Constructing Disease Trajectory Networks from Diagnosis and Prescription Patterns in Healthcare Data

Candidate Number: NHSC9[1]

MSc Data Science and Machine Learning

Supervisor: Prof. Anthony Hunter

September 2021

Abstract

Problematic polypharmacy is becoming an increasingly widespread issue, driven primarily by an ageing population. In this project we explore existing disease pathway modelling techniques and implement methods to construct disease trajectory networks using diagnosis and prescription data. In addition, we extend these methods by creating a novel algorithm which accounts for disease interactions by grouping multiple diagnoses/prescriptions into individual nodes in the networks. By revealing patterns of disease development in this way, our ultimate aim to assist with clinical decision making in the context of prescribing treatments to patients at risk of problematic polypharmacy. We also evaluate the stability of our networks and establish baseline results to be compared against in future work.

All code used in this project can be found at:
`https://anonymous.4open.science/r/disease-trajectory-modelling-D927/`

Acknowledgements

I would like to thank my supervisor, Prof. Tony Hunter, for his support and guidance throughout the course of this project. I would also like to express my gratitude to Prof. Reecha Sofat, Dr. Caroline Dale and Dr. Sandy Wright for their helpful comments and advice regarding the medical details of the project.

Contents

Chapter 1

Introduction

1.1 Motivation

In the UK the number of elderly people is growing rapidly. Within 50 years, the number of people in the UK aged 65 or older is expected to grow from 12.4 million, accounting for 18.5% of the population, to 19.8 million, accounting for 26.2% of the population [1]. This ageing population poses a number of challenges for the country, most notably with regard to the healthcare system. In particular, there are more and more people living with multiple long-term health conditions which necessitate numerous different treatments. This has meant there is an increasing prevalence of *polypharmacy*, which is generally defined as "the concurrent use of multiple medication items by a single individual" [2]. A study in Scotland [3] found that the proportion of adults receiving five or more drugs doubled between 1995 and 2010 and the proportion receiving 10 or more tripled. The study also revealed that the receipt of 10 or more drugs was "strongly associated" with increasing age.

Polypharmacy can in many cases be extremely beneficial but it can also be problematic, for example when medications are prescribed inappropriately. In these instances the treatments can often do more harm than good because there is an increased risk of hazardous drug interactions when using multiple medication items. This can put unnecessary strain on the healthcare system and even put lives at risk. Currently, a notable proportion of emergency hospital admissions are attributable to preventable adverse drug events [4, 5].

It is therefore crucial for clinicians to be aware of the trade-off between risks and benefits when prescribing multiple medications. However this can be a challenging task, particularly when the number of medications is large. From a computational perspective, the number of possible drug interactions is enormous and increases exponentially with the number of drugs. For this reason and others, very few clinical trials take into account polypharmacy or multi-morbidity and there is little medical research on these topics in a general setting. Thus the purpose of this work is to shed some light on key patient pathways in terms of their diseases and treatments by uncovering significant disease associations/trajectories in order to inform clinical decision making.

1.2 Objectives

Specifically, the aims of this project are to:

- Explore disease trajectory modelling techniques using diagnosis data and implement and evaluate methods to uncover paths of disease development at different levels
- Additionally apply these methods to a dataset of drug prescriptions to look at patterns in prescribing and to allow for comparison analysis between the diagnosis and medication trajectory models
- Extend these methods to incorporate disease/drug interactions by implementing a novel algorithm grouping multiple diagnoses/drugs into each 'node' of the network

These methods could ultimately help medical professionals to better assess risk when prescribing, particularly when a patient has numerous comorbidities or when there is a risk of problematic polypharmacy.

Chapter 2

Background and Problem Setup

In this chapter, we introduce and explain some key concepts such as network theory and relative risk which form major components of the methods used in this project. We then review previous relevant applications of these concepts to disease association and progression models. Finally, we provide a full description of our problem and how previous studies relate to it.

2.1 Networks

Significant technological advances over the last few decades have led to substantial increases in available computing power as well as the number and size of accessible datasets. This has allowed companies and researchers alike to model complex systems by recording and analysing connections between large numbers of entities. These increasingly complex and insightful networks have been produced across a wide array of disciplines from biology [6, 7] to social media [8].

But the basic structure of every network is the same. A *network* (or *graph* in mathematical contexts) can be defined as a collection of objects (called *nodes* or *vertices*) which are connected in some way through *edges* or *links* [9]. The main distinctions between different network types are:

- **Directed vs. undirected networks**: In directed networks, the connections between nodes have a specified direction which are usually represented visually using arrows. On the other hand, the edges in undirected networks have no specified direction.
- **Weighted vs. unweighted networks**: In weighted networks the edges have an additional attribute which can normally be represented by a single number. This can be interpreted as the *weight* or *strength* of the connection.

There are two main ways to fully specify a network with N total nodes. The first is using an *adjacency list* which is a collection of N lists, each of which contains the set of connected nodes (and weights for weighted networks) of a particular node in the network. The other way to represent a network mathematically is using an $N \times N$ *adjacency matrix* where the (i, j)-th entry is 1 (or the weight of the link in weighted networks) if there is a (directed) link between node i and node j and 0 otherwise.

2.2 Observational Studies in Medicine

The theory behind retrospective cohort studies plays a significant part in the methods employed in this project.

2.2.1 Observational Study Setup

Cohort studies, along with case-control studies and cross-sectional studies, are common types of medical observational study that yield 2×2 contingency tables [10]. For all three types of observational study, the setup, the way in which the data arises and the subsequent interpretation and analyses differ but the structure of the results table is the same. In particular, as depicted in Table 2.1, the results from such studies are reported as counts of observations at four different levels, representing the four different combinations of two binary variables. One of these variables is a *response* variable which often denotes resulting disease status. The other is an *explanatory* or *exposure* variable which indicates the presence or absence of some risk factor which is hypothesised to have an effect on the response variable. In clinical trials, for example, this can represent the application of treatment, or lack thereof, and so splits the patients into an intervention group and a control group. The goal is to determine the strength of the relationship between these two variables based on the distribution of counts across the four cells.

	Response variable	
Explanatory variable	Yes	No
Present	A	B
Absent	C	D

Table 2.1: Example 2×2 contingency table. A, B, C and D represent observed counts of patients across the four levels.

2.2.2 Cohort Studies

In cohort studies, unlike in clinical trials, the subjects of the study are not split into treatment and control groups prior to the administration of some treatment. Rather, the exposure variable represents the presence of some *pre-existing* condition or treatment and participants are followed up after a pre-defined amount of time and their disease status is recorded. At the start of the study, participants with the exposure variable present are often matched to other people using characteristics such as age, gender and/or ethnicity in order to reduce the possibility of effects from confounding variables.

The aim of a cohort study is to determine whether the risk of developing the disease of interest is impacted by the presence of the exposure variable.

Cohort studies can either be *prospective* or *retrospective*. The only difference between the two is the relative point in time at which the researcher begins the analysis. In prospective cohort studies, the cohorts are formed at the study inception and participants are followed up for disease status at a later date. Whereas in retrospective cohort studies the cohorts are assigned in the past

and the subsequent disease status is inferred from medical records produced before the start of the study.

2.3 Relative Risk

In the context of medical studies, *risk* is frequently used to denote the probability of some adverse outcome occurring [10]. In many cases the exact risk of an outcome is very difficult to acquire without data for the full population and often has to be estimated using sample data.

Relative risk (RR) is a form of *effect size* which measures the extent of association between the exposure variable and the response variable. It is defined as the ratio of the risk in exposed participants to the risk in unexposed participants.

The validity of using relative risk as a measure of effect size is highly dependent on the type of study conducted and has often been misinterpreted in medical literature [10]. In order to accurately calculate risk of disease for each of the exposed and unexposed groups, the proportion of participants with the disease compared to without the disease is required, by definition. However, in a case-control study, for example, where participants are sampled based on disease status and subsequently analysed for the presence of an exposure factor, the ratio of diseased to healthy participants is chosen by the researcher, by construction of the study. It is for this reason that the risks (and therefore the relative risk) cannot be calculated in a case-control study. On the other hand, since participants are selected based on exposure status in cohort studies, the relative risk can be correctly estimated in these types of study.

If the presence of exposure variable is denoted by E and the disease outcome is denoted by O, the relative risk is:

$$\mathrm{RR} = \frac{P(O|E)}{P(O|\neg E)}$$

In terms of estimating the RR from the results of a cohort study, using the same notation as in Table 2.1, the relative risk of this example is:

$$\mathrm{RR} = \frac{A/(A+B)}{C/(C+D)} = \frac{A(C+D)}{C(A+B)}$$

Relative risk can take values in $[0, \infty)$ and can be interpreted as follows:

- RR < 1 implies that the risk of the disease/outcome is reduced by the presence of the exposure factor
- RR $= 1$ implies that there is no relationship between exposure and disease/outcome
- RR > 1 implies that the risk of the disease/outcome is increased by the presence of the exposure factor

For example, a relative risk of 1.3 suggests that a participant with the exposure factor has 1.3 times the risk of developing the disease than a participant without the exposure factor.

2.4 Review of Related Works

Traditionally, epidemiological research and medical studies focused on analysing disease correlations are usually targeted in the sense that only a very small number of diseases are assessed and the temporal association between diseases is pre-defined. Only recently have more general studies taken place involving large numbers of diseases and where temporal correlations are uncovered empirically, taking into account the order in which the diseases are encountered in clinical care [11]. This is in part thanks to the increasing ability to construct large networks efficiently, as described in Section 2.1.

2.4.1 Disease Correlation Networks

There are a handful of papers which have studied this problem of constructing large disease networks in some detail [12, 13]. The approaches taken vary depending on a number of factors including computational power available, dataset size and type, study aim and the authors' preferences but the high-level structure of such models are similar. In essence, the problem can be separated into four sections:

1. Uncover correlations between diseases/conditions/diagnoses
2. Assign a direction to each correlation and filter out pairs without significant *directionality*
3. Construct network from resulting significant directed correlations
4. Analysis on the resulting network

In one paper, by Hidalgo et al. [13], one such approach that follows the above structure was taken. The aim was to further understanding of the complex origins of diseases and physiological failures from a phenotypic perspective to complement the numerous prior disease studies using genetic and proteomic datasets.

For the first step of their method (which corresponds to the first step of the above procedure), they introduce two measures to quantify the "distance" between two diseases. The first of which is a modified version of relative risk and the second is the ϕ-correlation (Pearson's correlation for binary variables). Both measures are functions of the number of patients affected by both diseases, the prevalence of each disease and the total number of patients in the dataset. Then, to determine the directionality of a connection between diseases i and j, they count the number of times disease i was diagnosed before disease j and vice versa in all patients who contracted each of the two diseases at some point. These counts are then normalised by the prevalence of each disease and a directionality variable introduced as the ratio of the logarithm of these normalised counts. Then, disease pairs with a directionality value of zero are defined as having no preferred direction but disease pairs with a directionality value either side of zero are defined as having a preferred direction. The further from zero this value is, the stronger this sense of directionality. The authors claim that the normalisation by prevalence is important as there is a large discrepancy in disease occurrences in the dataset used and the prevalence of disease i is proportional to the probability that it will be diagnosed before a different disease j.

In a more recent paper by Jensen et al. [12], a more comprehensive approach was taken. Their aim was to more intricately model disease progression patterns by forming full disease trajectories up to five diseases in length using 15 years of electronic health registry data from Denmark.

Their methodology for evaluating disease correlations has a stronger focus on temporally relevant associations but is more computationally complex than the approach taken by Hidalgo et al. as a result. To be precise, for each diagnosis pair D1 and D2, all patients with a discharge including a diagnosis of D1 are selected and counted. Then, the number of these patients who have a recorded discharge with a diagnosis of D2 within five years after the D1 discharge are counted. In order to account for disease prevalence, a number of comparison groups are formed by matching each D1 discharge with discharges of other patients which involve random diagnoses where they are matched on factors such as age, gender, ethnicity, discharge date and admission type. Then the number of these patients in each comparison group with a D2 discharge within five years of the relevant matched discharge is counted and averaged across all comparison groups. Therefore the correlation analysis for each diagnosis pair D1 and D2 is framed as a retrospective cohort study where the exposure variable is the presence of diagnosis D1 and the outcome variable is the appearance of diagnosis D2. And so a relative risk can be calculated for both directions of each diagnosis pair which accounts for its directed nature. Also, restricting the diagnoses to appear within five years of each other enforces a temporal correlation more directly and so makes it more likely that the resulting significant pairs are in some way related. A connection between two diagnoses occurring in the same patient but 50 years apart would not be picked up by this model, for example. In addition, a p-value is calculated as the proportion of comparison groups with a larger number of outcome-positive patients (ie. patients who have a diagnosis of D2 recorded within five years of the first diagnosis) than the original "exposure" group of patients with a D1 discharge.

Then, for diagnosis pairs with a relative risk greater than one and a significant p-value for either or both directions (D1 $\rightarrow$ D2 or D2 $\rightarrow$ D1), a directionality test is applied. Similarly to as in the method employed by Hidalgo et al., the number of times diagnosis D1 occurs before diagnosis D2 and vice versa is counted. But these two values are then each plugged into a binomial test with probability 0.5 and total number of trials $L_{\mathrm{D1}\rightarrow\mathrm{D2}} + L_{\mathrm{same}} + L_{\mathrm{D2}\rightarrow\mathrm{D1}}$, where $L_{\mathrm{D1}\rightarrow\mathrm{D2}}$ is the number of patients with D1 occurring first, $L_{\mathrm{D2}\rightarrow\mathrm{D1}}$ is the number of patients with D2 occurring first and L_{same} is the number of times they are both first assigned in the same discharge. P-values for this test can then be calculated and pairs are included in the model if the p-value for one of the directions is below a certain threshold (pairs can only be significant in one direction by construction of the binomial test). The potential downside to this directionality test is that disease prevalence is not accounted for.

In summary, the methods by Jensen et al. and Hidalgo et al. are structurally similar but do have notable differences. After the directionality tests and hence the formation of significant directed diagnosis pairs, both sets of authors constructed networks from these pairs. But beyond that the methods diverge: Hidalgo et al. studied questions such as how the network changes for patients of different ethnicities or genders and other questions related to the topology of the network such as whether diseases with more connections have higher mortality rates within a certain number of years. On the other hand, Jensen et al. looked at combining multiple pairs to form longer three-, four- and five-long disease trajectories to allow for analysis of longer disease progression patterns.

As for the methods used in this project, we drew on many of the approaches mentioned above and extended them further in a different direction. Full details of our methods can be found in Chapter 3.

2.4.2 Markov Models

Another popular class of models commonly used in disease progression studies is that of *Markov models*. Although ultimately we did not employ such models in our methods, they are related to the disease trajectory methods already discussed and their applicability is worth considering.

Markov models are state-based models characterised by a key property: the *Markov property*. The future behaviour of any process exhibiting this property is independent of past states and is solely dependent on the current state. This simplifying assumption dramatically reduces the computational cost of performing analysis on such models and as such it can be a desirable property for a model to possess.

One of the most frequently used type of Markov model found in epidemiological literature is the *hidden Markov model* (HMM). In hidden Markov models the states of interest are not themselves observable but instead there are a separate set of observable states which depend on the hidden states in some way. Such models are useful in modelling unobservable disease stages and estimating transition probabilities between them using some form of observable data. For example, Sukkar et al. [14] used hidden Markov models to model the progression of Alzheimer's disease in a more granular fashion than achieved previously by choosing six states of Alzheimer's disease of increasing severity but only measuring 4-dimensional data of brain readings. Similarly, Ceres et al. [15] used HMMs to model disease progression pathways in cows suffering from Johne's disease by recording faecal data.

This model framework is suited to the situation in which there is a single disease of which there are varying levels of severity. In this case the disease trajectory is linear. However, the Markov property is much less likely to apply in the more general setting where a large number of states from different diseases are included in the model. This is because any future disease path from a current state will probably be influenced by previous disease states. Therefore we do not consider Markov models as a viable method for the problem at hand, but they could certainly be utilised in future work which builds on the methods discussed in this report.

Next, we turn to describing the dataset available to us and explain the core challenges involved in this project.

2.5 Data

Before describing the dataset we used, we must first briefly explain the coding systems used to record diagnoses and prescriptions.

2.5.1 Diagnosis and Prescription Coding

ICD-10 Codes

In hospitals in England, ICD-10 codes are used to record diagnoses of diseases and symptoms in patients. The ICD (International Statistical Classification of Diseases and Related Health Problems) [16] is in essence a dictionary used to translate diagnoses into simple alphanumeric codes. These codes are arranged in a hierarchical structure, with each code falling within a *category* which falls within a *section* which falls within a *chapter*. The category is represented by the first three characters and additional characters reveal more granular details of the diagnosis. For example, for the code K35.2, in increasing levels of granularity,

- Codes K00-K93 represent Chapter XI: "Diseases of the digestive system"
- Codes K35-K38 represent Section: "Diseases of appendix"
- Code K35 represents Category: "Acute appendicitis"
- Code K35.2 represents: "Acute appendicitis with generalized peritonitis"

BNF Codes

Analogously to ICD-10 codes for diagnoses, there are BNF (British National Formulary) codes for drugs to be prescribed by GPs [17]. And similarly to ICD-10 codes, BNF codes are ordered hierarchically. In our dataset the BNF codes are eight characters in length, of which the first two identify the *BNF chapter*, the third & fourth identify the *BNF section* and the fifth & sixth identify the *BNF paragraph*.

2.5.2 Dataset

The dataset we used forms part of the Clinical Practice Research Datalink (CPRD) [18]. It is a large collection of anonymised primary care patient level medical records from general practitioners at practices all across the UK. There are records for more than 6.5 million patients who are roughly representative of the whole UK population in terms of their demographics. The dataset contains information on the medical histories of patients including all recorded symptoms and diagnoses at GPs as well as details of prescriptions in the form of BNF codes. For 4.6 million out of the 6.5 million patients in the dataset, there are also linked secondary care data records from the Hospital Episode Statistics (HES) database. These records contain information on hospital visits such as dates of admission and discharge, details and dates of certain procedures and recorded ICD-10 codes up to four characters in length. This dataset spans a 20-year period (1997-2017).

In particular, the dataset is split into a number of separate files. A summary of the most important files for our analyses is displayed in Table 2.2.

From a researcher's perspective, the size and breadth of this dataset is appealing as a large number of patients are accounted for and a rich picture of a patient's medical history can be built up using a number of metrics such as diagnoses, hospital visits and prescription items. Furthermore, the considerable temporal span of the data is impressive and allows for long-term studies to be conducted.

Title of CPRD table	**Total rows**	**Row-level data**	**Description**
Patient	6,529,382	One patient	Patient demographic information such as year of birth, gender and ethnicity.
Clinical	821,558,432	One medical event	Dates and details of recorded medical events at GP such as diagnoses, signs and symptoms.
Therapy	880,637,364	One prescription	Dates and details of prescribing data including treatment/drug type, quantity of prescribed product and number of days prescribed for.
Lookup_bnfcodes	2,367	One BNF code	Dictionary mapping prescription recordings in *Therapy* table to BNF codes.
Hes_patient	4,606,643	One patient	Basic patient data from hospital visits to link with *Patient* table.
Hes_hospital	27,621,427	One hospital spell	Details of hospital spells including admission and discharge dates, admission and discharge methods, admission source and discharge destination.
Hes_episode	31,721,865	One hospital episode	Dates and details of hospital episodes including corresponding hospital spell, order of episodes within spell, admission and discharge dates, discharge destination and method and specialty under which the consultant works.
Hes_diagnosis_epi	110,419,505	One recorded diagnosis	Recorded ICD-10 codes, each representing a diagnosis. Also includes corresponding episode and spell and the ordering of the diagnosis code within the relevant episode.
Hes_ons_death	831,181	One patient death	Data from the Office for National Statistics linking patient ID with date of death and cause(s) of death specified by ICD-10 code.

Table 2.2: Details of important data tables from the CPRD dataset used in this project. Includes name of table, total number of rows, what each row represents and description of table. Note that a hospital spell (or admission) refers to the entire continuous period of time a patient spends in a hospital trust. On the other hand an episode refers to the time spent under the care of a single consultant. Therefore one spell may be made up of multiple episodes.

However, the significant scope of the dataset posed an interesting challenge for us as the amount of processing power we had available was limited. We will discuss these limitations and how we overcame them shortly.

2.5.3 Data Safe Haven

Because of the sensitive nature of healthcare data, we were only able to access the CPRD dataset through UCL's "Data Safe Haven" (DSH)[1]. This is a "walled garden"-type technical solution used to store identifiable research data within a controlled environment. Access was granted to us through a remote desktop. All processing and analysing of the data had to also be carried out within the DSH and access to external websites or programs was prohibited.

We were able to access data science tools such as Python, R, MySQL and Microsoft Excel through the DSH but ultimately we were restricted in terms of the amount of available computing power.

2.6 Our Problem

The main challenge we faced was to create an efficient, reliable and robust disease correlation model subject to the constraints of both the dataset and the available compute in the DSH environment.

2.6.1 Computational Challenges

If we were to construct a model akin to those discussed in Section 2.4.1, then using the CPRD dataset we could form a directed weighted network with nodes as ICD-10 codes, arcs being the directed correlations between them and the arc weights being the relative risk of the ordered pair. Studying graph theory helped us to assess the feasibility of this. For any graph/network with N vertices, there are $\binom{N}{2} = \frac{1}{2}N(N-1)$ total pairs of vertices and hence also $\frac{1}{2}N(N-1)$ possible undirected arcs (not allowing for a vertex to connect to itself). If we allow for directed arcs then there are twice this many, $N(N-1)$, one for each of the two directions between a pair of vertices. Assuming any arc can be *on* or *off* in the network, that is it can either appear in the network or not, there are $2^{N(N-1)}$ different possible directed networks that can be formed using N vertices. This exponential growth is troublesome - see Table 2.3.

In the CPRD dataset, there are 11,699 unique ICD-10 codes. Therefore to form a directed network incorporating all these codes would require assessment of $11,699 \times 11,698 = 136,854,902$ connections. The difficulty comes in the assessment of these arcs. If we were to take the approach of Jensen et al., as described in Section 2.4.1, we would need to, for each of the 136,854,902 connections between diagnoses, firstly pick out the number of patients from the dataset who were diagnosed with the first diagnosis code in the pair at any point. The next step would be to count how many of these patients were diagnosed with the second diagnosis code in the pair within five years of the first diagnosis being recorded. Additionally, comparison patients matched on factors such age, gender, ethnicity and admission date would need to be found for each of the patients diagnosed with the first diagnosis code and this process of counting how many were diagnosed with the second diagnosis code within five years repeated for them. Finally, the test for directionality

[1] https://www.ucl.ac.uk/isd/services/file-storage-sharing/data-safe-haven-dsh

Number of vertices	Number of possible arcs	Number of possible graphs
2	2	4
5	20	1048576
10	90	$2^{90} \approx 10^{27}$
100	9900	$2^{4950} \approx 10^{2980}$
1000	999000	$2^{999000} \approx 10^{300729}$
10000	99990000	$2^{99990000} \approx 10^{30099989}$

Table 2.3: Number of possible directed arcs and number of possible directed graphs formed from certain numbers of vertices.

would need to be applied and the network constructed. The method by Hidalgo et al. is a similar but slightly simpler process. Unfortunately neither of these methods can be vectorised as they do not involve *straight-line* code. So even if this process could be completed in one second for each pair of diagnoses, completing the entire network would take roughly 792 days which is clearly far too impractical for our purposes.

2.6.2 Data Distribution

A further challenge with the dataset is that the distribution of prevalences of ICD-10 codes is highly asymmetric. Figure 2.1 shows the distribution of ICD-10 codes in terms of how often they appear in the *Hes_diagnosis_epi* table. For simplicity, these codes have been aggregated to category level (of which there are 2,047) by including only the first three characters of each code. The first 60 of these 2,047 ICD-10 categories account for 50% of all rows in the dataset. This severe imbalance reinforces the importance of accounting for disease prevalence in our methods.

Given that we had access to prescription data as well as diagnosis data, it makes sense that we might create an analogous model for drug prescriptions from GPs. In this way we created two separate networks which shed light on disease progression - one through the lens of diagnosed diseases/conditions and the other through the lens of prescribed drugs. It may be interesting to analyse the similarities and differences between the two graphs to check if one could reveal information about the other.

Prescribed drugs are identified by BNF codes in the dataset. We also encountered a similar phenomenon for BNF codes as we did for ICD-10 codes in that there is a large distribution imbalance - again a relatively small number of BNF codes represent a large proportion of the total prescriptions, as displayed in Figure 2.2.

2.6.3 Further Data Issues

Finally, there are a few issues we faced that are inherent to the data itself. Ultimately there is only so much information we can extract about disease progression from diagnosis and prescription data. The time at which a disease is diagnosed or medication is prescribed is unlikely to coincide perfectly with the actual contraction of the disease. In addition, there is no clear marker of the termination of a disease. In particular, the subsequent diagnosis of another ICD-10 code rarely

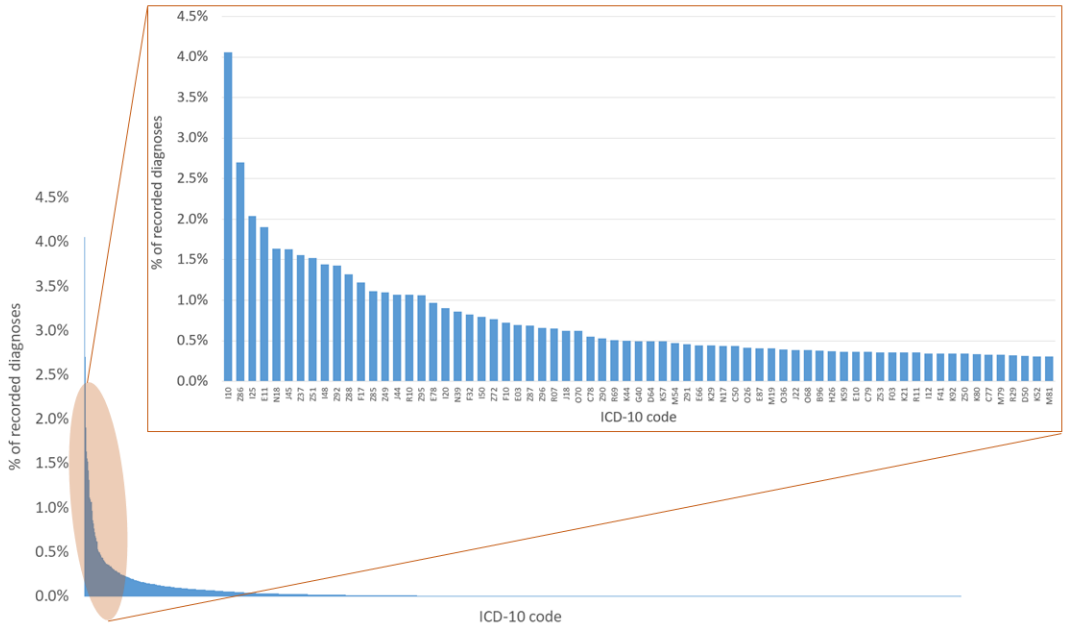


Figure 2.1: Distribution of ICD-10 codes appearing in the *Hes_diagnosis_epi* table in the CPRD dataset. Here the ICD-10 codes have been aggregated to category level (first three characters only).

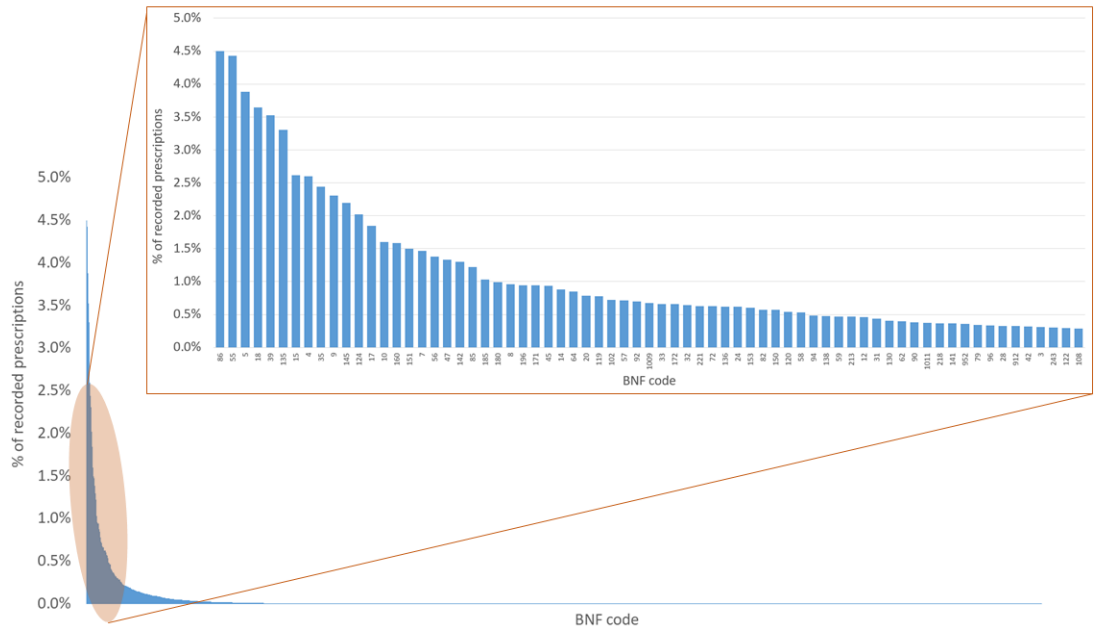


Figure 2.2: Distribution of BNF codes appearing in the *Therapy* table in the CPRD dataset.

indicates a definitive change of state of the patient from the first diagnosed disease/condition to the second, particularly since patients are usually diagnosed with several codes in a single episode. This reinforces that Markov models are unlikely to be applicable here. Further, there may be discrepancies in disease prevalence and diagnosis prevalence. For example, less serious diseases/conditions may rarely warrant hospital visits and so the diagnosis dataset may not pick up all incidences of these. Conversely, more serious conditions which almost certainly require hospital care such as cardiac arrest will be more likely to appear in the diagnosis dataset. This is similarly a problem in the prescribing dataset as patients with mild conditions may not always seek treatment. We can also expect there to be a few incidences of misdiagnosing or missing data as a result of human error, although the extent to which this occurs is purely speculative.

For some of these issues there was little we could do to alleviate them. But luckily they are mostly minor and in the next chapter we discuss methods to handle them and create robust diagnosis and drug correlation networks.

Chapter 3

Methods

In this chapter we describe the methods we chose to employ to evaluate disease correlations and drug correlations. This allowed us to construct trajectory networks and perform further analyses.

3.1 Software Used

Below are a list of the software and Python libraries we used, for reference:

- MySQL Workbench version 6.1.7.11891 build 1788
- Jupyter Notebook version 6.4.0 [19] running Python 3.8.5
- Anaconda version 4.8.3 [20]
- NumPy Python library version 1.20.2 [21]
- SciPy Python library version 1.6.2 [22]
- pandas Python library version 1.2.5 [23]
- pandasql Python library version 0.7.3 [24]
- NetworkX Python library version 2.5 [25]
- simple_icd_10_cm Python library version 1.0.4 [26]

3.2 Preprocessing

The first challenge we faced was that of deciding on the most pertinent data for our needs from the database and extracting it into a format which allowed us to process it.

3.2.1 Determining Appropriate Data Structure

We first focus on the model using ICD-10 codes and then follow-up with the preprocessing required for the BNF code model as it is similar. For the main model we chose to adapt the method in [12]

by Jensen et al. where the problem is framed as a retrospective cohort study in order to establish diagnosis pair correlations. Indeed, for this model, for any diagnosis pair D1 and D2, the first step involves counting the number of instances of patients being diagnosed with D2 within five years after being diagnosed with D1. We call this an instance of D1 $\rightarrow$ D2. However, before this, we needed to decide what in fact constitutes such an instance. There are a surprising number of possible approaches which produce different outcomes and invoke different medical interpretations. For example, consider the example diagnosis history of patients in Figure 3.1.

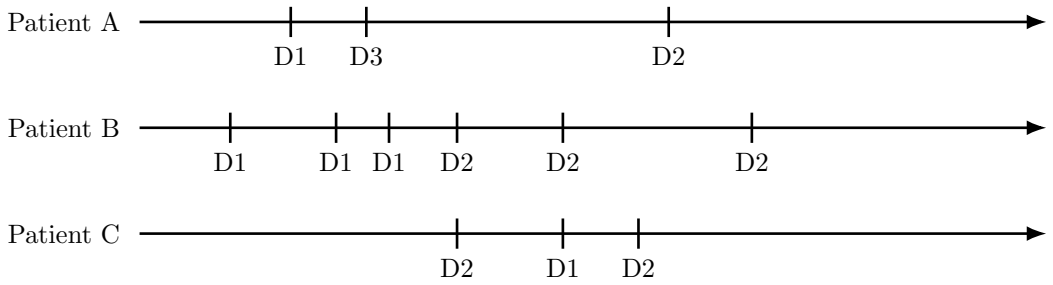


Figure 3.1: Example diagnosis histories for patients over a given five-year period. D1, D2 and D3 are recorded diagnoses in chronological order as they appear on the timelines from left to right.

For each of these patients in Figure 3.1, it is not clear how many instances of D1 $\rightarrow$ D2 should be counted. It is not sensible to enforce the stringent constraint that the D2 diagnosis must be the next recorded diagnosis after D1, not least because there are usually many diagnoses recorded in a single episode. So for patient A, there is a diagnosis of D2 within five years of a diagnosis of D1 so this counts as an instance of D1 $\rightarrow$ D2, despite the fact that a diagnosis of D3 was recorded in between.

For patient B, there are three diagnoses of D1 before three diagnoses of D2. Therefore the question is: how many instances of D1 $\rightarrow$ D2 should this count as? Counting every single pairing would result in $3 \times 3 = 9$ instances. However from a epidemiological perspective it is likely that, despite there being three reported diagnoses of each disease/condition, they only represent a single incidence of each disease/condition and so one could argue that this counts only as a single instance of D1 $\rightarrow$ D2. But for our method we would need to find comparison diagnoses occurring within a certain time period of the D1 diagnoses. So then if this were only to be counted as a single instance of D1 $\rightarrow$ D2, it is not immediately obvious which D1 diagnosis to choose of the three to match comparison diagnoses to.

And for patient C, there is a diagnosis of D1 within five years after a diagnosis of D2 and therefore this would count as an instance of D2 $\rightarrow$ D1. But there is another diagnosis of D2 after the diagnosis of D1. However, since in effect we are searching for directed disease pairs in the hope of uncovering causal relationships, it is perhaps not appropriate to count this also as an instance of D1 $\rightarrow$ D2 since it seems unlikely that D1 caused D2 in this patient given that a diagnosis of D2 was recorded first.

We first tried running our whole method (to be explained fully later in this chapter) by counting a D1 $\rightarrow$ D2 instance as whether one or more D2 diagnoses were present within five years of the *first* D1 diagnosis in each patient (if the patient is diagnosed with D1 at least once). However, this was suboptimal, for reasons to be discussed later in this chapter in Section 3.3.5.

Therefore we decided to filter the data by only including the first recorded instance of each diagnosis for each patient in the dataset. This is not a perfect solution by any means. If a patient contracted and was diagnosed with a disease, then recovered and subsequently contracted and was diagnosed with it again, the second disease occurrence is not found in the data and so is not picked up by the model. However, despite this, constructing the dataset in this way allows for reliable analysis by explicitly specifying the order in which diseases occur. Using this reasoning, from the diagnosis histories in Figure 3.1, patients A and B contribute one instance each of D1 $\rightarrow$ D2 but patient C does not because D2 first occurs before D1 and so the second D2 occurrence is removed from the dataset (patient C contributes an instance of D2 $\rightarrow$ D1 instead).

3.2.2 Preprocessing Pipeline

In order to select comparison patients/diagnoses, for every appearance of a diagnosis in the dataset, it is necessary to find other random diagnoses matched on patient year of birth, patient gender, patient ethnicity, type of hospital admission and date of diagnosis. For this task, we required tools to handle relational databases in order to connect the various tables, such as SQL. The approach we chose is to use MySQL Workbench to construct the necessary table and then export this as a CSV file to perform analysis in Python. The full preprocessing pipeline for the diagnosis-based model is displayed in Figure 3.2.

The preprocessing steps for the prescription-based model are very similar. For this model, we again used MySQL Workbench and took the *Therapy* table which contains the prescriptions data and grouped this by patient ID and BNF code, selecting the earliest prescription of each BNF code in each patient. Then, as in the diagnosis-based model, we joined the *Patient* table to extract patient demographic information (gender, year of birth and ethnicity). An additional difficulty with the prescription dataset is that the prescription codings in the *Therapy* table have to first be mapped to BNF codes using a lookup table. Next, we extracted the resulting table to CSV and opened it in Jupyter Notebook. We then used the BNF codes to create extra columns for BNF chapter, BNF section and BNF paragraph and filtered out certain BNF chapters which correspond to more specialised treatments rather than drugs. Finally, again we converted some textual columns to numeric codes in order for our models to run faster.

3.3 Disease Trajectory Models

Now that we have described how we extracted and formatted the data into a usable layout, we explain the details behind our models. First, we describe the diagnosis-based model before turning our attention to the prescription-based model.

As discussed already, we began by adapting the method in [12] by Jensen et al. However, our focus was different. Our ultimate goal was not to model and analyse longer three-, four- or five-long disease pathways but rather to extract significant directed correlations between diagnosis pairs and subsequently extend this model to include multiple diagnoses into each node of the network using a novel algorithm.

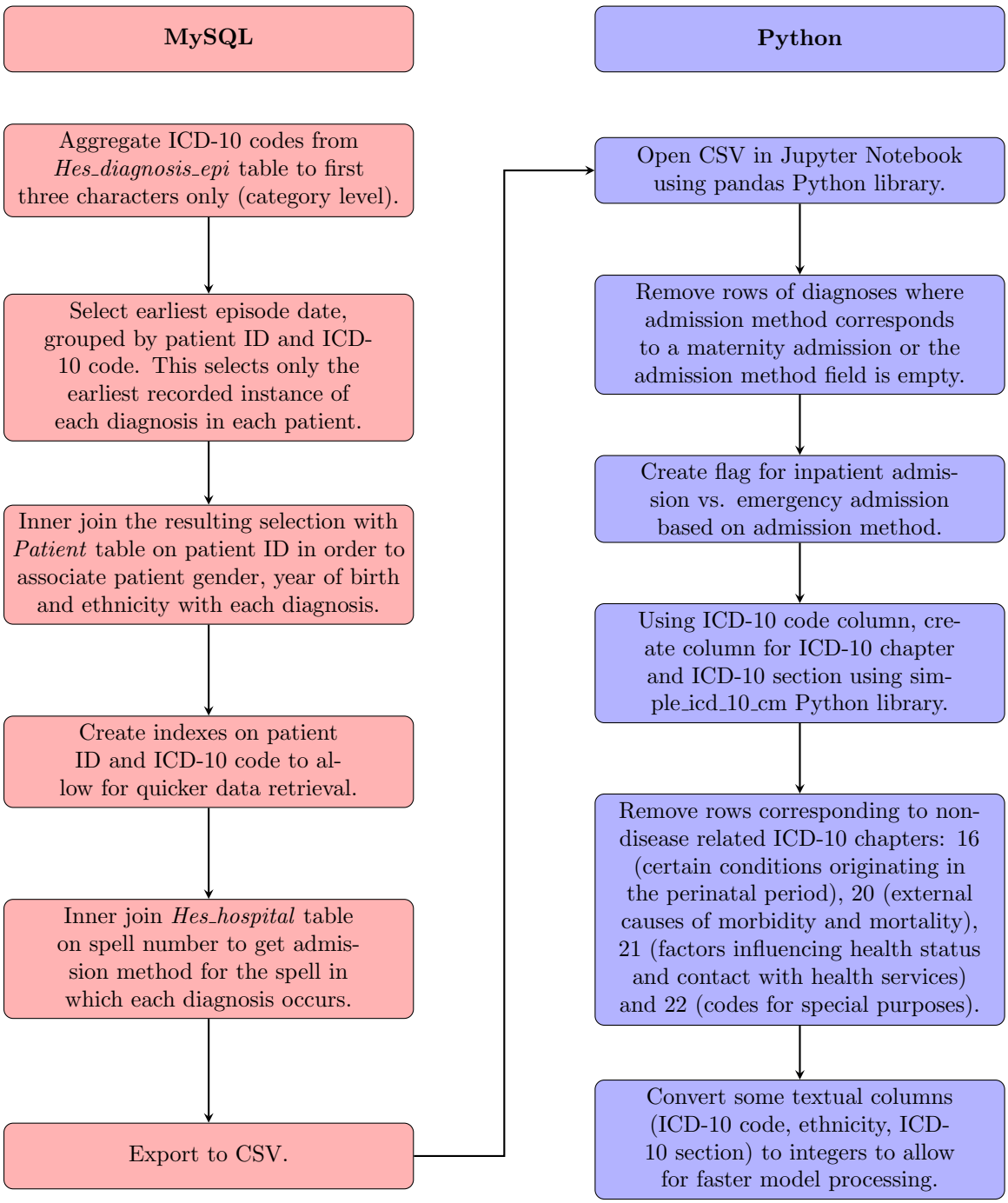


Figure 3.2: Full data preprocessing pipeline for the diagnosis-based model. The left-hand side of the flowchart (in red) is operations performed in MySQL Workbench. The right-hand side (in blue) is operations performed in Jupyter Notebook.

3.3.1 Adapting Data and Model to Handle Computational Constraints

Because of the significant technical limitations we faced as a result of working within the DSH, we were required to reduce the computational complexity of the problem in some way. Based on our analysis of network complexity in the previous chapter, we concluded that a network of 50 nodes strikes a good balance between providing opportunity for revealing interesting correlations from an epidemiological perspective and computational ease.

Next, we needed to decide what level of granularity to choose in terms of the ICD-10 codes. At the most granular level, using all four ICD-10 code characters, there are more than 10,000 different codes and so this was clearly too many - selecting only 50 out of these would not likely reveal many correlations of interest. The other options are ICD-10 category ($\sim 1,400$ unique codes), ICD-10 section (~ 200 unique codes) or ICD-10 chapter (18 unique codes). There is no 'correct' answer per se and ultimately it boils down to a question of disease ontologies, which is not the primary focus of this project. However we settled on ICD-10 section as the level of granularity is fine enough that most interactions between specific diseases would be identified by the model but also the size of each section is large enough that we could be sure that we would have enough data to run the model. The top 50 most common ICD-10 sections out of the 200 account for 74% of all recorded diagnoses, and so we were confident that the majority of meaningful disease correlations would be accounted for in our model. However, for the sake of completeness and in order to draw comparisons, we also ran our models using ICD-10 categories and ICD-10 chapters. All results and evaluations of our models can be found in Chapter 4.

Unfortunately, due to the large number of ICD-10 sections and the irregularities in the way they are coded, it was unfeasible to construct the mappings from ICD-10 code to section by hand. We therefore required the use of an external Python library. However, we could not find a library which provided such a mapping. We instead used the "simple_icd_10_cm" library which maps ICD-10-CM codes to their relevant sections. ICD-10-CM codes are slightly modified versions of ICD-10 codes used in the USA to record diagnoses. Most of the differences between ICD-10 and ICD-10-CM codes are at the most granular level and the differences between the mappings from the ICD categories to the most common 50 sections between ICD-10 and ICD-10-CM are very minimal. We therefore concluded that it was acceptable to proceed using these mappings.

By using ICD-10 sections, we distinguish our model further from the one created by Jensen et al. who used the more granular ICD-10 categories in their model. We hoped that our model might produce more interpretable correlations by not being so restrictive.

The final step required before we could feasibly run our model was to select only a subset of the data as we were unable to save CSV files anywhere near the size of the complete dataset and it would take an unworkable length of time to run our model using it. In the end we chose empirically a dataset of diagnosis records for 250,000 randomly selected patients. After preprocessing this data (following the steps in Figure 3.2), the resulting table contains nearly 2 million rows. The effectiveness of running our models using this number of patients is studied in Chapter 4.

3.3.2 Model Overview

As explained previously, we built on approaches already discussed in constructing directed disease correlations. We have already touched on parts of our method but now we explain it in full. Our

basic model can be broken down into two distinct parts:

1. Calculation of relative risk for all possible diagnosis pairs in both directions (arc weights in the network)
2. Establishing direction of pair, if there is one (arc directions)

3.3.3 Relative Risk Calculations

For the first step, we used theory from retrospective cohort studies to calculate relative risk. That is, for each directed diagnosis pair D1 → D2, we compared the proportion of D1 patients who were diagnosed with D2 within five years against the proportion of random patients who were diagnosed with D2 within five years. For example, the output we obtained by performing this calculation for ICD-10 sections I20-I25 (ischaemic heart diseases) → E70-E90 (metabolic disorders) is displayed in Table 3.1.

	E70-E90 diagnosis within five years?	
Patient Group	Yes	No
I20-I25 patients	5312	20368
Randomly matched patients	3707	20895

Table 3.1: Model results for I20-I25 → E70-E90

Therefore the relative risk for this pair in the direction specified is calculated as:

$$\text{RR} = \frac{5312/(5312 + 20368)}{3707/(3707 + 20895)} = 1.373$$

Therefore it appears that an I20-I25 diagnosis increases the risk of a subsequent E70-E90 diagnosis.

To construct Table 3.1 as well as the results for every other diagnosis pair, we used the "pandasql" Python library which allows you to run SQL-like code in Python. The procedure for obtaining the top row of the table is written in simplified pseudocode in Algorithm 1.

In order to calculate the proportion of randomly matched patients who were diagnosed with D2 within five years of their matched diagnosis, we first needed to obtain a comparison group of matched diagnoses. To do this, we iterated over each row in our dataframe and randomly selected another row, where possible, subject to the following conditions relative to the original diagnosis row:

- Same episode date +/- 30 days
- Same gender of patient
- Same year of birth as patient +/- 5 years
- Same ethnicity of patient
- Different ICD-10 section

```
D1_list := Top 50 most common ICD-10 sections;
D2_list := Top 50 most common ICD-10 sections;
for D2 in D2_list do
    Take whole dataframe and add flag for whether the patient in each row was diagnosed
     with D2 within five years after the episode date in the given row. Do this by using
     pandasql to join dataframe to itself on columns: patient ID, episode date (within five
     years) and ICD-10 section.;
    for D1 in D1_list do
        if D1 == D2 then
            Skip iteration;
        else
            D1_patients := Slice of dataframe where ICD-10 section == D1;
            D1_patients := Group D1_patients by patient ID and choose only earliest
             episode date for each;
            Total_D1_patients := len(D1_patients);
            Num_D1_to_D2 := len(D1_patients where D2 flag == 1);
        end
    end
end
```

Algorithm 1: Algorithm for calculating the top row of Table 3.1 for all possible diagnosis pairings (in both directions). For the example diagnosis pair, where D1 is I20-I25 and D2 is E70-E90, 5312 = Num_D1_to_D2 and 20368 = Total_D1_patients − Num_D1_to_D2.

- Different patient
- Same admission type (emergency or inpatient)

Matching on patient demographics ensures that unwanted effects from variations in disease prevalence across patient groups are minimised. We also stratified by episode date to reduce effects from seasonal diseases, as is done in [27]. Furthermore, the importance of matching on admission type is highlighted in [12] as the distribution of diagnoses differs significantly between inpatient and emergency admissions.

There were a small number ($\sim 1.4\%$) of rows where there were no possible matches, hence why the sum of the values in the bottom row of Table 3.1 is slightly smaller than the sum of the values in the top row. Where there was no possible match, we added a blank row so that each row in the original dataframe matches one-to-one with the matched dataframe (ie. row n in the original dataframe is matched with row n in the matched dataframe).

Then, in order to calculate the bottom row of Table 3.1 (and again for every other diagnosis pair), it was necessary to select the rows from the matched dataframe which correspond to the I20-I25 (or other relevant D1 diagnosis) patients and deduce how many were diagnosed with E70-E90 (or other relevant D2 diagnosis) within five years. To do this, we followed a similar procedure as in Algorithm 1 but with some modifications:

1. First, merge the original dataframe and the matched dataframe horizontally so that each row of the resulting combined dataframe contains the information for the diagnosis in the original dataframe as well as the information for the corresponding matched diagnosis.
2. Next, iterate over *D2_list* and for each diagnosis in this list, create a flag in the combined dataframe for whether the matched patient was diagnosed with D2 within five years after

the matched patient's episode date. Do this by joining the combined dataframe with the original dataframe where the diagnosis in the original dataframe is D2, patient ID in the original dataframe is joined to the matched patient ID in the combined dataframe and similarly episode date in the original dataframe is restricted to being within five years after the matched episode date in the combined dataframe.

3. Then, the rest of the procedure is more or less as before. Iterate over *D1_list* within the D2 for loop and select the slice of the dataframe where ICD-10 section for the original patients is D1. Then select the earliest episode per patient, filter out the rows with no matches and count the total number of patients and the number with the D2 flag.

In summary, the structure of the table created at each iteration through *D2_list* is shown in Figure 3.3.

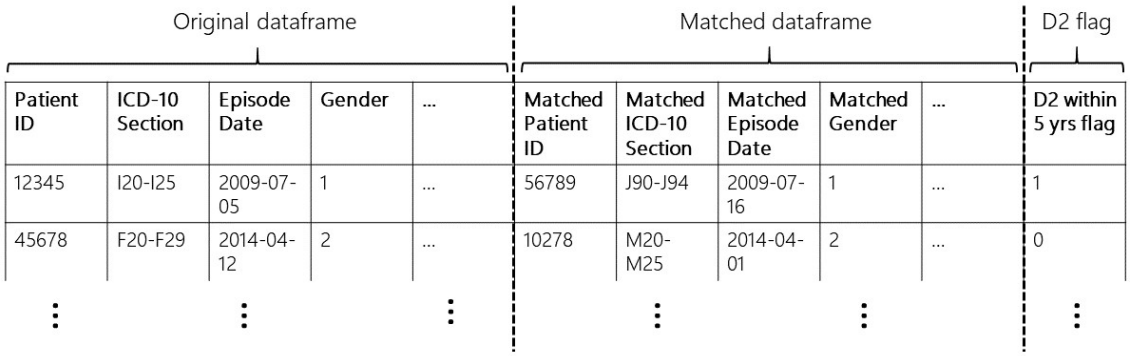


Figure 3.3: Structure of the table/combined dataframe created for each D2 diagnosis. The D2 flag column indicates whether the matched patient has a recorded diagnosis of the relevant D2 within five years of the matched episode date. Then the slice of the dataframe is taken where the 'ICD-10 Section' column is the relevant D1 diagnosis. So, for the example diagnosis pair in Table 3.1, D2 is E70-E90 and only the first row of the two in the above table would be selected because 'ICD-10 Section' is I20-I25. A similar table is created for the non-matched patients, as described in Algorithm 1, but there is no need for the matched dataframe and the D2 flag column refers to the patient in the original dataframe, not the matched dataframe.

This concludes the method for the relative risk calculations. We ended up with $50 \times 49 = 2,450$ values for the connections (arcs) which we then filtered down further by testing for directionality.

3.3.4 Directional Tests

At this point, using the methods discussed so far, we obtained values for relative risk for both directions of each diagnosis pair. These are useful in and of themselves to look at how one disease increases the risk of another disease. However, our aim was to go further and establish potential causal relationships between diseases. Further epidemiological analysis would need to be carried out on a case-by-case basis in order to prove such causal relationships but it is certainly possible to reveal strong indications of causality.

To do so, it is necessary to carry out directional tests. We described two such types of test in Section 2.4.1. We adapted one of these methods, from [12], which does not account for disease

prevalence but involves running two binomial tests for each pair, one in each direction. We proceeded as follows. For all diagnosis pairs D1 and D2, we calculated $L_{\text{D1}\to\text{D2}} + L_{\text{same}} + L_{\text{D2}\to\text{D1}}$ where $L_{\text{D1}\to\text{D2}}$ is the number of times D1 is diagnosed before D2 across all patients and vice versa and L_{same} is the number of times D1 and D2 are first diagnosed in the same episode. We then assumed the null hypothesis that each of D1 and D2 are equally likely to be diagnosed first. Each binomial test is one-sided with the 'number of trials' $N = L_{\text{D1}\to\text{D2}} + L_{\text{same}} + L_{\text{D2}\to\text{D1}}$ and 'probability of success' $p = 0.5$ and the values to be compared are the observed values of $L_{\text{D1}\to\text{D2}}$ and $L_{\text{D2}\to\text{D1}}$. Then if either of $L_{\text{D1}\to\text{D2}}$ or $L_{\text{D2}\to\text{D1}}$ were significantly large when compared to a binomial distribution with the aforementioned parameters, ie. we got a p-value below the significance level (where the p-value is defined as the probability of attaining a result equally as large or larger than the observed values from the given binomial distribution), then we rejected the null hypothesis that each of D1 and D2 are equally likely to be diagnosed first and assigned the pair the relevant direction. We expect that the assumption of independence required for a binomial test holds here as the diagnoses of one patient are highly unlikely to affect the diagnoses of another.

We chose a significance level of 0.05 as we felt this strikes an appropriate balance between providing enough statistical power and minimising the probability of a type I error (incorrectly rejecting the null hypothesis). However, because we ran many of these tests simultaneously, we encountered the problem of *multiple testing*. There are a number of methods to counteract this issue which involve controlling either the family-wise error rate (probability of making one or more type I errors) or the false discovery rate (proportion of rejections of the null which are wrong). We chose to use the Bonferroni correction which controls the family-wise error rate as it is the simplest to implement and is conservative. To implement this, we obtained the Bonferroni-corrected significance level by dividing the original significance level, 0.05, by the total number of tests we ran, 2,450. Then, after running the tests for all diagnosis pairs, the null hypothesis was rejected only for the tests which resulted in a p-value which was smaller than this corrected significance level.

For the sake of completeness, we also implemented the other type of directional test as described in Section 2.4.1 and compare the results of the two methods later in this report in Section 4.1.1 to further justify our choice.

3.3.5 Running Model on a Modified Dataset

We first tried running the methods explained so far on a different version of the dataset. Instead of only including the first of each diagnosis for each patient, we included all recordings of diagnoses. We still counted a D1 $\to$ D2 instance as whether a D2 diagnosis was present within five years after the first D1 diagnosis in each patient (if the patient is diagnosed with D1 at least once), but since all recordings of diagnoses are present in this dataset, the D2 diagnosis was not restricted to be the first appearance of that diagnosis in that patient's records. For example, patient C in Figure 3.1 contributes not only an instance of D2 $\to$ D1 but also an instance of D1 $\to$ D2 under this rationale.

However, upon evaluating the model using this dataset, we discovered that the average relative risk across all diagnosis pairs was 0.76 when in reality it should be approximately one. This is

because, for every diagnosis pair (D1 → D2) or (D3 → D4) or (D5 → D6) or ..., seemingly more comparison patients experienced a subsequent D2/D4/D6/... diagnosis on average than the exposure group of D1/D3/D5/... patients. We hypothesise that this is because for the first diagnosis in the pair (D1/D3/D5/...), we enforced that this was the first time the patient was diagnosed with that disease. However for the comparison diagnoses which we matched to each diagnosis, we simply selected another random diagnosis subject to some constraints (similar aged patient, similar date of diagnosis, same gender patient, same ethnicity patient, same admission type). But critically we did not restrict each comparison diagnosis to be the first diagnosis of that type in the comparison patient and so it is possible that comparison diagnoses were more often being selected in patients who were more ill and had more comorbidities. Therefore they were more likely to be diagnosed with the second diagnosis in the pair (D2/D4/D6/...) within five years after.

3.3.6 Completed Network and Visualisations

At this stage, now that we have detailed the methods behind our first model, we describe how we collated the results by selecting the diagnosis pairs which have a RR > 1 and pass the directionality test. We used the NetworkX Python library to visualise the network. The full network is shown in Figure 3.4. The arrows represent the directed arcs in the network and the relative risks are represented by the thickness of the arrows.

By visually inspecting this diagram, we notice that there are large differences in the number of arcs leading into or out of each node. Broadly speaking, it appears that most nodes fall into two categories: "source" and "sink" diagnoses. *Source* diagnoses are those where all, or almost all, arcs which are connected to that node point *out of* the node, whereas *sink* diagnoses are those where all, or almost all, arcs which are connected to that node point *into* the node.

By connecting significant diagnosis pairs together, we can form longer disease trajectories which can help to shed light on disease progression patterns. By doing so for a couple of example diagnoses, we can further illustrate the phenomenon of source and sink diseases, as displayed in Figures 3.5 and 3.6.

The materialisation of source and sink diseases from disease trajectory networks was first noted in [13]. In their model, Hidalgo et al. found that sink diseases tended to have higher "lethality" than source diseases, that is patients are more likely to die soon after being diagnosed with sink diseases. Such insights could be useful for clinicians deciding on treatment options for patients with unique diagnosis histories.

3.3.7 Prescription-Based Model

As discussed earlier, we also applied this methodology to construct analogous networks for prescription data. We chose to use the 50 most common BNF sections as nodes in the prescription-based model for similar reasons as to why we chose to construct the diagnosis-based model using ICD-10 sections as opposed to ICD-10 categories or ICD-10 chapters. That is, it is likely to be more useful and informative from an epidemiological perspective. But, again, we also ran this model using BNF paragraphs and BNF chapters as well to compare model performance and robustness, as explained in Chapter 4.

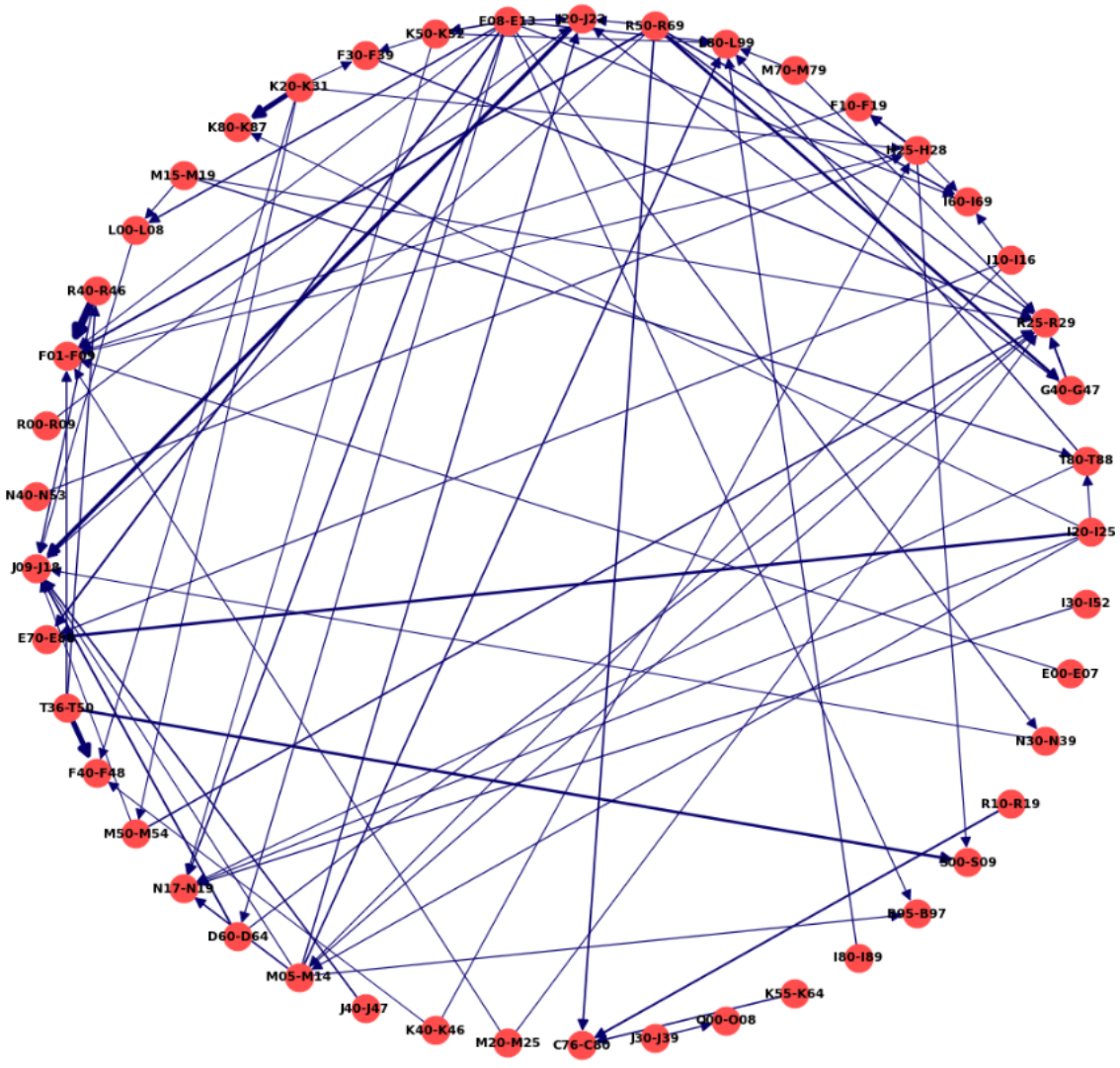


Figure 3.4: Full diagnosis-based model network. Nodes represent ICD-10 sections and arrows represent directed links between significant pairs which have RR > 1 and passed the directional test. The thickness of the arrows signify the relative risk of the directed pair. Nodes with no connecting arcs according to the above conditions not shown.

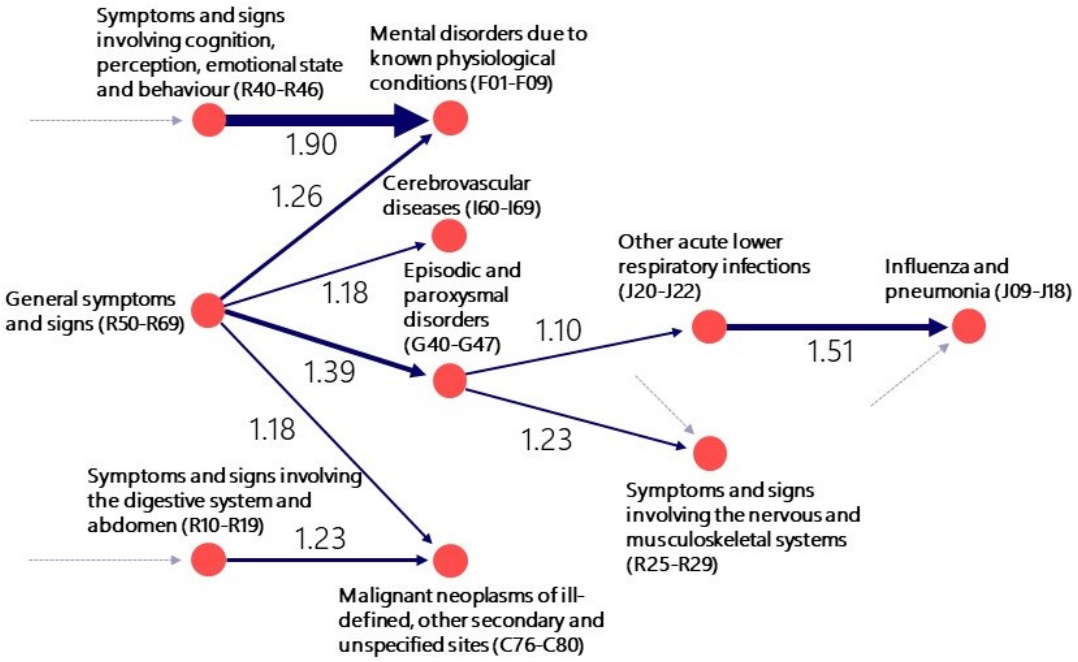


Figure 3.5: Subset of disease trajectories connected to R50-R69, a *source* diagnosis. The number beside each arc is the relative risk of the pair in that direction (also signified by the thickness of the arrow). Faded dotted arrows represent connections to/from elsewhere, not shown.

One of the only significant differences with the construction of this model compared to the diagnosis-based model is that there is one fewer condition on the matched patients since there is no counterpart to admission method in this dataset. However there are also further fundamental differences between the two datasets in that there are noticeably more prescriptions on average per patient than diagnoses. Therefore we chose this dataset to contain prescription data for 120,000 randomly selected patients, totalling over 1.7 million rows in length. The network for this model is displayed in Figure 3.7.

Compared with in the diagnosis-based model, in the prescription-based model there are more than three times as many pairs which passed the directional test. We hypothesise that this is because drugs are prescribed by health professionals by choice as opposed to diagnoses which occur naturally and hence not by choice. Therefore, drug prescription patterns are more likely to be structured. For this reason we obtained a number of much stronger connections, as can be seen by visually comparing the thickness of the lines in Figures 3.4 and 3.7. Note that in Figure 3.7 we only include arcs with a RR > 1.2 because including all arcs with a RR > 1 (as is done in Figure 3.4) makes it too difficult to distinguish individual arcs. We again observe the phenomenon of source and sink nodes in the prescription-based model, suggesting that these nodes are representative of source and sink diseases, as in the diagnosis-based model. This could be confirmed by further analysis.

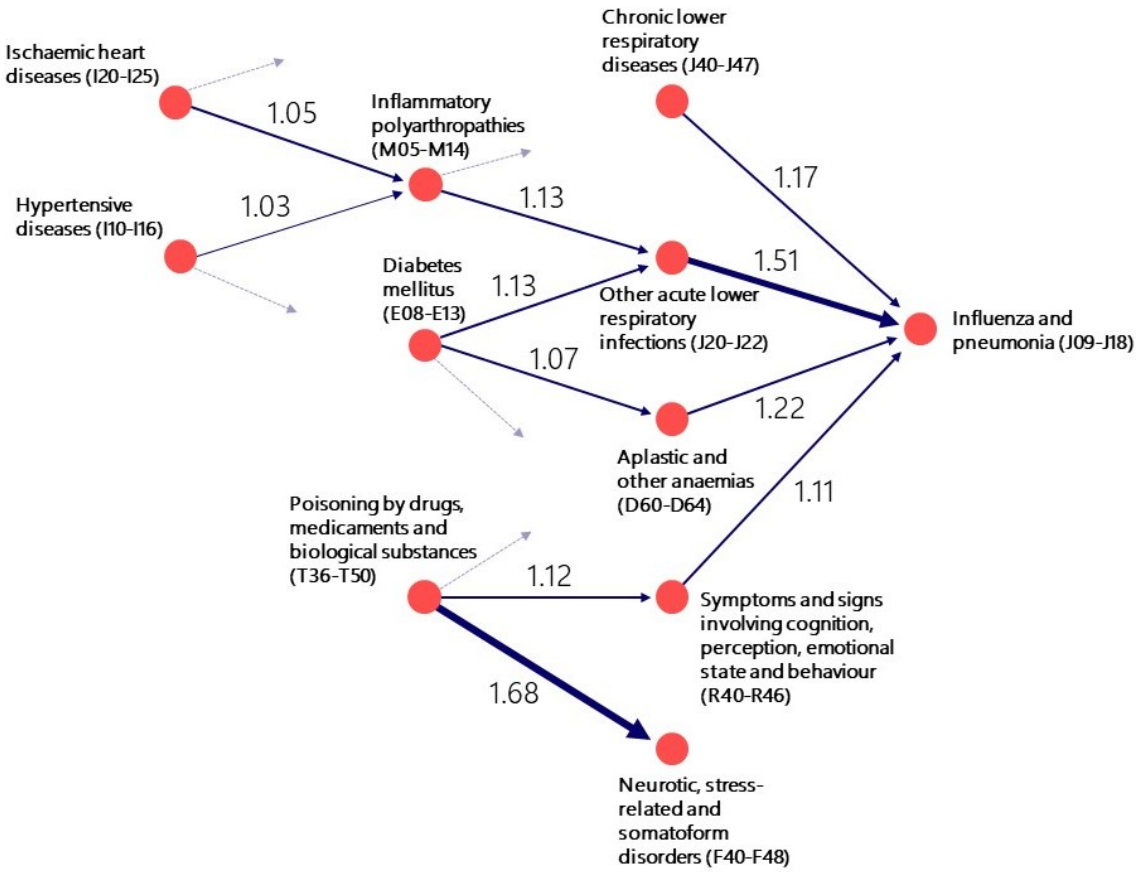


Figure 3.6: Subset of disease trajectories connected to J09-J18, a *sink* diagnosis. The number beside each arc is the relative risk of the pair in that direction (also signified by the thickness of the arrow). Faded dotted arrows represent connections to/from elsewhere, not shown.

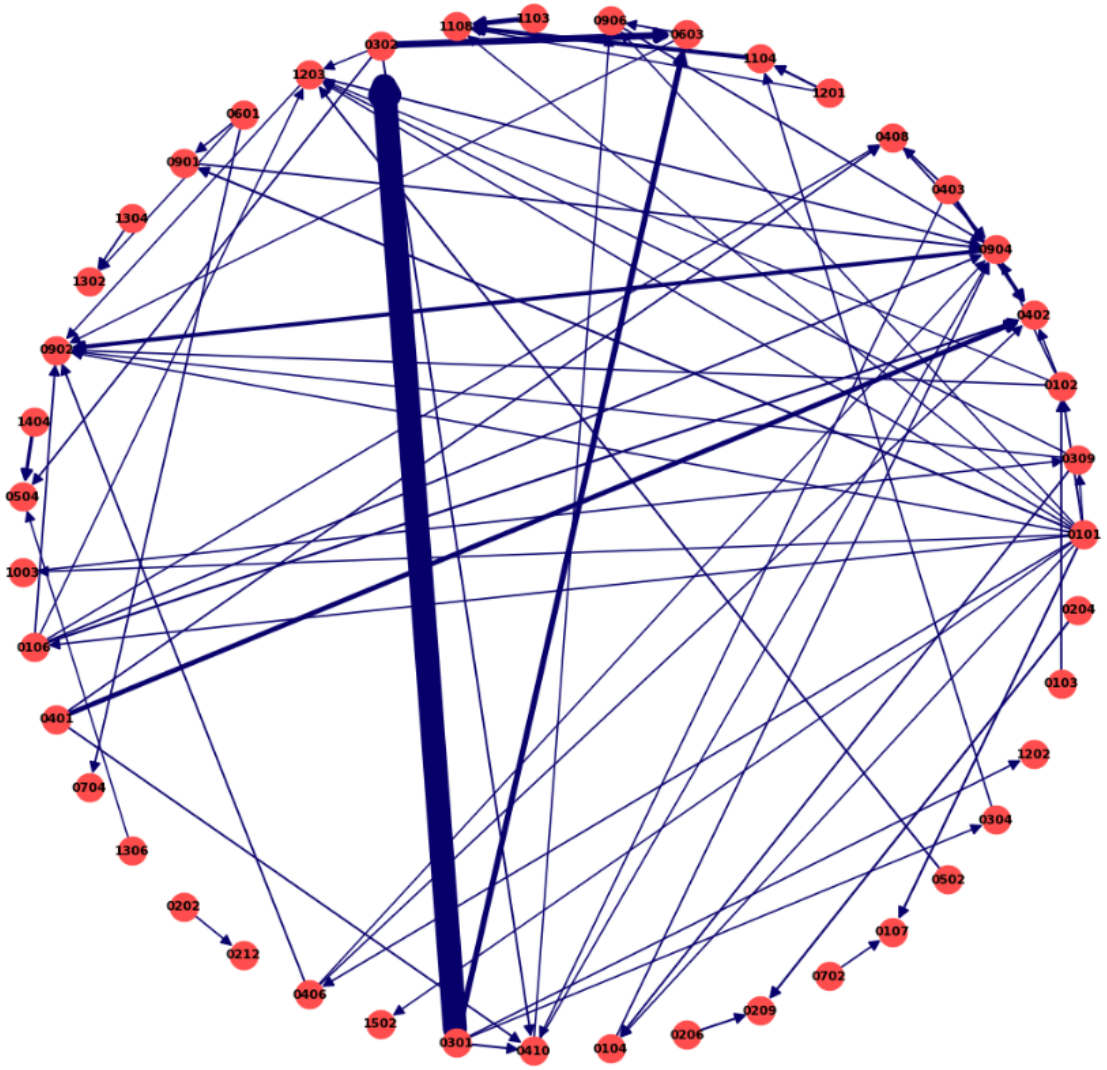


Figure 3.7: Full prescription-based model network. Nodes represent BNF sections and arrows represent directed links between significant pairs which have RR > 1.2 and passed the directional test. The thickness of the arrows signify the relative risk of the directed pair. Nodes with no connecting arcs according to the above conditions not shown. The thickest connection is from BNF section 0301 (bronchodilators) to BNF section 0302 (corticosteroids (respiratory)) with a RR of 4.58. This reflects a common treatment pathway for certain respiratory problems.

3.4 Multi-Diagnosis/Multi-Prescription Models

We now return to our diagnosis-based model. The main motivation for this project is the rising incidence of problematic polypharmacy. Therefore a focus on diagnosis or drug combinations is required. The problem of increasing complexity of graphical models as the number of nodes increases was discussed earlier in this report. However, if we want to look at combinations of diseases, that is combine two, three, four, or more diseases into a single node, the number of possible disease combinations grows quickly with the number of nodes, and so constructing such graphs is an even more difficult task. For a dataset of 50 different diseases, a model with x diseases per node has $\binom{50}{x}$ disease combinations of size x and hence $\binom{50}{x}$ nodes, as shown in Figure 3.8.

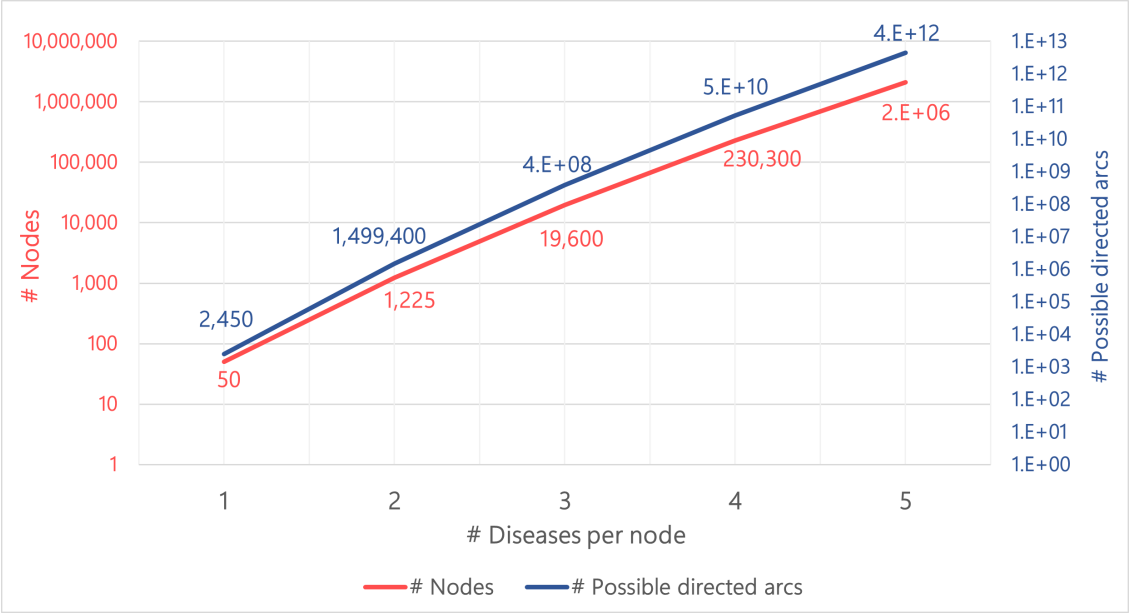


Figure 3.8: Graph comparing how network complexity scales with the number of diseases per node for a total of 50 diseases. Both y-axes use logarithmic scales base 10.

In addition, we would need to use a larger dataset to obtain enough instances of patients contracting each of the possible disease combinations. Therefore, creating such models under the constraints of our working environment was not feasible. Furthermore, these types of model are not so useful from a medical perspective. Rather than many-to-many correlations, it is arguably more valuable to study many-to-one correlations. For example, given that a patient has diseases D1, D2 and D3, calculating the risk of them subsequently contracting disease D4 is likely to be more insightful than calculating the risk of them contracting diseases D4, D5 and D6 simultaneously.

Taking all of this into account, we designed a novel algorithm which uses the single-diagnosis model as a filter to select the key correlations to extend into multi-diagnosis models. We chose a relative risk value to use as a cutoff in order to decide which arcs in the single-diagnosis model to extend to the multi-diagnosis models. Next, for each of the outgoing nodes connected to these arcs, we found the most common comorbidities, that is the most commonly diagnosed diseases in patients with at least one diagnosis of the disease in the outgoing node. Then, for the two-diagnosis model, if, for example, pair D1 $\rightarrow$ D3 passed the filtering step and D2 is one of the most common

comorbidities of D1, we performed the following procedure:

1. Select all instances of patients contracting D1 and D2 within a year either side of each other. If a patient contracted both diseases within this time frame more than once, then select the first instance. Do this using pandasql in Python by joining the diagnosis dataset to itself, joining on patient ID, ICD-10 sections D1 and D2 and episode dates within one year of each other.

2. Taking the patients selected above, use a similar method to that in the first step of Algorithm 1 to add a flag for if these patients were diagnosed with D3 within five years after the contraction of D1/D2 (whichever was first).

3. Select the diagnoses matched to the diagnoses in step 1. (match to the earlier diagnosis out of D1 and D2 for each).

4. Repeat step 2. for the matched patients/diagnoses ie. count how many of them were diagnosed with D3 up to five years after the matched diagnosis.

5. Calculate the relative risk of D1 + D2 $\rightarrow$ D3 in the usual way ie. the proportion of D1/D2 patients who were subsequently diagnosed with D3 within five years divided by the proportion of matched patients who were subsequently diagnosed with D3 within five years.

We iterated through this procedure for all possible triplets D1, D2 and D3, where D1 and D3 are respectively the outgoing and incoming nodes/diagnoses from the arcs which passed the filtering step (ie. the single-diagnosis model) and D2 is a common comorbidity of D1. And we easily extended this to three- and four-diagnosis models by respectively incorporating two or three comorbidities of D1. The only major difference is that in step 1., we selected patients who were diagnosed with all three or four diseases within one year instead of just the two diseases.

By performing this procedure, we provide more granular insights into the highest risk disease connections. But this is a general method to shed light on disease interactions which can be applied to any group of diagnoses or indeed BNF codes as we also applied the exact same methodology to the prescription-based model, but we leave out its description for the sake of brevity; the architectural differences between the two models are inconsequential.

In the models we implemented, we used a relative risk cutoff of 1.25 and chose the top five comorbidities for each outgoing diagnosis. It is difficult to visualise all the results at once so we instead show in Figure 3.9 the results obtained from extending one example diagnosis pair J20-J22 $\rightarrow$ J09-J18 to the multi-diagnosis models.

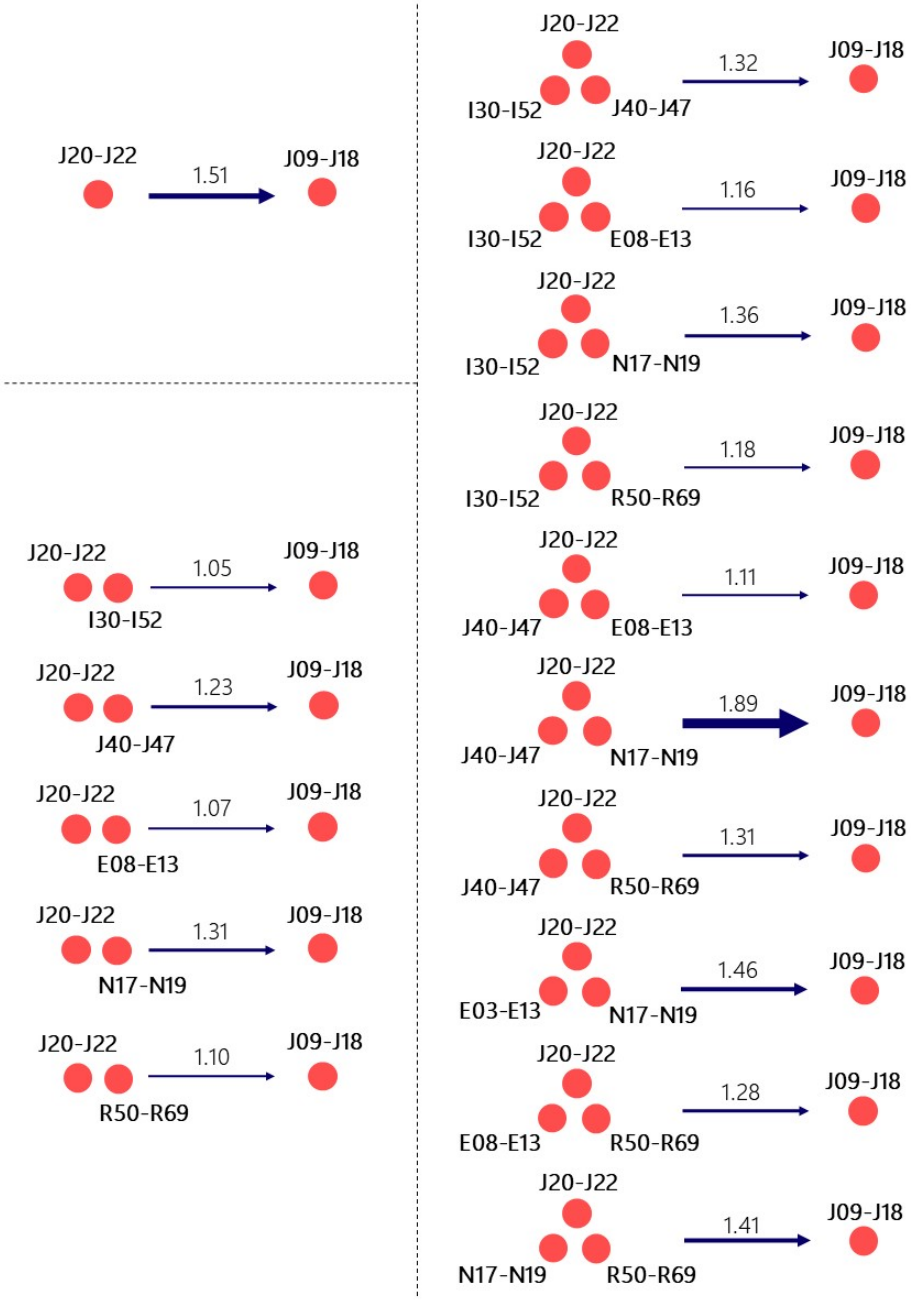


Figure 3.9: **Top left**: relative risk of J20-J22 → J09-J18. **Bottom left**: results from multi-diagnosis model with two diagnoses per node using all combinations of the five most common comorbidities of J20-J22. **Right**: results from multi-diagnosis model with three diagnoses per node using all combinations of the five most common comorbidities of J20-J22.

Chapter 4

Evaluation

Now that we have explained our methods in detail, we turn to evaluating their effectiveness and robustness. Because of the uniqueness of this project, there are no standard evaluation metrics for our models. This only reinforces the importance of evaluation for us, especially given the constraints of our working environment. Therefore, we both systematically compare our models with alternative models in the literature as well as devise a number of our own evaluation methods.

4.1 Comparisons with Alternative Methods

4.1.1 Alternative Directional Tests

In addition to the test for directionality we chose to implement in our models, as described in Section 3.3.4, we also applied another method, adapted from [13]. This involves defining a directionality variable,

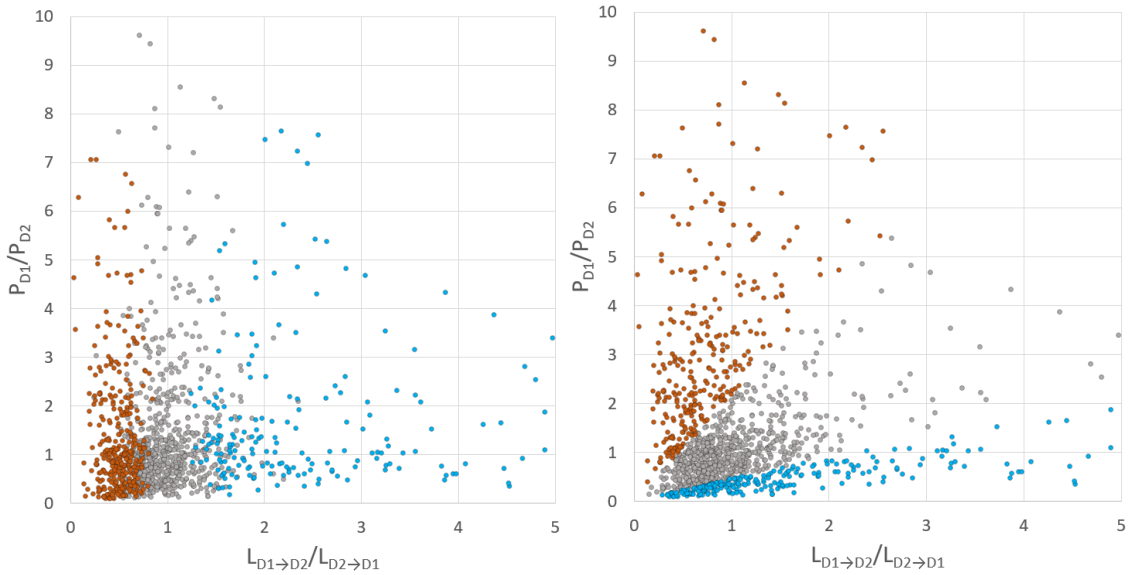
$$\lambda_{\mathrm{D1}\to\mathrm{D2}} \coloneqq \log_{10}\left(\frac{l_{\mathrm{D1}\to\mathrm{D2}}}{l_{\mathrm{D2}\to\mathrm{D1}}}\right)$$

where $l_{\mathrm{D1}\to\mathrm{D2}} \coloneqq L_{\mathrm{D1}\to\mathrm{D2}}/P_{\mathrm{D1}}$ and where $L_{\mathrm{D1}\to\mathrm{D2}}$ is the number of times D1 was diagnosed before D2 across all patients and P_{D1} is D1 prevalence (ie. the total number of appearances of D1 in the dataset). Then, the pair D1 and D2 is defined as having a significant direction if $|\lambda_{\mathrm{D1}\to\mathrm{D2}}| > x$, for some cutoff x, and the direction is defined by the sign of $\lambda_{\mathrm{D1}\to\mathrm{D2}}$.

The decision as to which directionality method to choose ultimately boiled down to whether we felt disease prevalence should be incorporated into this calculation. We already accounted for disease prevalence in the calculation of the relative risks due to the fact that we framed the process as a retrospective cohort study, but this does not necessarily imply that we should or should not do so for this directionality calculation.

The results from the two methods on the dataset of diagnoses are displayed in Figure 4.1.

From these results, we see that, as expected, in method A (the main method we implemented) the results of the directionality test are not dependent on diagnosis prevalence. This is in contrast to method B where there are a number of diagnosis pairs D1 and D2 where the ratio $L_{D1\to D2}/L_{D2\to D1}$ is below one (ie. D2 is diagnosed first more often out of the pair) and yet the significant direction is deemed to be D1 $\to$ D2 because D2 is more prevalent than D1. And



(a) Method A (implemented method)

(b) Method B (alternative method). Cutoff value of 0.33 for directionality variable chosen so that the same proportion of pairs pass the directionality test as in method A.

Figure 4.1: Comparison of results between different tests for directionality. Each point represents a diagnosis pair D1 and D2. The x-axis of each graph is the ratio of the number of times D1 is diagnosed before D2 to the number of times D2 is diagnosed before D1. The y-axis of each graph represents the ratio of diagnosis prevalences. Blue points represent diagnosis pairs which are deemed to have the specified direction D1 $\to$ D2 according that test. Orange points represent diagnosis pairs which are deemed to have the specified direction D2 $\to$ D1. Grey points represent diagnosis pairs which are not deemed to have a specified direction.

similarly vice versa for D2 → D1.

Ultimately, we deemed it unnecessary, and in some cases counterintuitive, to account for diagnosis prevalence at this stage. This was because diagnosis prevalence tends to overwhelm the proportion of times one disease is diagnosed before the other. For any disease/diagnosis pair, if one of the pair is much more prevalent than the other, then it is practically impossible for the directionality test to result in anything other than (less prevalent disease) → (more prevalent disease) being the significant direction. For example, for the disease pair E10-E14 (diabetes mellitus) and L00-L08 (infections of the skin and subcutaneous tissue), there are more than twice as many patients who were first diagnosed with E10-E14 before L00-L08 as opposed to the other way around. However, because E10-E14 is nearly five times more prevalent in the dataset, the sign of the directionality variable is such that it indicates that the direction of the pairing is reversed. However, it is well known that infections are in general more common and severe in patients with diabetes [28] and so it is more likely that E10-E14 → L00-L08 is the correct direction.

One argument for accounting for diagnosis prevalence is that some diseases are more difficult to diagnose and so may not be identified as often or as early as they should. However, the ratio between how often a disease is diagnosed and how often that disease occurs is very difficult to deduce without performing analysis on a case-by-case basis. Therefore we felt that the diagnosis of a disease before another is the best indication we have of directionality or possible causation, regardless of diagnosis prevalence. Therefore, we chose the binomial test method to assess directionality in our final model.

4.1.2 Alternative Matching Patient Methodology

The model used in [12] was run at a more granular level (ICD-10 category) than our diagnosis-based model and included a preprocessing step in order to highlight key diagnosis pairs to be fed into their full model. We adapted this preprocessing filter and used it instead as a method of evaluation to compare the results between our method and this method. The crucial difference between this method and our own is how the calculations involving the comparison patients/diagnoses were performed.

This method takes the form of a binomial test and frames the sampling of each matching diagnosis as a Bernoulli trial. For each diagnosis pair D1 → D2, rather than randomly selecting a single comparison diagnosis to match to *each* D1 diagnosis and estimating the relative risk from this, using this method we randomly selected a number, C, of D1 diagnoses and, for each of these C, counted the number of *possible* matching patients n_{match} as well as how many of these patients were diagnosed with D2 within five years, $n_{\text{match}\to\text{D2}}$. So for $i = 1, ..., C$ and for each D1 → D2 pair, we obtained the following probability of sampling a matching diagnosis with a D2 diagnosis within the five-year window:

$$P(D2)_i = \frac{n_{\text{match}\to\text{D2},i}}{n_{\text{match},i}}$$

We could then average these probabilities and use this as an estimate for the probability parameter in a one-sided binomial test:

$$P(D2)_{\text{test}} = \frac{1}{C}\sum_{i=1}^{C} P(D2)_i$$

For each diagnosis pair D1 $\rightarrow$ D2, we tested whether the number of D1 $\rightarrow$ D2 patients was significantly large given the number of D1 patients and the above probability parameter. This is equivalent to testing whether the proportion of D1 patients who were subsequently diagnosed with D2 within five years of their D1 diagnosis is larger than the proportion of matched patients who were subsequently diagnosed with D2 within five years of their matched diagnosis. This is effectively what we did in the method we employed in our main model but the difference between the two methods comes in this estimation of the proportion of matched patients who were subsequently diagnosed with D2 within five years of their matched diagnosis. We expect that the independence assumption required for binomial tests holds here as we assume that the diagnoses of each patient are independent of the diagnoses of other patients.

Before performing these binomial tests though, it was necessary to decide on a value for C. In order to choose C, we needed to balance run-time with statistical power. For each D1 $\rightarrow$ D2 pair, we were effectively trying to estimate the population mean using a sample mean, $P(D2)_{\text{test}}$ (taking C to be the total number of patients with a recorded D1 diagnosis would give us the population mean). The sample mean is an unbiased estimator of the population mean and the central limit theorem implies that sample means are normally distributed around the true mean for large enough sample sizes. In general, sample sizes greater than 30 are considered sufficient and so this is an approximate lower bound for C needed to use this normal approximation. We chose to create a 95% confidence interval for the population mean. In order to create such a confidence interval we needed to first approximate the population variance. We did so by choosing five random diagnosis pairs D1 $\rightarrow$ D2 and, for each of these five pairs, we selected 100 of the D1 patients and calculated the number of possible matching patients to each of the 100 and the number of these matching patients who were diagnosed with D2 within five years to calculate the required proportions, $P(D2)_i$. This gave us five samples of 100 proportions from which we could calculate the sample variance of each. Since we chose the same value of C for all diagnosis pairs for simplicity, we selected the largest sample variance σ_s^2 from the five which produces the most conservative confidence interval. Then, the z-score for a 95% two-tailed confidence level is 1.96. But using a Bonferroni correction to adjust for multiple testing of 2,450 pairs, it was necessary to adjust the level of the confidence interval to $100 \times (1 - \frac{0.025}{2450})\% = 99.999\%$. The resulting confidence interval is $+/- 4.26\sqrt{\sigma_s^2/C}$.

We found that the largest sample variance from our testing was 0.000162. We also found using $C = 50$ resulted in a acceptable running time. This resulted in a 95% confidence interval for the population mean of approximately $+/- 4.26\sqrt{0.000162/50} = +/- 0.00767$ which we concluded was reasonable from empirical testings. It is important to highlight that this is not a thoroughly comprehensive methodology but rather serves as a first pass to compare the two methods in case of unexpected differences before more extensive model assessments in the next section.

Now looking instead at the method we implemented in our models, we calculated relative risks rather than p-values, but it is straightforward to return probabilities instead, from which it is possible to obtain p-values. We did this by simply taking the denominator of the RR calculation (number of patients from the comparison group who were diagnosed with D2 within five years of

the comparison episode date divided by the size of comparison group). We could then similarly plug these into the binomial test as the probability parameter to deduce p-values which can be compared with the alternative method described above. We chose a conservative significance level of 0.001 in order to protect against false positives and used the Bonferroni correction for multiple testing. The results comparing the probability parameters obtained in the two methods are shown in Figure 4.2 and the resulting confusion matrix is shown in Table 4.1.

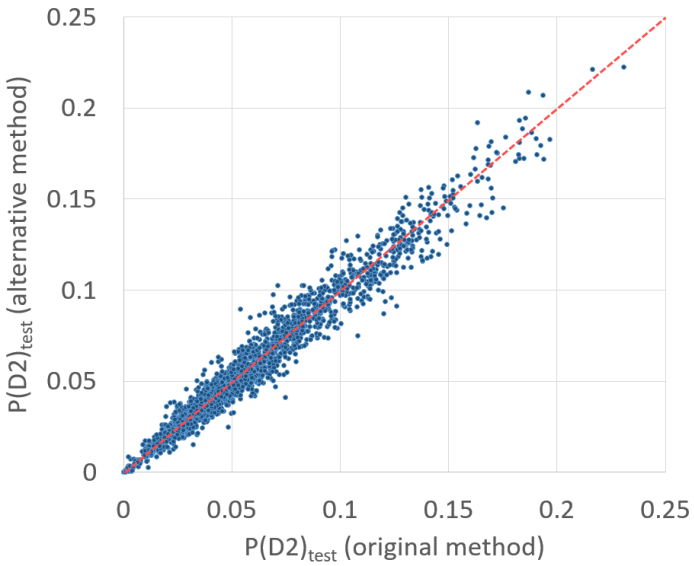


Figure 4.2: Graph comparing original and alternative methods for estimating the proportion of patients matched to a D1 diagnosis who subsequently go on to be diagnosed with D2, for each pair D1 → D2. Red dotted line represents the optimum where the two proportions are equal.

		Alternative method		
		Reject null	**Retain null**	Total
Original method	**Reject null**	**201**	**43**	244
	Retain null	**184**	**2022**	2206
	Total	385	2065	2450

Table 4.1: Matrix comparing results of one-sided binomial tests to test whether there is a significant difference in proportions of (D1 → D2) patients vs. (matched → D2) patients. The difference in the two methods comes from the calculation of (matched → D2) proportions. For both methods the null hypothesis is that there is no difference between the proportion of D1 patients who were subsequently diagnosed with D2 within five years of their D1 diagnosis and the proportion of matched patients who were subsequently diagnosed with D2 within five years of their matched diagnosis.

Upon visual inspection of Figure 4.2 it appears that the two methods are mostly in agreement, with some slight deviations from the red dotted line. This is confirmed by Table 4.1 which displays an "agreement" of 90.7%. However, we find that the original method is more conservative, rejecting the null hypothesis in 244 out of the 2,450 total diagnosis pairs compared to the alternative method

which rejects the null in 385 pairs. We decided to use the original method in our model because we felt using a more conservative method would be preferable as it would highlight only the more significant disease correlations and restrict the number of false positives. In addition, we can produce a relative risk for each pair using this method, giving us an idea of the strength of the connection.

Finally, analysing the results from both these methods highlights the importance of using a different (matched $\rightarrow$ D2) proportion for each (D1 $\rightarrow$ D2) pair to compare to the (D1 $\rightarrow$ D2) proportion. By using a single fixed (matched $\rightarrow$ D2) proportion, for example the average (matched $\rightarrow$ D2) proportion obtained across all diagnosis pairs using our original method, we obtained more than three times as many significant diagnosis correlations than when using our original method. Doing so in essence ignores disease prevalence and means that highly prevalent diseases are much more likely to be included as D2 or 'destination' nodes whereas rare diseases are less likely to be included. This is particularly important given the large discrepancies in diagnosis prevalence, as shown in Figure 2.1.

4.2 Stability Analysis

Now we display the results from a more comprehensive evaluation of our models. We ran our models on a number of datasets of different sizes and explored a number of metrics which inform us of the quality of the models as well as the stability/variance of the models by indicating how much the model would be expected to change were it run on a different subset of the data. We performed this analysis on both the prescription-based model and the diagnosis-based model. We also ran our models at different levels of granularity (ICD-10 category vs. section vs. chapter for the diagnosis-based model and BNF paragraph vs. section vs. chapter for the prescription-based model) in order to analyse whether performance varies noticeably at this level. The evaluations in this section apply to both the single-diagnosis/single-prescription based models and the multi-diagnosis/multi-prescription models.

4.2.1 Evaluation Metrics

The metrics we chose to use in our evaluations are detailed below.

Percentage of patients with a possible match

By calculating the percentage of patients in the dataset for which a matching patient could be found, we could evaluate whether we are using a large enough dataset to effectively run our models. Clearly the optimal value is 100%. By having too few matched/comparison patients, our results are less certain and the confidence intervals for the relative risks grow, as explained below.

Average RR 95% CI width

In order to evaluate how certain we are in our estimates of relative risk, we needed to create confidence intervals for them. We could not create such confidence intervals directly because the relative risk is not normally distributed. However, from [29], the natural logarithm of the relative risk is approximately normally distributed and hence we could construct confidence intervals for

RR through a logarithm transformation. Using the same notation as in Table 2.1, the standard error of $\log(RR)$ is:

$$SE(\log(RR)) = \sqrt{\frac{1}{A} - \frac{1}{A+B} + \frac{1}{C} - \frac{1}{C+D}}$$

The lower (L) and upper (U) bounds for the logarithm of the relative risk can then be calculated as follows:

$$L = \log(RR) - z_{1-\alpha/2} \times SE(\log(RR))$$
$$U = \log(RR) + z_{1-\alpha/2} \times SE(\log(RR))$$

where $z_{1-\alpha/2}$ is the z-score from the standard normal distribution for the $100 \times (1-\alpha)\%$ confidence level.

Then the lower and upper bounds for the $100 \times (1-\alpha)\%$ confidence interval for the relative risk are e^L and e^U, respectively. We again needed to adjust the confidence interval to account for multiple testing, as explained previously.

By reporting the average width of the 95% confidence interval across all diagnosis/prescription pairs, we get an idea of how certain we are in our estimates.

Percentage of RR CI's that do not contain 1

We complement our reporting of the average width of the relative risk confidence intervals with the proportion of these intervals which do not contain 1. We aimed to have as few confidence intervals as possible contain 1 so that we can be satisfied that the presence of the first diagnosis/prescription in each pair either increases or decreases the risk of the second diagnosis/prescription in the pair.

Average distance between networks

We also explored possible network comparison methods in order to establish the stability of the networks. The problem of comparing networks is strikingly non-trivial. Not only is it highly domain dependent, but also requires a trade-off between interpretability, computational efficiency, and effectiveness [30]. There are a vast number of different network comparison techniques, but these can be broadly separated into two categories:

- **Known node-correspondence (KNC) methods:** These methods require that the two networks have the same size and structure. That is, the two networks are comprised of the same set of nodes, or there exists a correspondence between nodes that is one-to-one.
- **Unknown node-correspondence (UNC) methods:** These methods do not enforce such strong constraints and allow for comparisons between any two graphs of any size or structure. Thus typically these methods compare higher-level differences in global structure.

Our goals were to compare multiple networks using the same node set but on different subsets of the data in order to test model stability. Hence, we decided to employ KNC methods. From [30], the decision as to which KNC methods are suitable also depends on whether the network is directed or not and whether it is weighted or not. For directed, weighted networks, as is our case,

the applicable methods all involve computing the difference between the adjacency matrices using a matrix norm. After evaluating a number of distance functions, we chose to use the *weighted Jaccard distance* [31] because of its interpretability as well as the fact that it does not scale with the size of the matrices, unlike in the case of Euclidean distance or Manhattan distance, for example.

We first define the *Jaccard distance* as it is simpler and leads naturally on to the weighted Jaccard distance. For two unweighted networks with adjacency matrices A and B, the Jaccard distance between the adjacency matrices $d_J(A, B)$ is:

$$d_J(A, B) = 1 - J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

where $J(A, B)$ is defined as the *Jaccard Similarity* between A and B which captures the amount of overlap between two sets by dividing the intersection of the two sets by their union. Adjacency matrices for unweighted networks contain only ones and zeroes and the intersection and union are performed element-wise on these values.

The weighted Jaccard distance $d_{WJ}(A, B)$ for two weighted networks with adjacency matrices $A = [a_{ij}]$ and $B = [b_{ij}]$ is defined instead as:

$$d_{WJ}(A, B) = 1 - J_W(A, B) = 1 - \frac{\sum_{i,j \in V} \min(a_{ij}, b_{ij})}{\sum_{i,j \in V} \max(a_{ij}, b_{ij})}$$

where V is the set of nodes in the networks and $J_W(A, B)$ is the *weighted Jaccard Similarity* between A and B. The weighted Jaccard distance takes values in $[0, 1]$. The weighted Jaccard Similarity is one for identical matrices and hence the weighted Jaccard distance is zero. Conversely, the greater the difference between each of the entries in A with the corresponding entries in B, the closer the weighted Jaccard distance comes to one.

Percentage overlap between top 3/5 highest RR connections

The most pertinent outputs from our models are the highest RR values we obtained as these represent the strongest and most meaningful connections between nodes. We therefore include an additional metric which evaluates the level of agreement between the networks we ran in terms of the highest RR connections for each node. Between two networks of interest, we looked at the average percentage of overlap between the sets of the top three or five highest RR connections for each node.

For example, if for the ICD-10 category model, the top three highest RR connections to some node were G40, E08 and J09 in any order and upon running the model on a different dataset the highest RR connections to that same node are E08, N80 and J09 then the overlap between the two sets is $\frac{2}{3}$ or 66.7%. This is averaged over all nodes.

4.2.2 Results

Main Results

The results from our experiments and the reported values for all but the last of the metrics listed above can be found in Table 4.2 for the diagnosis-based model and Table 4.3 for the prescription-

based model. For each version of our model listed in the tables, we ran it four additional times on separate datasets of the same size and calculated the average weighted Jaccard distance between these networks and the relevant original network to provide us with a clear indication of the stability of our models. This is the metric reported in the final column of each of these tables.

As expected, performance of our models improves across all metrics as dataset size increases. It is important to note that the first column, 'Total dataset size', is solely dependent on which patients are included in the dataset and indeed the number of patients selected (ie. is independent of the granularity level). The differences across granularity levels in this column are due to the fact that the models were run on different subsets of the data (ie. using different patients). Furthermore, the percentages in the 'Percentage of patients with a possible match' column tend to decrease slightly as granularity decreases. This is because, for the diagnosis-based model for example, the ICD-10 categories/sections/chapters for the matched diagnoses were restricted to be from different categories/sections/chapters respectively to the diagnoses to which they were matched. Finding a matched diagnosis from a different chapter is more restrictive than finding a matched diagnosis from a different section, for example. Therefore, if both models were run on the same dataset, the number of possible matches for the ICD-10 section model would be an upper bound on the number of possible matches for the ICD-10 chapter model.

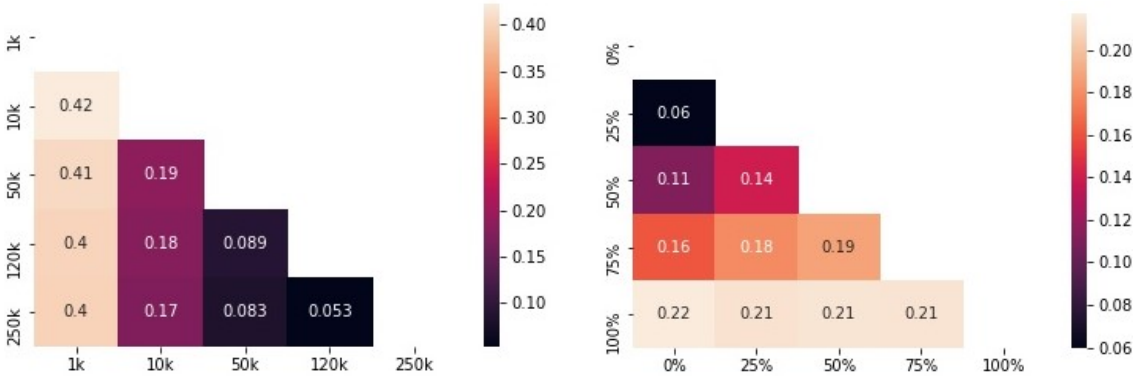
The 'chapter' models for both the diagnosis-based and prescription-based models contain fewer nodes and hence fewer arcs than the other models. There are 18 ICD-10 chapters which passed our preprocessing steps and are included in our models and only 15 BNF chapters. In comparison, all other networks contain 50 nodes. As a result, for a given dataset size, there is more data available to calculate the relative risks across each arc in the network for the 'chapter' models. For example, for the dataset of 250k patients for the diagnosis-based model, there are, on average, 3.7 times as many patients per chapter than per section (for the top 50 sections) and 81.4 times as many patients per chapter than per category (for the top 50 sections). This results in confidence intervals being much tighter for the ICD-10 chapter model.

From these results tables, we see that using only a dataset of 1,000 patients to run our models is not sufficient to produce reliable networks. In particular, as for the diagnosis-based model using 1k patients, only marginally more than 50% of rows in the dataset had a possible match. This resulted in the average 95% RR confidence interval width being larger than 1 which is far too large for our purposes since we expect the vast majority of arcs in the network to have RR values between 0.5 and 1.5. All other dataset sizes we ran our models on resulted in more than 90% of rows with available matches which is a significant improvement. However, only the ICD-10 chapter model run on datasets of 120k and 250k patients and the BNF chapter model run on a dataset of 120k patients resulted in more than 50% of confidence intervals not containing 1.

Furthermore, in Figures 4.3 and 4.4 for the ICD-10 section and BNF section model respectively, we compare the weighted Jaccard distances between these models run on different sized datasets as well as with a proportion of the rows randomly permuted, for reference. We discover that for the ICD-10 section model with 250k patients, the weighted Jaccard distance between this network and the same network with 50% of the rows randomly permuted is just 0.11. And the weighted Jaccard distance between the network and the same network with 100% of the rows randomly permuted is 0.22. In comparison, from Table 4.2, the average weighted Jaccard distance between the ICD-10 section model run on separate datasets of 250k patients is surprisingly only slightly smaller than

ICD-10 code granularity and number of patients	Total dataset size	Percentage of patients with a possible match	Average RR 95% CI width	Percentage of RR CI's that do not contain 1	Average distance* between networks of this type
Category 1k	8111	50.10%	7.337	0.082%	0.4667
Category 10k	78095	92.03%	2.483	1.347%	0.3749
Category 50k	393794	95.38%	1.245	3.837%	0.2366
Category 120k	943257	97.39%	0.819	7.184%	0.2012
Category 250k	1817711	98.60%	0.586	13.347%	0.1889
Section 1k	8209	50.12%	5.589	0.286%	0.5050
Section 10k	80949	91.85%	1.913	0.980%	0.3074
Section 50k	395511	95.19%	0.948	4.735%	0.2311
Section 120k	944704	97.36%	0.623	15.796%	0.2176
Section 250k	1817711	98.56%	0.457	27.469%	0.2066
Chapter 1k	8436	50.53%	2.834	0.327%	0.4130
Chapter 10k	77944	91.69%	1.021	2.614%	0.1553
Chapter 50k	389146	95.08%	0.504	29.412%	0.0704
Chapter 120k	944580	97.22%	0.325	50.000%	0.0442
Chapter 250k	1817711	98.49%	0.234	58.170%	0.0416

Table 4.2: Results from diagnosis-based models run using different numbers of patients and at different levels of ICD-10 code granularity (category, section, chapter in order of decreasing granularity).
*'distance' here refers to weighted Jaccard distance.



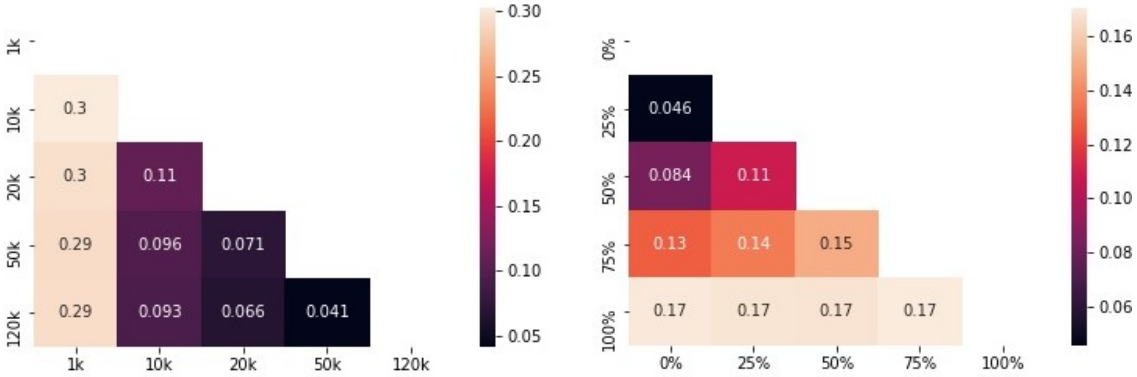
(a) Networks constructed using different numbers of patients.

(b) 250k patient network with a percentage of the rows randomly permuted.

Figure 4.3: Matrices comparing weighted Jaccard distances between networks constructed using ICD-10 sections data.

BNF code granularity and number of patients	Total dataset size	Percentage of patients with a possible match	Average RR 95% CI width	Percentage of RR CI's that do not contain 1	Average distance* between networks of this type
Paragraph 1k	15515	83.99%	3.117	1.061%	0.4141
Paragraph 10k	155329	95.73%	1.117	2.939%	0.1951
Paragraph 20k	310060	97.24%	0.786	5.306%	0.1734
Paragraph 50k	776492	98.63%	0.496	11.102%	0.1619
Paragraph 120k	1724782	99.34%	0.332	20.490%	0.1508
Section 1k	15805	85.28%	2.994	0.939%	0.4125
Section 10k	155124	95.47%	1.059	2.694%	0.2041
Section 20k	311563	97.29%	0.761	5.878%	0.1817
Section 50k	780146	98.63%	0.482	14.939%	0.1682
Section 120k	1724782	99.32%	0.325	27.755%	0.1575
Chapter 1k	15098	84.01%	1.312	0.000%	0.1731
Chapter 10k	156529	95.85%	0.410	6.191%	0.0686
Chapter 20k	307211	97.27%	0.294	13.810%	0.0710
Chapter 50k	778900	98.53%	0.183	40.000%	0.0604
Chapter 120k	1724782	99.29%	0.123	56.191%	0.0621

Table 4.3: Results from prescription-based models run using different numbers of patients and at different levels of BNF code granularity (paragraph, section, chapter in order of decreasing granularity).
*'distance' here refers to weighted Jaccard distance.



(a) Networks constructed using different numbers of patients.

(b) 120k patient network with a percentage of the rows randomly permuted.

Figure 4.4: Matrices comparing weighted Jaccard distances between networks constructed using BNF sections data.

this at 0.21. This indicates that the model is not particularly robust to variations in the data. We notice a similar trend in the prescription-based model. There, we see from Figure 4.4b that the weighted Jaccard distance between the BNF section model with 120k patients and the same model with all rows permuted is 0.17. Whereas, from Table 4.3, the average weighted Jaccard distance between the BNF section model run on separate datasets of 120k patients is only slightly lower at 0.16. In comparison, both 'chapter' models using 250k patients for the diagnosis-based model and 120k patients for the prescription-based model appear to be much more stable, with an average weighted Jaccard distance between networks of these types of 0.04 and 0.06 respectively.

Our analysis so far has suggested that the models implemented using ICD-10 and BNF sections are not stable enough to provide reliable results consistently. However, these metrics take into account all possible arcs in the network. Instead, it is perhaps more logical to focus on the most important arcs in the network.

Analysis of Highest RR Connections

We calculated the percentage overlap between the top 3/5 highest RR connections for each node as described earlier between each version of our model and the networks created in four additional runs of the same model on different subsets of the data. The results of this are displayed in Table 4.4. This informs us from a slightly different angle of how robust the networks are to being run on different datasets, for a number of dataset sizes. And we also calculated this metric between the networks run on different sized datasets for the ICD-10 and BNF section models in Figures 4.5a and 4.5b respectively.

From these results, we again see, as expected, that the networks run using only 1,000 patients perform poorly with regard to these metrics. However, as dataset size increases, performance improves dramatically. For our main models of interest, the ICD-10 and BNF section models with 250k and 120k patients respectively, we find that there is nearly 75% agreement on average across these models when run on different datasets in terms of the top 3/5 RR connections between nodes. This suggests that, despite relatively poor performance in terms of the metrics described in the previous section, these models do exhibit some degree of consistency in as much as they agree on the most important correlations the significant majority of the time.

4.2.3 Results Summary

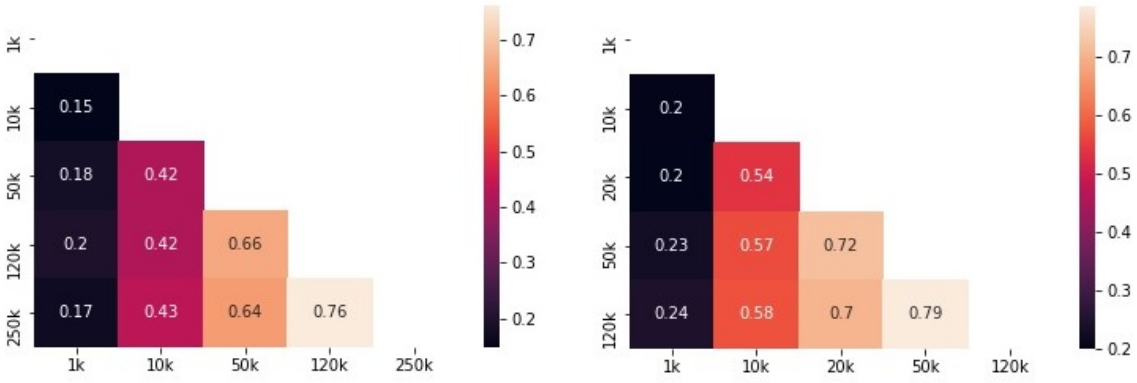
Overall, our results have shown that, despite demonstrating reasonable stability on specific metrics, our 'section' models (which are our models of interest) would ultimately need to be run on a significantly larger dataset in order to provide consistently reliable results. The 'chapter' models, on the other hand, exhibit much better performance across all metrics. This is likely due to the fact that there are more patients per chapter than per section. Therefore, we estimate that a diagnosis dataset of approximately 925k patients would provide the same number of patients per ICD-10 section on average as there are patients per ICD-10 chapter in the diagnosis dataset of 250k patients. And a prescription dataset of 540k patients would provide the same number of patients per BNF section as patients per BNF chapter in the prescription dataset of 120k patients.

These results may serve as baselines for future models run on the same or even different datasets or datasets comprising more patients.

ICD-10 code granularity and no. patients	% top 3 overlap	% top 5 overlap
Cat. 1k	14.0%	18.0%
Cat. 10k	12.0%	21.2%
Cat. 50k	41.3%	45.2%
Cat. 120k	57.3%	58.8%
Cat. 250k	69.3%	67.6%
Sec. 1k	6.67%	12.8%
Sec. 10k	28.7%	30.8%
Sec. 50k	56.7%	54.8%
Sec. 120k	71.3%	66.8%
Sec. 250k	76.7%	72.4%
Chap. 1k	25.9%	38.9%
Chap. 10k	40.7%	50.0%
Chap. 50k	64.8%	75.6%
Chap. 120k	77.8%	83.3%
Chap. 250k	83.8%	84.2%

BNF code granularity and no. patients	% top 3 overlap	% top 5 overlap
Para. 1k	7.33%	12.4%
Para. 10k	45.3%	42.0%
Para. 20k	52.7%	50.4%
Para. 50k	64.0%	60.0%
Para. 120k	72.1%	69.8%
Sec. 1k	10.0%	17.2%
Sec. 10k	38.6%	38.8%
Sec. 20k	51.3%	57.6%
Sec. 50k	66.0%	63.2%
Sec. 120k	74.8%	70.2%
Chap. 1k	40.0%	44.0 %
Chap. 10k	55.6%	70.7%
Chap. 20k	62.2%	69.3%
Chap. 50k	80.0%	72.0%
Chap. 120k	77.8%	78.6 %

Table 4.4: Results from running each network five times and, between each of the five networks, comparing the percentage of overlap between the top 3/5 highest RR connections for each node. **Left**: diagnosis-based model. **Right**: prescription-based model



(a) Diagnosis-based model (ICD-10 sections)

(b) Prescription-based model (BNF sections)

Figure 4.5: Matrices comparing the percentage of overlap between networks run using different numbers of patients between the top 5 highest RR connections for each node.

Chapter 5

Conclusion

5.1 Summary

In this project, we sought to develop methods to reveal patterns of disease progression across a large number of diseases. We achieved this by finding correlations between recorded diagnoses and correlations between recorded prescriptions from large datasets of anonymised primary and secondary care patient level medical records from general practitioners and hospitals across the UK. From these we established significant directions for each pairing and constructed networks containing these directed links. In addition, we described and implemented a method for accounting for disease interactions in our models by grouping multiple diagnoses/prescriptions into each 'node' of the network.

We have shown that our models are generalisable in many respects and can be run at many different levels and indeed on different datasets. This generality ensures that our methods could be applicable and adaptable to a variety of different styles of healthcare data records. The results we present in Chapter 4 may serve as baselines for the results obtained when applying further modified versions of our algorithms and/or applying them at larger scales.

The usefulness of such models stems from the growing need to address the issues brought about by problematic polypharmacy. There is a rising number of elderly patients with a variety of long-term health conditions for which they are prescribed a number of treatments, some of which may be unnecessary or potentially harmful. By establishing relationships between diseases and exploring disease trajectories and disease interactions, we can provide healthcare professionals with further insights into the *risks* of disease progression in order to assist them with treatment decisions.

5.2 Model Limitations

However, the methods we described in this report have a number of limitations. Firstly, we have no guarantee that disease correlations and trajectory patterns are indicative of causal relationships. For example, it is possible that the node associated with the ICD-10 section J40-J47 (chronic lower respiratory diseases) is a surrogate for smoking and is hence the reason why this node is strongly correlated with a number of other respiratory diseases [12]. Such relationships would likely need

to be explored in more detail on a case-by-case basis in order to establish causal patterns. In addition, in these models we have assumed that recorded diagnoses and prescriptions correspond with disease incidence. However, this is unlikely to be the case in many situations. There is a lag between disease incidence and diagnosis or the prescribing of treatment and the rate of disease diagnosing varies depending on a number of factors including disease severity and the inherent nature of the disease itself. For example, some diseases may be much easier to diagnose than others. Furthermore, the limits on computing power already discussed in this report restricted the complexity of our models.

5.3 Future Work

Finally, there are a number of possible of directions for future work. We already discussed the emergence of *source* and *sink* nodes in our models. Upon informal inspection of our resulting networks, it appears that highly connected sink nodes tend to involve more serious conditions/advanced disease stages in the diagnosis-based model and stronger drugs in the prescription-based model. These can be reached through many different paths through the network. By linking mortality data, it would be possible to analyse the extent to which disease connectivity (the number of arcs leading to or from a node) is related to risk of mortality [13].

It would also be insightful to study in depth the relationship between the diagnosis-based and prescription-based models and whether they reflect the same disease patterns. Direct network comparisons using known node-correspondence graph comparison methods are not applicable in this setting because there is not a bijection between the two networks. For example, one drug may be used to treat a number of different diseases, or, conversely, there may be a number of different treatments usually prescribed to treat a single disease, but at different stages of severity. However, this lack of a one-to-one correspondence between nodes in the two networks only reinforces the value in producing both graphs and future work could involve utilising unknown node-graph comparison methods to measure the commonalities between the two models.

Bibliography

[1] ONS Analytical Impact Team. *Overview of the UK population: January 2021*. 2021. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/overviewoftheukpopulation/january2021 (visited on 07/27/2021).

[2] Martin Duerden, Tony Avery, and Rupert Payne. "Polypharmacy and medicines optimisation". In: *Making it safe and sound. London: The King's Fund* (2013).

[3] Bruce Guthrie et al. "The rising tide of polypharmacy and drug-drug interactions: population database analysis 1995–2010". In: *BMC medicine* 13.1 (2015), pp. 1–10.

[4] Munir Pirmohamed et al. "Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients". In: *Bmj* 329.7456 (2004), pp. 15–19.

[5] Rachel L Howard et al. "Which drugs cause preventable admissions to hospital? A systematic review". In: *British journal of clinical pharmacology* 63.2 (2007), pp. 136–147.

[6] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. "Network medicine: a network-based approach to human disease". In: *Nature reviews genetics* 12.1 (2011), pp. 56–68.

[7] Iman Habibi, Effat S Emamian, and Ali Abdi. "Advanced fault diagnosis methods in molecular networks". In: *PloS one* 9.10 (2014), e108830.

[8] Evelien Otte and Ronald Rousseau. "Social network analysis: a powerful strategy, also for the information sciences". In: *Journal of information Science* 28.6 (2002), pp. 441–453.

[9] Nykamp DQ. *Network definition*. URL: http://mathinsight.org/definition/network (visited on 07/29/2021).

[10] Christopher L Sistrom and Cynthia W Garvan. "Proportions, odds, and risk". In: *Radiology* 230.1 (2004), pp. 12–19.

[11] David Westergaard et al. "Population-wide analysis of differences in disease progression patterns in men and women". In: *Nature communications* 10.1 (2019), pp. 1–14.

[12] Anders Boeck Jensen et al. "Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients". In: *Nature communications* 5.1 (2014), pp. 1–10.

[13] César A Hidalgo et al. "A dynamic network approach for the study of human phenotypes". In: *PLoS computational biology* 5.4 (2009), e1000353.

[14] Rafid Sukkar et al. "Disease progression modeling using hidden Markov models". In: *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2012, pp. 2845–2848.

[15] Kristina M Ceres, Ynte H Schukken, and Yrjö T Gröhn. "Characterizing infectious disease progression through discrete states using hidden Markov models". In: *PloS one* 15.11 (2020), e0242683.

[16] World Health Organization. *The International Statistical Classification of Diseases and Health Related Problems ICD-10: Tenth Revision. Volume 1: Tabular List*. Vol. 1. World Health Organization, 2004.

[17] Joint Formulary Committee and Royal Pharmaceutical Society of Great Britain. *British national formulary*. Vol. 64. Pharmaceutical Press, 2012.

[18] Emily Herrett et al. "Data resource profile: clinical practice research datalink (CPRD)". In: *International journal of epidemiology* 44.3 (2015), pp. 827–836.

[19] Thomas Kluyver et al. "Jupyter Notebooks - a publishing format for reproducible computational workflows". In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Ed. by Fernando Loizides and Birgit Scmidt. Netherlands: IOS Press, 2016, pp. 87–90. URL: https://eprints.soton.ac.uk/403913/.

[20] *Anaconda Software Distribution*. Version Vers. 2-2.4.0. 2020. URL: https://docs.anaconda.com/.

[21] Charles R. Harris et al. "Array programming with NumPy". In: *Nature* 585 (2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2.

[22] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[23] Wes McKinney et al. "Data structures for statistical computing in python". In: *Proceedings of the 9th Python in Science Conference*. Vol. 445. Austin, TX. 2010, pp. 51–56.

[24] Greg Lamp. *pandasql Python library*. Version 0.7.3. Aug. 7, 2021. URL: https://github.com/yhat/pandasql/.

[25] Aric Hagberg, Pieter Swart, and Daniel S Chult. *Exploring network structure, dynamics, and function using NetworkX*. Tech. rep. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[26] Stefano Travasci. *simple_icd_10_CM Python library*. Version 1.0.4. Aug. 6, 2021. URL: https://github.com/StefanoTrv/simple_icd_10_CM.

[27] Troels Siggaard et al. "Disease trajectory browser for exploring temporal, population-wide disease progression patterns in 7.2 million Danish patients". In: *Nature communications* 11.1 (2020), pp. 1–10.

[28] Juliana Casqueiro, Janine Casqueiro, and Cresio Alves. "Infections in patients with diabetes mellitus: A review of pathogenesis". In: *Indian journal of endocrinology and metabolism* 16.Suppl1 (2012), S27.

[29] Julie A Morris and Martin J Gardner. "Statistics in medicine: Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates". In: *British medical journal (Clinical research ed.)* 296.6632 (1988), p. 1313.

[30] Mattia Tantardini et al. "Comparing methods for comparing networks". In: *Scientific reports* 9.1 (2019), pp. 1–19.

[31] Sergey Ioffe. “Improved consistent sampling, weighted minhash and l1 sketching”. In: *2010 IEEE International Conference on Data Mining*. IEEE. 2010, pp. 246–255.