting point to extract all data.

The initial patient group for analysis and modelling was then defined as:

- adult patients with a diagnosis of depression (ICD-10 codes F32 or F33)
- referral to SLaM must have been accepted
- first depressive episode in SLaM started 1st January 2010 or later
- depression diagnosis must have occurred during their episode at SLaM
- discharged to either GP or Home

This initial group of patients would then be classed into 'relapse' and 'non-relapse' groups based on who returned to the trust following their initial depressive episode within the timeframe to be defined.

Picklist values were extracted from the database to analyse the categorical data. Some of these were presented to the clinician and used to define features. For instance, Face-to-face events were defined as those with the event type of 'Face To Face' or 'Face-to-face' and an event outcome of 'Attended', 'Attended on time/before HCP ready', 'Arrived late but was seen' or 'Arrived late after HCP available but seen'. A similar exercise was performed for 'Did Not Attend' (DNA) events.

The data requirements were defined based on the need for further analysis and to be used in the models. Therefore, two sets of requirements were defined: one for non-sequential extracts to be used for analysis and one for sequential extracts to be used for event and timeframe analysis and modelling.

As patients can have multiple HoNOS scores from different dates, the HoNOS closest to the discharge date was used when preparing data for further analysis. This was not done when including HoNOS scores in the sequential extracts, as multiple HoNOS scores could be included as part of the sequence. Of the six HoNOS scales, only four were used. HoNOSCA was excluded, as it is used for children and adolescents. HoNOS-LD was excluded, as it uses different questions and there were only 13 patients with this data populated.

Area level deprivation is reflected in the Index of Multiple Deprivation (IMD) scores. In CRIS, data from the census can be used to link the geographic level from the IMD to the patient address, and then a score can be assigned to a patient. The most recent years were used for both the IMD scores and the census data, 2015 and 2011, respectively. As patients can have multiple addresses which could result in multiple IMD scores, an average IMD score was used where multiple addresses were active during the patient referral timeframe. This was not done when including IMD scores in the sequential extracts, as multiple IMD scores could be included as part of the sequence.

Specifications were produced to document the requirements defined for the CRIS data extracts. These are discussed in more detail in Chapter 4. The activities performed in this section fulfilled objective O2, except for the definition of the relapse timeframe.

3.2.3 Collect Data

I wrote several Structured Query Language (SQL) queries and extracted data directly from the CRIS SQL Server database for all but one extract. The symptoms extracted from clinical notes using NLP were not accessible by me. Therefore, I compiled a list of depressive symptoms and provided this to the SME who then built an extract. It contained patients and corresponding depression symptoms using one-hot encoding for each symptom.

First, extracts were produced based on feedback from the clinician with static or non-temporal analysis in mind. Next, additional sequential extracts were produced in order to perform analysis on the temporal data and prepare the data for a sequential model.

3.2.4 Prepare Data

Several steps were performed to prepare the data for exploratory analysis in order to fulfil objective O3. The sequential data was prepared separately from the non-sequential data in order to perform different analyses.

First, the .csv files were loaded into Python and joined together. Medication columns with only 0s were removed, and some categorical features with missing values were imputed with the word ‘Missing’. Numerical features with missing values were imputed with -1 for the sequential data. This was to distinguish missing values from values set to 0.

Some categorical fields, like Marital Status and Gender, had similar values, so these were combined where it made sense.

Within the non-sequential data, multiple HoNOS scores were addressed by keeping the HoNOS with the most recent date. Some patients had a HoNOS record for multiple HoNOS types, e.g. HoNOS and HoNOS65+, which is how the multiple scores were produced. Missing date fields were imputed with 01-jan-1900. Numerical features for non-sequential data were imputed with 0, as most of these fields represented counts.

Some data like the HoNOS Totals and Age were on larger scales, which can affect the results of neural networks. Therefore, the continuous data was normalised so that all data was between 0 and 1.

3.2.5 Analyse Data

The goal of the Exploratory Data Analysis (EDA) was to understand the distributions of the data and look for correlations between features and the target. This task fulfilled objective O4 and the remaining part of objective O2, defining the relapse timeframe.

The following questions were posed as part of EDA:

- How often are people relapsing / being re-referred?
- How long do relapse and remission last?
- What do the other timeframes look like?
- How is the categorical data distributed?
- How does categorical data vary between relapse and non-relapse groups?
- How is the numerical data distributed?
- How does numerical data vary between relapse and non-relapse groups?
- How are features correlated with each other and with the target?
- What do the event types and event counts look like?
- How do the event types and event counts vary for relapse and non-relapse groups?
- What do events look like for different time frequencies?
- What are the minimum and maximum gaps between events?

3.2.6 Select Algorithm

Hidden Markov Models (HMMs) and LSTMs were considered as potential algorithms for modelling the sequential event data.

The literature on HMMs did not provide much detail on how data was prepared for the models, whereas several papers using LSTMs described their data preparation. Some R packages were investigated for building HMMs, but none of them were appropriate for multi-variate sequence modelling. However, the Keras library in Python is capable of handling this. There is also evidence that probabilistic models do not perform as well as RNNs in sequence classification of clinical data, as found in a study done by Hasan *et al.* (2017). In addition, my interest in learning and gaining experience in neural networks and deep learning was considered.

After building and testing a successful prototype in Python and considering the points above, LSTM was chosen as the sequential algorithm.

LSTM was proposed by Hochreiter and Schmidhuber (1997) as a method for solving the vanishing gradient problem found in RNNs and feedforward networks. The LSTM cell, according to Chollet (2018, p. 204), is designed to 'allow past information to be reinjected at a later time, thus fighting the vanishing gradient problem.' An LSTM unit consists of a memory cell with different gate units. An input gate unit is used to filter out irrelevant inputs and an output gate unit is used to prevent irrelevant memory from affecting other units (Hochreiter and Schmidhuber, 1997). Later, a forget gate was added by Gers, *et al*. (2000) which lets the network reset itself to avoid the network breaking down when using continuous input streams.

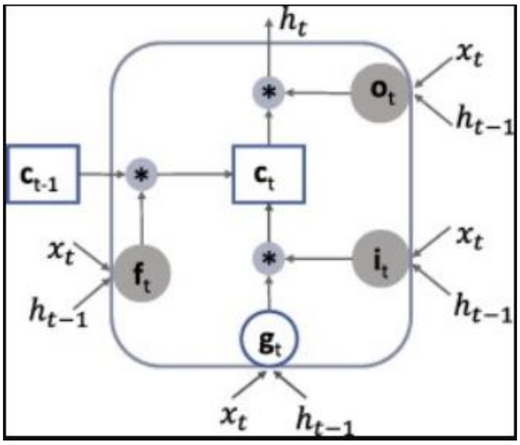


Figure 3.7. An LSTM unit (Pham et al., 2017).

Figure 3.7 shows an LSTM unit with inputs x at time t and h at time t-1. The input gate, output gate and forget gate at time t are represented by i_t, o_t and f_t, respectively. The memory cell is shown as c_t and the output is h_t (Pham *et al*., 2017). The network architecture used in the original LSTM paper was one input layer, one hidden layer, and one output layer with the hidden layer holding the memory cells and gate units.

3.3 Build and Evaluate Phase

3.3.1 Select Evaluation Metrics

The problem was defined as a binary sequence classification problem. After being trained on a historical sequence of patient events, the model would categorise each patient 'relapse' or 'non-relapse'. Given that the dataset was heavily unbalanced, i.e. only 12% relapsing in two years, classification accuracy alone was not a good metric to measure the success of the model. Therefore, additional binary classification metrics were used including area under the receiver operating characteristic curve (AUC), F1-score, precision, recall (sensitivity), and a confusion matrix. More details on these can be found in the glossary. The code for the metrics was provided by Jason Brownlee (2019) in 'How to Calculate Precision, Recall, F1, and More for Deep Learning Models' and by Scikit-learn API Reference (2019).

3.3.2 Build Model

The LSTM prototype code in Python was adapted so that it could be used with the full dataset. The binary cross entropy loss function and the Adam optimization function from the prototype were used. The loss function is used to calculate the loss or error during the training of the model, and the goal is to minimise this (Bishop, 2006, p. 41). Bishop (2006, p. 235, as cited in Simard *et al*., 2003) also states that cross entropy is a better loss function than sum-of-squares for binary classification problems, as it can improve generalisation and reduce training times. The optimization function, or optimizer, is used to minimise the loss during the training of the model. The Adam optimization function introduced by Kingma and Ba (2015) is computationally efficient, requires little memory and is suitable for large data sets. Accuracy was used as the metric to monitor during the model training and prediction.

The LSTM layer in Keras requires that all sequences are the same length. Therefore, event sequences were truncated and padded to the same number for all patients. Initially, 29 events was chosen as the sequence length, because this covered 80% or more of the patients in the relapse and non-relapse groups. However, this meant many patients would have extra data generated and add more sparsity to an already sparse matrix. Longer sequence lengths would also cause the models to run longer. Early testing of the model with 29 events and 9 events (the median) produced similar results. Therefore, the median number of events was used in all models.

Chollet (2018, pp. 219-220) suggested reversing the direction of the sequences, i.e. reverse chronological order, to see if it has any impact on the model. This was not attempted, but it did lead me to another idea which was to pad the shorter sequences at the beginning of the sequence rather than at the end. This approach could also lead to better results as the more recent events would contain the valid data, while the older events would contain the generated data.

The dataset was highly imbalanced with only 12% of the patients in the relapse group. To address this imbalance, two techniques were investigated as suggested by Jason Brownlee (2015): 1) under-sampling the training data so that both classes had the same number of records and 2) applying a class weight to the model so that the minority class was given more importance. Both techniques produced similar initial results. Therefore, the class weight was used, as it allowed the use of the full dataset instead of removing data from the majority class.

Grid search is a technique used to tune a model by evaluating each combination of hyperparameters to see which set performs best. A limited grid search was built to identify the best model hyperparameters. The grid search code was adapted from a blog post by Jason Brownlee (2019) called 'How to Grid Search Hyperparameters for Deep Learning Models in Python With Keras'. The grid search was limited in that not all combinations were evaluated at once due to restrictions accessing the data for longer than one business day. Therefore, the grid search was broken into sections, e.g. the first two parameters were evaluated in one grid search and the results from that were used in the next grid search evaluating two other parameters. This allowed each grid search to be completed onsite during business hours.

Dropout is a technique used to avoid overfitting. Overfitting is when the model does not generalise well with unseen data. It involves removing connections between some neurons in the network. In the case of LSTM, dropout can be used in the input layer and in the recurrent layers (Chollet, 2018, pp. 217-218).

The table below shows the parameters evaluated during the grid search.

Hyperparameter	Definition
Number of epochs	An epoch is one iteration through the full training data. This sets the number of iterations.
Batch size	The batch size is the number of sequences processed before the network weights are updated.
Dropout rate	The percentage of units left out of the input layer of the network.
Recurrent dropout rate	The percentage of units left out in the recurrent layers of the network.
Number of hidden neurons	The number of units in the hidden layer of the neural network.

Table 3.6 Grid Search parameters (Chollet, no date) and values

The loss and accuracy were plotted for each epoch to monitor the training and validation performance and check for overfitting using code provided by Chollet (2018, pp. 74-75). Early stopping was used to prevent overfitting. Early stopping is a technique used to monitor and stop the training of the network once the validation loss starts to increase, indicating the model is overfitting. The early stopping code was provided by Chollet (2018, p. 250).

The steps performed in this section fulfilled objective O5, although specific results of the model are discussed further in Chapter 4.

3.3.3 Build Additional Models by Gender

The next stage of modelling included splitting the data into male and female patients, tuning new models for each gender, and training both models to compare the performance to the first model. Six patients were left out of these models as their gender was not identified as male or female.

As above, the sequences were truncated and padded to the same length using the median number of events for the corresponding data set. For both models, this was set to 9.

A separate grid search was used for each model to tune hyperparameters, and class weights were applied to the imbalanced data. The weights were calculated based on the amount of data found in each class for the corresponding data set. For the female model, the weights were set to 8.3 for the minority class and 1 for the majority class. For the male model, the weights were set to 7.94 for the minority class and 1 for the majority class. Early stopping was used to prevent overfitting.

The steps performed in this section fulfilled objective O6, although specific results of the models are discussed further in Chapter 4.

3.3.4 Build Additional Model using Filtered Data

The final stage of modelling included filtering the data, tuning a new model, and training it to compare performance to the first model. The data was filtered in this manner:

- All rows representing referral accepted events were removed, as all patients had these events. In addition, no additional measurement was included with this event type as compared to discharge, which everyone also had but could be categorised as either GP or Home.
- Patients with very large maximum gaps (defined as greater than 1.5 times the interquartile range above the 75th percentile) between subsequent events were treated as outliers and their sequences were removed.
- All events occurring on the same date for the same patient were consolidated into the same row, and the maximum value for each feature was kept. As timestamps were not recorded in the source data, this did not result in any sequential information being lost.
- HoNOS detailed questions were removed.

As above, the sequences were truncated and padded to the same length using the median number of events for the data set, or 6.

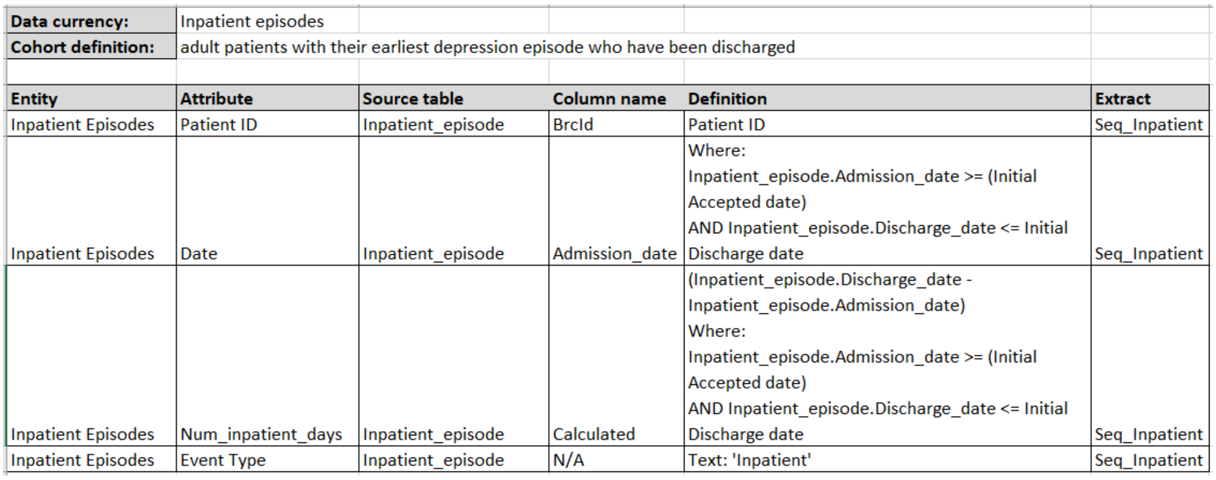
A grid search was used for this model to tune hyperparameters, and class weights were applied to the imbalanced data. The weights were calculated based on the amount of data found in each class for the data set. The weights were set to 8.64 for the minority class and 1 for the majority class. Early stopping was used to prevent overfitting.

The steps performed in this section fulfilled objective O7, although specific results of the model are discussed further in Chapter 4.

Chapter 4 – Results

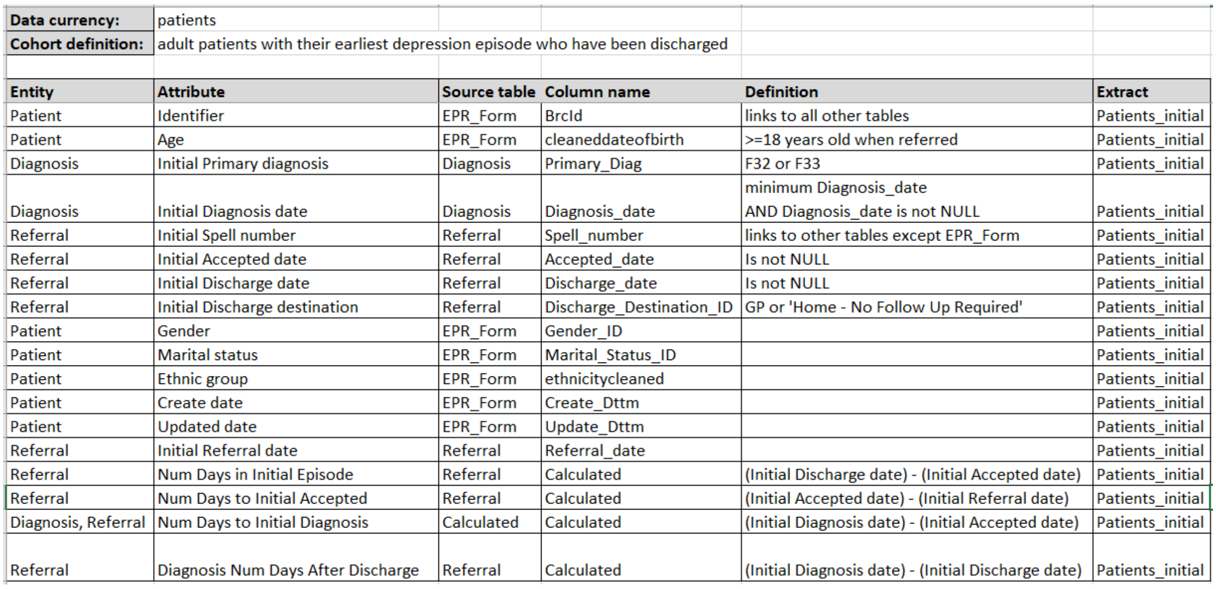
4.1 Data Specifications

Following the interviews with the SME and clinician, I produced data extract requirements specifications. These were in the format of spreadsheets and described each entity and attribute and the rules for extracting them from the CRIS SQL database. One specification was produced for sequential extracts and one was produced for non-sequential extracts. Below is an excerpt from each specification. The full specifications can be found in Appendix D.



Data currency:	Inpatient episodes				
Cohort definition:	adult patients with their earliest depression episode who have been discharged				
Entity	**Attribute**	**Source table**	**Column name**	**Definition**	**Extract**
Inpatient Episodes	Patient ID	Inpatient_episode	BrcId	Patient ID	Seq_Inpatient
Inpatient Episodes	Date	Inpatient_episode	Admission_date	Where: Inpatient_episode.Admission_date >= (Initial Accepted date) AND Inpatient_episode.Discharge_date <= Initial Discharge date	Seq_Inpatient
Inpatient Episodes	Num_inpatient_days	Inpatient_episode	Calculated	(Inpatient_episode.Discharge_date - Inpatient_episode.Admission_date) Where: Inpatient_episode.Admission_date >= (Initial Accepted date) AND Inpatient_episode.Discharge_date <= Initial Discharge date	Seq_Inpatient
Inpatient Episodes	Event Type	Inpatient_episode	N/A	Text: 'Inpatient'	Seq_Inpatient

Table 4.1. Inpatient episode details from the sequential extract specification.



Data currency:	patients				
Cohort definition:	adult patients with their earliest depression episode who have been discharged				
Entity	**Attribute**	**Source table**	**Column name**	**Definition**	**Extract**
Patient	Identifier	EPR_Form	BrcId	links to all other tables	Patients_initial
Patient	Age	EPR_Form	cleaneddateofbirth	>=18 years old when referred	Patients_initial
Diagnosis	Initial Primary diagnosis	Diagnosis	Primary_Diag	F32 or F33	Patients_initial
Diagnosis	Initial Diagnosis date	Diagnosis	Diagnosis_date	minimum Diagnosis_date AND Diagnosis_date is not NULL	Patients_initial
Referral	Initial Spell number	Referral	Spell_number	links to other tables except EPR_Form	Patients_initial
Referral	Initial Accepted date	Referral	Accepted_date	Is not NULL	Patients_initial
Referral	Initial Discharge date	Referral	Discharge_date	Is not NULL	Patients_initial
Referral	Initial Discharge destination	Referral	Discharge_Destination_ID	GP or 'Home - No Follow Up Required'	Patients_initial
Patient	Gender	EPR_Form	Gender_ID		Patients_initial
Patient	Marital status	EPR_Form	Marital_Status_ID		Patients_initial
Patient	Ethnic group	EPR_Form	ethnicitycleaned		Patients_initial
Patient	Create date	EPR_Form	Create_Dttm		Patients_initial
Patient	Updated date	EPR_Form	Update_Dttm		Patients_initial
Referral	Initial Referral date	Referral	Referral_date		Patients_initial
Referral	Num Days in Initial Episode	Referral	Calculated	(Initial Discharge date) - (Initial Accepted date)	Patients_initial
Referral	Num Days to Initial Accepted	Referral	Calculated	(Initial Accepted date) - (Initial Referral date)	Patients_initial
Diagnosis, Referral	Num Days to Initial Diagnosis	Calculated	Calculated	(Initial Diagnosis date) - (Initial Accepted date)	Patients_initial
Referral	Diagnosis Num Days After Discharge	Referral	Calculated	(Initial Diagnosis date) - (Initial Discharge date)	Patients_initial

Table 4.2. Initial patient group details from the non-sequential extract specification.

4.2 Data Extracts

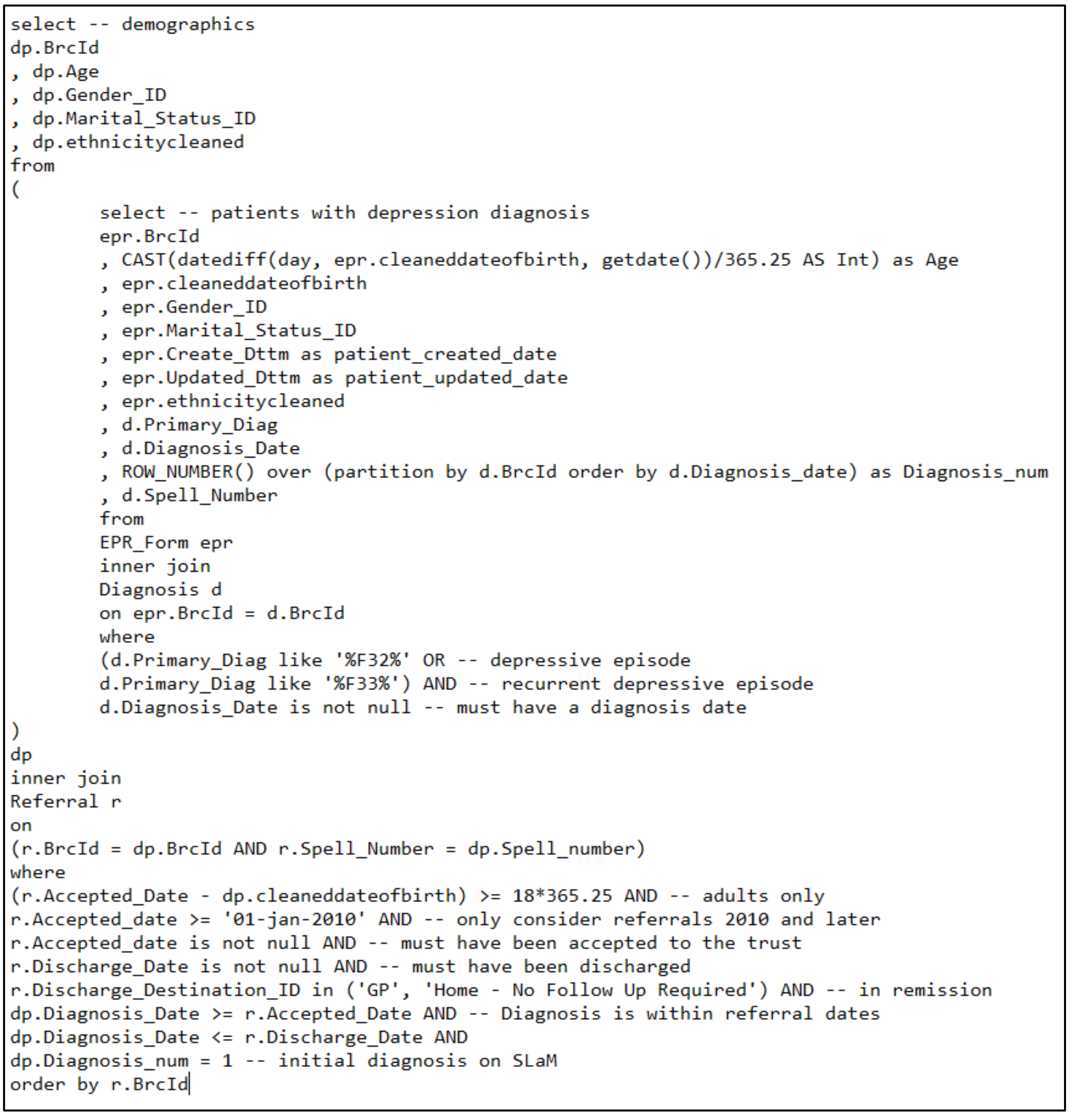
The specifications were used as the instructions to build the data extracts in SQL. The full list of extracts I built is shown below. The Appendix where each SQL query is located, i.e. E.24, is shown for each extract.

Extracts (Appendix)	**Description**
Area Level Deprivation (E.1)	Patients and average Index of Multiple Deprivation (IMD) scores
Demographics (E.2)	Patients and age, gender, ethnicity and marital status
Discharge Destinations (E.3)	List of unique discharge destinations and patient counts
Face to Face Contact (E.4)	Patients and number of face-to-face events in last 6 months
History of Bipolar (E.5)	Patients with any history of Bipolar affective disorder (ICD-10 code F31)
History of Manic (E.6)	Patients with any history of Manic episode (ICD-10 code F30)
History of Organic (E.7)	Patients with any history of Organic disorders (ICD-10 code F0)
History of Schizophrenia (E.8)	Patients with any history of Schizophrenia (ICD-10 code F2)
History of Substance (E.9)	Patients with any history of Mental or behavioural disorders due to substance use (ICD-10 code F1)
HoNOS (E.10)	Patients and most recent HoNOS scores
HoNOS abi (E.11)	Patients and most recent HoNOS-ABI scores
HoNOS Secure (E.12)	Patients and most recent HoNOS-secure scores
HoNOS 65+ (E.13)	Patients and most recent HoNOS65+ scores
Inpatient Days (E.14)	Patients and number of days as an inpatient in last 6 months
Medication (E.15)	Patients and anti-depressant medication
Medication (one-hot encoded) (E.16)	Patients and anti-depressant medication in one-hot encoded format
Number of DNAs last 6 months (E.17)	Patients and number of Did Not Attend (DNA) events in last 6 months
Patients initial episode (E.18)	Patients and initial referral, diagnosis and discharge details. **This is the main patient group used to define all other extracts.**
Patients relapse (E.19)	Patients and relapse referral, diagnosis and discharge details
Psychotic Symptoms (E.20)	Patients with psychotic symptoms in last 12 months
Rereferral count (E.21)	Number of re-referrals for each patient
Risk Assessment New (E.22)	Patients with any history of self-harm or suicide ideation on the new Risk Assessment form
Risk Assessment Old (E.23)	Patients with any history of self-harm or suicide ideation on the old Risk Assessment form

*Symptoms	Patients and depressive symptoms
Sequential Accepted events (E.24)	Patients and Referral Accepted events
Sequential Area Level Deprivation events (E.25)	Patients and Index of Multiple Deprivation (IMD) scores for each patient address change
Sequential Comorbidities events (E.26)	Patients and events of other diagnoses (Bipolar, Manic, Organic, Schizophrenia or Substance use)
Sequential Diagnosis events (E.27)	Patients and Diagnosis events
Sequential Discharge events (E.28)	Patients and Discharge events
Sequential DNAs events (E.29)	Patients and Did Not Attend events
Sequential Face to Face events (E.30)	Patients and Face-to-Face events
Sequential HoNOS events (E.31)	Patients and HoNOS events
Sequential HoNOS abi events (E.32)	Patients and HoNOS-ABI events
Sequential HoNOS Secure events (E.33)	Patients and HoNOS-secure events
Sequential HoNOS 65+ events (E.34)	Patients and HoNOS65+ events
Sequential Inpatient events (E.35)	Patients and Inpatient events
Sequential Medication events (E.36)	Patients and Medication start events

Table 4.3. Full list of data extracts produced. *Symptoms extract was produced by the CRIS SME.

Due to the restrictions with CRIS data, the data extracts cannot be published in this report. However, the SQL queries can be shown. An example query used to build the demographics extract is shown below in Figure 4.1. The full set of SQL query files is included in Appendix E.



```
select -- demographics
dp.BrcId
, dp.Age
, dp.Gender_ID
, dp.Marital_Status_ID
, dp.ethnicitycleaned
from
(
        select -- patients with depression diagnosis
        epr.BrcId
        , CAST(datediff(day, epr.cleaneddateofbirth, getdate())/365.25 AS Int) as Age
        , epr.cleaneddateofbirth
        , epr.Gender_ID
        , epr.Marital_Status_ID
        , epr.Create_Dttm as patient_created_date
        , epr.Updated_Dttm as patient_updated_date
        , epr.ethnicitycleaned
        , d.Primary_Diag
        , d.Diagnosis_Date
        , ROW_NUMBER() over (partition by d.BrcId order by d.Diagnosis_date) as Diagnosis_num
        , d.Spell_Number
        from
        EPR_Form epr
        inner join
        Diagnosis d
        on epr.BrcId = d.BrcId
        where
        (d.Primary_Diag like '%F32%' OR -- depressive episode
        d.Primary_Diag like '%F33%') AND -- recurrent depressive episode
        d.Diagnosis_Date is not null -- must have a diagnosis date
)
dp
inner join
Referral r
on
(r.BrcId = dp.BrcId AND r.Spell_Number = dp.Spell_number)
where
(r.Accepted_Date - dp.cleaneddateofbirth) >= 18*365.25 AND -- adults only
r.Accepted_date >= '01-jan-2010' AND -- only consider referrals 2010 and later
r.Accepted_date is not null AND -- must have been accepted to the trust
r.Discharge_Date is not null AND -- must have been discharged
r.Discharge_Destination_ID in ('GP', 'Home - No Follow Up Required') AND -- in remission
dp.Diagnosis_Date >= r.Accepted_Date AND -- Diagnosis is within referral dates
dp.Diagnosis_Date <= r.Discharge_Date AND
dp.Diagnosis_num = 1 -- initial diagnosis on SLaM
order by r.BrcId
```

Figure 4.1. SQL query for Demographics extract.

4.3 Prototype Results

The prototypes were run and timed to test for feasibility on the SLaM hot desks after having concerns that the computers may not be powerful enough to run deep learning models. Not knowing at this stage how many epochs would be needed to train the model, the prototypes were tested for different numbers of epochs.

The Python LSTM model with 152 patients, 20 epochs and 100 hidden neurons ran in 22 seconds. Increasing this to 200 epochs took 3 minutes and 22 seconds. For 15,000 patients and 20 epochs, this was estimated to run in approximately 36 minutes. Increasing this to 200 epochs was estimated to run in just over 6 hours.

The R RNN model with 15,000 patients, 20 epochs and 100 hidden neurons ran in 4 hours. Increasing this to 200 epochs was estimated to take 40 hours.

Table 4.4 below shows results and estimates for both prototypes.

Prototype (language - package)	**20 Epochs**	**200 Epochs**	**20 Epochs**	**200 Epochs**
	152 patients	**152 patients**	**15,000 patients**	**15,000 patients**
Python - LSTM	22 seconds	3 minutes 22 seconds	*36 minutes	*6 hours 5 minutes
R - RNN	*2 minutes	*24 minutes 19 seconds	4 hours	*40 hours

Table 4.4. Prototype run times. * denotes an estimate.

The RNN model in R was considered too time-consuming, and therefore this approach was abandoned. The longer run time may have been partly due to R running on a virtual machine, whereas Python was running directly from a hot desk.

4.4 Exploratory Data Analysis Results

The EDA produced answers to the questions posed in section 3.2.5. Each question is answered below, including some data visualisation to explain further. Full EDA results are included in the zip file submitted separately.

How often are people relapsing / being re-referred?

Generally, people are relapsing once, but some are relapsing between two or more times, as shown in Figure 4.2.

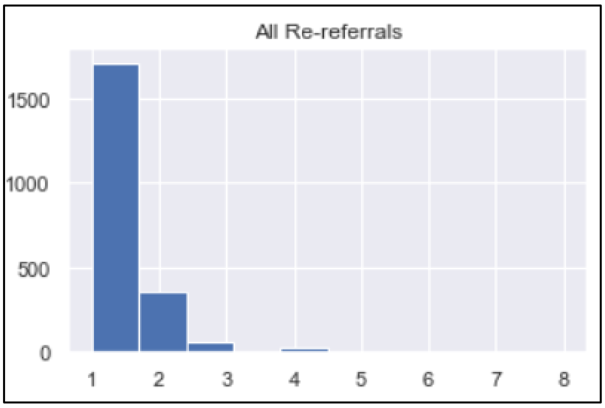


Figure 4.2 Relapse occurances

How long do relapse and remission last?

The median relapse period is around four months (130 days). The median remission period is around ten months (294 days).

State	Count	Mean	Std	Min	25%	50%	75%	Max
Remission*	1,174	547	632	1	83	294	791	3383
Relapse	1,174	251	347	0	34	130	315	2640

Table 4.5 Timeframes for patients accepted by January 2010 and discharged in or before July 2016. *Only includes patients who have relapsed. Patients who did not relapse are assumed to still be in remission.

Table 4.6 shows the percentage of patients from the relapse group and from the entire patient group for different time periods of remission.

Relapse timeframe	% of All relapse patients returning	% of All patients
1 year from discharge	55%	8.2%
2 years from discharge	**73%**	**10.8%**
3 years from discharge	83%	12.3%

Table 4.6 Relapse timeframes for patients accepted by January 2010 and discharged in or before July 2016.

Defining the Target

After assessing the relapse timeframes, a target timeframe of 2 years was chosen. This allowed more relapse patients to be included than in the 1-year relapse group, but it was assumed to be more clinically valuable, i.e. to be able to predict who would relapse sooner, than a 3 year relapse period.

The target values were created by combining the initial patient group and relapse group datasets. The relapse_in_24M field created in the relapse group dataset was used to identify those in the relapse group, with a 1 representing relapse and a 0 representing non-relapse.

All further analysis was performed with a target timeframe of 2 years. All numbers below include target values for a 2-year relapse period and patients were excluded from further analysis and modelling if they were discharged after July 2017 to allow for a 2-year prediction window. This produced a patient group of 11,971 patients with 1,467 (12.25%) in the relapse group and 10,504 (87.75%) in the non-relapse group.

What do the other timeframes look like?

The median initial episode period is about four months (118 days).

State	Count	Mean	Std	Min	25%	50%	75%	Max
Initial Episode	11,971	233	316	0	34	118	293	2640

Table 4.7 Initial episode timeframes for patients accepted from January 2010 to July 2017.

The distribution for the number of days in the initial episode is highly skewed. The same is shown with the number of days between having a referral accepted and being diagnosed.

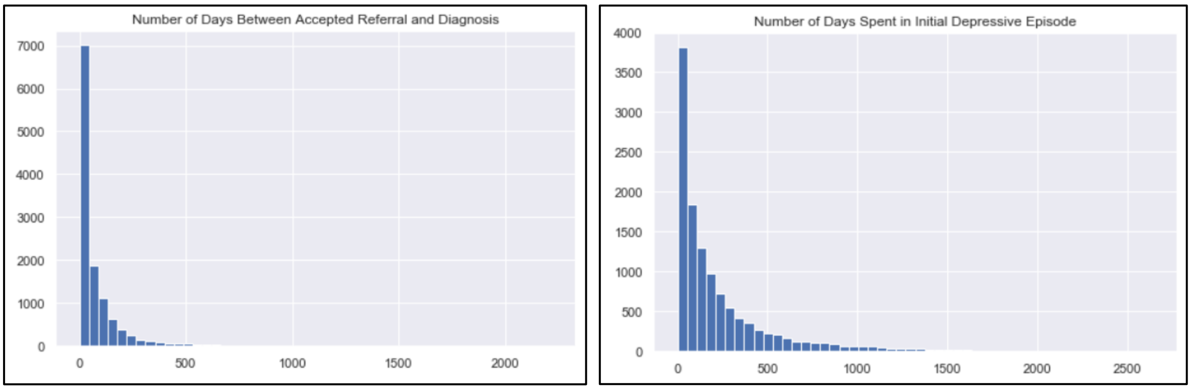


Figure 4.3 Days between Accepted and Diagnosis and Days in Initial Episode

How is the categorical data distributed?

The following results are for the entire patient group, i.e. those who relapse and those who do not.

Most diagnoses are F32.1, which is defined as moderate depressive episode. The next highest occurrence is F32.0, which is mild depressive episode. This was surprising, as the SME initially mentioned that most people with depression in CRIS were likely to be severe, treatment-resistant or have multiple conditions. However, the patients with lower severity diagnoses could still be treatment-resistant or have multiple conditions.

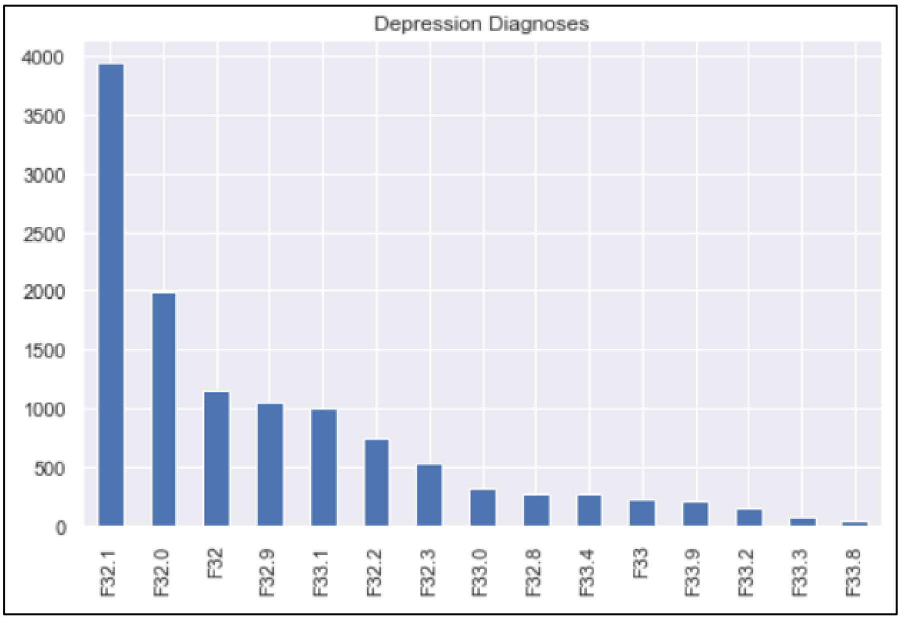


Figure 4.4 Depression diagnoses

The top three anti-depressants used are Mirtazapine, Sertraline, and Citalopram.

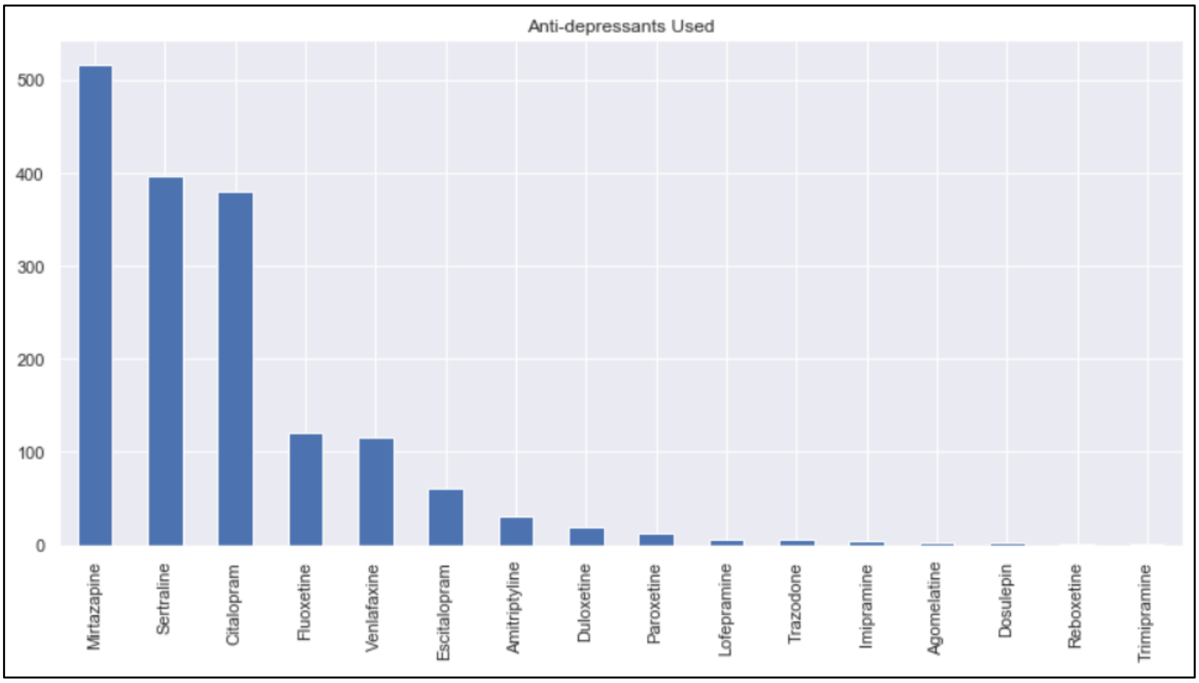


Figure 4.5 Anti-depressant use for depression

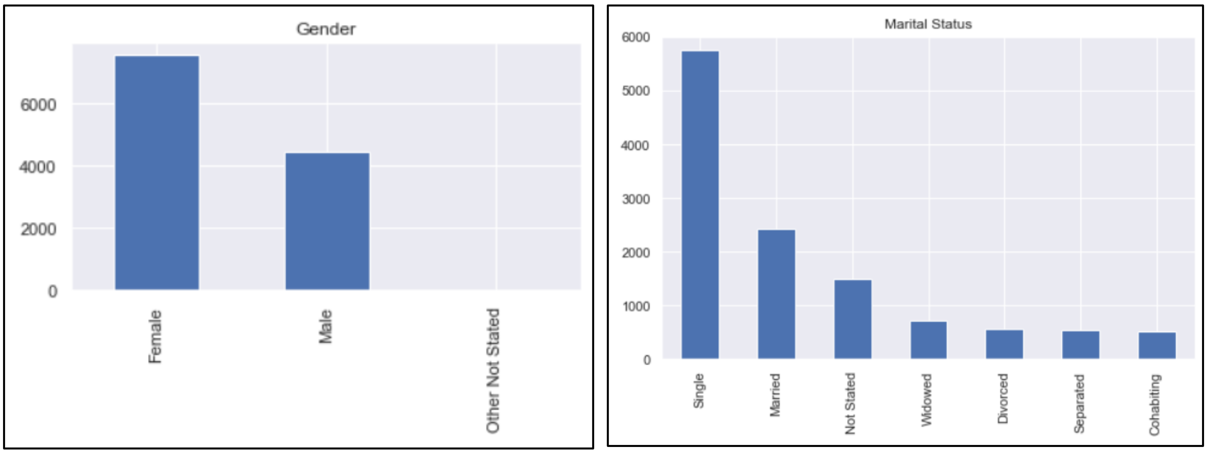


Figure 4.6 Depression by gender and marital status

Most depression patients at SLaM are female and single.

How does categorical data vary between relapse and non-relapse group?

The highest proportion of patients who relapsed within 2 years had a diagnosis of F32.3, which is severe depressive episode with psychotic symptoms. This is followed closely by F33, which is recurrent depressive disorder. The sub-category X.0 represents mild, X.1 represents moderate and X.2 represents severe without psychotic symptoms (World Health Organization, 2016). The proportion of patients in the relapse group is higher where depression is more severe, as circled in red in Figure 4.7. Unsurprisingly, F33 (recurrent depressive disorder) has a higher relapse proportion than F32 (depressive episode).

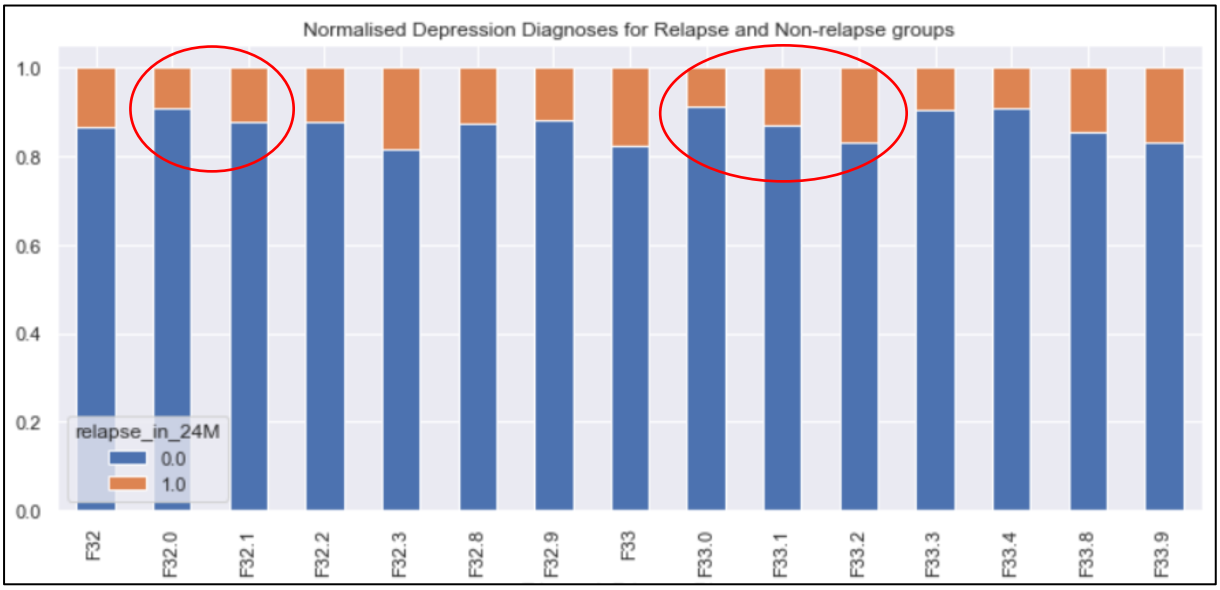


Figure 4.7 Normalised diagnoses. Increased severity is associated with an increase in relapse proportion.

Agomelatine and Trazadone have the highest proportion of relapse patients compared to other anti-depressants. However, there are very few patients with this medication recorded. Similarly, Dosulepin, Reboxetine and Trimipramine have no relapse patients, but these also have very few users.

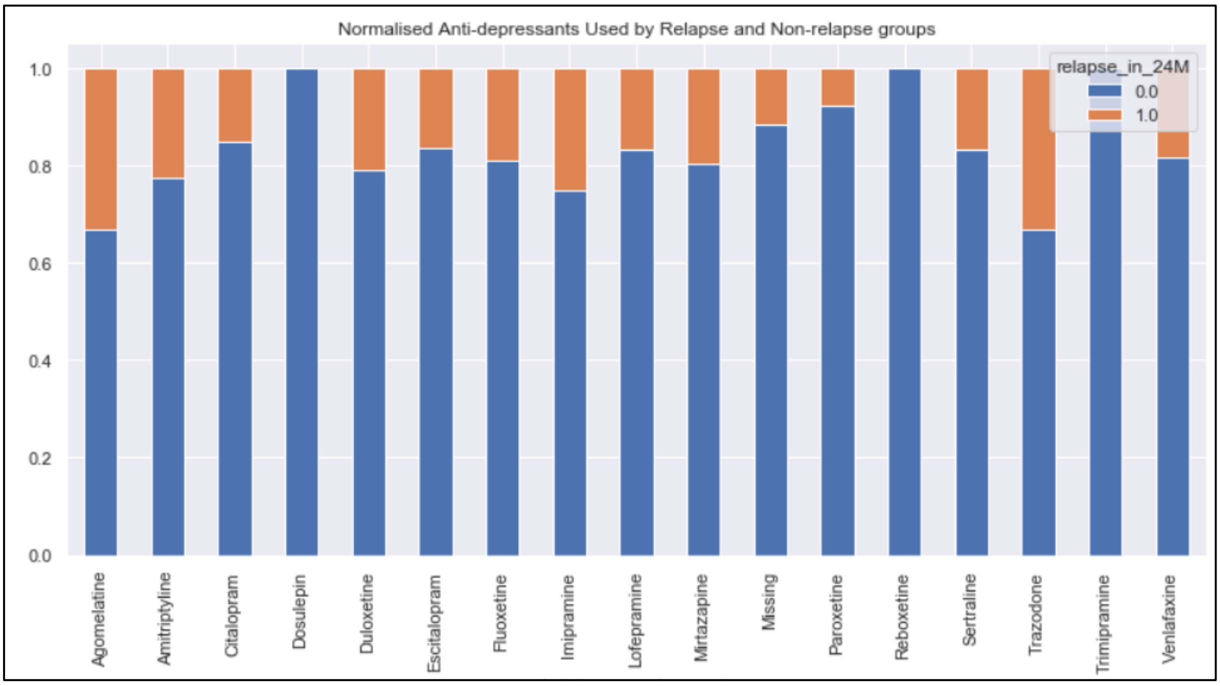


Figure 4.8 Normalised anti-depressant use

Regarding gender, although there are far more females at SLaM with depression, the proportion of males and females relapsing in two years is about the same.

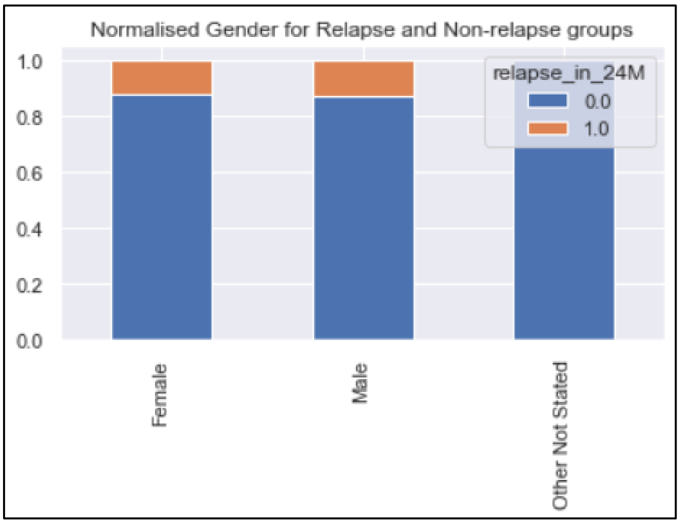


Figure 4.9 Normalised depression by gender

Divorced patients have the highest proportion of relapses, closely followed by separated, married, widowed, and single patients. Cohabitating patients have the smallest proportion of relapses of those whose marital status was stated.

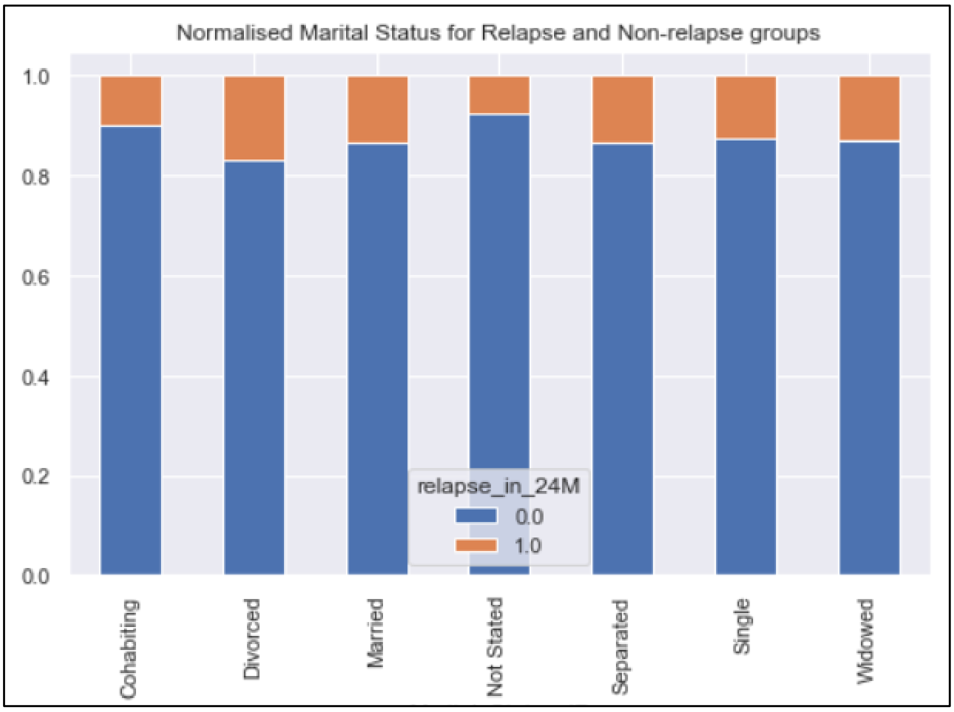


Figure 4.10 Normalised depression by marital status

How is the numerical data distributed?

The IMD scores look roughly normal centred around 30. The HoNOS Total Scores are right skewed with most patients having total scores between the bins for 5 and 10. Age is right skewed with a peak around 30 and a gradual decrease after that.

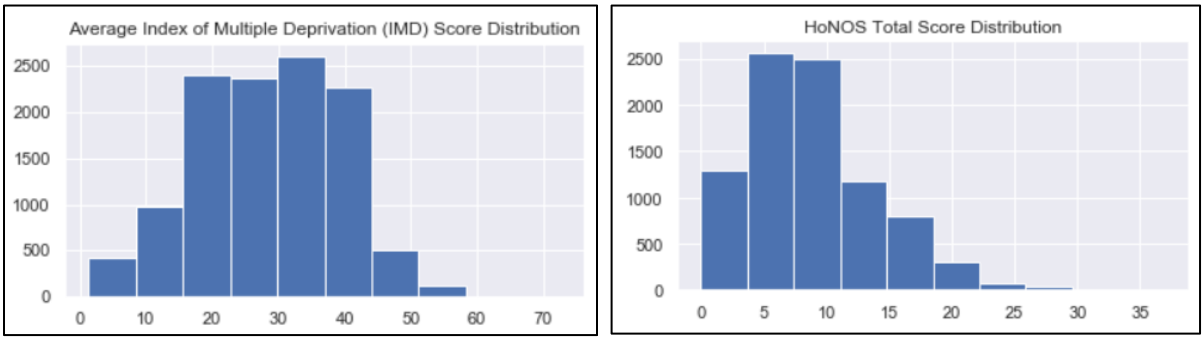


Figure 4.11 IMD scores and HoNOS Total score distributions

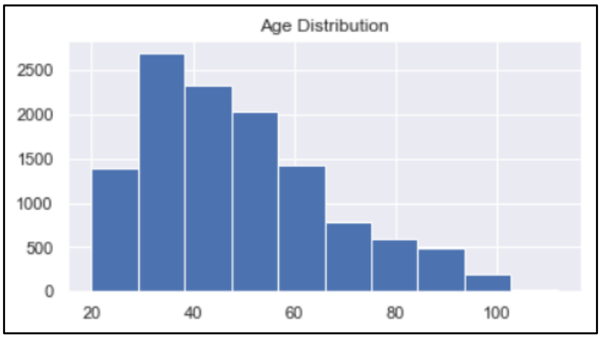


Figure 4.12 Age distribution

How does numerical data vary between relapse and non-relapse group?

The IMD scores are more concentrated between about 15 and 45 for the non-relapse group and the range is higher than the relapse group. Similar patterns can be found between the HoNOS Total scores for relapse and non-relapse groups.

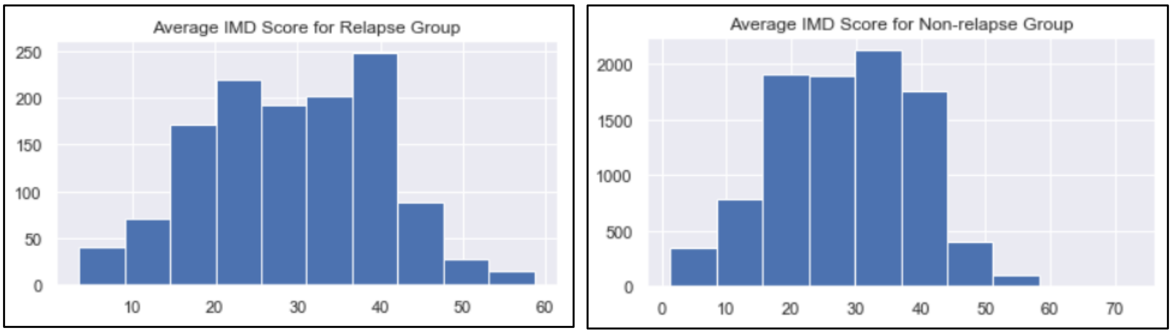


Figure 4.13 IMD scores for relapse and non-relapse groups

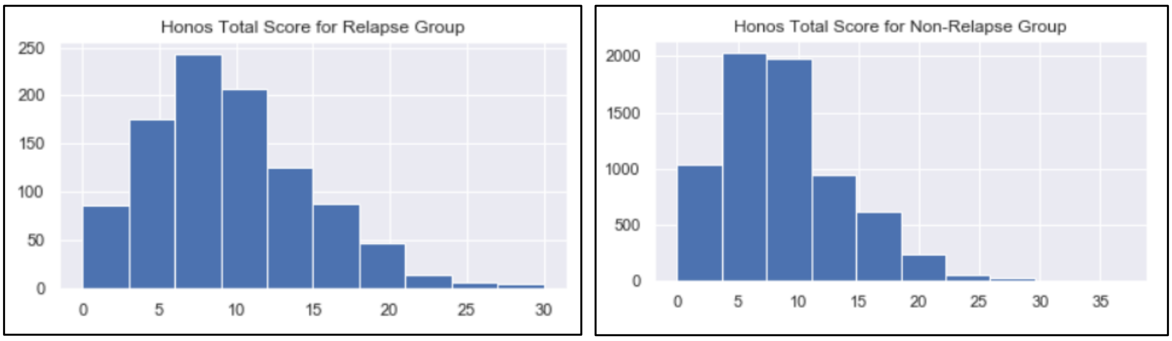


Figure 4.14 HoNOS Total scores for relapse and non-relapse groups

How are features correlated with each other and with the target?

A heatmap of a correlation map was generated for all features and the different relapse time periods. It is not presented here, as it is unreadable because of the large number of features.

Unfortunately, it did not show any strong correlations between any features and the different relapse periods. There were positive correlations among groups of similar fields, such as among the symptoms.

What do the event types and event counts look like?

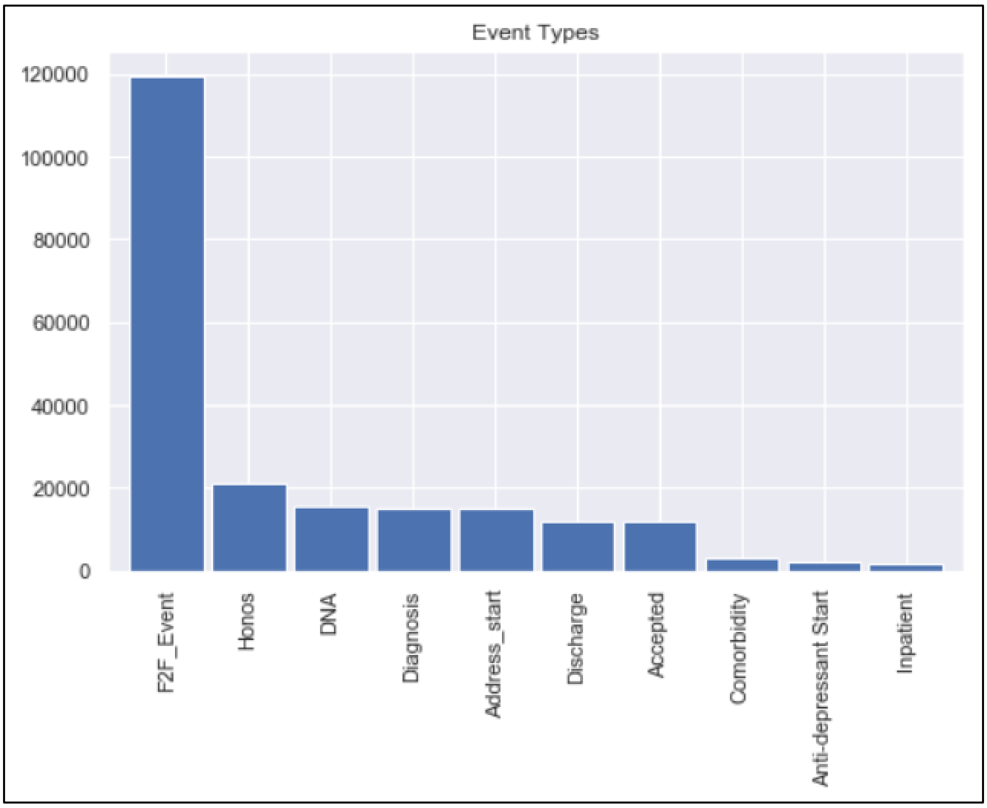


Figure 4.15 Different event types

By far, the most common type of event is a face-to-face event. Relative to other types, there are very few inpatient, anti-depressant starts and comorbidity events. Based on the rules to define the patient group, accepted, diagnosis and discharge events are held by all patients. The median number of events per patient is 9.

Count	Mean	Std	Min	25%	50%	75%	Max
216,563	18	26	2	6	9	19	687

Table 4.8 Number of patient events.

How do the event types and event counts vary for relapse and non-relapse groups?

Anti-depressant starts and Inpatient episodes have a slightly higher proportion of relapse patients than other event types.

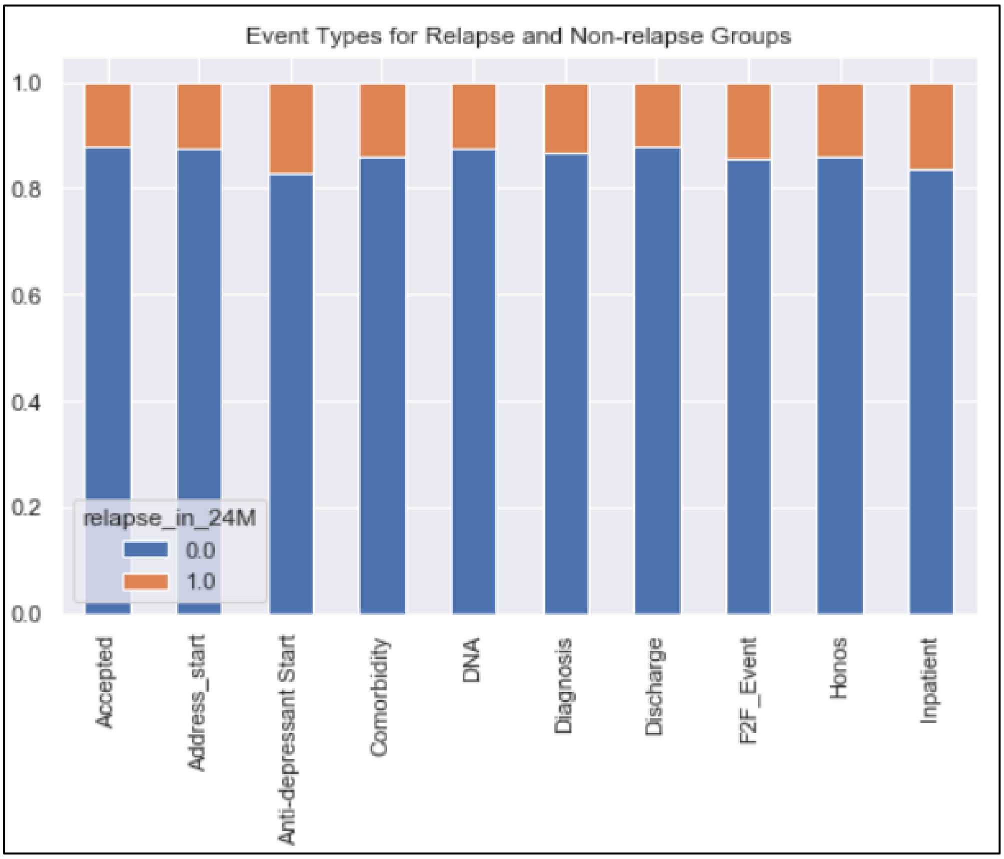


Figure 4.16 Normalised event types for relapse and non-relapse groups

What do events look like for different time frequencies?

Frequency	Count	Mean	Std	Min	25%	50%	75%	Max
Yearly	22,015	10	13	1	3	6	11	217
Quarterly	40,895	5	7	1	2	3	6	100
Monthly	70,668	3	4	1	1	2	3	59
Weekly	121,316	2	2	1	1	1	2	20
Daily	160,555	1	1	1	1	1	1	13

Table 4.9 Number of patient events for different frequencies.

What are the minimum and maximum gaps (in days) between events?

Gap	Mean	Std	Min	25%	50%	75%	Max
Maximum	301	873	0	26	68	193	40,849
Minimum	0	0	0	0	0	0	0

Table 4.10 Biggest and smallest gaps between consecutive events in days.

EDA Conclusion

As no strong correlations were found in the heat map, all of the data types considered in the analysis were used in the model. As mentioned previously, Zhao *et al.* (2019) found that performance improved on several models when incorporating temporal features. The EDA supported the idea that patient events could be modelled as sequences, so the data was modelled sequentially. The event dates were **not** considered; only the sequences of the events were modelled.

The EDA also showed a higher proportion of females in SLaM with depression. Gender was suggested by the clinician as an attribute for which separate models could be considered, so the second set of sequential models was split by gender.

Lastly, the temporal analysis suggested there were some outliers with many events and large gaps between consecutive events. The final sequential model was built using data with outliers and other event data filtered out or aggregated.

The EDA was also used to define the features to use in the models, as shown in Table 4.11.

Features included in sequential models	**Number of features**	**Field description**
BrcId	1	Unique ID of patient
Discharge events	2	One-hot encoded Discharge Destination: GP and Home
Area Level Deprivation events	1	Normalised IMD score
Comorbidities events	5	One-hot encoded Comorbidity event (Diagnoses: F30, F31, F0, F1, F2)
Diagnosis events	1	One-hot encoded Diagnosis event
DNAs events	1	One-hot encoded Did Not Attend event
Face to Face events	1	One-hot encoded Face-to-Face event
HoNOS events	14	HoNOS scores for each question (values: 0 – 4) and Normalised Total and Adjusted Total
Inpatient events	1	Normalised Number of Inpatient days
Medication events	1	One-hot encoded Medication event
Demographics	27	Normalised Age and One-hot encoded Gender, Marital Status and Ethnicity
Risk Assessment New	1	One-hot encoded Risk event (new form)
Risk Assessment Old	1	One-hot encoded Risk event (old form)
Symptoms	20	One-hot encoded Symptoms

Table 4.11 Features included in models.

Table 4.12 shows a sample of the structure of the data imported into the model using synthetic data. Each group of rows with the same BrcId represents a sequence. Missing data is represented by -1. The values in bold in the table indicate the event represented for that row. For example, the first patient events in order are: Area Level Deprivation (represented by IMD_Score), Diagnosis, and Discharge (represented by GP).

BrcId	GP	Home	IMD_Score	F30	F31	…	Diagnosis	Anti-depressant
XXXXXXX1	-1	-1	**0.35**	-1	-1		-1	-1
XXXXXXX1	-1	-1	-1	-1	-1		**1**	-1
XXXXXXX1	**1**	0	-1	-1	-1		-1	-1
XXXXXXX2	-1	-1	-1	**1**	0		-1	-1
XXXXXXX2	-1	-1	-1	-1	-1		**1**	-1
XXXXXXX2	-1	-1	-1	-1	-1		-1	**1**
XXXXXXX2	0	**1**	-1	-1	-1		-1	-1
XXXXXXX3	-1	-1	-1	-1	-1		**1**	-1
XXXXXXX3	-1	-1	**0.512**	-1	-1		-1	-1
XXXXXXX3	**1**	0	-1	-1	-1		-1	-1

Table 4.12 Structure of features showing synthetic data.

4.5 Model Results

Model 1 resulted in a high accuracy of 89%, but all patients were being assigned to the non-relapse class, as shown in the confusion matrix in Figure 4.17. The data is highly imbalanced with only 12% in the minority class.

		Actual	
		Positives	Negatives
Predicted	Positives	TP: 0	FP: 0
	Negatives	FN: 389	TN: 3204

Figure 4.17 Model 1 confusion matrix.

A grid search was run for Model 1 to find the best hyperparameters. The table below shows the parameters that were evaluated and the best values found during the search.

Hyperparameter	Values Tested	Best Value
Number of epochs	2, 5, 10, 20	2
Batch size	1, 10, 20	10
Dropout rate	0.0, 0.2, 0.4, 0.6, 0.8	0.0
Recurrent dropout rate	0.0, 0.2, 0.4, 0.6, 0.8	0.0
Number of hidden neurons	25, 50, 75, 100, 125	25

Table 4.13 Grid Search parameters and results for Model 1

Model 2 applied a class weight to the minority class, which effectively balanced the two classes. The accuracy went down to 80%, but the model did not assign everything to the non-relapse class. Recall, F1 score and AUC all improved, but not by much. Also, the loss of the validation data started to increase after about epoch 20, indicating the model was overfitting (Figure 4.18). The fluctuation in both the loss and accuracy were also assumed to be caused by overfitting.

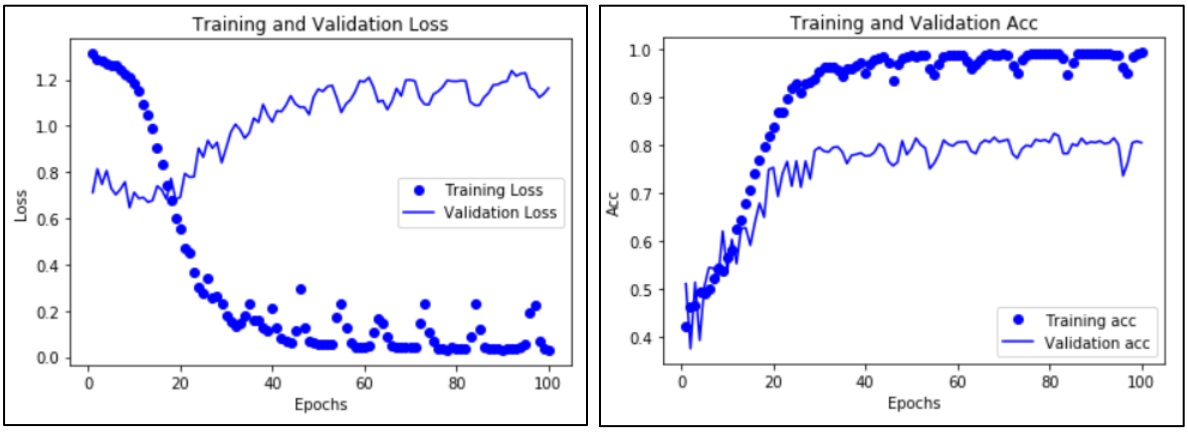


Figure 4.18 Model 2 with class weights.

		Actual	
		Positives	Negatives
Predicted	Positives	TP: 57	FP: 370
	Negatives	FN: 332	TN: 2834

Figure 4.19 Confusion Matrices for Model 2

To address the overfitting, early stopping was added to produce Model 3. Accuracy was drastically reduced to 59% after this, but all other metrics improved with recall improving significantly to 52%.

		Actual	
		Positives	Negatives
Predicted	Positives	TP: 201	FP: 1291
	Negatives	FN: 188	TN: 1913

Figure 4.20 Confusion Matrices for Model 3

A grid search was run for Models 2 and 3 to find the best hyperparameters. The table below shows the parameters that were evaluated and the best values found during the search.

Hyperparameter	**Values Tested**	**Best Value**
Number of epochs	10, 20, 50, 100	100
Batch size	10, 20, 30	30
Number of hidden neurons	50, 100, 150, 200, 250, 300	300

Table 4.14 Grid Search parameters and results for Models 2 and 3

Table 4.15 shows the models and metrics.

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**AUC**
Model 1 – base model	**89.17%**	0%	0%	0%	0.50
Model 2 – class weights	80.46%	13.34%	14.65%	13.97%	0.52
Model 3 – class weights and early stopping	58.84%	**13.47%**	**51.67%**	**21.37%**	**0.56**

Table 4.15 Model 1-3 results

Chollet (2018, pp. 217-218) suggested adding dropout and recurrent dropout to the LSTM layer to address overfitting. However, as seen in the grid search, Model 1 actually performed better without dropout. This was also found to be the case with earlier versions of grid searches for Models 1 - 3, so dropout and recurrent dropout were removed from future grid searches to save time.

4.6 Additional Model by Gender Results

Next, the data was split by gender into two datasets for male and female. Each dataset was run through a grid search to find the best parameters. The models were then run with those parameters using class weights and early stopping. The results for both are below.

Model	Accuracy	Precision	Recall	F1 Score	AUC
Model 4 - Female	**68.47%**	13.64%	39.91%	20.34%	0.56
Model 5 - Male	58.09%	**15.41%**	**54.32%**	**24.01%**	**0.56**

Table 4.16 Model 4 and 5 results

		Actual	
		Positives	Negatives
Predicted	Positives	TP: 91	FP: 576
	Negatives	FN: 137	TN: 1457

Figure 4.21 Confusion Matrices for Model 4

		Actual	
		Positives	Negatives
Predicted	Positives	TP: 88	FP: 483
	Negatives	FN: 74	TN: 684

Figure 4.22 Confusion Matrices for Model 5

A grid search was run for Models 4 and 5 to find the best hyperparameters. Table 4.16 below shows the parameters for Model 4, which had the same best results as Model 2 and 3. Table 4.17 shows Model 5 grid search parameters and results.

Hyperparameter	Values Tested	Best Value
Number of epochs	10, 20, 50, 100	100
Batch size	10, 20, 30	30
Number of hidden neurons	50, 100, 150, 200, 250, 300	300

Table 4.16 Model 4 Grid Search parameters and results

Hyperparameter	Values Tested	Best Value
Number of epochs	10, 20, 50, 100	100
Batch size	10, 20, 30	20
Number of hidden neurons	50, 100, 150, 200, 250, 300	250

Table 4.17 Model 5 Grid Search parameters and results

4.7 Additional Model using Filtered Data Results

Lastly, the data was filtered as described in section 3.3.4. The dataset was run through a grid search to find the best parameters. The model was then run with those parameters using class weights and early stopping. The results are below.

Model	Accuracy	Precision	Recall	F1 Score	AUC
Model 6 – Filtered Data	59.81%	12.2%	46.03%	19.28%	0.54

Table 4.18 Model 6 results

		Actual	
		Positives	Negatives
Predicted	Positives	TP: 145	FP: 1044
	Negatives	FN: 170	TN: 1662

Figure 4.23 Confusion Matrices for Model 6

Hyperparameter	Values Tested	Best Value
Number of epochs	10, 20, 50, 100	100
Batch size	10, 20, 30	20
Number of hidden neurons	50, 100, 150, 200, 250, 300	300

Table 4.19 Model 6 Grid Search parameters and results

4.8 Summary

The table below shows all model results in one table for comparison. The best accuracy, 89%, was for Model 1. However, the best model for all other metrics was Model 5.

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**AUC**
Model 1 – base model	**89.17%**	0%	0%	0%	0.50
Model 2 – class weights	80.46%	13.34%	14.65%	13.97%	0.52
Model 3 – class weights and early stopping	58.84%	13.47%	51.67%	21.37%	0.56
Model 4 – Female	68.47%	13.64%	39.91%	20.34%	0.56
Model 5 – Male	58.09%	**15.41%**	**54.32%**	**24.01%**	**0.56**
Model 6 – Filtered Data	59.81%	12.2%	46.03%	19.28%	0.54

Table 4.20 All model results

Chapter 5 – Discussion

This section discusses how the objectives were met and how they relate to the results. It then provides an answer to the research question. Finally, it discusses the limitations of the study.

5.1 Objectives

O1. To research sequence modelling algorithms and related literature

Several studies were reviewed to find examples of sequence modelling algorithms used in healthcare. I narrowed the focus to Hidden Markov Models and Recurrent Neural Networks, although there was limited information in the literature about how data was prepared for HMMs. The studies with RNNs used different methods for preparing the sequences such as considering sequences only or aggregated data into time windows. It would have been interesting to try these different techniques and compare the performance. Min *et al*. (2019) found better results grouping patient events by month compared to using sequences alone. Grouping into time windows may have produced better results for this project.

There were some HMM and RNN examples available in both R and Python on different websites and one book. These were used to understand how to prepare the data, build prototypes and eventually build the actual models. Although there were lots of examples for time series predictions, there were fewer examples of sequence classification.

O2. To define attributes and thresholds from the source data to be used for analysis

I reviewed the CRIS data dictionary during the project planning and used information from the literature review to define an initial set of data attributes. I then used semi-structured interviews with the CRIS clinician and SME to help define attributes and thresholds from the source data. Presenting the clinician with picklist values for discharge destination and event types and outcomes was useful to further refine definitions of features and remission.

One change that may have improved model results would be to define a minimum time for remission. For example, a patient in remission for only one day who then returns would have been considered part of the relapse group. However, a more likely explanation is that they weren’t truly in remission. Another change would be to use the free text in the discharge notes to better define remission. The definition of remission, i.e. discharged to home or a GP, while straightforward and easy to implement for a short project, may not mean the patient is truly in remission. They may be managing treatment with their GP. However, using the discharge notes to define remission would have been more complicated and taken longer to implement.

In an attempt to include only patients with initial depressive episodes on SLaM, the starting date of 1 January 2010 was used. However, this may not have been the true first episode if they had moved from a different trust or had only been treated by their GP. Also, patients with recurrent depressive disorder (F33) were included in the patient group which conflicted with this approach. Taking data from 2007 regardless of whether it was the patient's initial depressive episode and leaving the F33 diagnoses in the dataset would have been more consistent, and it would have provided three more years of patient data. This could have improved model performance as it would have given the models more data to use for training.

Depression was only considered if it was listed as the patient's primary diagnosis using the ICD-10 codes, so this study could have missed patients who only had diagnoses recorded only in clinical notes.

Only anti-depressant medication was included, so any other medications the patient was taking was not used. Also, non-mental health medication was not generally available in CRIS.

O3. To cleanse and prepare the data for further analysis and use in modelling

Data was cleansed and prepared using Python so that analysis could be done prior to modelling the data. In some cases, where it was more straightforward, the preparation was done in SQL. An example of this is the additional medication extract showing the medication in one-hot encoded format. My approach was to extract as much data as needed from the source as well as getting it in the right shape when extracted, where possible. However, this may have resulted in more SQL extracts than necessary.

The data entities were stored in many tables in the database. Preparing all the data to analyse each entity and to analyse the data as a set of patient events took a few iterations to get right. The order of loading each entity as well as how it was joined to other entities was important and required checks at each step to ensure data was not inadvertently truncated or duplicated. Working with sequential data added an extra complication as it contained a third dimension, i.e. time. However, the data required was prepared into the right shape for the sequential model.

O4. To perform exploratory data analysis (EDA) to better understand the data and define data for modelling

Through the EDA, I was able to answer several questions about the data and make decisions about what data to include and how to define the target for the model.

The timeframe for relapse was defined as 2 years based on analysis of patient relapse counts overall and for different timeframes. The relapse group was defined based on rules from the clinician, as well as my analysis. A more robust approach would have been to have the patient records reviewed and labelled by a clinician. However, this would not have been feasible in the project timeframe.

Dates of symptoms are the recorded dates, so this does not capture start and end dates or the duration the patient was experiencing the symptom. The symptoms were derived from clinical notes, and there is no indication if it was a current or prior symptom. For these reasons, symptoms were not treated as sequential events. Had the start and end dates been present and used as part of the sequence, the results may have been different.

Risk events were not treated as sequential events. The question on the new risk assessment form asks about any current or historical risk of self-harm or suicide and was therefore not appropriate to add as an event. The old form had a mix of questions related to historical and current behaviour, so for simplicity this was also treated as non-sequential.

No events occurring after the patient's initial discharge (remission) were included. While there were some events recorded for some patients following discharge, such as address changes, a decision was made to stop all sequences at the point of discharge for consistency. It is possible events occurring between remission and relapse may have been important in predicting relapse.

O5. To build a machine learning model that determines if a patient will experience a relapse of depression

An LSTM model was built, although the performance was poor when considering all metrics defined. While accuracy of the first version of the model was high (89%), the imbalanced nature of the data meant the model classified all patients into the non-relapse group. This resulted in 0% for recall, precision, and F1 score and only 0.50 AUC. A revised version of the model using class weights to balance the classes improved the other metrics some, but they were still very low. Lastly, early stopping was added to stop overfitting. This improved recall significantly and somewhat improved F1 score and AUC, but accuracy dropped significantly as well.

O6. To investigate whether building separate machine learning models for different patient groups will improve model performance or provide additional understanding of the illness

Two additional models were built for females (Model 4) and males (Model 5). These both had similar AUC to Model 3, and Model 5 outperformed Models 1 – 4 in all metrics accept accuracy.

O7. To investigate whether placing limits on the data included in the model will improve model performance or provide additional understanding of the illness

A final model was built using filtered data (Model 6). This model performed worse than Model 3 in all areas and only improved accuracy by 1%.

5.2 Answering the Research Question

To what extent can machine learning be used with anonymised electronic health records to determine the likelihood a patient will experience a relapse/recurrence of depression within a given time frame following remission?

Using the specific data that was collected and the LSTM algorithm, machine learning can be used to a limited extent to determine patients who will experience a depression relapse.

Although the performance of all models was poorer than anticipated, Model 5 using only data for male patients had the best performance for all metrics except accuracy. Males have a slightly higher proportion in the relapse group than females (12.59% vs 12.05%) and other patients (12.25%), which may be one reason for this.

Model 5 had the highest AUC (0.56), recall (54%), F1 score (24%) and precision (15%). While its overall accuracy is the lowest at 58%, Model 5 was better at identifying patients in the relapse group (true positives) than the other models. Model 1 with the highest accuracy of 89% was not able to identify any patients in the relapse group. It simply predicted that no patient would have a relapse, which is as good as having no model at all.

AUC is a common performance metric. When comparing with the literature on depression, there were two studies with similar, although still higher, AUC measurements. Kessler *et al.* (2016) reported an AUC of 0.63 when predicting high chronicity (number of years) of depression. Dinga *et al.* (2018) reported an AUC of 0.66 when predicting MDD diagnosis at a 2-year follow-up when studying the naturalistic course of depression. Results from Dinga *et al.* (2018) with a higher AUC were considered too low to be useful in a clinical setting. Based on this, I assume the model performance in this project is too low for clinical use.

5.3 Limitations

There were several limitations of this study, some of which likely contributed to the poor performance of the models.

- Apart from symptoms, no details from clinical notes or discharge summaries were extracted and used in this study. Information relevant to depression or depression relapse captured in notes may have been missed if it was not also recorded in structured fields set up specifically to record that information.
- Some data is limited. It is unclear whether this is a true reflection of the occurrence or if it is not being recorded in all cases. The following entities had data populated for fewer than 20% of the patient group: comorbidities, psychotic episodes, inpatient episodes, risk events, DNA events, and medication. This may be related to the point above if data is being recorded in notes rather than specific entities or fields designed to capture the information.
- CRIS data does not capture records from GPs or A&Es. This means patients being treated only by a GP for depression and people only attending A&E for depressive episodes are not included in this study. For patients in SLaM, the data available may show a limited representation of the

actual clinical contact without records from the GP and A&E. A patient having a relapse who seeks treatment urgently through A&E would not have been included in the relapse group.

- Patients who moved out of the area covered by the trust during the 2-year relapse period may or may not have experienced a relapse. Without any knowledge of movers, by default they would have been added to the non-relapse group.
- The severity of the symptoms was not available. A previous study by Dinga *et al.* (2018) indicated symptom severity to be the most predictive factor of depression trajectory, so this may have been important in predicting relapse.
- Data access was limited to hot desks onsite and remote access via a VPN. This was a limitation in terms of software and computational power. It's possible a better combination of hyperparameters could have been found with a comprehensive grid search. However, it would not have been feasible to run this on a hot desk, as it would have taken multiple days. The hot desk could be shut down or another researcher could log in, which would have stopped the process running. Also, Python libraries, e.g. Keras, were not available via VPN and could not be installed. So, a multi-day process was also not possible via VPN. Even on the hot desk, it took several days to resolve an issue with the Keras installation, which almost ruled RNNs with Python out for this project. Had the data been available to use on a personal device with a dedicated GPU, this may have sped up processing as well as allowed processes to run for long periods of time without the risk of interruption. This also could have opened up other deep learning options requiring more processing power.

Many factors may contribute to a patient relapsing. It is possible that even if the desired clinical data were well-populated or features were extracted from clinical notes, there may be other reasons not available in an EHR that someone relapses from depression.

Chapter 6 – Evaluation, Reflections, and Conclusions

6.1 Evaluation and Reflections

Having used Scrum throughout the project, I felt it would be appropriate to end the project using the same questions used during a Scrum Retrospective: 1) What went well? and 2) What could be improved? I also include some recommendations for future work.

6.1.1 What went well?

Scrum was a good way to organise the work by priority, monitor progress, and adapt to change. Taiga proved to be a useful free tool to keep track of all the project artefacts and manage my work. The one-day CRIS SQL training was quite useful, as it provided information on the CRIS database structure and nuances with the data. Having access to the clinician and SME was also invaluable in understanding the domain, which data was important and getting answers to questions about the data. I was pleased to put my SQL skills to use to build the data extracts. It gave me more exposure to the data, which meant I could ask more intelligent questions during the interviews with the SME and clinician. This also proved to be a better option than requesting extracts from the CRIS team, as I made a number of changes as I learned more about the data and how to shape it for the sequential algorithm. I learned a great deal about mental health, NHS clinical data and previous research on depression. I also gained more experience with Python programming and the Keras deep learning library. I now have a better understanding of recurrent neural networks, especially LSTM.

6.1.2 What could be improved?

The literature and algorithm research took much longer than anticipated. This in combination with feedback from the clinician led to an assumption about modelling the data as non-sequential. Early work on the extracts were based on this assumption. These extracts were still used during EDA, but I may have approached them differently had I completed the literature and algorithm research first, understood how the data needed to be shaped and then started to build the extracts.

Later in the project, I ran into an issue installing additional Python libraries on hot desks. I lost some time troubleshooting and almost had to abandon LSTM with Python. During this time, I looked at alternatives like HMM and RNN in R. The issue was finally resolved with help from someone in the office by setting proxy server details. It would have saved some time had I tried installing additional libraries when I installed Anaconda early in the project.

Towards the end of the project, I discovered that discharge dates had not been restricted to allow for a full prediction window. In other words, patients being discharged up until the extracts were run could be included in the dataset. This was problematic, because patients who had been discharged from SLaM within the last 2 years, i.e. after July 2017, would not have had a full 2-year window within which to relapse. This was addressed by restricting the discharge dates, re-running the EDA, grid searches and models. The results did not change much, but it was important to correct this.

Finally, having reviewed the model results and noticing the apparent trade-off between accuracy and F1 score, it seems F1 score should also have been evaluated during the grid search and the parameters with the highest F1 score *and* accuracy should have been chosen to build the models. Also, including different class weight values in the grid search may have helped. The values chosen were to balance the classes equally, but this may not have given the best performance. Lastly, dropout and recurrent dropout values were evaluated in early grid searches and the highest accuracy always resulted from no dropout. Trying dropout with F1 score and different class weights in the grid search may have been a better alternative to reduce overfitting compared to early stopping.

6.1.3 Future Work

First, modelling this problem using aggregated non-sequential data is one option that could be considered for comparison to the sequential modelling. It would be interesting to see whether the same poor performance arises. If performance were improved, it would likely indicate that the sequences of events are not important and therefore a sequential algorithm is not appropriate.

Next, assuming sequential modelling is still being considered, the EDA of the temporal data suggests there are wide variations in the time gaps between events. The maximum event gap has a median of 68 days and the minimum event gap has a median of 0 days. A patient with the same sequence of events spread out over much different time scales could have a different outcome. As mentioned above, Min *et al*. (2018) showed that a model of the sequence alone performed worse than a model including the temporal data. The EDA also suggests the sequence of events could still be captured if the data were regrouped by day, and in at least half of the patients if it were regrouped by week. Some of the sequence could be captured if the data were regrouped by month. A potential future study would be to reshape the data into weekly or monthly time windows to see if this improves the performance of the sequential models.

Lastly, another option is to predict the relapse date or month for a patient instead of classifying patients into relapse and non-relapse groups. Although, considering the performance of a binary classification with a set timeframe was poor, it is difficult to imagine achieving better performance in predicting an actual date or month.

6.2 Conclusions

Even though the model results were disappointing, I hope some of the findings from the analysis are useful to the CRIS team and clinicians within SLaM interested in depression recurrence. Potentially the findings from the EDA, as well as Model 5 (males only) having the best performance metrics, excluding accuracy, could be valuable.

This project also shows that it is feasible to build a deep learning model in Python using the SLaM BRC hot desks for any future researchers wishing to do something similar using CRIS data. However, given the difficulties mentioned above, it would be best to know exactly what needs to be installed early in the project to allow for any issues to be resolved or alternatives to be identified.

Glossary

Term	Full Name	Description
AUC	Area Under the Receiver Operating Characteristic curve	The area under the curve when plotting the true positive rate (recall) against the false positive rate.
Batch size	Batch size	When training a neural network, this is the number of sequences processed before the network weights are updated.
Dropout	Dropout	A technique used to avoid overfitting which involves removing connections between some neurons in the network
Epoch	Epoch	When training a neural network, this is one iteration through the training data.
Early stopping	Early stopping	A technique used to monitor and stop the training of a neural network once the validation loss starts to increase, indicating the model is overfitting
F1-score	F1-score	2*true positives / (2*true positives + false positives + false negatives)
HoNOS	Health of the Nation Outcome Scales	A set of questions used to measure mental health
ICD-10	International Classification of Diseases, version 10	Clinical diagnostic coding standard (version 10)
One-hot encoding	One-hot encoding	Each categorical value is represented as a column and the columns are populated with a 1 if it is true or a 0 if it is not true for each observation.
Overfitting	Overfitting	When a model does not generalise well with unseen data
PHQ-9	Patient Health Questionnaire	Questionnaire used to measure severity of depression
Precision	Precision	True positives / (true positives + false positives)
Recall	Recall	True positives / (true positives + false negatives)

References

Ayabakan, S., Bardhan, I. and Zheng, Z. (2016) 'What Drives Readmission? A New Perspective from Hidden Markov Model Analysis', *International Conference on Information Systems*, Dublin, Ireland, 11-14 December. Available at: https://www.researchgate.net/publication/329589065_What_Drives_Readmission_A_New_Perspective_from_Hidden_Markov_Model_Analysis (Accessed: 3 June 2019).

Bishop, C. (2006) *Pattern Recognition and Machine Learning*. New York: Springer Science + Business Media, LLC.

Brownlee, J. (2015) '8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset', *Machine Learning Mastery*, 19 August. Available at: https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/ (Accessed 10 Aug 2019).

Brownlee, J. (2016) 'Sequence Classification with LSTM Recurrent Neural Networks in Python with Keras', *Machine Learning Mastery*, 26 July. Available at: https://machinelearningmastery.com/sequence-classification-lstm-recurrent-neural-networks-python-keras/ (Accessed 25 July 2019).

Brownlee, J. (2019) 'How to Calculate Precision, Recall, F1, and More for Deep Learning Models', *Machine Learning Mastery*, 29 March. Available at: https://machinelearningmastery.com/how-to-calculate-precision-recall-f1-and-more-for-deep-learning-models/ (Accessed 29 August 2019).

Brownlee, J. (2019) 'How to Normalize and Standardize Time Series Data in Python', *Machine Learning Mastery*, 28 August. Available at: https://machinelearningmastery.com/normalize-standardize-time-series-data-python/ (Accessed 4 September 2019).

Brownlee, J. (2019) 'How to Grid Search Hyperparameters for Deep Learning Models in Python With Keras', *Machine Learning Mastery*, 19 August. Available at: https://machinelearningmastery.com/grid-search-hyperparameters-deep-learning-models-python-keras/ (Accessed 28 August 2019).

Cao, B., Zheng, L., Zhang, C., Yu, P., Piscitello, A., Zulueta, J., et al. (2017) 'DeepMood: modeling mobile phone typing dynamics for mood detection', *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax, Canada, 13-17 August. pp. 747–755. doi: 10.1145/3097983.3098086.

Che, Z., St. Sauver, J., Liu, H., Liu, Y. (2017) 'Deep Learning Solutions for Classifying Patients on Opioid Use', *AMIA Annual Symposium Proceedings*, Washington, USA, 3-8 November. pp. 525–

534. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977635/ (Accessed 23 July 2019).

Chollet, F. (2018) *Deep Learning with Python*. Shelter Island: Manning Publications Co.

Chollet, F. *et al*. (no date) *Keras*. Available at: https://keras.io (Accessed: 25 July 2019).

City, University of London Department of Computer Science (2019) *Research Ethics*. Available at: https://www.city.ac.uk/department-computer-science/research-ethics (Accessed: 26 March 2019).

Cooper, J., Lewis, J., Lord, J. (2019) *Healthcare expenditure, UK Health Accounts: 2017*. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthcaresystem/bulletins/ukhealthaccounts/2017. (Accessed: 27 September 2019).

Dawson, C. (2015) *Projects in Computing and Information Systems: A Student's Guide*. Harlow: Pearson Education Limited.

Dinga, R., Marquand, A., Veltman, D., Beekman, A., Schoevers, R., van Hemert, A., Penninx, B., Schmaal, L. (2018) 'Predicting the naturalistic course of depression from a wide range of clinical, psychological, and biological data: a machine learning approach', *Translational Psychiatry*, 8. doi: 10.1038/s41398-018-0289-1.

Doryab, A., Dey, A., Kao, G. and Low, C. (2019) 'Modeling Biobehavioral Rhythms with Passive Sensing in the Wild: A Case Study to Predict Readmission Risk after Pancreatic Surgery', *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*. 3(1). doi: 10.1145/3314395.

Durstewitz, D., Koppe, G., Meyer-Lindenberg, A. (2019) 'Deep neural networks in psychiatry', *Molecular Psychiatry*, ISSN 1476-5578. doi: 10.1038/s41380-019-0365-9.

Eaton, W., Shao, H., Nestadt, G., Lee, B., Bienvenu, O., Zandi, P. (2008) 'Population-Based Study of First Onset and Chronicity in Major Depressive Disorder', *Arch Gen Psychiatry*, 65(5), pp. 513–520. doi:10.1001/archpsyc.65.5.513.

Evans, R. (2016) 'Electronic Health Records: Then, Now, and in the Future', *Yearbook of Medical Informatics*, Suppl 1, pp. S48-S61. doi:10.15265/IYS-2016-s006.

Gers, F., Schmidhuber, J., Cummins, F. (2000) 'Learning to Forget: Continual Prediction with LSTM', *Neural Computation*. 12(10), pp. 2451-2471. doi: 10.1162/089976600300015015.

Hasan, M., Kotov, A., Carcone, A., Dong, M., Naar, S. (2017) 'Predicting the Outcome of Patient-Provider Communication Sequences using Recurrent Neural Networks and Probabilistic Models', *AMIA Joint Summits on Translational Science Proceedings*, San Francisco, USA, 27-30 Mar. pp. 64–73. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5961827/ (Accessed 5 August 2019).

Hayes, R., Chang C., Fernandes, A., Begum, A., To, D., Broadbent, M., Hotopf, M., Stewart, R. (2012) ‘Functional Status and All-Cause Mortality in Serious Mental Illness’, *PLoS ONE*, 7(9): e44613. Available at: https://doi.org/10.1371/journal.pone.0044613 (Accessed: 6 March 2019).

Hayes, R., Chang, C., Fernandes, A., Begum, A., To, D., Broadbent, M., Hotopf, M., Stewart, R. (2012) 'Associations between symptoms and all-cause mortality in individuals with serious mental illness', *Journal of Psychosomatic Research*, 72(2), pp. 114-119. doi: 10.1016/j.jpsychores.2011.09.012.

Hochreiter, S., Schmidhuber, J. (1997) ‘Long Short-Term Memory’, *Neural Computation*, 9(8), pp. 1735-1780, doi: 10.1162/neco.1997.9.8.1735.

Kawamoto, R., Nazir, A., Kameyama, A., Ichinomiya, T., Yamamoto, K., Tamura, S., Yamamoto, M., Hayamizu, S. and Kinosada, Y. (2013) 'Hidden Markov Model for Analyzing Time-Series Health Checkup Data', *Studies in health technology and informatics*, Copenhagen, Denmark, 19-22 August. pp. 491-495. doi: 10.3233/978-1-61499-289-9-491.

Kessing, L. (2008) 'Severity of depressive episodes during the course of depressive disorder', *The British Journal of Psychiatry*, 192(4), pp. 290-293. doi: 10.1192/bjp.bp.107.038935.

Kessler, R., van Loo, H., Wardenaar, K., Bossarte, R., Brenner, L., Cai, T., Ebert, D., Hwang, I., Li, J., de Jonge, P., Nierenberg, A., Petukhova, M., Rosellini, A., Sampson, N., Schoevers, R., Wilcox, M., Zaslavsky, A. (2016) ‘Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports’, *Molecular Psychiatry*, 21(10), pp. 1366–1371. doi: 10.1038/mp.2015.198.

Kingma, D. and Ba, J. (2015) ‘Adam: A Method for Stochastic Optimization’, *International Conference for Learning Representations*, San Diego, USA, 7-9 May. pp. 1-15. Available at: https://arxiv.org/pdf/1412.6980v8.pdf (Accessed 20 September 2019).

Lin, Y., Huang, S., Simon, G., Liu, S. (2016) 'Analysis of depression trajectory patterns using collaborative learning', *Mathematical Biosciences*, 282, pp. 191-203, doi: 10.1016/j.mbs.2016.10.008.

Lin, Y., Huang, S., Simon, G., Liu, S. (2018) ‘Data-based Decision Rules to Personalize Depression Follow-up’, *Scientific Reports*, 8. doi: 10.1038/s41598-018-23326-1.

Mahmoud, S., Lotfi, A., Langensiepen, C. (2013) 'Behavioural pattern identification and prediction in intelligent environments', *Applied Soft Computing*, 13(4), pp. 1813-1822, doi: 10.1016/j.asoc.2012.12.012.

McIntosh, A., Stewart, R., John, A., Smith, D., Davis, K., Sudlow, C., Corvin, A., Nicodemus, K., Kingdon, D., Hassan, L., Hotopf, M., Lawrie, S., Russ, T., Geddes, J., Wolpert, M., Wölbert, E., Porteous, D. (2016) 'Data science for mental health: a UK perspective on a global challenge', *The Lancet Psychiatry*, 3(10), pp. 993-998, doi: 10.1016/S2215-0366(16)30089-X.

Melfi, C., Chawla, A., Croghan, T., Hanna, M., Kennedy, S., Sredl, K. (1998) 'The Effects of Adherence to Antidepressant Treatment Guidelines on Relapse and Recurrence of Depression', *Arch Gen Psychiatry*, 55(12), pp. 1128–1132. doi:10.1001/archpsyc.55.12.1128.

Mic (2016) ‘Plain vanilla recurrent neural networks in R: waves prediction’, *R-bloggers*, 5 August. Available at: https://www.r-bloggers.com/plain-vanilla-recurrent-neural-networks-in-r-waves-prediction/ (Accessed: 1 August 2019).

Min, X., Yu, B. and Wang, F. (2019) 'Predictive Modeling of the Hospital Readmission Risk from Patients’ Claims Data Using Machine Learning: A Case Study on COPD', *Scientific Reports*, 9. doi: 10.1038/s41598-019-39071-y.

Mind (2019) *Anti-depressants A-Z.* Available at: https://www.mind.org.uk/information-support/drugs-and-treatments/antidepressants-a-z/ (Accessed: 11 July 2019).

Mitchell, T. (1997) *Machine Learning*. Singapore: McGaw-Hill.

Nie, Z., Gong, P. and Ye, J. (2016) ‘Predict Risk of Relapse for Patients with Multiple Stages of Treatment of Depression’, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16*, San Francisco, USA, 13-17 August. pp. 1795–1804. doi: 10.1145/2939672.2939870.

Oates, B. (2006). *Researching Information Systems and Computing*. London: SAGE Publications Ltd.

Perera, G., Broadbent, M., Callard, F., Chang, C., Downs, J., Dutta, R., Fernandes, A., Hayes, R., Henderson, M., Jackson, R., Jewell, A., Kadra, G., Little, R., Pritchard, M., Shetty, H., Tulloch, A., Stewart, R. (2016) ‘Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource.’ *BMJ Open*, 6: e008721. doi:10.1136/bmjopen-2015-008721.

Perlis, R., Iosifescu, D., Castro, V., Murphy, S., Gainer, V., Minnier, J., Cai, T., Goryachev, S., Zeng, Q., Gallagher, P., Fava, M., Weilburg, J., Churchill, S., Kohane, I. and Smoller, J. (2012) 'Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model', *Psychological Medicine*, 42(1), pp. 41–50. doi: 10.1017/S0033291711000997.

Pham, T., Tran, T., Phung, D., Venkatesh, S. (2017) ‘Predicting healthcare trajectories from medical records: A deep learning approach’, *Journal of Biomedical Informatics*, 69, pp. 218-229. Available at: https://doi.org/10.1016/j.jbi.2017.04.001 (Accessed: 26 March 2019).

Royal College of Psychiatrists (2019) *Health of the Nation Outcome Scales (HoNOS).* Available at: https://www.rcpsych.ac.uk/events/in-house-training/health-of-nation-outcome-scales (Accessed: 16 July 2019).

Rumshisky, A., Ghassemi, M., Naumann, T., Szolovits, P., Castro, V. McCoy, T. Perlis, R. (2018) 'Predicting early psychiatric readmission with natural language processing of narrative discharge summaries', Translational Psychiatry, 6: e921. doi: 10.1038/tp.2015.182.

Sakurai, H., Suzuki, T., Yoshimura, K., Mimura, M., Uchida, H. (2017) ‘Predicting relapse with individual residual symptoms in major depressive disorder: a reanalysis of the STAR*D data’, *Psychopharmacology*, 234(16), pp. 2453–2461. doi: 10.1007/s00213-017-4634-5.

Scikit-learn API Reference (2019) Available at: https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics (Accessed: 29 August 2019).

Scrum.org (2019) *What is Scrum?* Available at https://www.scrum.org/resources/what-is-scrum (Accessed: 16 August 2019).

Scrum Guides (2017) *The Scrum Guide*. Available at: https://www.scrumguides.org/scrum-guide.html (Accessed: 16 August 2019).

Shickel, B., Tighe, P., Bihorac, A., Rashidi, P. (2018) 'Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis', *IEEE Journal of Biomedical and Health Informatics*, 22(5), pp. 1589-1604. doi: 10.1109/JBHI.2017.2767063.

Singh, A. (2019) ‘A Hands-On Introduction to Time Series Classification (with Python Code)’, *Analytics Vidhya*, 7 January. Available at: https://www.analyticsvidhya.com/blog/2019/01/introduction-time-series-classification/ (Accessed on 30 July 2019).

South London and Maudsley NHS Foundation Trust (2013) *CRIS Security Model*. Available at: https://www.maudsleybrc.nihr.ac.uk/media/112184/cris-security-model.pdf (Accessed: 26 March 2019).

South London and Maudsley NHS Foundation Trust (2019) *Clinical Record Interactive Search (CRIS)*. Available at: https://www.maudsleybrc.nihr.ac.uk/facilities/clinical-record-interactive-search-cris/ (Accessed: 4 January 2019).

Suhara Y., Xu Y., Pentland A. (2017) 'Deepmood: forecasting depressed mood based on self-reported histories via recurrent neural networks', *Proceedings of the 26th International Conference on World Wide Web*, Perth, Australia, 3-7 April. pp. 715–724. doi: 10.1145/3038912.3052676.

Taiga (2019) *Taiga*. Available at: https://taiga.io/ (Accessed: 24 June 2019).

Thomas, C., Morris, S. (2003) ‘Cost of depression among adults in England in 2000’, *British Journal of Psychiatry*, 183, pp. 514-519. doi: 10.1192/bjp.183.6.514.

Verma, A., Powell, G., Luo, Y., Stephens, D., Buckeridge, D. (2018). 'Modeling disease progression in longitudinal EHR data using continuous-time hidden Markov models', *Machine Learning for Health (ML4H) Workshop at NeurIPS*, Montreal, Canada, 3-8 December. Available at: https://www.researchgate.net/publication/329388277_Modeling_disease_progression_in_longitudinal_EHR_data_using_continuous-time_hidden_Markov_models (Accessed: 3 June 2019).

Wang, J., Patten, S., Sareen, J., Bolton, J., Schmitz, N., MacQueen, G. (2014) ‘Development and Validation of a Prediction Algorithm for use by Health Professionals in Prediction of Recurrence of Major Depression’, *Depression and Anxiety*, 31(5), pp. 451–457. doi: 10.1002/da.22215.

Wing, J., Beevor, A., Curtis, R., Park, S., Hadden, S., Burns, A. (1998) 'Health of the Nation Outcome Scales (HoNOS). Research and development’, *British Journal of Psychiatry*. 172(1), pp. 11-18. Available at: https://0-www-cambridge-org.wam.city.ac.uk/core/services/aop-cambridge-core/content/view/55BD9DCA7F2A95AEC649AD202D031363/S0007125000149098a.pdf/health_of_the_nation_outcome_scales_honos.pdf (Accessed: 16 July 2019).

World Health Organization (2016) *ICD-10 Version:2016*. Available at: https://icd.who.int/browse10/2016/en (Accessed on: 2 July 2019).

World Health Organization (2018) *Depression Key Facts*. Available at: https://www.who.int/en/news-room/fact-sheets/detail/depression (Accessed on: 29 July 2019).

World Health Organization (2019) *Classification*. Available at: https://www.who.int/classifications/icd/en/ (Accessed on: 25 September 2019).

Zhao, J., Feng, Q., Wu, P., Lupu, R., Wilke, R., Wells, Q., Denny, J., Wei, W. (2019) 'Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction', *Scientific Reports*, 9. doi: 10.1038/s41598-018-36745-x.