Chapter 1

Measuring Multimorbidity

1.1 Introduction

As Chapter **??** showed, multimorbidity can be defined as the presence of two or more chronic health conditions within an individual (NICE, 2016). It is associated with higher mortality (Gijsen *et al.*, 2001), increased use of health care (Gijsen *et al.*, 2001; Salisbury *et al.*, 2011), psychological distress (Fortin *et al.*, 2006), worse quality of life (Fortin *et al.*, 2004, 2005), and worse functional status (Kadam and Croft, 2007). It not only affects older people but has been observed in greater absolute numbers in those under the age of 65 and affects those with lower socioeconomic status disproportionately (Barnett *et al.*, 2012). As the proportion of older people in the population increases, multimorbidity is expected to affect increasing numbers of people in the future (OECD, 2011; Imison, 2012).

Many of the negative outcomes associated with mulitmorbidity are due to the structure of healthcare delivery which tends to concentrate treatment goals towards single diseases (OECD, 2011, Starfield *et al.* (2005)). As a result healthcare for individuals with multimorbidity can, at best, preferentially treat one disease to the detriment of others or, at worst, cause harm and affect patient safety e.g. through medication interactions (May *et al.*, 2009).

Epidemiological research into multimorbidity has tended to count numbers or report proportions of health conditions. There is, however, marked heterogeneity in the population considered and the number of diseases included in measurement. As a result, prevalence rates of multimorbidity have been shown by systematic reviews to vary widely in the general population from 13.1% - 71% (Fortin *et al.*, 2012) and 12.9% - 95.1% (Violan *et al.*, 2014). There has been little research into the prevalence of combinations of diseases and non-random, co-occurrence of diseases partly due to the high number of theoretical combinations meaning large samples and complex calculations are required (van-den-Akker *et al.*, 2001; Cornell *et al.*, 2009). More recent research has applied a variety of statistical techniques to overcome this problem (Prados-Torres

et al., 2014).

Identification of non-random co-occurrence of health conditions is important for a number of reasons; a) to gain a better understanding of the complicated nature of multimorbidity, b) to help assess the impact multimorbidity has on health outcomes, c) to help assess the sociodemographic differences in prevalence of multimorbidity which could have implications for health policy and the delivery of health services, and d) unexpected non-random associations could prompt further research into possible causal mechanisms for the association.

The objective of this chapter was to apply a novel two-way clustering framework to a large dataset based on the Scottish population, The framework aims to identify clusters of the most significant non-random co-occurrence of health conditions and multimorbidity patterns among individuals (Ng, 2015). The specific aims were to a) Identify non-random associations of health conditions from the dataset, b)identify if meaningful, homogeneous sub-groups of individuals according to groups on non-random conditions can be formed, c) assess the sociodemographic make-up of identified clusters.

Add para describing purpose of this chapter to overall thesis

1.2 Background

Previous clustering attempts in the field of multimorbidity have employed a number of different statistical techniques including; factor analysis, cluster analysis, the observed to expected ratio, multiple correspondence analysis (Prados-Torres *et al.*, 2014), principal component analysis, and latent class analysis (Islam *et al.*, 2014). Each of these techniques is a variant of latent variable modelling where clusters are deemed unobserved, or latent, variables that can be measured indirectly depending on the values of two or more observed variables (Collins and Lanza, 2013). In their systematic review of clustering techniques, Prados-Torres et al (2014) found marked heterogeneity in the numbers of diseases included in analyses, populations studied, and resulting clusters of conditions. The authors recommended that future attempts at clustering should be conducted using large numbers of diseases and in population-based datasets as opposed to sub-groups of populations.

An important decision in any clustering research is which statistical technique to apply to a given dataset. Collins & Lanza (2013) argue that when the latent variable (cluster) and the observed variables used to identify the latent variable are categorical in nature, Latent Class Analysis (LCA), a version of a finite mixture model, is an appropriate statistical technique to employ. Other techniques such as factor or cluster analysis rely on latent and/or observed variables being continuous in nature (Collins and Lanza, 2013). The aim of this chapter is to identify whether homogeneous sub-groups of individuals can be identified from a dataset of diseases represented by binary, categorical data.

Such sub-groups would also be categorical in nature suggesting LCA is an appropriate technique to apply.

The limitation to LCA is that as the number of observed variables increases it becomes more difficult for the LCA model to be well identified (Collins and Lanza, 2013). This is because the first step of LCA is to create a contingency table of all possible combinations of outcomes. For example, in a simple dataset with two binary Yes/No variables there are $2^2 = 4$ potential response outcomes; No/No, Yes/No, No/Yes, and Yes/Yes. In the example of the dataset used in the current chapter with 40 health conditions recorded as binary variables the number of cells in the contingency table will contain 2^{40} (over 1 trillion) cells. This makes good model identification highly unlikely. Ideally the ratio of n/W where n denotes the sample size and W denotes the number of potential response outcomes should be as high as possible (Collins and Lanza, 2013). When this ratio is very small there are only two options to overcome the problem, "…increase the amount of known information or reduce the amount of unknown information." (Collins and Lanza, 2013 :93)

In response to the "high-dimensional" problem created when trying to identify latent variables from health datasets with large numbers of diseases, Ng (2015) proposed a two-way model to identify multimorbidity clusters. The first step of this method is to "clump" diseases in the dataset into statistically correlated groups of conditions thereby reducing the number of variables and therefore the size of contingency table of potential responses. This first step is described in more detail in Ng et al (2012). The second step of the method then aims to identify latent groups of individuals based on response patterns to these "clumped" groups with a finite mixture model similar to LCA (Ng, 2015). This technique was applied to an Australian national health survey which contained self-reported responses of the presence of 24 physical and mental health conditions with full results detailed in Ng (2015). The aim of this chapter is to identify whether this technique is valid in a much larger dataset of administrative health data.

1.3 Methods

1.3.1 Data

In Scotland, much multimorbidity research has been informed by the Scottish Programme for Improving Clinical Effectiveness in Primary Care (SPICE-PC) dataset (Elder *et al.*, 2007; Barnett *et al.*, 2012; McLean *et al.*, 2015; Agur *et al.*, 2016). Diagnostic data from the year 2007 is available on 1,754,133 people in Scotland drawn from 314 general practices. The anonymised dataset has information on the presence of 32 physical and 8 mental health conditions (?? add box with list of diseases??) in addition to age, gender and deprivation index data. Diagnostic data is derived from codes entered into IT systems in General Practices and prescription data. A full de-

scription of the 40 included diseases and the methods used to classify them is available as supplementary information in Barnett et al (2012). The analysis for this chapter was restricted to adults over the age of 18 resulting in n=1,426,196. (10% sample of this so far!!) Ethical approval for secondary analysis of the SPICE-PC dataset using LCA was granted by the Research Ethics Committee of the College of Social Sciences at the University of Glasgow on 29/04/2016.

1.3.2 Analysis

The two-way method proposed by Ng (2015) was applied to the above dataset of 40 health conditions. Firstly groups of conditions that co-occur are identified and then latent groups of individuals are identified using a mixture model approach. To identify the groups of conditions that co-occur the method aims to calculate the significant pairwise multimorbidity between all conditions. The asymmetric Somer's D statistic was used to quantify the degree of random co-morbidity with all pairs of health conditions. Significance of the Somer's D statistic for each pair of conditions was calculated using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) to control the false discovery rate (FDR) with $\alpha = 0.001$. Diseases were then clustered into overlapping groups using a "clumping" procedure based on the technique described by Jardin & Robson (1968) and fully described as equation 11 in Ng et al (2012). The strength of multimorbidity in each cluster was calculated using the average pairwise Somer's D statistics of disease within the group. From these overlapping groups, non-overlapping groups of diseases were created using an amended version of the algorithm specified in Ng (2015):-

1. Name the cluster with the highest strength as the first group and then remove its member health conditions in all subsequent clusters with smaller strength. Each member in a cluster must have the condition with which the pairwise concordance statistic is maximum in the same cluster;
2. Repeat (1) for the next cluster and name it as a group if it is not 'singular' (singular cluster is defined as a cluster consisting of a single health condition);

3. If a group is formed in (2), remove its member health conditions in all subsequent clusters with smaller strength or singular clusters;
4. Repeat (2) and (3) until all clusters are visited;
5. Put the condition in a singular cluster into a predefined group where more than half of the member conditions are significantly comorbid with the condition;
6. Name those remaining singular clusters as a singular group.

A matrix was then created assigning a score to individual observations for each identified non-overlapping group. Scores were assigned as:-

- 1 for having no diseases within the group
- 2 for having one of the diseases in the group
- 3 for having 2 or more diseases within the group.

In the second part of the analysis a mixture-model of multivariate generalised Bernoulli distributions was applied to the matrix of these scores in order to identify latent groups of individuals according to response patterns to the score matrix. The most appropriate number of latent groups to fit the dataset was determined by the lowest Bayesian Information Criterion (BIC) and substantive theoretical analysis. Models with 3 to 9 latent groups were compared to find best model fit. Each observation was then assigned to the latent group for which the posterior probability was highest. Descriptive analysis of sociodemographic difference in latent group was then employed to identify patterns in groupings.

All statistical analysis was completed using R version 3.3.3 (R-Core-Team, 2017). Two-way clustering was conducted with amended code provided by Ng(2015). Data manipulation and visualisation was conducted using tidyverse packages (**???**) and the corrplot package.

1.4 Results

1.4.1 Pairwise correlation

With α controlled at 0.001, 143 of a possible 780 pairwise correlations proved to be significant. A matrix showing all statistically significant correlations is shown in Figure 1.1. The number of expected false positives for 143 pairs is less than 1.

1.4.2 Grouping diseases

Using the "clumping" procedure detailed in Ng (2015), Thirty-eight overlapping groups of conditions were found as shown in Table 1.1

Six diseases; Epilepsy, Learning Disability, Sinusitis, Crohns, Anorexia, Psoriasis/Eczema did not have strong enough pairwise correlations to be included in any of the 38 groups. Thirteen non-overlapping groups were derived from the results of the clumping method using the algorithm described above and named according to the characteristics of the member disease in the groups as shown in Table 1.2.

A further four diseases; Diverticular disease, Prostate, IBS, and Dyspepsia were excluded from the non-overlapping groups. These four conditions did not appear in any groups with the diseases with which they had strongest pairwise correlation, a condition that had to be met in the first stage of the algorithm described above.

Table 1.1: Overlapping groups of diseases derived from "clumping" procedure

Group	Diseases Included	Strength
1	CHD, CKD, AtrialFib, Hypertension, HeartFailure, Rheu-Arthritis, PVD	0.2833
2	CHD, CKD, AtrialFib, Diabetes, TIA-Stroke, Hypertension, HeartFailure	0.2806
3	CHD, CKD, Hypertension, HeartFailure, Rheu-Arthritis, Pain	0.2653
4	CHD, Hypertension, Prostate	0.2622
5	CHD, CKD, Diabetes, TIA-Stroke, Hypertension, HeartFailure, Pain	0.259
6	CHD, TIA-Stroke, Hypertension, HeartFailure, Laxat-Constip, Pain	0.2525
7	Mental-Alcohol, LiverDisease, Depression, Anxiety	0.2512
8	CHD, Hypertension, HeartFailure, Bronchitis, Pain	0.2487
9	CHD, CKD, Diabetes, TIA-Stroke, Hypertension, PVD, Pain	0.2348
10	CHD, CKD, Hypertension, PVD, Rheu-Arthritis, Pain	0.2331
11	Mental-Alcohol, Mental-Psycho, ViralHepatitis, Depression, Anxiety	0.2324
12	Mental-Psycho, Depression, Schiz-Bipolar, Anxiety	0.2304
13	CHD, Hypertension, Diverticular, Laxat-Constip, Pain	0.2294
14	Depression, Migraine, Pain	0.2278
15	CKD, ThyroidDisorders, Hypertension, Pain	0.2262
16	CHD, Blindness, Glaucoma, Hypertension	0.2236
17	Depression, Dyspepsia, Laxat-Constip, Pain, Anxiety	0.218
18	MS, Depression, Laxat-Constip, Pain	0.2178
19	CHD, Diabetes, TIA-Stroke, Blindness, Hypertension, Pain	0.2171
20	ActiveAsthma, Bronchitis, Bronchiectasis	0.2166
21	CHD, TIS-Stroke, Blindness, Hypertension, Laxat-Constip, Pain	0.2138
22	CHD, Hypertension, PVD, Bronchitis, Pain	0.2114
23	CHD, Dementia, Parkinsons, TIA-Stroke, Hypertension, Laxat-Constip, Pain	0.2068
24	Dementia, Parkinsons, Depression, Laxat-Constip, Pain, Anxiety	0.2068
25	CHD, Dementia, Hypertension, Rheu-Arthritis, Pain	0.204
26	Diabetes, Hypertension, LiverDisease, Pain	0.1964
27	Dementia, Mental-Psycho, Depression, Pain, Anxiety	0.1939
28	Hypertension, Diverticular, Dyspepsia, Laxat-Constip, Pain	0.1938
29	CHD, Dementia, Parkinsons, Hypertension, Laxat-Constip, Pain, Anxiety	0.1922
30	Hypertension, Bronchitis, Bronchiectasis, Pain	0.1917
31	Blindness, HearingLoss, Hypertension	0.179
32	Dementia, Parkinsons, TIA-Stroke, Depression, Laxat-Constip, Pain	0.173
33	HearingLoss, Hypertension, Prostate	0.1729
34	AnyCancer-Last5Yrs, Hypertension, Pain	0.1707
35	Hypertension, LiverDisease, Dyspepsia, Pain, Anxiety	0.1593
36	Depression, IBS, Pain	0.1563
37	CHD, Dementia, Mental-Psycho, Pain, Anxiety	0.1385
38	Diverticular, IBS, Pain	0.1226

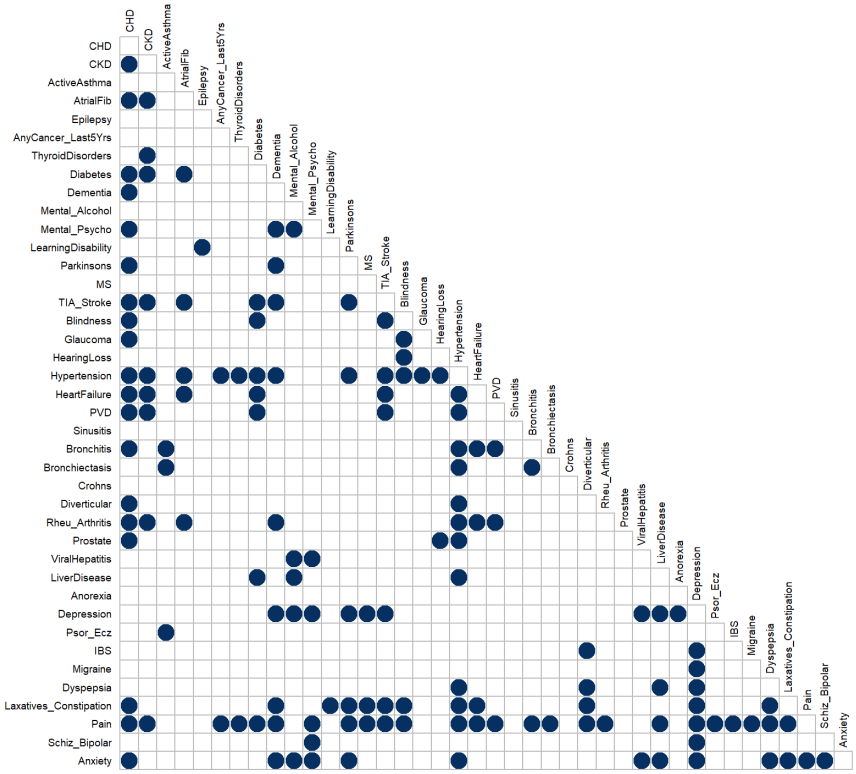

Figure 1.1: Significant Pairwise Correlations

Table 1.2: Non-overlapping groups of diseases derived from Ng algorithm

Disease group name	**Diseases included**
Cardiovascular	CHD, CKD, AtrialFib, Hypertension, HeartFailure, Rheu-Arthritis, PVD
Diabetes/Stroke	Diabetes, Stroke
Pain/MS	Laxatives/Constipation, Pain, MS
Mental Health/Liver	Mental/Alcohol, Liver Diseases, Depression, Anxiety
Psychosis	Mental-Psycho, Viral Hepatitis
Sciz/Bipolar	Sciz-Bipolar
Migraine	Migraine
Thyroid	Thyroid disorders
Blindness	Blindness, Glaucoma
Respiratory	ActiveAsthma, Bronchitis, Bronchiectasis
Neurodegenerative	Dementia, Parkinson's
Hearing Loss	Hearing Loss
Cancer	Any Cancer in Last 5 years

1.4.3 Grouping individuals

Individuals were assigned a score depending on the number of diseases they had in each of the non-overlapping groups according to the criteria described above. The finite mixture model of multivariate generalised Bernoulli distributions was then applied to this dataset in order to identify latent groups of individuals depending on their scores for each of the 13 groups, based on presence of 30 diseases. BIC scores suggested the model for nine latent groups had best fit to the data. The nine-group model was compared with the model with second-lowest value of BIC, that for eight latent groups. However, despite offering a more parsimonious solution, the eight-group model did not provide a more substantive theoretical fit to the data. Thus, the nine-group model was deemed best fit and analysed further.

Item-response probabilities for the nine-group model are shown in Figure 1.2.

Each latent group was labelled according to these item response probabilities as follows;

- Latent Group 1. **Well**. High probabilities of having no diseases in any of the groups.
- Latent Group 2. **Cardiovascular only**. Item response in this Latent group have more than 50% chance of having at least 1 condition from the Cardiovascular group. High probabilities of having no diseases in any of the other groups.
- Latent Group 3. **Cancer**. Individuals in this group have almost 100% probability of having had cancer in last 5 years. Weak probabilities across some other groups.
- Latent Group 4. **Mental Health/Pain** High probability of having at least one diseases from Pain/MS group and at least one disease from Mental Health/Liver group.
- Latent Group 5. **Mental and Physical Multimorbidity** High probabilities of having at least 1 cardiovascular, diabetes/stroke, Pain/MS, and Mental health/Liver disease. This group also has the highest probability across latent groups for having 1 respiratory disease.
- Latent Group 6 **Physical Multimorbidity** High probability of a least one cardiovascular disease and one of Diabetes/Stroke.
- Latent Group 7. **Physical multimorbidity (Strong Cardio)**. Similar to Latent group 6 but individuals in this group are more likely to have 2 or more cardiovascular diseases.
- Latent Group 8. **Dementia with Mental/Physical MM** Individuals in this latent group have almost 100% probability of having dementia or Parkinson's. Also highly likely to have at least one Cardiovascular disease and mildly raised probability of having mental health or psychosis diseases.
- Latent Group 9. **Mental Health only**. High probability of having a least 1 of the Mental Health/Liver group diseases. Also mildly raised probabilities in the Psychosis and Sciz_Bipolar groups.

Figure 1.2: Item response probabilities for nine latent group model

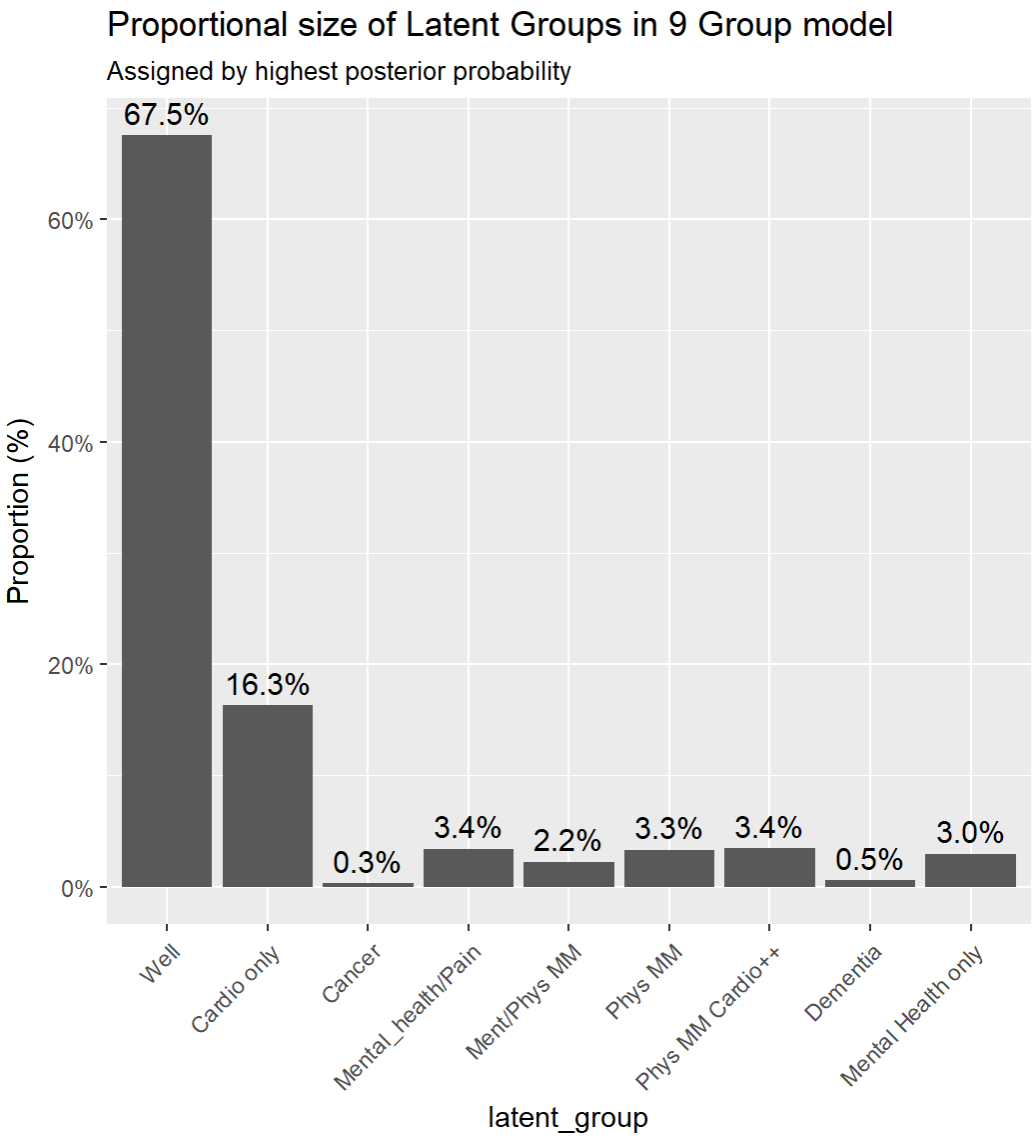

Figure 1.3: Latent Group Proportions

Proportionate size of each of the assigned latent variable groups is shown in Figure 1.3.

The "Well" latent group was by far the largest group with almost 70% of individuals within it. The other groups formed much smaller proportions of the data.

1.4.4 Sociodemographic breakdown of latent groups

Each of the Histograms of Age for each of the latent groups shown in Figure 1.4 show variation in distribution. The "Well" and "Mental Health only" groups have a much younger age distribution compared to the other groups. The "Dementia" group has a much older distribution.

Figure 1.5 shows higher proportions of males in the "Physical MM" latent group. The "Well" and "Physical MM high Cardio" groups have even splits by gender. All other groups have higher proportions of females.

Figure 1.6 shows distribution of deprivation across latent groups by Carstairs Decile.

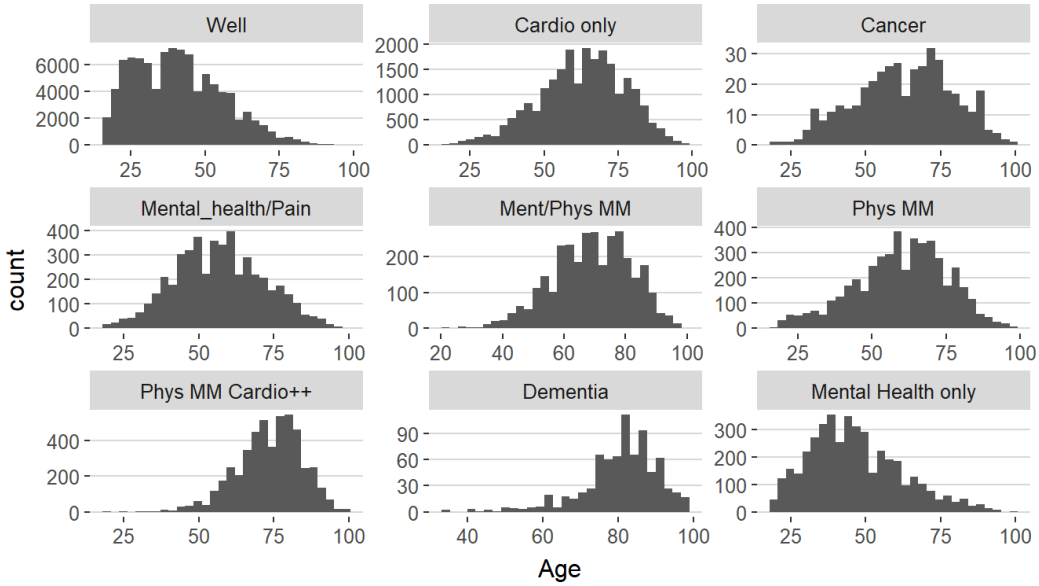

Figure 1.4: Histograms of Age, by Latent Group

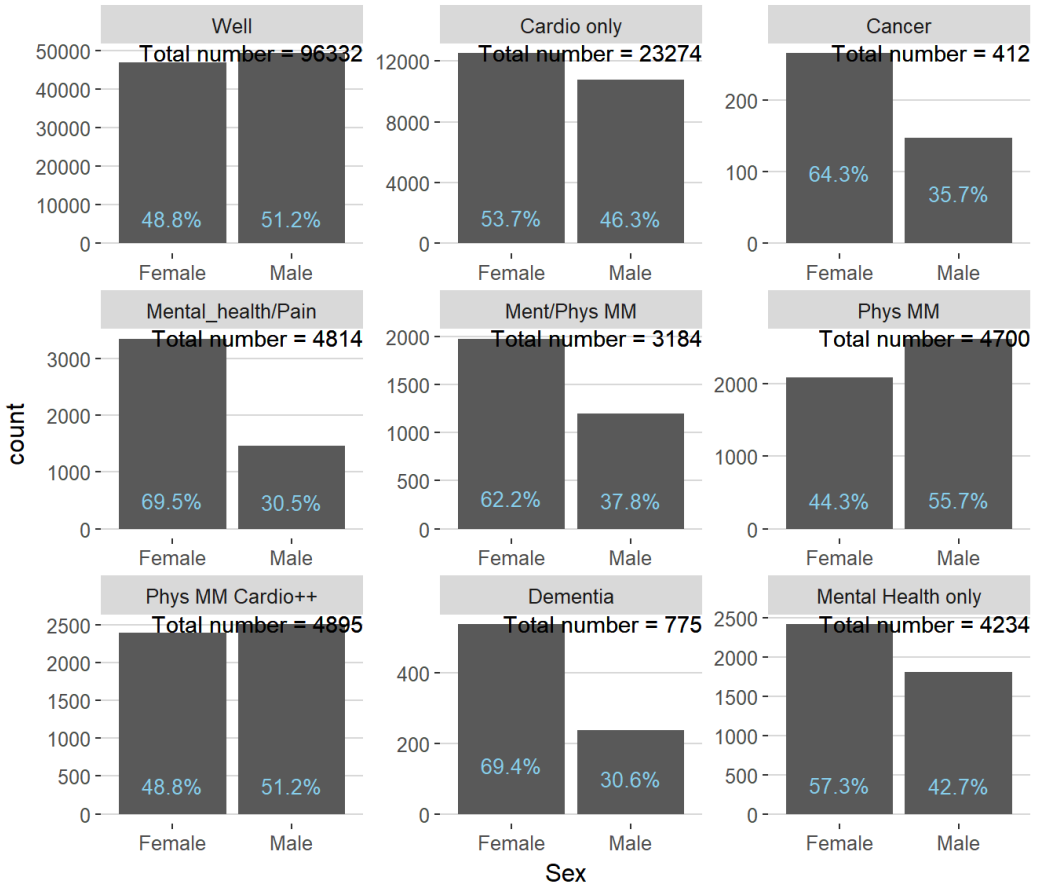

Figure 1.5: Count of Sec, by Latent Group - with Proportions

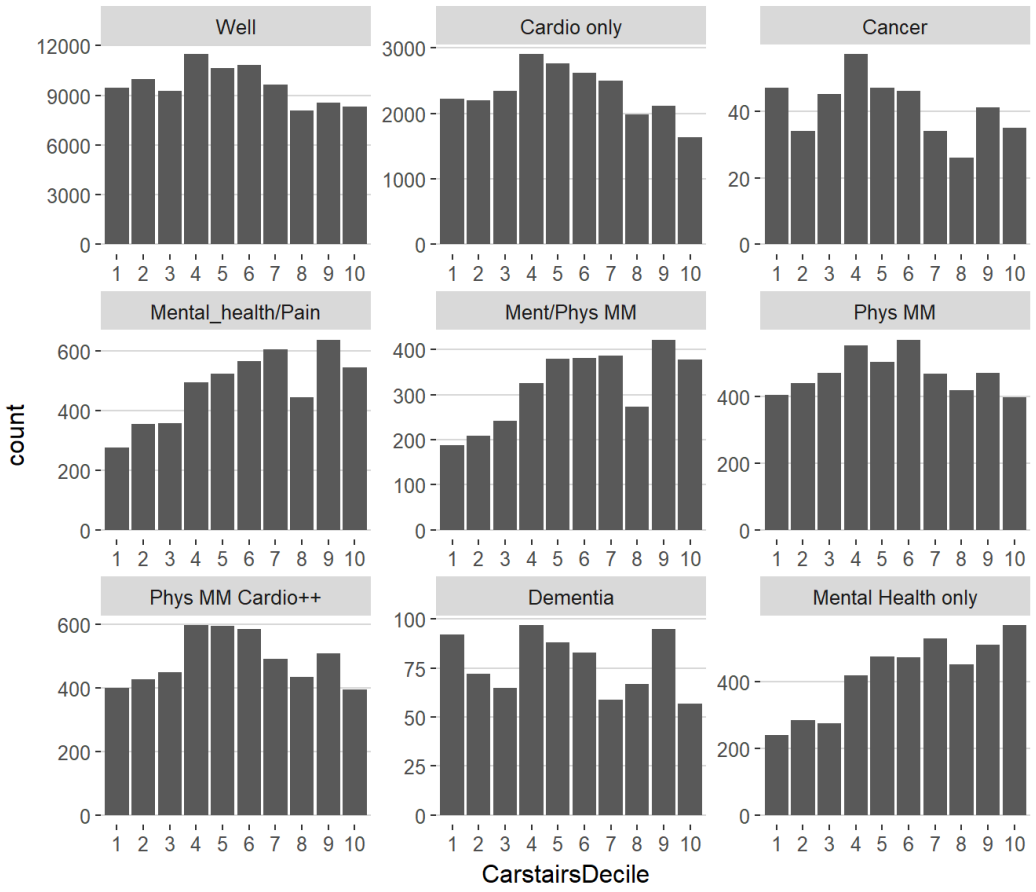

Figure 1.6: Count of Carstairs Decile, by Latent Group

There are clear gradients in the "Mental Health & Pain", "Ment/Phys MM", and "Mental Health only" groups with much higher numbers of people in Decile 10 (most deprived) areas.

1.5 Discussion

When proposing the two-way clustering method, Ng (2015) applied the technique to Australian survey data of 24 self-reported diseases from n=8841 respondents. These were "clumped" into nine non-overlapping groups of diseases with four latent groups of individuals being identified by the mixture model. Given the different classification of diseases, the nature of how data were collected, and the different populations, it is unsurprising that results from the present study are markedly different. What is of interest is whether the technique provided a meaningful representation of the data which can be used to improve the understanding of multimorbidity within this population.

The first step of the two-way clustering technique resulted in 13 non-overlapping groups being formed. As a result of this process, ten of the diseases recorded in the dataset were excluded from the final groups. Six diseases; Epilepsy, Learning Disability, Sinusitis, Crohns, Anorexia, and Psoriasis/Eczema did not have strong enough pairwise Somer's D correlation statistics with any other disease to be included in the non-overlapping

"clumping" procedure. A further four diseases; Diverticular, Prostate, IBS, and Dyspepsia did not appear in any non-overlapping group with the disease with which they had the strongest Somer's D correlation. Identifying which diseases do not co-occur at non-random levels is a particularly interesting finding and adds to the debate about which diseases should be included in any multimorbidity measure.[be explicit – do you think this suggests we shouldn't count these 10 in any MM measure and, if so, why? -NB] Of the 40 diseases present in the SPICE-PC dataset, in this analysis, only 30 diseases co-occurred with enough statistical power to be added to non-overlapping disease groups and considered in the mixture model.

Past research has highlighted associations between; mental health and thyroid problems, mental health and pain, and diseases associated with the metabolic syndrome (Prados-Torres *et al.*, 2014). As shown in figure 1, there is no pairwise correlation between any mental health diseases and thyroid disorders in the present study. Pain is associated with a large number of conditions, including mental health diseases. Diseases associated with metabolic syndrome such as, hypertension, coronary heart disease, diabetes, stroke, and heart failure all show correlations with each other to varying degrees. Prados-Torres et al (2014) also identified associations between Chronic Obstructive Pulmonary Disease (COPD) and Gastroesophageal reflux disease (GORD) with mental health diseases. No associations with between any respiratory disease and mental health diseases is apparent in figure 1. The SPICE-PC dataset does not record GORD, however Dyspepsia may be considered a similar diagnosis and has pairwise correlations with both Depression and Anxiety.

The non-overlapping disease groups were named according to the characteristics of the diseases present in each, however many diseases appeared in groups that may not come from similar aetiology. The groups reflect diseases that commonly co-occur and therefore do not always fit into clinically recognisable groups. Individuals are scored as to the number of diseases they have from each group. This results in a loss of some information but the advantage of this method is that it reduces the number of variables in the dataset making the likelihood of identifying a good mixture model much more likely.

Applying a mixture model of multivariate generalised Bernoulli distributions to these 13 non-overlapping disease groups identified nine latent groups as described above. These groups are clinically recognisable with some showing presence of diseases from only one disease group and others illustrating a more multimorbid population. Clear distinctions between mental and physical disease are made enabling individuals to have diseases from both groups. In their systematic review of clustering studies of multimorbidity, Prados-Torres et al (2014) found that included studies reported between three and twenty diseases clusters. From these they found three most common groups of diseases; cardiovascular, mental health, and musculoskeletal. Latent groups in the current study also contain cardiovascular and mental health elements, although they do not reflect

any groups that could be identified as musculoskeletal.

Clear sociodemographic patterns were identified in the latent groups. Those classified as belonging to a latent group with mental health involvement such as; mental health/pain, mental and physical multimorbidity, or mental health only, were more likely to be female and from lower deprivation deciles. Two similar latent groups; physical multimorbidity and physical multimorbidity high cardiovascular, had clear age differences with the former more likely to have younger male individuals assigned to it. These findings are similar to simple analyses of the same dataset which identified those in the most deprived areas being more likely to develop physical and mental health multimorbidity 10-15 years earlier than their more affluent peers (Barnett *et al.*, 2012).

1.5.1 Limitations

The two-step method proposed by Ng (2015) enables reduction of dimensions in datasets with large numbers of disease variables making model good identification of mixture models more likely. This, however, results in a loss of detail making interpretation of results more difficult. Report of the item-response probabilities shown in figure 4 identifies that individuals have a probability of having none, one, or two of the diseases in any group. It is impossible to identify exactly which diseases. Hypertension with 13.4% of individuals is the highest reported disease in the SPICE-PC . To what degree this accounts for individuals having one disease in the cardiovascular group across latent classes is unknown.

This is particularly a problem as the non-overlapping groups formed by the two-way clustering method do not always follow recognised clinical groupings. For example, Rheumatoid arthritis is found in the cardiovascular group despite being a connective tissue disorder. It is the only disease in the group that does not directly affect the cardiovascular system. The reason it is found in this group may be due to the association of increased NSAID use with cardiovascular outcomes (ref) resulting in frequent co-occurrence. As some latent groups are classified as "Cardiovascular only", there is a possibility that small numbers of people with only Rheumatoid Arthritis are misclassified. [This is could be down to nomenclature and I maybe need to revisit group names. It does not get rid of the problem that for the group with high probs for the cardio/rheum disease we can't distinguish which disease the individual has]

When clustering individuals, fitting a finite mixture model of multivariate generalised Bernoulli distributions makes the assumption that the indicator variables used to identify latent groups are independent. In a dataset containing medical conditions such as e.g. Hypertension and Coronary Heart Disease, or Atrial Fibrillation and Stroke, this assumption is clearly violated. However, ignoring the independence assumption for multivariate categorical variables often results in better fit than when more complicated techniques are applied to account for non-independence of indicator variables

(Hand and Yu, 2001; Topchy *et al.*, 2005; Ng, 2015).

Results presented here account for 10% of the SPICE-PC dataset due to the heavy computational requirements of fitting the finite mixture model. Confirmation of results on the whole dataset is required.

1.5.2 Future research

Sociodemographic comparison in the current study was limited to visualisation of histograms. More detailed reports of measures of central tendency and comparison of latent groups with sociodemographic variables such as Carstairs decile with logistic regression is warranted. SPICE-PC data also contains variables on lifestyle factors such as smoking status and alcohol intake. These variables should be included in further comparisons.

A further calculation offered in the two-way clustering technique (Ng, 2015) is the calculation of a multimorbidity score to each latent group and to each individual in the dataset based on their disease profile and posterior probabilities. Such as score would be continuous in nature and would offer the benefit of comparison with sociodemographic variables such Carstairs score or Age. Such comparisons would be amenable to well-recognised regression techniques.

Part 1 of the two-step method involves reducing the dimensions of the dataset to be amenable to mixture modelling. Using the outcome of this first step, the non-overlapping groups, Ng's technique for clustering individuals should be compared to Latent Class Analysis which is also a suitable technique for the nature of the data (Collins and Lanza, 2013). In supplementary material, NG (2015) compared his two-step method with Latent Profile Analysis and found it inferior to the two-step method. Latent Profile Analysis is a mixture model for continuous indicator variables (Collins and Lanza, 2013) and would not be suitable for the SPICE-PC dataset.

1.6 Conclusion

A novel two-step method for clustering health conditions and individuals was applied to large, population representative, dataset of administrative primary care data. The method found 10 of the 40 conditions in the dataset did not co-occur with other diseases strongly enough to included in further analysis. Thirty conditions were "clumped" into 13 groups of commonly co-occurring conditions and individuals in the dataset were assigned a score depending on the number of disease within each group they had. From this information nine latent groups of individuals were identified with a finite mixture model. These groups reflected varying degrees of physical, mental and physical/mental multimorbidity and showed sociodemographic patterning.

References

Agur, K. *et al.* (2016) 'How does sex influence multimorbidity? Secondary analysis of a large nationally representative dataset', *International journal of environmental research and public health*, 13(4), p. 391. Available at: http://www.mdpi.com/1660-4601/13/4/391/htm.

Barnett, K. *et al.* (2012) 'Epidemiology of multimorbidity and implications for health care, research, and medical education: A cross-sectional study', *Lancet*, 380(9836), pp. 37–43. doi: 10.1016/S0140-6736(12)60240-2.

Benjamini, Y. and Hochberg, Y. (1995) 'Controlling the false discovery rate: A practical and powerful approach to multiple testing', *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300. Available at: http://www.stat.purdue.edu/~doerge/BIOINFORM.D/FALL06/Benjamini%20and%20Y%20FDR.pdf.

Collins, L. M. and Lanza, S. T. (2013) *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. New Jersey: John Wiley; Sons.

Cornell, J. E. *et al.* (2009) 'Multimorbidity clusters: Clustering binary data from multimorbidity clusters: Clustering binary data from a large administrative medical database', *Applied Multivariate Research*, 12(3), pp. 163–182.

Elder, R. *et al.* (2007) *Measuring quality in primary medical services using data from spice*. Report. Available at: http://www.em-online.com/download/medical_article/36074_spice_report_july_2007.pdf.

Fortin, M. *et al.* (2006) 'Psychological distress and multimorbidity in primary care', *The Annals of Family Medicine*, 4(5), pp. 417–422. Available at: http://www.annfammed.org/content/4/5/417.short.

Fortin, M. *et al.* (2005) 'Comparative assessment of three different indices of multimorbidity for studies on health-related quality of life', *Health and Quality of Life Outcomes*, 3(1), p. 74. Available at: https://hqlo.biomedcentral.com/articles/10.1186/1477-7525-3-74.

Fortin, M. *et al.* (2004) 'Multimorbidity and quality of life in primary care: A systematic review', *Health and Quality of life Outcomes*, 2(1), p. 51. Available at:

https://hqlo.biomedcentral.com/articles/10.1186/1477-7525-2-51.

Fortin, M. *et al.* (2012) 'A systematic review of prevalence studies on multimorbidity: Toward a more uniform methodology', *The Annals of Family Medicine*, 10(2), pp. 142–151. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3315131/pdf/0100142.pdf.

Gijsen, R. *et al.* (2001) 'Causes and consequences of comorbidity: A review', *Journal of clinical epidemiology*, 54(7), pp. 661–674. Available at: http://ac.els-cdn.com/S0895435600003632/1-s2.0-S0895435600003632-main.pdf?_tid=108170ce-6dc1-11e5-8a3e-00000aacb35f&acdnat=1444311347_f32307220b48c8d4a7d7674af602

Hand, D. J. and Yu, K. (2001) 'Idiot's bayes—not so stupid after all?', *International statistical review*, 69(3), pp. 385–398. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1751-5823.2001.tb00465.x/full.

Imison, C. (2012) *Future trends: Overview*. Report. Available at: https://www.kingsfund.org.uk/sites/files/kf/field/field_publication_summary/future-trends-overview.pdf.

Islam, M. M. *et al.* (2014) 'Multimorbidity and comorbidity of chronic diseases among the senior australians: Prevalence and patterns', *PloS one*, 9(1), p. e83783. Available at: http://www.plosone.org/article/fetchObject.action?uri=info:doi/10.1371/journal.pone.0083783&representation=PDF.

Jardine, N. and Sibson, R. (1968) 'The construction of hierarchic and non-hierarchic classifications', *The Computer Journal*, 11(2), pp. 177–184. Available at: http://biocomparison.ucoz.ru/_ld/0/60_jardine_constru.pdf.

Kadam, U. and Croft, P. (2007) 'Clinical multimorbidity and physical function in older adults: A record and health status linkage study in general practice', *Family practice*, 24(5), pp. 412–419. Available at: http://fampra.oxfordjournals.org/content/24/5/412.full.pdf.

May, C. *et al.* (2009) 'We need minimally disruptive medicine', *BMJ: British Medical Journal (Online)*, 339. Available at: http://search.proquest.com/openview/548fdf667b7583537cd7c76e7f48cb8a/1?pq-origsite=gscholar&cbl=2043523.

McLean, G. *et al.* (2015) 'General practice funding underpins the persistence of the inverse care law: Cross-sectional study in scotland', *British Journal of General Practice*, 65(641), pp. e799–e805. doi: 10.3399/bjgp15X687829.

Ng, S. (2015) 'A two-way clustering framework to identify disparities in multimorbidity patterns of mental and physical health conditions among australians', *Statistics in medicine*, 34(26), pp. 3444–3460.

Ng, S. K. *et al.* (2012) 'Identifying comorbidity patterns of health conditions via cluster analysis of pairwise concordance statistics', *Statistics in medicine*, 31(27), pp.

3393–3405.

NICE (2016) *Multimorbidity: Clinical assessment and management. multimorbidity: Assessment, prioritisation and management of care for people with commonly occurring multimorbidity: NICE guideline ng56:* Report. Available at: https://www.nice.org.uk/guidance/ng56.

OECD (2011) 'Multimorbidity: The impact on health systems and their development', in OECD (ed.) *Health reform: Meeting the challenge of ageing and multiple morbidities*. OECD Publishing. Available at: http://dx.doi.org/10.1787/9789264122314-9-en.

Prados-Torres, A. *et al.* (2014) 'Multimorbidity patterns: A systematic review', *Journal of Clinical Epidemiology*, 67(3), pp. 254–266. doi: 10.1016/j.jclinepi.2013.09.021.

R-Core-Team (2017) *R: A language and environment for statistical computing. r foundation for statistical computing*. Report. Available at: http://www.R-project.org.

Salisbury, C. *et al.* (2011) 'Epidemiology and impact of multimorbidity in primary care: A retrospective cohort study', *British Journal of General Practice*, 61(582), pp. e12–e21. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3020068/pdf/bjgp61-e12.pdf.

Starfield, B. *et al.* (2005) 'Contribution of primary care to health systems and health', *Milbank quarterly*, 83(3), pp. 457–502. Available at: http://onlinelibrary.wiley.com/wol1/doi/10.1111/j.1468-0009.2005.00409.x/full.

Topchy, A. *et al.* (2005) 'Clustering ensembles: Models of consensus and weak partitions', *IEEE Transactions on pattern analysis and machine intelligence*, 27(12), pp. 1866–1881. Available at: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1524981.

van-den-Akker, M. *et al.* (2001) 'Problems in determining occurrence rates of multimorbidity', *Journal of clinical epidemiology*, 54(7), pp. 675–679.

Violan, C. *et al.* (2014) 'Prevalence, determinants and patterns of multimorbidity in primary care: A systematic review of observational studies', *Plos One*, 9(7). doi: 10.1371/journal.pone.0102149.