Studying Bayesian Multimorbidity Networks: The Impact of Methodological Decisions regarding Social Demographics

Charlotte Mitchell

Master of Science
Data Science
School of Informatics
University of Edinburgh
2024

Abstract

Multimorbidity, the presence of multiple long term health conditions in one individual, is a growing public health concern that reduces quality of life and strains healthcare systems. When using Bayesian network theory to learn the structure of multimorbidity networks, there are many methodological decisions that impact the network's structure, which in turn impacts the insights drawn from the network.

This dissertation utilises a dataset of 1.75 million Scottish General Practise patients, and their recorded long-term health conditions. It verifies that the chosen structure learning algorithm; choice to study an entire population or only those with multimorbidity; and the applied technique for discretising continuous variables in the data all meaningfully impact the properties of Bayesian multimorbidity networks. Thus, the risk of inferring knowledge from networks for clinical multimorbidity research is highlighted, especially when methodological decisions are not justified nor their impacts interrogated, as commonly seen in the literature.

It is also shown that the properties of networks produced by stratifying a population by social demographic factors (namely age, sex, urbanity and social deprivation) are heavily biased by the size of the stratified sub-populations. This limits their usefulness for comparing the impacts of these factors on disease-disease interactions. Instead, including social demographic factors as network nodes allows for a more straightforward assessment of their impacts, and clearly demonstrates that these factors mediate apparent connections between many of these diseases.

Research Ethics Approval

This project obtained approval from the Informatics Research Ethics committee.
Ethics application number: 137110
Date when approval was obtained: 2024-04-27

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Charlotte Mitchell*)

Acknowledgements

I am very grateful to my supervisor, Dr Guillermo Romero Moreno, for his guidance over the course of this project. I have learned so much this summer, and it has been a pleasure to work with him. I am grateful too to all of the lecturers whose fascinating courses not only provided me with background knowledge for my dissertation but also helped me to become a data scientist. Being a part of the School of Informatics over the last two years has been so inspiring and an a true privilege.

Lastly, I am so grateful to my friends, to my mum, and to Jake for their endless support and for enduring my raving monologues on network structure learning algorithms.

Table of Contents

Chapter 1

Introduction

Multimorbidity is the presence of more than one long term health condition in an individual. The presence and degree of multimorbidity is increasing globally, not simply due to aging populations, but also lifestyle changes and increased urbanisation [1]. As health conditions have historically been studied and treated in isolation, multimorbidity is not well understood and is currently a topic of extensive research [2]. It presents multiple negative impacts, not only to individuals (such as increased healthcare costs, increased medication use, and reduced quality of life) but also to healthcare systems, including increased hospitalisation and medical appointments, increased medication use and increased costs [1]. Essentially, in both cases, multimorbidity causes great strain, and understanding it better is essential to facilitating better treatment and to successfully implementing policies for disease prevention and personalised medical care [3].

A common approach to the study of multimorbidity is to use large datasets of Electronic Healthcare Records (EHR) to build graphical networks, with nodes representing conditions and edges representing some relationship between them. Many studies have used pairwise correlations between diseases to build edges [4], [5]. An important drawback of such methods is that they do not account for confounding factors – that is, those whose presence may cause pairwise associations between others. Foundational work comparing pairwise and Bayesian networks revealed that direct connections between diseases are much rarer than previous works with pairwise connection suggested, and thus highlighted the importance of a Bayesian approach to the study of multimorbidity [6].

Bayesian networks are defined as Directed Acyclical Graphs (DAGs) made up of nodes that represent random variables, and edges that represent conditional relationships between them. Each node has an associated conditional probability table that defines

the probability distribution of the node, given the values of its parent nodes. To build Bayesian networks to study multimorbidity using only EHR datasets, an algorithm is required to learn the network structure from the dataset. Parameters can then be learned from the network and dataset via several methods [7]. Where Social Demographic Factors (SDFs) such as age, sex, race and others are known, two approaches for analysing their impacts can be explored. Stratification of the data by these factors to generate and compare sub-networks, which is a common approach for non-Bayesian networks [3], [8], [9] is has also been used for Bayesian ones [10], [11]. Incorporating these factors as nodes in the networks instead better aligns to Bayesian network theory, but the impacts of doing so are not always discussed [6] [10]. The impacts of discretising continuous variables into categorical ones have also been highlighted as important and understudied [12]. Several other decisions, such as how diseases are defined and what population datasets are used, must also be made or will be a constraint of a preexisting dataset.

The choices made across all of these steps can result in vastly different multimorbidity networks and associated interpretations, even if based on the same underlying data. This variability poses a significant risk where these models are used by clinicians aiming to draw conclusions regarding connections between diseases and SDFs. Therefore, it is crucial to thoroughly understand the impacts of different methodological decisions on the structure and validity of Bayesian multimorbidity networks.

The first aim of this dissertation is to interrogate the process of developing Bayesian networks, in the context of muiltmorbidity, in order to provide a clearer overview of the impacts of decisions made in this process than the literature on multimorbidity networks has provided to date. This is achieved by studying the impacts on network structure of four structure learning algorithms; utilising data from a general versus multimorbid patient population; and various approaches for variable discretisation. The second aim is to investigate two methods of representing the impacts of factors that may confound disease-disease relationships, namely four SDFs: age, sex, urbanity and social deprivation. Stratification and node-addition approaches are investigated in order to determine their effects on network structure and the implications of drawing clinical insight from them.

These aims have been achieved by use of the data initially collated by the University of Aberdeen's Primary Care Clinical Informatics Unit and enhanced by Barnet et al. in their 2012 study [2]. It is hereafter referred to as the 'PCCIU dataset', and consists of 1.75M EHRs from Scottish GP practices, and has been used in this work to build

multiple Bayesian networks using a variety of methodological decisions. For clarity, interpreting any clinically relevant findings for multimorbidity research is not an aim for this work, as doing so would require input from clinical experts. However, this work intends to from a contribution to clinical multimorbidity research by outlining the most appropriate methods for assessing the impacts of SDFs on multimorbidity.

Following this introduction, the necessary background information to support understanding of the subsequent work is provided in Chapter 2. Chapter 3 then outlines the methodology followed through the course of the dissertation period, specifically how the networks were generated and analysed. In Chapter 4, the generated networks generated are presented, along with analysis and comparison of their attributes. Finally, in Section 5, conclusions are drawn and suggestions for future work are made.

Chapter 2

Background

2.1 Bayesian Networks

2.1.1 Overview

Bayesian networks fall within the broader category of graphical networks, which are used to represent and infer insights from complex systems. Unlike other networks, Bayesian network edges convey conditional dependence relationships between variables, and the absence of an edge also implies conditional independence. To utilise Bayesian networks correctly for causal inference, the following assumptions must hold [13]:

- **Stable Unit Treatment Value Assumption (SUTVA):** Each variable (node) does not causally interact with others outside its defined relationships.
- **Causal Markov Assumption:** Given its parents (its direct causes), each variable is conditionally independent of all other variables.
- **Causal Faithfulness Assumption:** The model includes all relevant variables.

Of these assumptions, Causal Faithfulness can be particularly challenging to achieve, especially in multimorbidity analysis. For instance, factors which may not be measured, such as a patient's diet or smoking status, may influence outcomes.

2.1.2 Markov Equivalence Classes

Graphs which have the same set of edges but without or with differing directions encode the same conditional independence relationships between nodes and are said to belong to the same Markov Equivalence Class (MEC) [14]. Often, it can be sufficient to

evaluate the properties of a DAG's MEC, rather than the DAG itself. This is true for multimorbidity networks, for which the research aim is to study the present edges, rather than their directions (as clinical insight is usually much better placed to define these).

2.1.3 Learning Networks from Data

2.1.3.1 Structure Learning

The placement of edges within a network is known as its structure. There are several classes of structural learning algorithms through which Bayesian networks can be learned from a dataset. These are necessary because as the number of nodes (variables) increases, the number of potential structures explodes, rendering exhaustive search infeasible. Structure learning algorithms can use constraint-based, score-based or hybrid search methods. Given a dataset, constraint-based algorithms determine the conditional independence relationships between the variables using statistical tests, whilst score-based algorithms will use a goodness-of-fit score (such as Bayesian Information Criterion) to rank potential DAGs [15]. Hybrid models combine score- and constraint-based methods but have been noted to be no better performing in terms of speed or accuracy and so are not explored further herein [16]. The same work also notes that there is no algorithm that is consistently best performing for structure learning. The two models used in this work are described as follows:

Constraint-based: The PC-stable algorithm is a variant of the foundational PC algorithm that is stable against column order permutations of the input dataset. It starts with a fully connected, undirected network and performs conditional independence tests to determine which can be removed, before determining edge directions [17].

Score based: The Hill-Climbing search with Tabu list (hereafter referred to as the Tabu algorithm) is a common score based algorithm. It begins with unconnected nodes and adds edges one by one. In each iteration, the algorithm generates a subset of possible structures with only one change and selects the one with the best score. A 'tabu' list is updated at each iteration to record recently visited structures to prevent cycling back to them. The algorithm will end either when the change in score is below a given threshold, or if the maximum number of iterations has been reached [18].

The two conditional independence tests (the Chi-squared (χ^2) test and G test) used by the PC-stable algorithm and the two scores scores used by the Tabu algorithm (the Bayesian Information Criterion (BIC) and the K2 score) are defined in Appendix A.

2.1.3.2 Variable Discretisation

Whilst disease presence tends to be recorded as binary variables, continuous variables may be present depending on the underlying data and how it is processed. However, many algorithms and associated scores and tests can only be applied to fully discrete or fully continuous datasets. Discretisation is also necessary when stratifying a dataset by a continuous variable in order to compare the resulting sub-networks.

There are several methods of discretising continuous variables into discrete ones. These include manual discretisation using expert knowledge; and creating categories either of equal width or of equal instance count [19]. A new method is proposed in this work, based on maximising the combined structural differences between stratified sub-networks.

2.1.3.3 Parameter Learning

The parameters of a Bayesian network are the set of probability functions for each node, conditioned on its parents. As with structure, parameters can be learned from a dataset. The most common method when working with complete data is Maximum Likelihood Estimate (MLE) [7]. These probabilities can be very useful in that they quantify the strength of relationships between nodes. However, they are not considered further in the scope of this work.

2.2 Evaluating and Comparing Networks

2.2.1 Evaluation

The larger a network is, in terms of both nodes and edges, the more difficult it is to draw insights from studying graph visualisations alone. Node-node adjacency matrices with cell values that represent edges are commonly used to show graph structures more systematically [11]. A non-zero value in cell $[i, j]$ of the matrix indicates the presence an edge from the node in row i to the node in column j. Additionally, the following are standard metrics for evaluating networks [20]. Variations of the below and other such metrics are available, but have not been required to fulfil the aims of this work.

- **Edge count:** The number of edges within a network – a basic measurement of its complexity.

- **Connectivity:** A network is said to be 'fully connected' if all nodes are connected to each other node. A network is 'connected' if there is a path of edges from each node to every other.
- **Degree:** The degree of an edge is the count of its neighbours – that is, the nodes it is connected to. In- and out-degrees respectively measure the number of edges leading to and from a node in a DAG.
- **Clustering Coefficient:** For each node, this is the observed number of edges that interconnect its neighbours, as a portion of the possible total. In aggregate, it indicates how tightly-knit a network's nodes are.
- **Assortativity:** This measures the tendency for nodes of similar degree to connect. It is calculated as the Pearson correlation coefficient between the degrees of each node pair in the network. Positive assortativity therefore indicates that nodes with high degrees (hubs) tend to connect to other hubs. Conversely, negative assortativity indicates that hubs are more likely to be connected to low-degree nodes.

2.2.2 Network Comparison

The Structural Hamming Distance (SHD) is a useful score for comparing the structure of two graphs – especially when 'ground truth' networks, against which other such scores like precision and recall can be measured, are not known. SHD was initially defined as the number of operations required to make the input graph equal to another 'target' graph, with possible operations being addition, removal or redirection of an edge [21]. Other approaches have since assigned lower importance to or omitted differences in edge directions in SHD scores [22]. This latter approach acknowledges that missing and extra edges are more serious errors which can have knock-on effects, and essentially measures the difference between one MEC and another.

2.3 Related Work

There is only one prior work which applies network science to the PCCIU dataset [23]. In this work, a Bayesian inference framework is used to examine associations between diseases and focus on the population aged 90 upwards. The framework incorporates

uncertainty and provides a more cautious and reliable estimation of associations compared to traditional pairwise measures, which is of particular benefit to the study of multimorbidity in small sub-populations.

Within the limited literature regarding Bayesian multimorbidity networks, a variety of structure learning algorithms are applied. These include the Tabu algorithm [11], a Markov Chain Monte Carlo algorithm [6], [12] and Maximum Weight Spanning Tree search [24]. The majority of these do not discuss the implications of the chosen method. Studies have also been noted to combine the results of multiple algorithms into a single network, which risks compounding the biases inherent in different structure learning approaches [25], [15]. All of the mentioned algorithms require that any continuous variables, such as age and blood pressure, are discretised into categories. Only one work was noted to have discussed that this too can impact network structures [12].

Lastly, studies which stratify Bayesian networks by multiple SDFs do so in contradiction of the Causal Faithfulness Assumption [11], [10], as it implies that all known variables must be included in order to avoid misleading network structures. The Causal Markov Assumption also requires that SDFs should be represented as single nodes and not multiple binary nodes. This latter practice has not been observed in the literature, but the Assumption is noted here as a constraint on the methodology in Chapter 4.7.

Given these gaps and inconsistencies in the existing research, an assessment of the impact of methodological decisions is a necessary contribution to the field of multimorbidity in Bayesian networks.

Chapter 3

Methodology

3.1 Exploring the PCCIU Dataset

All networks and associated analysis relating to this work have been generated from an augmented version of the PCCIU dataset, which consists of 1.75 million rows, each representing a patient of a Scottish GP Practice and their medical information, as recorded in 2007. It is considered to be representative of the Scottish population. The columns used herein are 40 binary indicators for long term health conditions (diseases), which were generated based on the data in preexisting columns by the clinicians among the authors of Barnett et al.'s 2012 PCCIU dataset analysis [2]. These 40 diseases are deemed to be the most important ones to consider in the multimoribity context, and are listed in Tables B.2 and B.3 in Appendix B. Their 'short names' used for figures, and their prevalence amongst all and multimorbid patients are also listed.

The dataset also includes each patient's age and biological sex, as well as their Carstairs Score (and associated variants) and their urbanity (how rural or urban their postcode is deemed) as measured by the Scottish Government's Urban Rural Classification [26]. The six-fold classification categories are presented in Table B.1 in Appendix B, with 1 being the most urban and 6 being most remote areas. The Carstairs Score is a measure of social deprivation, based on four factors (lack of car ownership, low occupational social class, overcrowded households and male unemployment) and assessed at postcode area level via Census data. The higher the Carstairs Score, the more deprivation is associated with the postcode area. Per recommendations from Public Health Scotland, the Carstairs Decile variant (numbers from one to five, with five being most deprived) was chosen to discretise deprivation [27]. Similary, Barnett et al. discretised the population into five age categories which have been adopted as

default herein. These categories are: 0-24, 25-44 , 45-64, 65-84 and 85-100 [2].

Further variables in the PCCIU dataset include patient registration details, and records of vaccinations and prescriptions, which were noted to be either irrelevant or unusable without input from clinicians. Measurements including weight, height, smoking status and alcohol intake, were also present. These were noted to be potentially useful factors for the multimorbidity networks, but were not included as all had at least 25% missing data.

Following variable selection, a minimal exploratory analysis of the dataset was required due to the prior work of Barnett et al, who noted that only 23% of the patients exhibited multimorbidity [2]. Accordingly, a basic set of bar charts and histograms to characterise the dataset's population in terms of morbidities and the four SDFs were produced, using the pandas and matplotlib Python packages.

3.2 Network Methods

3.2.1 Network Generation

Initially, the causallearn Python package was selected for structure learning, due to its minimal requirement for disk space relative to other options. It's PC-stable algorithm took approximately 8 hours to run on 30,000 rows of PCCIU data, which meant that bash scripts were required to prepare multiple samples and learn their structures in parallel, with the intention of bootstrapping these into a single network. To cover the whole dataset, 59 samples were required. Scripts were generated for this and the 18 network permutations required for this work. However, it was found that the outputs were not actually stable (as they were dependent on the order of columns in the input data) and so this package and all associated results were abandoned.

Ultimately, networks were generated from the PCCIU dataset using the PC-stable and Tabu algorithm functions from from the pgmpy Python package (both of which were confirmed to be stable, as advertised). For the PC-stable algorithm with both the χ^2 and G tests, the default significance threshold of 0.05 was used. For the two versions of the Tabu algorithm (with BIC score and K2 score), a tabu list length of 100 was used, along with a stopping criteria of either 1 million iterations or a change of score less than 0.0001. These four algorithms are hereafter referred to as PC-χ^2 , PC-G, Tabu-BIC and Tabu-K2. The outputs of these algorithms were a Bayesain model object that could be used as in input to build a directed graph in the networkx Python package, and a binary

adjacency matrix encoding directed edges between nodes.

3.2.2 Network Analysis

The networkx graph objects generated by the structure learning algorithms were used to calculate the network metrics outlined in Section 2.2.1 and to create visualisations of the networks. In the visualisations, the sizes of disease nodes were configured to represent their prevalence within the full population. The 'spring' layout was also used (whereby nodes 'repel' each other and edges 'pull' connected nodes closer), which tends to gives hubs a more central position.

To compare networks by their MEC, adjacency matrices were adapted by mirroring any values within the bottom diagonal onto to the top diagonal instead (effectively yielding a matrix of undirected edges). These matrices were used to calculate the SHD between two networks, by adding up the number of non-matched cell values between the two matrices.

3.2.3 Continuous Variable Discretisation

Four methods were adopted to discretise the age variable into five alternative categories (bins) to those defined by the Barnett et al. clinicians. The first two are equal width (20 years) and equally sized splits of the population (subsequently referred to as 'equal count'). The third and fourth are an exhaustive and a greedy search algorithm, provided in Appendix C. Both aim to find the set of bin boundaries that maximise the combined SHD between all pairings in the set of five age-group sub-networks associated with each bin (using Algorithm 1). The exhaustive search is outlined in Algorithm 2 and calculates the combined SHD for all bin boundary sets that obey the minimum bin width and bin width increment constraints. The greedy search (outlined in 3) obeys the same constraints, but adds the bin boundaries one at a time, by adding the boundary that maximises the combined SHD of the set of bins created from the already established boundaries and each new candidate. Naturally, the intention of the greedy algorithm is to reduce the time required to find the optimum set of bin boundaries.

Chapter 4

Analysis

4.1 Overview

The first of the subsequent sections in this chapter stands apart from those that follow it as it is intended to provide a basic analysis of the PCCIU dataset in order to support the later sections. Four structure learning algorithms are then compared. One of these is used to generate the 'Full' network from the whole dataset, which forms the baseline against which all other networks are compared. It is first compared to a network drawn from the multimorbid sub-population, then to 18 sub-networks stratified by each category of the four SDFs. Finally, networks incorporating these SDFs as nodes are analysed, along with methods for discretising the sole continuous variable (age) that they incorporate.

4.2 Insights from the PCCIU Dataset

The plots in Figure 4.1 show how the dataset's population and associated multimorbidity rates are distributed across each SDF. There is an almost equal split between male and female patients, with multimorbidity being 30.5% more prevalent in females. Social deprivation is normally distributed, although with more patients in the two lowest (least deprived) quintiles than the two highest, and a clear relationship between social deprivation and multimorbidity (Pearson R value: 0.937, p-value: 0.019). The vast majority of patients live in urban areas (Urbanity categories 1 and 2), where the prevalence of multimorbidity is slightly lower than in most rural areas.

In the lower half of Figure4.1, a very strong relationship between age and multimorbidity is also demonstrated (Pearson R value: 0.940, p-value: 0). Less than 0.1% of

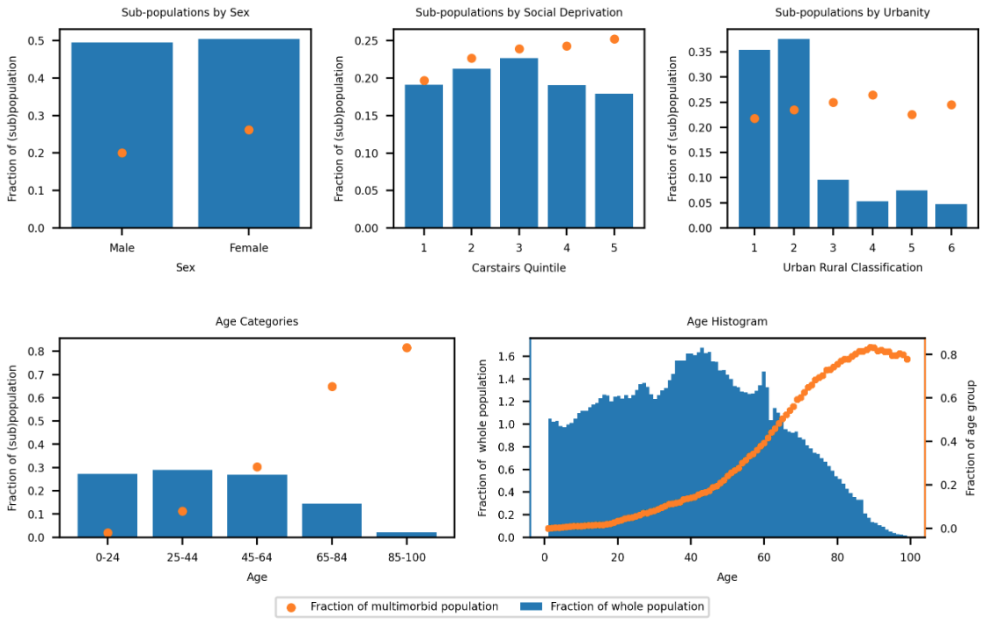

Figure 4.1: Breakdown of whole and multimorbid populations by Sex, Social Deprivation, Urbanity Age Category and Age.

those below age 25 exhibit multimorbidity, but this increases with age up to 80% for those over 84. Additionally, the distribution in age is neither flat or normal, with the population shrinking as age and multimorbidity increase, and with the top 15% of ages being represented by only 2% of the population. Of course, the histogram reveals more information than the bar chart, demonstrating the loss of information that occurs when continuous variables are discretised into categories.

Figure 4.2 shows how disease prevalence for each sub-population relating to one SDF category differs from the full population. Here, and in subsequent figures, age categories are coloured blue, start with 'a' and are numbered youngest to oldest; Sex ('s') is orange with 1 for males and 2 for females; Urbanity ('u') is green; and Social Deprivation ('c' for 'Carstairs Quintile') is red. All have at least a 95% Pearson R correlation with the prevalence of diseases in the full population, except for the age categories (other than a3). Similarly, all but these have a maximum disease prevalence that is close to the full population's (of 13.4% for Hypertension), with the youngest and oldest patients exhibiting lower and (markedly) higher disease prevalence.

Whilst these insights can be anticipated even without clinical knowledge, they help to characterise the population represented by the dataset to support subsequent analysis. They also suggest that age is likely to be the most important SDF to consider in terms of multimorbidity research and, given that it is the sole continuous variable, methodology.

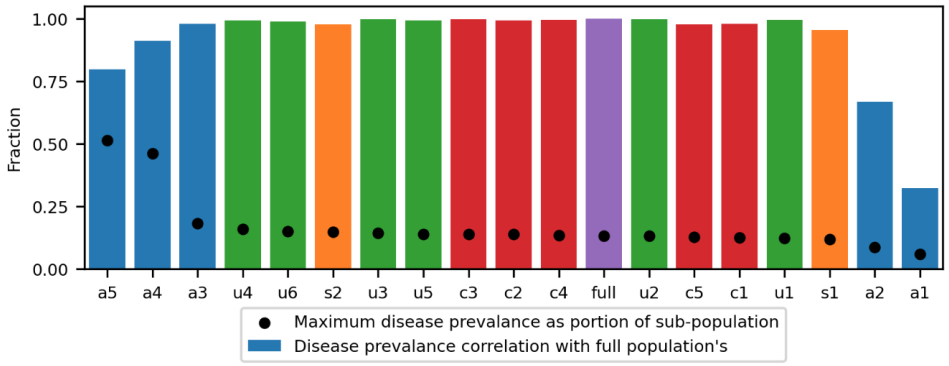

Figure 4.2: Bar chart with each sub-population's Pearson R correlation for prevalence of each disease with full population's. Scatter points are the maximum prevalence of any disease in the sub-populations.

4.3 Comparing Structure Learning Algorithms

In order compare their differences and select one to use for subsequent analysis, the two versions of the PC-stable and and the Tabu algorithms (PC- χ^2 , PC-G, Tabu-BIC and Tabu-K2) were tested. For each algorithm, and for five sample sizes up to 10,000 patients, the same five random samples were used as the input dataset. Results are given in Figure4.3, and plot (a) shows that the two PC-stable algorithms have runtimes that are significantly greater than the Tabu algorithms, taking an average of 83 minutes (PC-χ^2) and 41 minutes (PC-G) to run on the 10,000 row samples. Conversely, the Tabu algorithms take an average of 8 seconds (Tabu-BIC) and 18 seconds (Tabu-K2) on the same samples. Whilst these runtime are of course related to the processing power of the machine used, the pattern would hold for another more or less powerful machine. The trends in runtimes for the PC-stable algorithms suggest that these would not resolve within a reasonable time frame for the full 1.75M row dataset. Indeed, PC-χ^2 was tested on a 1% sample and took 3.95 hours to resolve.

The algorithms were also compared on the basis of the network structures they produced. Figure 4.3 (b) illustrates that the two Tabu algorithms consistently generate more than double the edges that the PC-stable algorithms do, and that all algorithms except Tabu-K2 exhibit an increase in the number of edges generated with sample size (although this increase is less pronounced for the two PC-stable algorithms). The similarity between graphs (ignoring edge directions, so in fact the similarity between MECs) is represented in Figure 4.3 (c) by the combined SHD between each network and those generated from other algorithms from the same sample, averaged over the

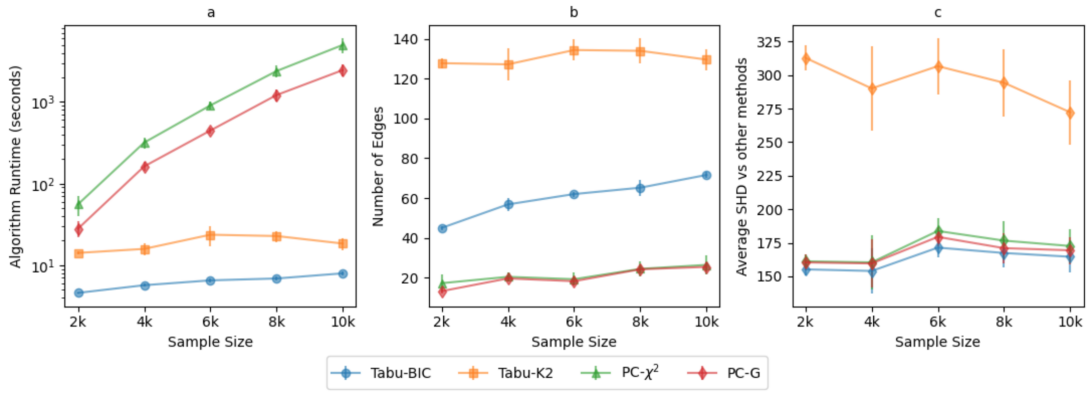

Figure 4.3: Comparison of (a) Algorithm Runtime, (b) Number of Edges Generated, and (c) Average Structural Hamming Distance between networks produced by other methods for four structure learning algorithms. Average values across 5 random samples of each size are plotted, with error bars for standard deviation.

sampling iterations. Tabu-K2 score exhibits a consistently higher combined SHD than the other three, with Tabu-BIC being slightly lower than the PC-stable algorithms.

Figure 4.4 provides comparative adjacency matrices to visualise the different edges generated by the algorithms. Although it is only for one 10,000 row sample, it is consistent with the SHD distribution in Figure 4.3. It shows that Tabu-K2 captures all the edges that both PC-stable algorithms do (no blue cells), and Tabu-BIC captures almost all of them (1-3 blue cells). However, the two PC-stable algorithms differ from each other considerably, whilst Tabu-K2 captures all but one of the edges that Tabu-BIC does. These findings suggest that the choice of independence test greatly impacts the structure defined by a PC-stable algorithm, whereas the choice of scoring metric for the Tabu algorithm impacts the threshold above which edges will stop being generated. Of course, further scores and independence tests would need to be assessed to confirm if these findings generalise.

It is important to note that without input from clinicians, it is not possible to determine which algorithm generates the multimorbidity network structure 'best'. However, it was decided to use Tabu-BIC for generating subsequent networks in this work, due to its comparatively fast runtime and as it provides a middle ground between the edge generation tendencies of the other algorithms.

To further explore the characteristics for the Tabu-BIC, five rounds of sampling were run at five scales between 10% and 100% of the dataset. The runtimes and resulting edge counts are shown in Figure 4.5, which illustrates that both of these increase with sample size. This trend occurs despite the fact that the smaller samples accurately

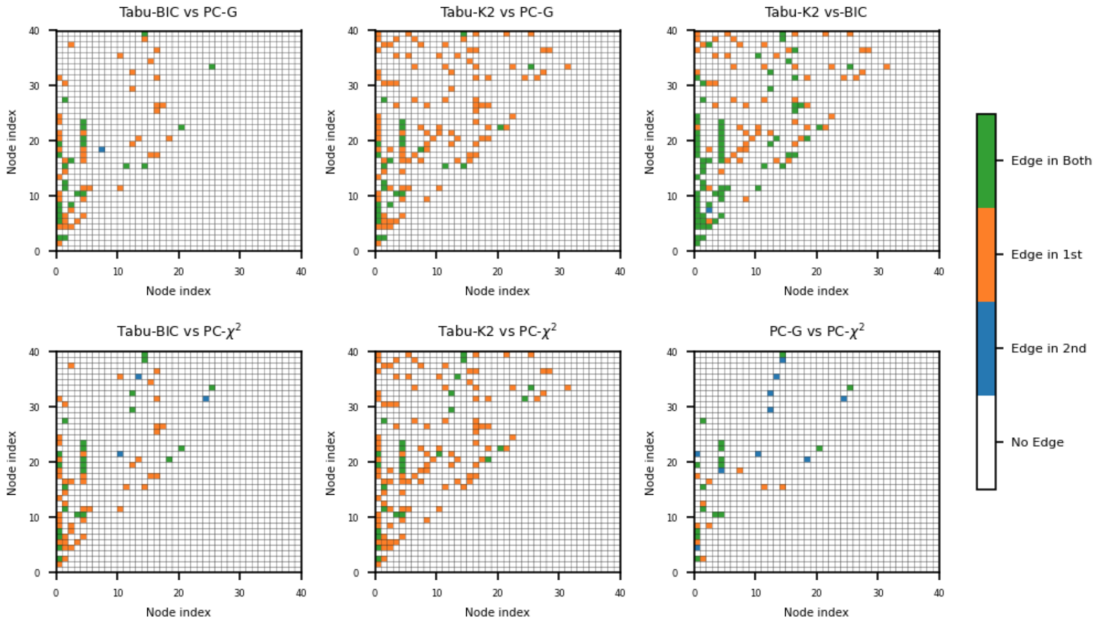

Figure 4.4: Edge presence comparison between all paired combinations of the two PC-stable and two Tabu algorithms, for a random 10,000 row sample of the PCCIU dataset.

reflect the proportions of healthy and multimorbid patients in the full dataset. However, larger samples most likely better detect the true disease dependencies, as noise and variability are reduced, allowing the algorithm to identify the less common relationships. Therefore, as sample size grows, the network becomes more connected.

From Figure 4.5 (c), it is also observed that the average SHD increases with network size initially, because the networks themselves have a limited number of connected nodes. The average SHD then peaks for the 70% sample (although far below the maximum possible SHD of 780), before reaching zero for the full network. This indicates that the structures of the larger samples are similar, but do not contain enough data to represent all of the relationships within the full network.

4.4 Characterisation of 'Full' Network

A graph of the network produced from Tabu-BIC for all 1.75M patients in the dataset is shown in Figure 4.6, and its degree distributions are shown in Figure 4.7. It is a connected network with 214 edges, giving it a density of 0.137. The minimum and the most common node degree is 4, and the degree distribution has an average of 10.7 with a standard deviation of 7.0.

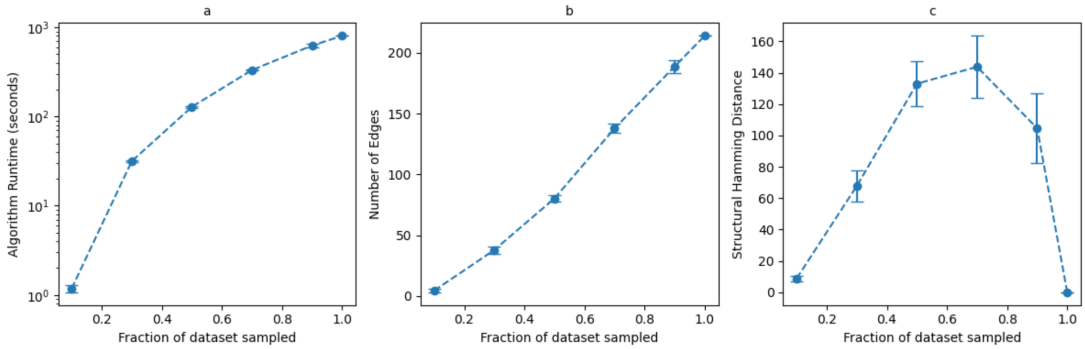

Figure 4.5: Averaged results with error bars for standard deviation for 5 sets of Tabu-BIC algorithm results against fraction of dataset sampled, for (a) runtime, (b) number of edges generated and (c) for the Structural Hamming Distance between networks from one run and another.

The graph indicates that the more prevalent diseases (i.e. the larger nodes) generally appear to have higher degree. This is confirmed from the degree distribution in Figure 4.7, where a positive correlation is shown between degree and prevalence (Pearson R value: 0.830, p-value: 0). A weaker negative correlation also exists between node in- and out-degrees (Pearson R value: -0.414, p-value: 0.008). As such, it appears that the prevalence differential between connected nodes influences the direction of the connection, and that direction should not be taken as an indication of causality. Hence, edge directions are not discussed further in this work.

4.5 Impact of Patient Subset

When studying multimorbidity networks, the choice of data to work with can be limited in terms of both availability and accessibility [28]. Whilst the PCCIU dataset contains a majority of patients with no recorded conditions, studies often use datasets from 'sick' populations such as those making health insurance claims [29], [5]. To examine the impact of this, a network of only multimorbid patients was generated, representing 23% of the patients in the full dataset. The Table 4.1 compares descriptive metrics between the full network and the network generated from only multimorbid patients (MM). The total number of edges in the MM network is slightly lower by 11 edges, but this reduction is very small compared to the trend in plot (b) of Figure 4.5. This, along with the fact that for both networks around 40% of edges are unique (i.e., not found in the other), emphasises that filtering out non-multimorbid patients yields a network that

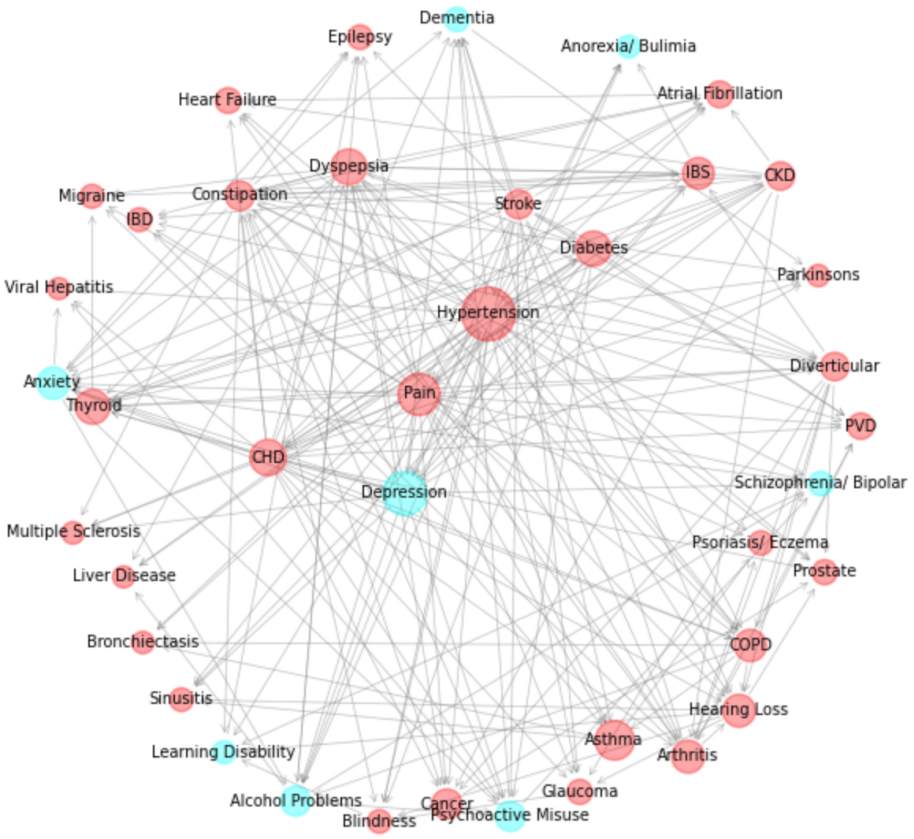

Figure 4.6: Graph for full network where node size indicates prevalence, and colour indicates physical (red) or mental (blue) disease classification.

captures different, rather than just less information. The SHD in the table is simply the sum of the sum of the two sets of unique edges.

In line with the reduction in edges, the average degree is slightly lower in the MM network. However, the degree standard deviation is slightly higher, indicating more variability in how nodes are connected. This is also shown in Figure 4.8, which maintains a very strong relationship between prevalence and degree (Pearson R value: 0.857, p-value: 0). The change in average clustering coefficient is proportional to the reduction in edges in the MM network, but MM does exhibit marginally higher variance in clustering. Interestingly, assortativity is higher in the MM network, indicating a more pronounced tendency for nodes with similar degree to connect in MM. However, in both networks, the strength of assortativity (which has a maximum of 1) is very small.

The adjacency matrix for edges that occur in the two networks is shown in Figure 4.9, alongside the nodes with changes in degree. There is no linear correlation between changes in node prevalence and degree (Pearson R value: 0.007, p-value: 0.967). Clearly filtering out the patients with less than two conditions changes the disease

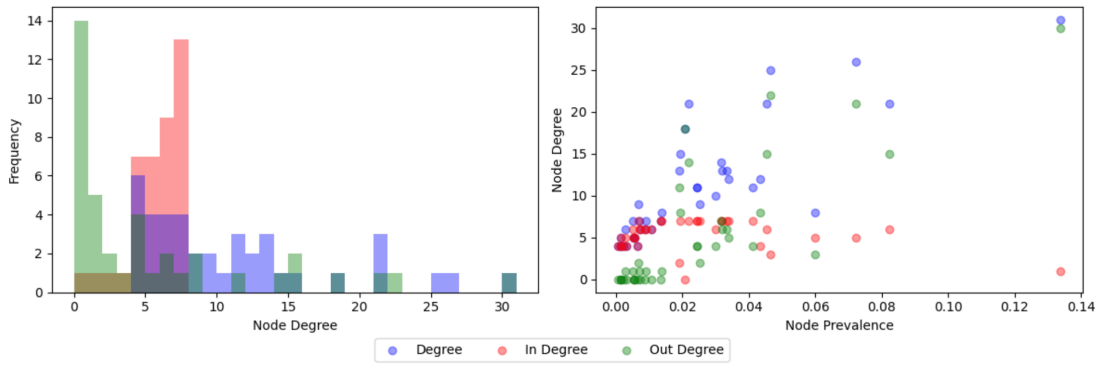

Figure 4.7: Full network degree distribution (left) and node degree versus node prevalence (right).

Network	Full	MM	Difference
Total Edges	214	203	-11
Unique Edges	93 (43.46%)	82 (40.39%)	-11
Average Degree	10.7	10.15	-0.55
Degree Standard Deviation	6.962	7.234	+0.272
Average Clustering	0.315	0.302	-0.013
Clustering Standard Deviation	0.098	0.108	+0.010
Assortativity	0.005	0.061	+0.056
Common Edges	121		-
Structural Hamming Distance	175		-

Table 4.1: Comparison of Full and Multimorbid (MM) Network Edges

interactions significantly. Of the 40 disease nodes, 20 decrease in degree (Dyspepsia most so by 10, followed by Pain by 7) and 11 increase in degree (most so Asthma by 14 and then Depression by 11). These nodes have in common that they all increase in prevalence relative to their neighbours (per Figure 4.10, where the nodes are ordered by most to least prevalent in the Full network), and are the only nodes to do so other than Hypertension (which is the most prevalent condition in both networks and doesn't change in degree) and Coronary Heart Disease (which decreases by 2).

These findings suggest that disease prevalence influences edge creation, but it does not solely explain the structural differences between the two networks, which has not been fully discerned. A better understanding of the differences may be achieved from a combination of input from clinicians and further feature engineering to identify other

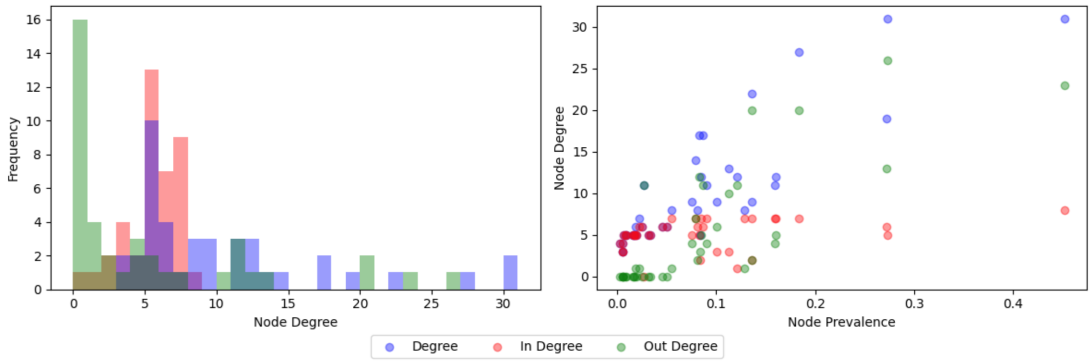

Figure 4.8: Full network (a) degree distribution and (b) degree against node prevalence.

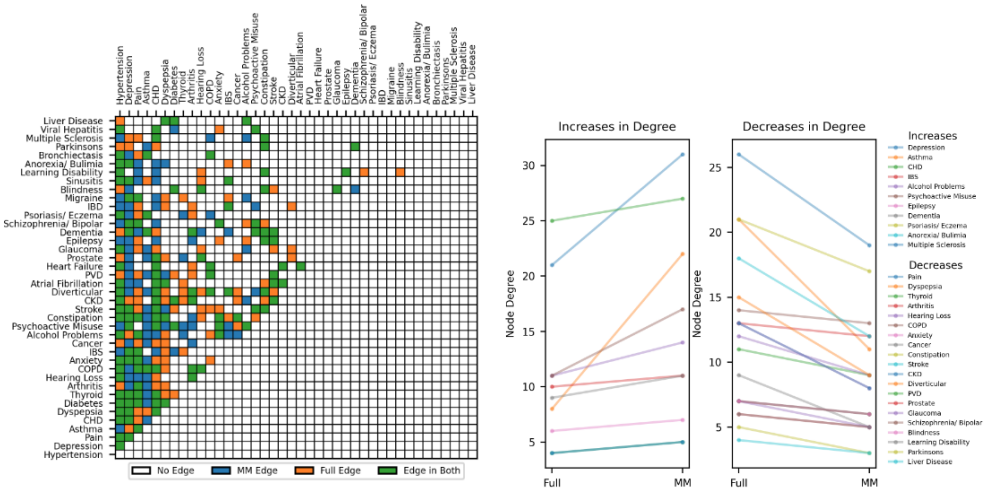

Figure 4.9: Comparison of Full and Multimorbid (MM) Networks via (a) Edge Adjacency Matrix and (b) Nodes with Changes in Degree between Networks

patterns among the affected nodes.

4.6 Impacts of Stratification

The dataset was stratified into 18 patient subsets, with one per category for each of the four SDFs. From each of these, a new 'sub-network' was generated. Figure 4.11 examines the relationship between both average degree and SHD from the Full network against the portion of the Full network that each subset represents. The SHDs indicate that the sub-networks exhibit various differences from the Full network, particularly the age networks – which (apart from a3) are the only sub-networks to have at least one unconnected node. Indeed, without the age networks, a strong linear relationship is observed between SHD from and proportion of the Full dataset (Peason-R value:

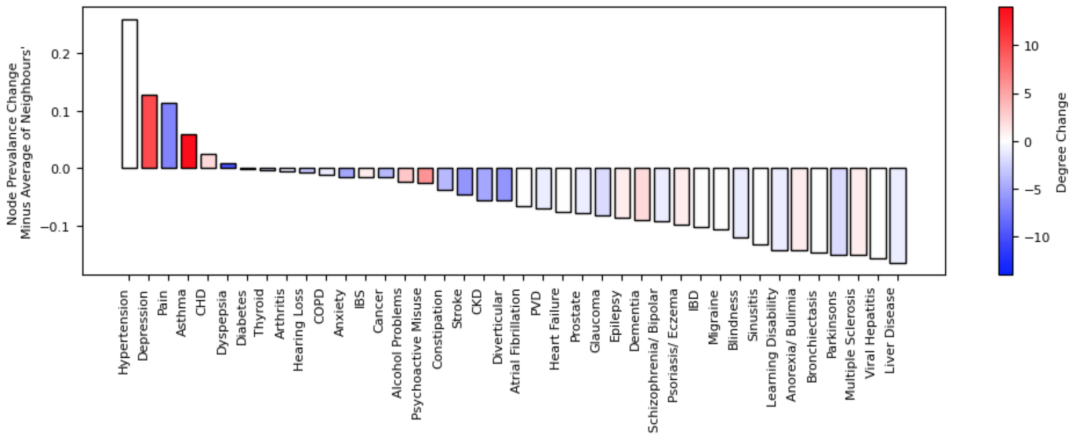

Figure 4.10: Node Prevalance Change from Full to Multimorbid Network, minus Average of Neighbours' Prevalence Changes, ordered by prevalence in the full dataset.

-0.809, p-value: 0). With the age sub-networks, the relationship is much more muted (Peason-R value: -0.414, p-value: 0.088).

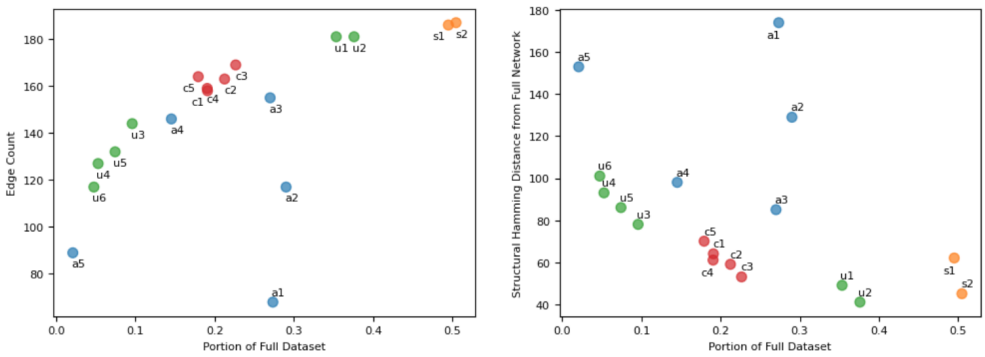

Figure 4.11: Edge Counts (left) and SHD from Full network (right) for the sub-networks representing each category of Age (blue points), Sex (orange), Urbanity (green) and Social Deprivation (red). For Sex, 's1' represents males and 's2' females.

A similar pattern is observed between average degree and portion of the Full network. Across all stratified networks, a moderate positive correlation is observed (Pearson-R value: 0.5821, p-value: 0.0113) . However, this relationship also becomes considerably stronger when the age networks are excluded (Pearson-R value: 0.9230, p-value: 0). This nearly perfect positive correlation underscores the consistent relationship between degree and network proportion in the non-age sub-networks, and suggests that the size of the network is impacting node degrees more so than the characteristics of its sub-population. This agrees with the pattern in Figure 4.5 (b), where number of edges (which is directly proportional to average degree) is almost linear with sample size.

These observations clearly reflect that the age subsets contain very different information from each other, and from the full population. The youngest and oldest populations (a1,a2 and a5) are the most and least multimorbid respectively, and and are the greatest outliers in Figure 4.11. Although this underscores the importance of studying the impact of age in multimorbidity networks, it appears that variability in subset size would bias such analysis. This is exemplified by the Urbanity sub-networks, which, despite all correlating almost exactly with the disease prevalence patterns of the Full network (as shown in Figure 4.2), are grouped in Figure 4.11 based on their size. To check that this was not unique to Tabu-BIC, the urbanity sub-networks were created using Tabu-K2 and the same linear trend was shown again between size and edge count (Pearson R value: 0.986, p-value: 0).

Exploring how the size of the sub-networks biases their structures is not a straightforward task, as there is no obvious benchmark for comparison. In an attempt to devise one, each stratification dataset was sampled 100 times with samples equal to the size of the smallest dataset, a5 (with 36,569 rows). For each stratification, the sample networks were then bootstrapped to create network containing all edges that appeared across the sample networks. Edges were given weights between 0 and 100, depending on their appearance frequency.

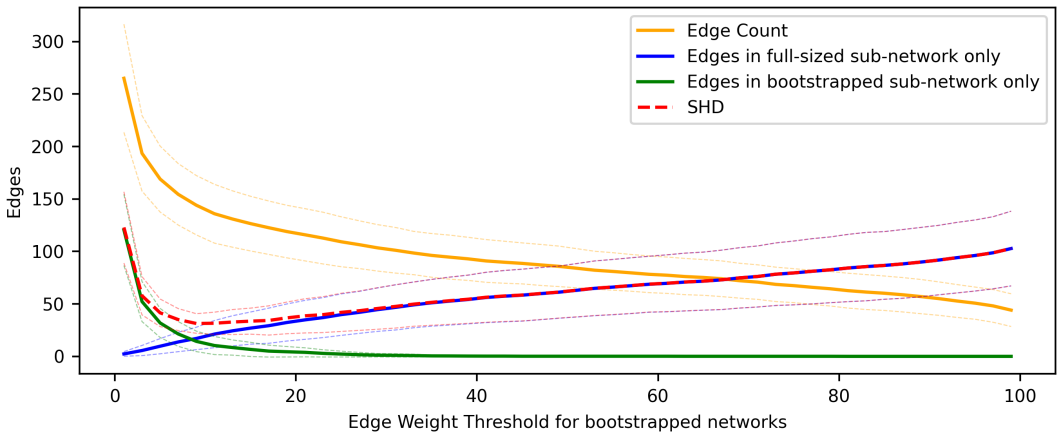

Figure 4.12: Averaged impact of edge weight thresholds for differences between original and bootstrapped sub-networks. Faint lines indicate standard deviation across the sub-networks.

The differences in edge generation between the bootstrapped and original sub-networks were explored across all edge-weight thresholds. These is shown in Figure 4.12 for averaged values with standard deviation across all sub-networks. When all

edges in the bootstrapped networks are included (i.e., a threshold of 1), the SHD is predominantly influenced by edges absent in the full network. At this low threshold, the networks are more susceptible to noise, leading to potentially spurious edges. As the threshold increases, the number of unique edges in the bootstrapped network decreases, approaching zero, while the number of unique edges in the original sub-network increases. This causes a reduction in the overall edge count until the average across bootstrapped sub-networks drops to 48. At this point, the majority of information within the stratified populations is lost.

Although selecting a threshold that minimises SHD between the bootstrapped and original sub-networks might seem a promising trade-off, this approach was ultimately rejected. All threshold options appeared to be somewhat arbitrary and risked losing valuable information about the variability of the networks generated from the samples.

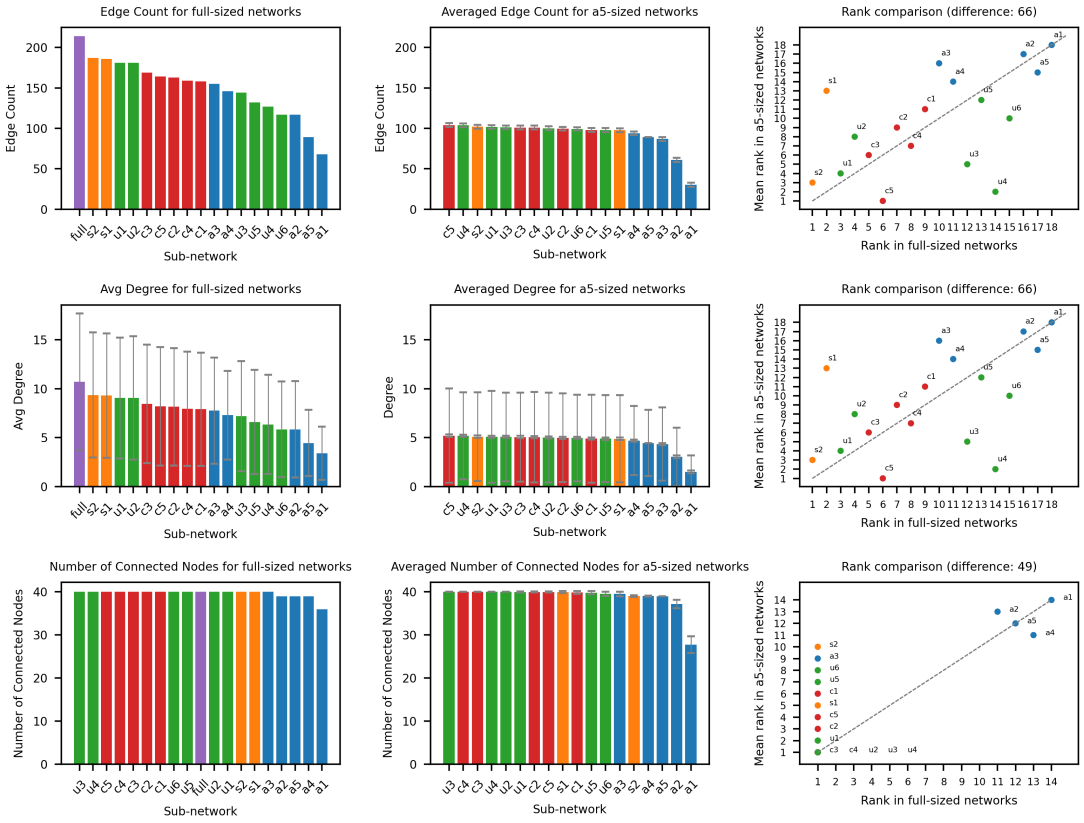

Figure 4.13: Comparison of sub-network characteristics and sub-network rank for Edge Count, Number of Connected Edges and Average Degree. For the rank plots, the 'difference' is the sum of all absolute changes in rank.

Instead, the averaged characteristic metrics of the sample networks were compared to the full sized ones, as shown in Figures 4.13 and 4.14. From the scatter plots

comparing the ranks between the sub-network sets for each metric, it is immediately clear that sample size as well as the sub-population sampled impacts the characteristics of the sub-networks. Generally, the small error bars for the a5-sized sub-network metrics (in the middle column) indicate that variance is consistently low and suggest that the rankings of these sub-networks are sufficiently stable to be used for comparison in this context.

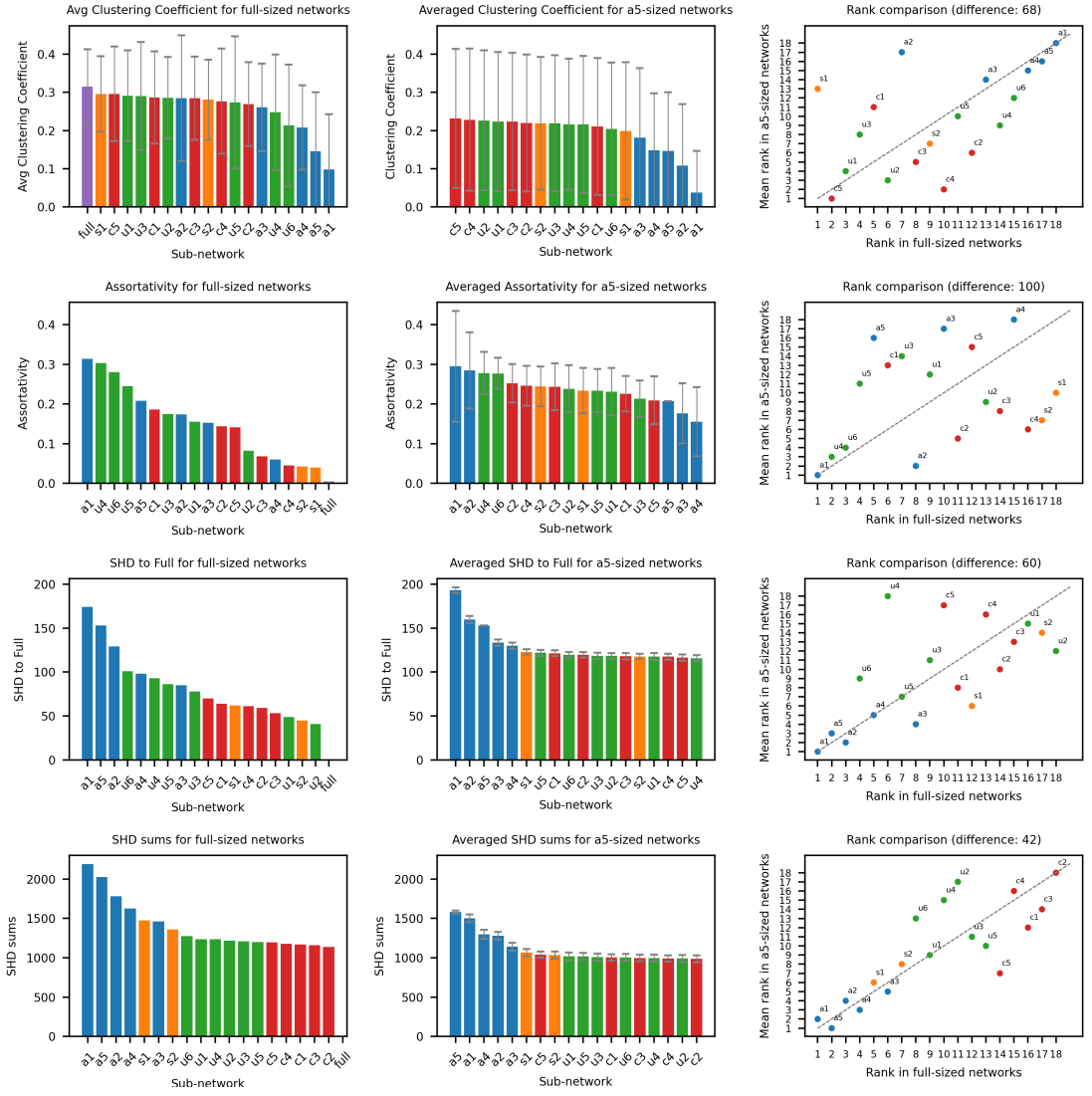

Figure 4.14: Comparison of sub-network characteristics and sub-network rank for Average Clustering Coefficient, Assortativity, Structural Hamming Distance to Full network, and Sum of Structural Hamming Distances to all other sub-networks. For the rank plots, the 'difference' is the sum of all absolute changes in rank.

As expected, the edge counts (and therefore there average degrees) for the a5-sized networks are significantly lower than the original sub-networks, except a5 which

is of course consistent between the two sets. The age sub-networks (other than a3) consistently have more lone nodes, which is due to low comorbidities of certain age-related diseases. For example, Parkinson's consistently has no edges in a1 because it appears 4 times in this subset, but not with comorbidities. However, the averaged number of connected nodes falls below 40 for 9 of the a5-sized sub-networks, as numerous samples do not contain sufficient evidence of the relationships found in the original sub-networks.

Across the two sub-network sets, clustering coefficients are naturally lowest for those with fewer edges. Similarly, the a5-sized networks (all with fewer edges) have lower clustering coefficients versus their original counterparts. All sub-networks are positively assortative, and significantly more so than the Full network. Unlike clustering coefficients, assortativity has a significant and negative correlation with sub-population size (Pearson R value: -0.587, p-value: 0.011). Whilst assortativity ranks are the most different of all metrics between the two sets, a1 is consistently most assortative, likely due to the fact that is has the lowest range in edge degrees in both sets.

As previously noted, the SHDs between the original sub-networks and the Full network are heavily influenced by sub-population size. When this size bias is mitigated in the a5-sized sub-networks, the structural differences related to the age-based networks (particularly a1,a2 and a5) are further emphasised. the This is also the case for the SHD sums (the combined distance of each sub-network from all others), where the two sex sub-networks are also consistently more different from the others. However, the inclusion of multiple age groups within the other sub-networks may blur out more subtle patterns within them. Further stratification (say into 30 sub-networks for each age and urbanity category) may appear to offer a more granular analysis, but the imbalances in the subset sizes would persist, meaning that the same bias would be encountered in analysis.

4.7 Adding Social Demographic Factors as Nodes

To examine how adding SDFs as single nodes impacted the full population multimorbidity network structure, four networks with one additional node representing an SDF were generated. These are referred to collectively as the SDF networks and as Age-1, Sex-1 etc. independently. A network with all of the SDF nodes (referred to as All-4 and shown in Figure 4.15) was also generated.

Clinical insights should be drawn from All-4 rather any of the SDF networks because

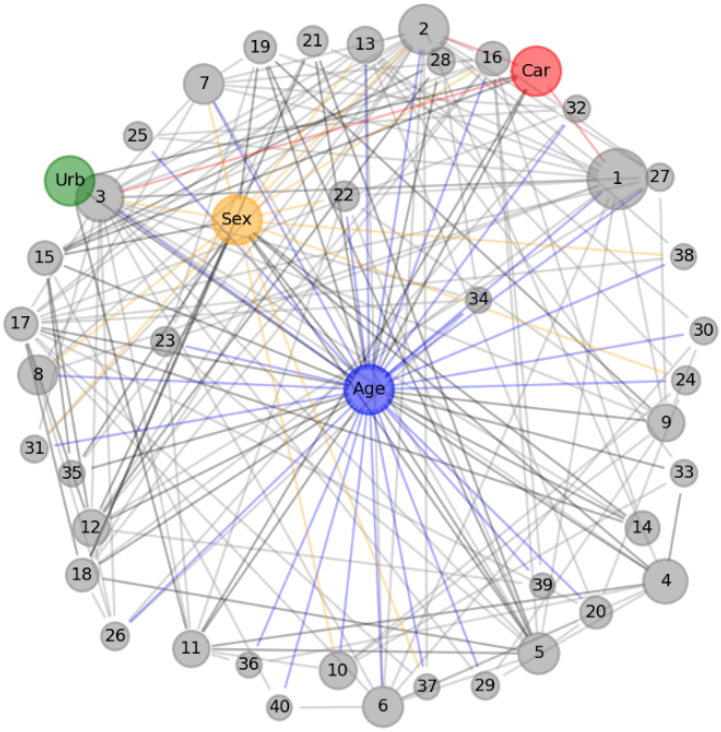

Figure 4.15: Visualisation of the All-4 network. Disease nodes are numbered per the prevalence ranks in Table B.2, and edges from SDF nodes are shown in the SDF's associated colour.

unlike these, it obeys the Causal Faithfulness Assumption for Bayesian networks by including all known variables. This means that the additional mediating impacts of any SDFs absent in the SDF networks are present in All-4. However, before focusing solely on All-4, it is useful to compare it to the SDF networks and the Full network, to better understand the influence of each of the SDFs.

Network	**Age-1**	**Sex-1**	**Urb-1**	**Car-1**
Age-1	-	7	8	7
Sex-1	88	-	12	10
Urb-1	91	14	-	0
Car-1	92	14	2	-

Table 4.2: Edge differences between SDF networks (considering disease edges only). Value $[i, j]$ is the number of edges in row i's network that are not in column j's network. Values $[i, j]$ and $[j, i]$ sum to give the SHD between networks i and j.

Table 4.2 is a matrix of the edge differences between the common (i.e. disease)

nodes of the SDF networks. It clearly shows Age-1 to be the outlier amongst these. It is the smallest network (as it all others contain many more edges than it), but all except 7-8 of the edges in Age-1 are present in the other SDF networks. The Car-1 and Urb-1 networks are remarkably similar, with their SHD comprising just two additional edges in Car-1. Sex-1 is shown to differ from Urb-1and Car-1 to very similar extents, and is slightly more similar to Age-1 than the other two. These inter-SDF patterns agree with the SHD sums between the different stratified sub-networks in Figure 4.14, which is to be expected.

Measure	**Full**	**Age-1**	**Sex-1**	**Urb-1**	**Car-1**	**All-4**
Total Edges	214	186	224	216	218	192
Disease-Disease edges	214	128	209	211	213	121
Disease-Disease edges not in full	-	7	10	2	2	6
Full edges present	-	57%	93%	98%	99%	53%
SDF-SDF edges	0	0	0	0	0	4
SDF-Disease edges	-	40	15	5	5	-
SDF-Disease edges in All-4	-	40	20	1	6	67
Shared SDF-Disease edges	-	40	13	1	4	-
Shared portion	-	100%	59%	20%	57%	-
SHD from All-4	105	13	102	104	104	-
SHD from Full	-	105	25	7	5	105

Table 4.3: Summary of SDF network characteristics. SHDs are calculated based on disease-disease edges only. 'Shared SDF-Disease edges' are those which appear in both the given SDF and All-4 networks.

Table 4.3 compares the SDF networks to the Full and All-4 networks. It is clear that the presence of SDF variables reduces the number of disease-disease edges compared to the Full network. This is a clear sign that, as anticipated, the disease-disease connections observed in the full network are being mediated by the SDFs. These effects are most notable in Age-1, which differs most significantly in structure from the Full network, mediating away the majority of its edges. Conversely, low counts of SDF-Disease edges and low SHDs demonstrate that Urb-1 and Car-1 barely differ from the Full network, indicating that their impact on disease-disease connections is minimal. Sex-1 has a higher number of SDF-disease edges than these two, but does not mediate away many of the disease-disease edges that are present in the Full network when compared to

Age-1.

The extent to which the SDF networks' node relationships are preserved in All-4 varies across the different SDFs. Age retains all of its 40 disease neighbours, suggesting these relationships are stable and relevant even when other factors are accounted for. In contrast, for the other SDFs, particularly urbanity, there is more difference between their sets of SDF-disease nodes and those in All-4. This indicates that their influences on diseases are largely mediated by the other factors (and namely Age).

In Figure 4.16, the range in each node's degree across the SDF networks is compared to those in All-4 and Full. It is clear that Urbanity, Carstairs Quintile and (slightly less so) Sex follow the degree distribution of Full closely. This is in alignment with their similarity to (i.e., low SHDs from) Full. Conversely, the close correlation between the degree distributions of All-4 and Age-1 again demonstrate how Age dominates the network structure. The disease nodes are ordered from left to right by decreasing prevalence. From this, it is evident that the decrease in degree caused by Age is higher for more prevalent nodes (Pearson R value: 0.596, p-value: 0). The edge adjacency

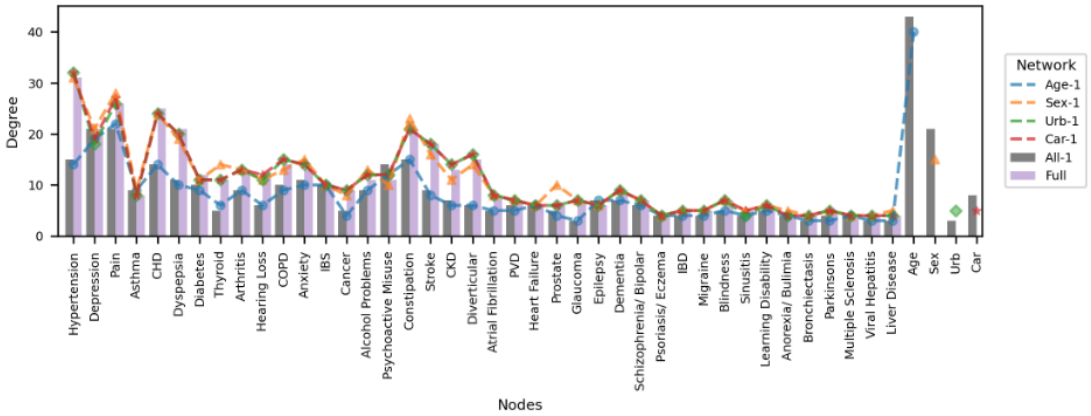

Figure 4.16: Comparison of node Degrees in SDF-1, All-4 and Full networks.

matrix in Figure 4.17 illustrates the cause for the drop in degree of the disease nodes clearly. In linking to all diseases (as shown by the brown cells in the Age row) the Age node mediates away many disease-disease connections found in the Full network (blue cells).

In terms All-4's other characteristics, it is connected and has an Average Clustering Coefficient of 0.331 (with standard deviation of 0.11) and assortativity of -0.447. Given its reduction of edges versus the Full network of 40%, this lower Average Clustering Coefficient is expected. The negatively assortativity is also unsurprising, given that the Age node has a degree of 43 and the next highest is 21. It is more interesting to note

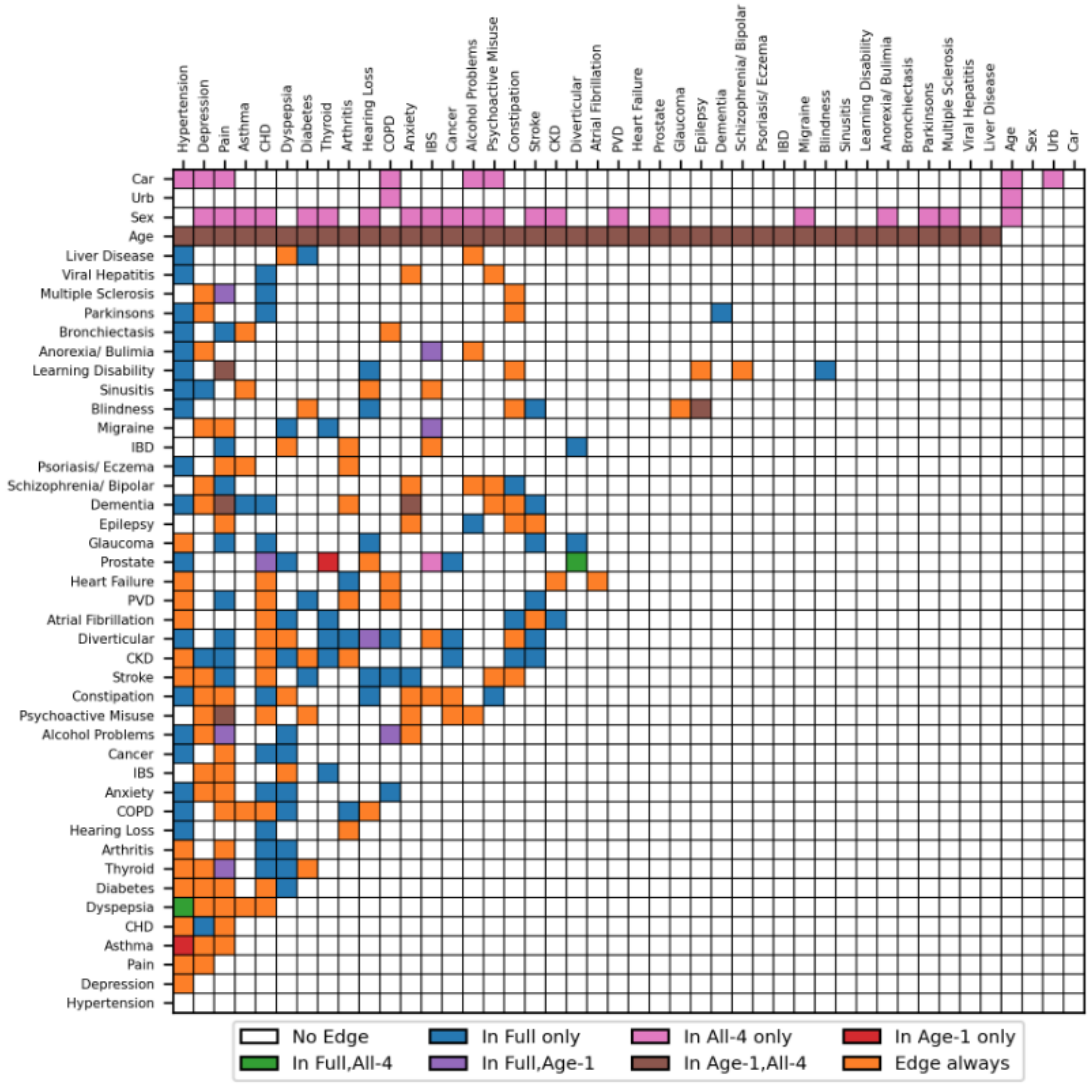

Figure 4.17: Adjacency matrix for comparing edges in Age-1, All-4 and Full Networks

that it is the only negatively assortative network generated in this work.

4.8 Impacts of Variable Discretisation

4.8.1 Discretisation Methods

All prior analysis of stratified networks and those with SDF nodes has been conducted using the manual discretisation of age. The impact of other discretisation methods is now explored.

Figure 4.18 shows the range in combined SHD sums across the age sub-networks resulting from the age bins considered by the exhaustive search algorithm. The results

for the other methods are also plotted, and show the success of the greedy search algorithm (it finds the best set of age bins). The results from the other methods are all shown to fall in the lower half of the ranking.

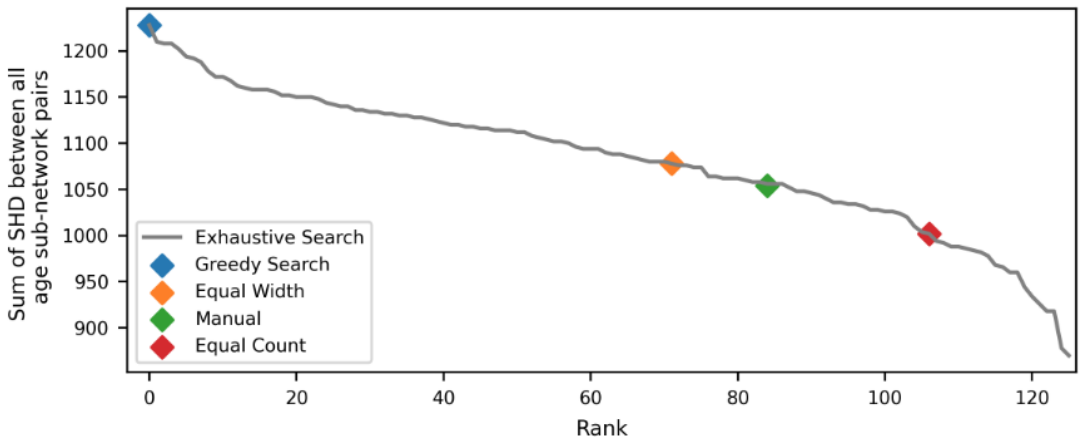

Figure 4.18: Enter Caption

The top and bottom Age bin sets are shown in Figure 4.19. The top 10 bins are consistent in that they all have a bin for Ages 0-9 and 90-100. The variability in the middle three bins further emphasises the importance of the outer two. Naturally, the bottom 10 bins make the same case, as their low combined SHDs appear to be caused by large outer bins. These findings emphasise the differences between the oldest and youngest patients relative to the rest of the population.

One benefit of the greedy search algorithm is that it finds the optimum bins in order of importance. It first finds a boundary at age 10, then at 60, 90 and 30. This confirms that multimorbidity patterns in the youngest 10% of patients differ more from the rest of the population than the top 10%. The second benefit of the greedy approach is that it is, of course, much faster than the exhaustive search. For the case explored with 5 bins and with bin boundaries being multiples of 10, the exhaustive search requires 630 structure-learning algorithm runs, whereas the greedy search requires 100 (only 16% of the exhaustive runs). However, the exhaustive search would require 19,380 runs if boundaries in multiples of 5 were considered instead, whilst the greedy search would only require 1.2% of this at 240 runs.

4.8.2 Network Structure Impacts

The impact of using the SHD-maximising age bins (i.e., the top bin set from Figure 4.19) to form stratified networks is essentially discernible from the objective function of

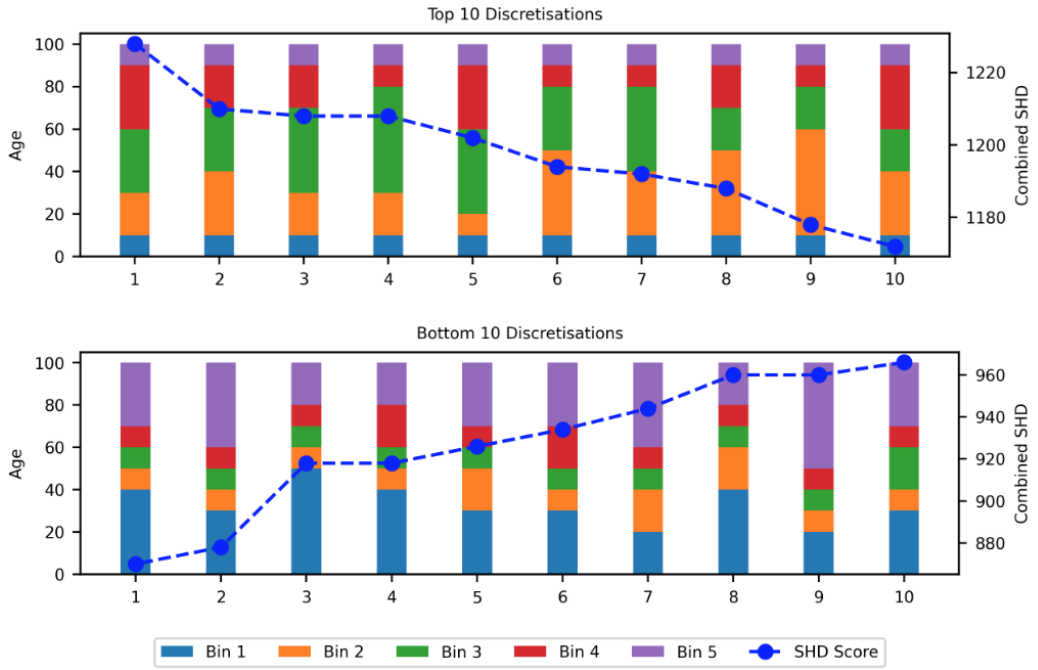

Figure 4.19: Top and Bottom 10 SHD-maximising Age bins found from exhaustive search (with boundaries constrained to multiples of 10).

the two algorithms. That is, the differences between the stratified age sub-network are maximised. Their impact on the All-4 network is explored in Table 4.4 and in Figure 4.20. To demonstrate the impact of more granular discretisation, an All-4 network was also created using 10 equal width bins with 10-year age ranges. This is compared to the manually-defined bins in Figure 4.21.

Clearly, discretisation of age is important and impacts the network structure, including the other SDFs' interactions with diseases. The impact of the 'optimal' bins is not overtly positive, in that it does not reduce the total number of edges. Clinicians would be better placed to comment on which of the 5-bin methods yields a structure that is more reflective of real world interactions, but clearly, the chosen bin boundaries for discretised nodes such as age are influential on network structures.

The effect of using 10 bins, which provides a more granular discretisation, is a significant reduction in both disease-disease and other SDF-disease edges. This is expected given the reduced information loss from the underlying data associated with having more bins. As such, more granular information with respect to age allows for more would-be edges to be mediated by age. Of course, the more granular an SDF category is made, the more complex the network's conditional probability distributions, which convey the strength of inter-node relationships, become to analyse. Therefore,

Measure	Manual bins	'Optimal' bins	10 bins
Total Edges	192	194	160
Disease-Disease edges	121	127	98
Disease-Disease edges not in full	6	1	2
Full edges present	53%	59%	44%
SDF-SDF edges	4	4	4
Age-Disease edges	40	40	40
Other SDF-Disease edges	27	23	18
Edges not in Manual-bin All-4	-	16	3
Edges not in Optimal-bin All-4	14	-	7
Edges not in 10-bin All-4	35	41	-
SHD from Full	105	89	120

Table 4.4: Summary of All-4 network characteristics with three discretisation variants for age bins. SHDs from Full are calculated based on disease-disease edges only.

it is important to consider how many bins are appropriate depending on the intended interpretation of the associated networks.

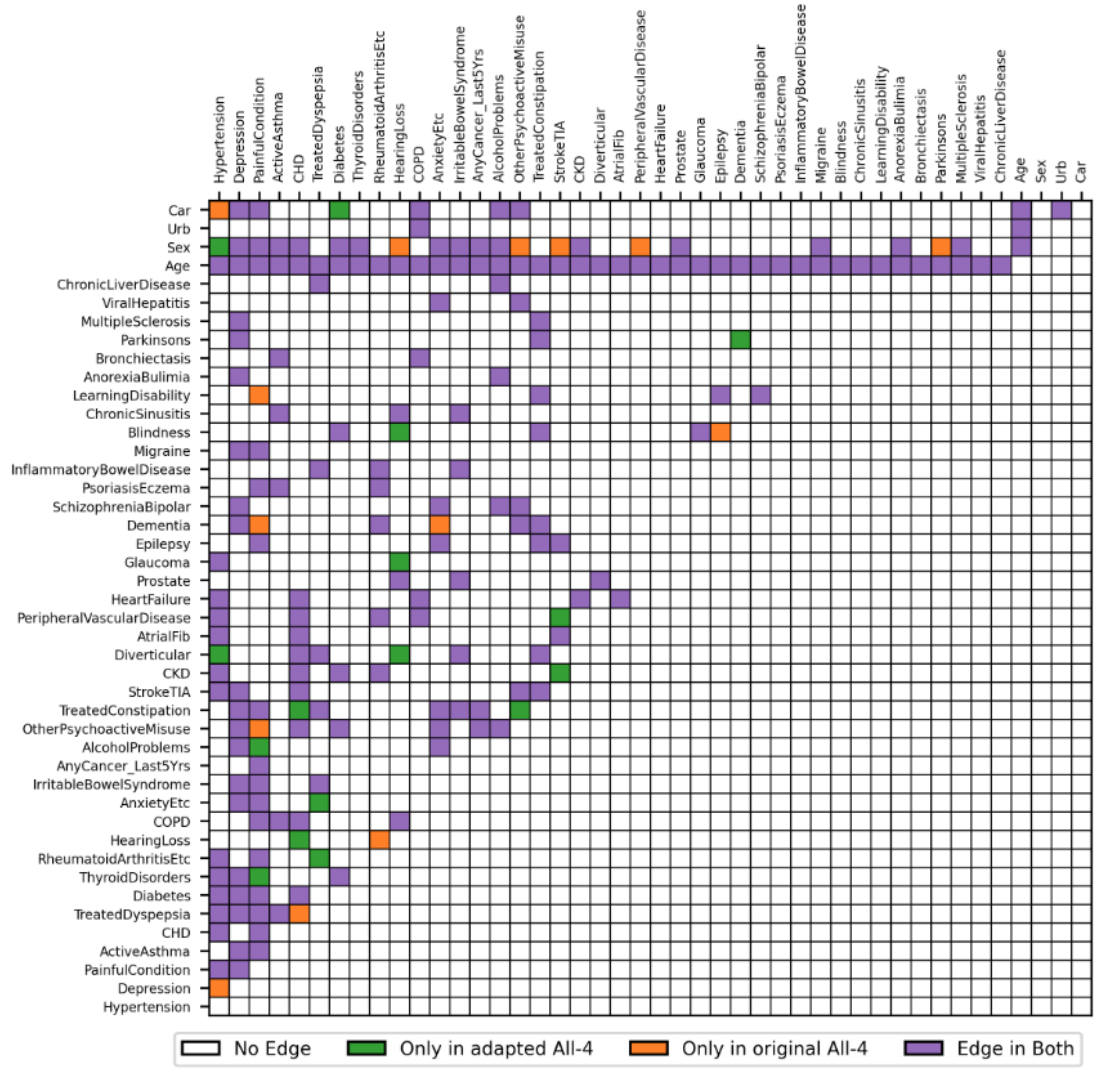

Figure 4.20: Comparison of the edges in variants of the All-4 network with age discretised manually (original network) and discretised 'optimally' using the SHD-maximising algorithms (adapted network).

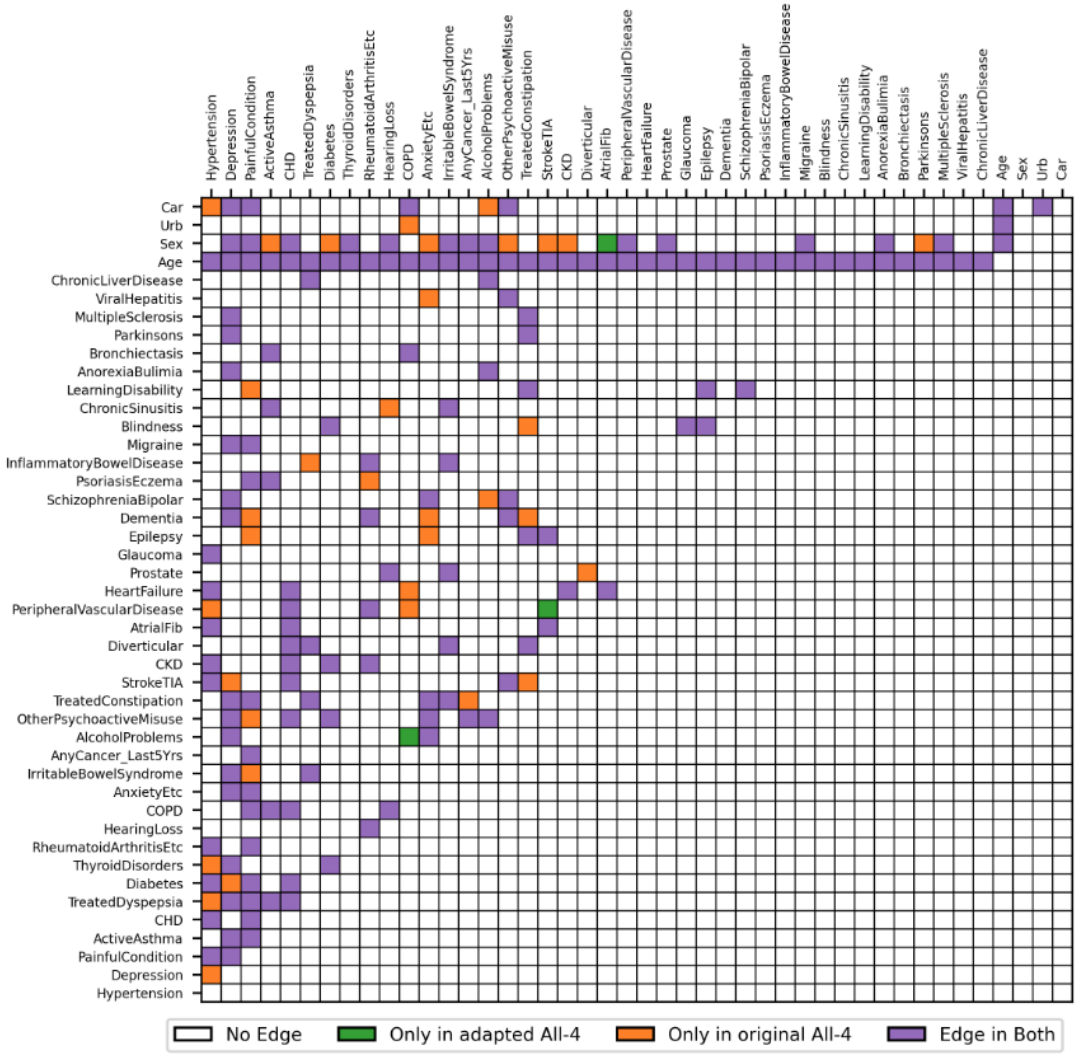

Figure 4.21: Comparison of the edges in variants of the All-4 network with age discretised manually (original network) and discretised into 10 equal-width categories (adapted network).

Chapter 5

Conclusions

5.1 Key Findings

Bayesian networks are a valuable tool to support clinical understanding of multimorbidity, which is a crucially important topic in medical and public health research. This work is motivated by the many methodological decisions that are required to produce Bayesian networks, and the various approaches to these in the existing academic literature. By examining the impact of methodology on the resulting network structures, several important insights have been provided, which highlight that caution should be applied when generating and drawing clinical insights from Bayesian multimorbidity networks.

As this study uses data from over a third of the Scottish population, and uses 40 diseases deemed by clinicians to most most suitable for multimorbidity studies, the findings herein can be expected to generalise beyond the dataset itself. An initial analysis of the dataset revealed that the sub-populations across five age categories exhibited significant deviations in the multimorbidity and disease prevalence trends across the whole population and in sex, deprivation and urbanity sub-populations. Although not surprising, this insight explains many of the patterns found in the sub-networks drawn from these sub-populations.

One of the most important methodological decisions for Bayesian networks is the algorithm used to learn their structure from the given data. A study of two variants of score-based (Tabu) and constraint-based algorithms (PC-stable) revealed that different algorithms can produce vastly different counts of edges, with the Tabu algorithms favouring more edges. Indeed even using different conditional independence tests for the PC-stable algorithm led to meaningful variations in edge generation. Whilst the

Tabu algorithm with BIC score was deemed to be the most suitable of the four variants for further use, it should be noted that all conclusions drawn from networks generated by it are influenced by this algorithm's tendencies.

When examining the Full network, drawn from the whole population, it was determined that the direction of edges is predictable in that the more prevalent disease will be the parent node. This justifies a limited analysis of edge directions in multimorbidity networks and that they should not be used to infer causality between diseases. Comparing the Full network to one drawn from only multimorbid patients revealed major differences in characteristics between the two, and highlighted that caution should be applied when comparing studies based on different patient datasets.

A key finding of this work is the strength of which sub-population size biases comparisons between associated stratified sub-networks. Only the characteristics of the age sub-networks, and in some cases sex sub-networks, were found to deviate from the trend associated with population size. Several approaches were considered in an attempt to explore the stratified networks without the influence of sub-population size, which raised drawbacks associated with averaging, sampling and bootstrapping. Although it is a popular technique, this is a major drawback of stratification as a method of analysing the influence of social demographics in multimorbidity networks.

Adding these social demographic factors as network nodes instead aligns better with Bayesian network theory, and is able to directly demonstrate that these factors, but mostly age, do influence disease manifestation and mediate many apparent disease-disease connections. Another advantage of incorporating SDFs as nodes is that only one network is required to consider their impacts. If stratification was used to study say, sex and age, then 10 networks (or generally twice the number of age categories applied) would need to be compared. However, as with stratification, the chosen number of and boundaries of categories that continuous variables are discretised into will impact network structures too. A method for discretising variables based on maximising the difference between resulting sub-networks was presented and applied to age. However clinical input is required to determine if there are cases where this method may be more valuable versus manual categorisation.

Based on all of these findings, it is recommended to use nodes instead of stratification when considering SDFs and other influential variables, and to acknowledge and justify all decisions made when generating Bayesian multimorbidity networks. If a particular method (such as discretisation of a variable like age) not somehow justified, then multiple approaches should be contrasted in order to draw robust rather than incidental

insights.

5.2 Limitations and Future Work

There are a number of limitations to this work that should be mentioned. Firstly, there is no treatment of Bayesian network parameters (conditional probability tables for connected nodes) in the analysis, other than to acknowledge that these become larger when more granular discretisation of continuous variables is applied. Demonstrating these impacts, and using the tables to answer questions such as 'What is the strength of influence of sex on diseases X and Y?' would further clarify the value of assessing SDF as nodes instead of via stratification.

Similarly, whilst Bayesian networks do not naturally have edge weights, it is possible to include these to represent additional information. A comparison of edge weighting approaches could help clarify the risks in utilising the network parameters for this purpose, as has been done in at least one case [11].

Although four structure learning approaches are applied in this work, the conclusions made on this topic are specific to those four and even more so to the parameters of these networks. Similarly, no hybrid algorithms were applied, nor were algorithms that can be applied to mixtures of continuous and discrete variables. Although no such algorithms were found to have implementations in Python, analysis could be conducted using R or proprietary software such as Tetrad [30] to examine how retaining naturally continuous variables impacts network structures. In this case, raw Carstairs scores could also be used (instead of Quintiles), allowing for a more thorough investigation of social deprivation impacts.

The SHD-maximising discretisation methods presented in this work are understood to be novel, but a more thorough investigation of their potential value would be useful. Given that they define bin boundaries based on maximising differences between sub-networks, they may be more applicable to stratification (in network science generally, rather than Bayesian networks) rather than for discretising networks' continuous nodes. The impact of sub-population size on the found bins should be studied, and it is possible that the algorithms could be adapted to balance the differences in structures and sub-population sizes.

Lastly, many networks have been generated in this work, but with analysis focusing primarily on their high-level structures and differences. Extending this analysis with input from clinicians could clarify which methodological decisions are the most and

least important, and to better assess the overall value of the findings herein to networks-based multimorbidity research.

Bibliography

[1] Luciana Pereira Rodrigues, Andréa Toledo de Oliveira Rezende, Letícia de Almeida Nogueira e Moura, Bruno Pereira Nunes, Matias Noll, Cesar de Oliveira, and Erika Aparecida Silveira. What is the impact of multimorbidity on the risk of hospitalisation in older adults? a systematic review study protocol. *BMJ open*, 11(6):e049974, 2021.

[2] Karen Barnett, Stewart W Mercer, Michael Norbury, Graham Watt, Sally Wyke, and Bruce Guthrie. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *The Lancet*, 380(9836):37–43, 2012.

[3] Eugene Jeong, Kyungmin Ko, Seungbin Oh, and Hyun Wook Han. Network-based analysis of diagnosis progression patterns using claims data. *Scientific reports*, 7(1):15561, 2017.

[4] Babak Fotouhi, Naghmeh Momeni, Maria A Riolo, and David L Buckeridge. Statistical methods for constructing disease comorbidity networks from longitudinal inpatient data. *Applied network science*, 3:1–34, 2018.

[5] Barret A Monchka, Carson K Leung, Nathan C Nickel, and Lisa M Lix. The effect of disease co-occurrence measurement on multimorbidity networks: a population-based study. *BMC Medical Research Methodology*, 22(1):165, 2022.

[6] Peter Marx, Peter Antal, Bence Bolgar, Gyorgy Bagdy, Bill Deakin, and Gabriella Juhasz. Comorbidities in the diseasome are more apparent than real: what bayesian filtering reveals about the comorbidities of depression. *PLoS computational biology*, 13(6):e1005487, 2017.

[7] Zhiwei Ji, Qibiao Xia, and Guanmin Meng. A review of parameter learning methods in bayesian network. In *Advanced Intelligent Computing Theories and*

Applications: 11th International Conference, ICIC 2015, Fuzhou, China, August 20-23, 2015. Proceedings, Part III 11, pages 3–12. Springer, 2015.

[8] Valerie Kuan, Spiros Denaxas, Praveetha Patalay, Dorothea Nitsch, Rohini Mathur, Arturo Gonzalez-Izquierdo, Reecha Sofat, Linda Partridge, Amanda Roberts, Ian CK Wong, et al. Identifying and visualising multimorbidity and comorbidity patterns in patients in the english national health service: a population-based study. *The Lancet Digital Health*, 5(1):e16–e27, 2023.

[9] Elma Dervić, Johannes Sorger, Liuhuaying Yang, Michael Leutner, Alexander Kautzky, Stefan Thurner, Alexandra Kautzky-Willer, and Peter Klimek. Unraveling cradle-to-grave disease trajectories from multilayer comorbidity networks. *npj Digital Medicine*, 7(1):56, 2024.

[10] Tamas Nagy, Bence Bruncsics, and Peter Antal. Bayesian network multimorbidity models in covid-19 mortality. In *Proceedings of the 28th Ph.D. Minisymposium Dept. Meas. Inf. Syst.*, pages 44–47, 2022.

[11] Zhilin Yong, Li Luo, Yonghong Gu, and Chunyang Li. Bayesian comorbidity network and cost analysis for asthma. *IEEE Journal of Biomedical and Health Informatics*, 26(9):4714–4724, 2022.

[12] Gabor Hullam, Peter Antal, Peter Petschner, Xenia Gonda, Gyorgy Bagdy, Bill Deakin, and Gabriella Juhasz. The ukb envirome of depression: from interactions to synergistic effects. *Scientific reports*, 9(1):9723, 2019.

[13] Marloes Maathuis, Mathias Drton, Steffen Lauritzen, and Martin Wainwright. *Handbook of Graphical Models*. CRC Press, 2018.

[14] Judea Pearl. *Causality*. Cambridge university press, 2009.

[15] Giovanni Briganti, Marco Scutari, and Richard J McNally. A tutorial on bayesian networks for psychopathology researchers. *Psychological methods*, 2022.

[16] Marco Scutari, Catharina Elisabeth Graafland, and José Manuel Gutiérrez. Who learns better bayesian network structures: Accuracy and speed of structure learning algorithms. *International Journal of Approximate Reasoning*, 115:235–253, 2019.

[17] Diego Colombo, Marloes H Maathuis, et al. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15(1):3741–3782, 2014.

[18] Stefano Beretta, Mauro Castelli, Ivo Gonçalves, Roberto Henriques, and Daniele Ramazzotti. Learning the structure of bayesian networks: A quantitative assessment of the effect of different algorithmic schemes. *Complexity*, 2018(1):1591878, 2018.

[19] Themistoklis Christos Mouliakos. *Bayesian Network Approach for Modelling and Inference of Communication Networks*. PhD thesis, Chalmers University of Technology, 2019.

[20] Márton Pósfai and Albert-László Barabási. *Network Science*. Citeseer, 2016.

[21] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65:31–78, 2006.

[22] Eric Perrier, Seiya Imoto, and Satoru Miyano. Finding optimal bayesian network given a super-structure. *Journal of Machine Learning Research*, 9(10), 2008.

[23] Guillermo Romero Moreno, Valerio Restocchi, Jacques D Fleuriot, Atul Anand, Stewart W Mercer, and Bruce Guthrie. Multimorbidity analysis with low condition counts: a robust bayesian approach for small but important subgroups. *EBioMedicine*, 102, 2024.

[24] Syed Hasib Akhter Faruqui, Adel Alaeddini, Carlos A Jaramillo, Jennifer S Potter, and Mary Jo Pugh. Mining patterns of comorbidity evolution in patients with multiple chronic conditions using unsupervised multi-level temporal bayesian network. *PloS one*, 13(7):e0199768, 2018.

[25] Mandana Rezaeiahari, Clare C Brown, Mir M Ali, Jyotishka Datta, and J Mick Tilford. Understanding racial disparities in severe maternal morbidity using bayesian network analysis. *Plos one*, 16(10):e0259258, 2021.

[26] The Scottish Government. Scottish government urban rural classification 2020. Technical report, The Scottish Government, 2022.

[27] Public Health Scotland. Public health scotland deprivation guidance for analysts, 2020. Accessed on 16/6/2024.

[28] Christopher Boulton and J Mark Wilkinson. Use of public datasets in the examination of multimorbidity: opportunities and challenges. *Mechanisms of Ageing and Development*, 190:111310, 2020.

[29] César A Hidalgo, Nicholas Blumm, Albert-László Barabási, and Nicholas A Christakis. A dynamic network approach for the study of human phenotypes. *PLoS computational biology*, 5(4):e1000353, 2009.

[30] Richard Scheines, Peter Spirtes, Clark Glymour, Christopher Meek, and Thomas Richardson. The tetrad project: Constraint based aids to causal model specification. *Multivariate Behavioral Research*, 33(1):65–117, 1998.

[31] Friedman Koller. Probabilistic graphical models: Principles and techniques, adaptive computation and machine learning, 2009.

[32] John H McDonald. *Handbook of Biological Statistics*, volume 2. Sparky House Publishing Baltimore, MD, 2009.

Appendix A

Definitions for Structure Learning Algorithms

A.1 Scoring Functions

A.1.1 BIC Score

The Bayesian Information Criterion (BIC) Score is a common function in data science that uses a complexity term to penalise the log likelihood of a model (a measure how well the model describes the underlying data). For Bayesian networks, it therefore aims to represent the relationships in the dataset with a penalty on the number of edges. In A.1 below, the first term is the log likelihood, and the second term is the penalty factor [31]. It is calculated as follows:

$$\begin{aligned} \mathrm{BIC}(\mathcal{G}, D) = &\sum_{i=1}^{n} \left(\sum_{\mathrm{pa}_i} \sum_{x_i} n(x_i, \mathrm{pa}_i) \log \frac{n(x_i, \mathrm{pa}_i)}{n(\mathrm{pa}_i)} \right) \\ &- \sum_{i=1}^{n} \left(\frac{1}{2} \log(N) \times |\mathrm{Pa}(X_i)| \times (|\mathrm{States}(X_i)| - 1) \right) \end{aligned} \tag{A.1}$$

Where:

- $\mathcal{G}$ is the graph structure of the Bayesian network.
- D is the dataset.
- X_i is the i-th variable in the dataset.
- $\mathrm{Pa}(X_i)$ is the set of parents of X_i in the network $\mathcal{G}$.

- $n(x_i, \text{pa}_i)$ is the count of data points where $X_i = x_i$ and its parents take the values pa_i.
- $n(\text{pa}_i)$ is the count of data points where the parents of X_i take the values pa_i.
- N is the total number of observations in the dataset.
- $|\text{Pa}(X_i)|$ is the number of possible parent state combinations.
- $|\text{States}(X_i)|$ is the number of possible states of the variable X_i.

A.1.2 K2 Score

The K2 Score is similar to BIC in that it uses the log-likelihood, but with a different penalty factor. [31]. It is calculated as follows:

$$\text{K2}(G,D) = \log(P(G)) + \sum_{i=1}^{n} \sum_{j=1}^{q_i} \left(\log\left(\frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \right) + \sum_{k=1}^{r_i} \log(N_{ijk}!) \right) \tag{A.2}$$

where:

- G is the graph structure of the network.
- D is the dataset.
- $\log P(G)$ is the log-likelihood of the structure G.
- n is the number of variables in the Bayesian network.
- q_i is the number of possible parent configurations for variable X_i.
- r_i is the number of possible states of variable X_i.
- N_{ij} is the sum of counts N_{ijk} for all states k of X_i given the parent configuration j.
- N_{ijk} is the count of instances where variable X_i is in state k and its parents are in configuration j.

A.2 Conditional Independence Tests

A.2.1 χ^2 Test

The χ^2 test, shown in Equation A.3, determines if there is a significant association between two variables. For Bayesian networks, it evaluates whether the observed frequencies of a variable's states are independent of the states of its conditioning set, based on the expected frequencies calculated when independence is assumed . It requires a 'contingency table' in which rows represent the different possible states of one variable and columns represent the states of the other [32]. It is calculated as follows:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{A.3}$$

where:

- O_{ij} is the observed frequency for the cell in the i-th row and j-th column.
- E_{ij} is the expected frequency for the cell in the i-th row and j-th column under the null hypothesis of independence.
- r is the number of rows in the table.
- c is the number of columns in the table.

A.2.2 G Test

The G test, shown in Equation A.4 is another statistical test for conditional independence between variables. Like the χ^2 test, it uses a contingency table, and compares the observed frequencies to the expected frequencies under the assumption of independence. However, it is based on likelihood ratio, and is calculated as follows:

$$G = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} O_{ij} \log \left(\frac{O_{ij}}{E_{ij}} \right) \tag{A.4}$$

where variables are the same as above for the χ^2 test.

Appendix B

Tables for PCCIU Dataset

Table B.1 provides the verbatim descriptions of the urbanity categories in the 6-fold Rural Urban Classification, as defined by the Scottish Government [26].

Category	**Classification**
1	Large urban areas: Settlements of over 125,000 people
2	Other urban areas: Settlements of 10,000 to 125,000 people
3	Accessible small towns: Settlements of between 3,000 and 10,000 people and within 30 minutes drive of a settlement of 10,000 or more
4	Remote small towns: Settlements of between 3,000 and 10,000 people and with a drive time of over 30 minutes to a settlement of 10,000 or more
5	Accessible rural: Settlements of less than 3,000 people within 30 minutes drive to a settlement of 10,000 or more
6	Remote rural: Settlements of less than 3,000 people and with a drive time of over 30 minutes to a settlement of 10,000 or more

Table B.1: The Six-fold Urban Rural classification categories

In Tables B.2 and B.3, the percentage prevalence and associated rank of each disease node in the dataset are given for the Full and Multimorbid populations. The 'short names' used for plotting are also provided.

Rank (Full)	**Name**	**Short name**	**% in Full**	**% in MM**	**Rank (MM)**
1	Hypertension	Hypertension	0.134	0.4517	1
2	Depression	Depression	0.082	0.2728	2
3	Painful Condition	Pain	0.072	0.2722	3
4	Active Asthma	Asthma	0.06	0.1361	8
5	CHD	CHD	0.047	0.1833	4
6	Treated Dyspepsia	Dyspepsia	0.045	0.1593	6
7	Diabetes	Diabetes	0.043	0.1608	5
8	Thyroid Disorders	Thyroid	0.041	0.1365	7
9	Rheumatoid Arthritis Etc	Arthritis	0.034	0.1217	10
10	Hearing Loss	Hearing Loss	0.034	0.1004	12
11	COPD	COPD	0.032	0.1133	11
12	Anxiety Etc	Anxiety	0.032	0.1289	9
13	Irritable Bowel Syndrome	IBS	0.03	0.0903	13
14	Any Cancer (Last 5 Yrs)	Cancer	0.025	0.0842	16
15	Alcohol Problems	Alcohol Problems	0.024	0.0789	19
16	Other Psychoactive Misuse	Psychoactive Misuse	0.024	0.0835	17
17	Treated Constipation	Constipation	0.022	0.0872	14
18	Stroke TIA	Stroke	0.021	0.085	15
19	CKD	CKD	0.019	0.0812	18
20	Diverticular	Diverticular	0.019	0.076	20

Table B.2: Name, prevalence and rank of the first 20 of 40 disease nodes in the Full dataset and the Multimorbid (MM) subset.

Rank (Full)	**Name**	**Short name**	**% in Full**	**% in MM**	**Rank (MM)**
20	Diverticular	Diverticular	0.019	0.076	20
21	Atrial Fib	Atrial Fibrillation	0.014	0.0553	21
22	Peripheral Vascular Disease	PVD	0.013	0.0505	22
23	Heart Failure	HeartFailure	0.011	0.0454	23
24	Prostate	Prostate	0.009	0.0315	25
25	Glaucoma	Glaucoma	0.009	0.0336	24
26	Epilepsy	Epilepsy	0.008	0.0223	28
27	Dementia	Dementia	0.007	0.0273	26
28	Schizophrenia Bipolar	Schizophrenia/ Bipolar	0.007	0.0252	27
29	Psoriasis Eczema	Psoriasis/ Eczema	0.007	0.0199	29
30	Inflammatory Bowel Disease	IBD	0.006	0.0167	32
31	Migraine	Migraine	0.006	0.0177	31
32	Blindness	Blindness	0.005	0.0184	30
33	Chronic Sinusitis	Chronic Sinusitis	0.005	0.0156	33
34	Learning Disability	Learning Disability	0.003	0.0091	34
35	Anorexia Bulimia	Anorexia/ Bulimia	0.003	0.0087	35
36	Bronchiectasis	Bronchiectasis	0.002	0.0061	38
37	Parkinsons	Parkinsons	0.002	0.0062	37
38	Multiple Sclerosis	Multiple Sclerosis	0.002	0.0069	36
39	Viral Hepatitis	Viral Hepatitis	0.001	0.0026	40
40	Chronic Liver Disease	Liver Disease	0.001	0.0059	39

Table B.3: Name, prevalence and rank of the second 20 of 40 disease nodes in the Full dataset and the Multimorbid (MM) subset.

Appendix C

Algorithm Pseudocode

Algorithm 1 below is the 'get_SHD_sum' function used by Algorithms 2 and 3. It takes a dataset and a set of age category boundaries as inputs, and first creates a new column for age category. It then runs the structure learning algorithm separately for the disease columns corresponding to the rows in each age category. It returns the cumulative sum of the structural hamming distances between each pair of age-category networks.

Algorithm 1 get_SHD_sum

Inputs: ($data$, bin_set)

append $age\_category$ column to $data$ using bin_set
$networks \leftarrow$ empty list
for $category$ in $age\_category$ **do**
$strat\_data \leftarrow$ disease columns of $data$ for all ages within $age\_category$
$network \leftarrow$ **call** learn$_s tructure$(strat_data) **append**networktonetworks

$total\_shd \leftarrow 0$
for each pair $(net1, net2)$ in $networks$ **do**
$total\_shd \leftarrow total\_shd + shd(net1, net2)$
end for

return $total\_shd$

Algorithm 2 finds the SHD-Maximising age bin boundaries using an exhaustive search. For the given age range and required number of bins, it finds all possible bin boundary sets that obey the minimum bin width and bin width increment constraints. It then uses Algorithm 1 to find the SHD sums for all valid sets, and returns all sets, ranked by their associated SHD sums.

Algorithm 2 SHD-Maximising age bins: exhaustive search

Inputs: (*data*, *age_min*, *age_max*, *min_width*, *bin_inc*, *num_bins*)

Boundaries ← list from $range(age\_min, age\_max, bin\_inc)$
All_combos ← list of all combinations of $num\_bins - 1$ items from *Boundaries*
Valid_combos ← empty list
for *combo* in *All_combos* **do**
    bin_set ← list from *age_min*, *combos*, *age_max*
    bin_widths ← list of $bins[i+1] - bins[i]$ for *i* in $range(len(bins) - 1)$]
    if a **then**ll *widths* in *bin_widths* > *min_width*:
        append *bin_set* to *valid_combos*
    end if
end for
Bins_SHDs ← empty list
for *bin_set* in *valid_combos* **do**
    shd_score ← **call** get_SHD_sum(*data*, *candidate_bins*)
    append ($total\_shd$, *candidate_bins*) to *Bins_SHDs*
end for
return *Bins_SHDs* sorted by $total\_shd$

Algorithm 3 finds the SHD-Maximising age bin boundaries using a greedy search. It starts with the maximum and minimum ages as the boundary set, and considers each possible new boundaries (that obeys the minimum bin width and bin width increment constraints). It then uses Algorithm 1 to find the SHD sums for the potential new boundary set containing each valid boundary, and adds that which maximises the SHD sum. This is repeated til the required number of boundaries is found.

Algorithm 3 SHD-Maximising age bins: greedy search

Inputs: ($data$, age_min, age_max, min_width, bin_inc, num_bins)

```
current_bounds ← [age_min, age_max]
while len(current_bounds) − 1 < num_bins do
    best_score ← −∞
    best_bound ← None
    for new_bound in range(age_min + bin_inc, age_max, bin_inc) do
        if new_bound in current_bounds then
            continue                                  ▷ go to next bound in range
        end if
        candidate_bins ← sorted list of current_bounds and new_bound
        bin_widths ← list of (b1 − b2) for consecutive (b1, b2) in candidate_bins
        if all widths in bin_widths ≥ min_width then
            shd_score ← call get_shd_score(data, candidate_bins)
            if shd_score > best_score then
                best_score ← shd_score
                best_bound ← new_bound
            end if
        end if
    end for
    append best_bound to current_bounds
    current_bins ← sorted current_bins
end while
return current_bins, best_score
```